



# FERRAMENTA PARA TRATAMENTO DE INFORMAÇÕES SENSÍVEIS EM BANCOS DE DADOS

Varlen Pavani Neto

Projeto de Graduação apresentado ao Curso de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Heraldo Luis Silveira de Almeida

Rio de Janeiro  
Fevereiro de 2020

# FERRAMENTA PARA TRATAMENTO DE INFORMAÇÕES SENSÍVEIS EM BANCOS DE DADOS

Varlen Pavani Neto

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO  
DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA PO-  
LITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO  
PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU  
DE ENGENHEIRO ELETRÔNICO E DE COMPUTAÇÃO

Autor:

---

Varlen Pavani Neto

Orientador:

---

Prof. Heraldo Luis Silveira de Almeida

Examinador:

---

Prof xxxxx

Examinador:

---

Prof xxxxx

Rio de Janeiro

Julho de 2020

## Declaração de Autoria e de Direitos

Eu, *Varlen Pavani Neto* CPF 142.722.117-06, autor da monografia *FERRAMENTA PARA TRATAMENTO DE INFORMAÇÕES SENSÍVEIS EM BANCOS DE DADOS*, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na disciplina de Projeto de Graduação da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.
2. Excetua-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.
3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.
4. O autor pode, excepcionalmente, encaminhar à Comissão de Projeto de Graduação, a não divulgação do material, por um prazo máximo de 01 (um) ano, improrrogável, a contar da data de defesa, desde que o pedido seja justificado, e solicitado antecipadamente, por escrito, à Congregação da Escola Politécnica.
5. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.
6. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.
7. Por ser verdade, firmo a presente declaração.

---

Varlen Pavani Neto

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Escola Politécnica - Departamento de Eletrônica e de Computação

Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária

Rio de Janeiro - RJ CEP 21949-900

Este exemplar é de propriedade da Universidade Federal do Rio de Janeiro, que poderá incluí-lo em base de dados, armazenar em computador, microfilmear ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es).

## DEDICATÓRIA

TODO

## AGRADECIMIENTO

TODO

## **RESUMO**

Inserir o resumo do seu trabalho aqui. O objetivo é apresentar ao pretense leitor do seu Projeto Final uma descrição genérica do seu trabalho. Você também deve tentar despertar no leitor o interesse pelo conteúdo deste documento.

Palavras-Chave: Proteção de Dados, Privacidade, Anonimização, Bancos de Dados, Segurança da Informação.

## ABSTRACT

Insert your abstract here. Insert your abstract here. Insert your abstract here.  
Insert your abstract here. Insert your abstract here.

Keywords: Data Protection, Privacy, Anonymization, Databases, Information Security.



## SIGLAS

UFRJ - Universidade Federal do Rio de Janeiro

GDPR - *General Data Protection Regulation*

LGPD - Lei Geral de Proteção de Dados

OLAP - *Online Analytics Processing*

OLTP - *Online Transaction Processing*

ETL - *Extract-Transform-Loading*

PII - *Personally Identifiable Information*

NER - *Named Entity Recognition*

# Sumário

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>Introdução</b>  | <b>1</b> |
| 1.1      | Tema . . . . .   | 1        |
| 1.2      | Delimitação . . . . .  | 2        |
| 1.3      | Justificativa . . . . .  | 2        |
| 1.4      | Objetivos . . . . .  | 2        |
| 1.5      | Metodologia . . . . .  | 2        |
| 1.6      | Descrição . . . . .  | 3        |
| <b>2</b> | <b>Fundamentação</b>   | <b>4</b> |
| 2.1      | Privacidade . . . . .  | 4        |
| 2.1.1    | Privacidade e o Indivíduo . . . . .                            | 4        |
| 2.1.2    | A Sensibilização sobre privacidade . . . . .                   | 5        |
| 2.1.3    | A <i>General Data Protection Regulation</i> europeia . . . . . | 5        |
| 2.1.4    | A Lei Geral de Proteção de Dados brasileira . . . . .          | 7        |
| 2.2      | Bases de Dados . . . . .                                       | 8        |
| 2.2.1    | Dados Estruturados e Não-Estruturados . . . . .                | 8        |
| 2.2.2    | Tipos, Atributos e Registros . . . . .                         | 9        |
| 2.2.3    | Consulta . . . . .   | 10       |
| 2.2.4    | OLTP e OLAP . . . . .  | 10       |
| 2.2.5    | <i>Data Warehouses</i> . . . . .                               | 11       |
| 2.2.6    | Ferramentas e Processos de ETL . . . . .                       | 11       |
| 2.3      | Anonimização e Reidentificação . . . . .                       | 12       |
| 2.3.1    | Anonimização . . . . .   | 12       |
| 2.3.2    | Reidentificação . . . . .                                      | 13       |
| 2.4      | Técnicas de Anonimização . . . . .                             | 13       |

|          |   |           |
|----------|---|-----------|
| 2.4.1    | Aleatorização e adição de ruído . . . . .               | 14        |
| 2.4.2    | Privacidade Diferencial . . . . .                       | 15        |
| 2.4.3    | Agregação . . . . .                                     | 15        |
| 2.4.4    | k-anonimização . . . . .                                | 15        |
| 2.4.5    | l-diversidade . . . . .                                 | 15        |
| 2.4.6    | t-proximidade . . . . .                                 | 15        |
| 2.4.7    | Avaliação . . . . .                                     | 15        |
| 2.4.8    | Pseudonimização . . . . .                               | 15        |
| 2.4.9    | Comparativo das técnicas de anonimização . . . . .      | 15        |
| <b>3</b> | <b>Metodologia</b>                                      | <b>16</b> |
| 3.1      | Descoberta de Ferramentas . . . . .                     | 16        |
| 3.1.1    | Anonymizer . . . . .                                    | 16        |
| 3.1.2    | Data::Anonymization . . . . .                           | 17        |
| 3.1.3    | ARX - Open Source Data Anonymization Software . . . . . | 18        |
| 3.1.4    | Presidio . . . . .                                      | 19        |
| 3.2      | Visão Geral . . . . .                                   | 21        |
| <b>4</b> | <b>Implementação</b>                                    | <b>22</b> |
| 4.1      | Requisitos . . . . .                                    | 22        |
| 4.2      | Arquitetura Proposta . . . . .                          | 22        |
| 4.3      | Desenvolvimento . . . . .                               | 22        |
| <b>5</b> | <b>Validação</b>  | <b>23</b> |
| 5.1      | Dados de Teste . . . . .                                | 23        |
| 5.2      | Resultados . . . . .                                    | 23        |
| <b>6</b> | <b>Conclusão</b>  | <b>24</b> |
|          | <b>Bibliografia</b>                                     | <b>25</b> |

# Lista de Figuras

# Lista de Tabelas

|     |  |   |
|-----|--|---|
| 2.1 | Exemplos de registros para ficha cadastral de pacientes em uma clínica | 9 |
|-----|--|---|

# Capítulo 1

## Introdução

O presente capítulo tem como objetivo apresentar brevemente o escopo do trabalho desenvolvido assim como sua motivação, enumerando conceitos pertinentes a área de conhecimento.

### 1.1 Tema

Este projeto tem como tema o estudo de técnicas para anonimização e dessensibilização em conjuntos de dados. Especificamente, serão avaliadas diferentes técnicas e suas implementações em software de código aberto.

Este é um trabalho majoritariamente de Engenharia de Software, contemplando um ciclo de vida para planejamento, pesquisa e implementação da solução proposta.

É possível dizer que se trata de um trabalho da Área de Engenharia de Dados, dada a natureza das entidades que serão manipuladas. Também serão vistos conceitos da área de Segurança Digital, especificamente Direito Digital e Privacidade.

Assim, este projeto implementa uma nova ferramenta de anonimização e dessensibilização de código aberto englobando mais funcionalidades do que outras ferramentas existentes atualmente.

## 1.2 Delimitação

Este trabalho se limita a estudar técnicas de anonimização, garantia de privacidade e dessensibilização, criadas até o presente momento de sua concepção, que permitam compatibilizar um banco de dados contendo informações sensíveis a uso por terceiros. As implementações destas técnicas, quando existentes, serão estudadas a partir de softwares de código aberto.

## 1.3 Justificativa

TODO

Apresentar o porquê do tema ser interessante de ser estudado. Cuidado, não é a motivação particular. Devem ser apresentadas razões para que alguém deva se interessar no assunto, e não quais foram suas razões particulares que motivaram você a estudá-lo (tamanho do texto: livre).

## 1.4 Objetivos

O objetivo geral deste trabalho é implementar um software livre que permita processar um conjunto de dados tornando-o dessensibilizado, sendo desta forma possível utilizá-lo para apoiar atividades de desenvolvimento e testes de sistemas sem a preocupação com vazamento de dados por parte de terceiros. Especificamente, o software deve: (1) Remover informações que permitam associar indivíduos com um conjunto de dados; (2) Permitir a substituição de dados reais sensíveis por dados gerados a partir de estatísticas; (3) Respeitar e manter a estrutura do modelo de dados existente, alterando somente o seu conteúdo.

## 1.5 Metodologia

TODO Como é a abordagem do assunto. Como foi feita a pesquisa, se houve validação, etc. Em resumo, você deve explicar qual foi sua estratégia para atender ao objetivo do trabalho (tamanho do texto: livre).

## 1.6 Descrição

No capítulo 2 será feita a apresentação introdutória sobre privacidade, conceitos associados e suas relevâncias no contexto socioeconômico contemporâneo.

O capítulo 3 apresenta um panorama de ferramentas de anonimização cuja licença garanta a abertura do código-fonte com o intuito de prospectar as funcionalidades fornecidas por esses programas.



# Capítulo 2

## Fundamentação

O presente capítulo propõe-se a introduzir o leitor ao entendimento existente sobre o conceito de privacidade, principalmente no que contempla a área de Tecnologia da Informação, o respectivo impacto social e a consequente legislação, com foco nos pontos de interesse deste trabalho, principalmente com relação ao tratamento de informações pessoais.

Inicia-se pelo ponto de vista mais amplo e filosófico da relação entre o indivíduos e privacidade, de onde se derivam suas consequências sociotécnicas.

A parte final do capítulo dedica-se a eventual familiarização do leitor ao domínio de implementação deste trabalho, apresentando definições técnicas relativas a armazenamento de dados.

### 2.1 Privacidade

#### 2.1.1 Privacidade e o Indivíduo

De acordo com Goffman[1], quando uma pessoa se apresenta diante de outras, esta terá motivações diversas para controlar a impressão que causa, moldando sua identidade social, através da disponibilização controlada de informações sobre si, de acordo com a audiência.

Assim, a revelação indesejada de informações pode levar a identidade construída ao descrédito, sendo potencialmente prejudicial ao indivíduo.

Um exemplo concreto deste conceito pode ser observado na definição de profissionalismo apresentada por Abril et al [2] que implica que o profissionalismo requer a segregação entre a persona profissional e privada do indivíduo, dado que existem comportamentos ditos socialmente aceitáveis que não são permitidos no ambiente de trabalho.

A privacidade, definida por Warren e Brandeis como “direito à reserva de informações pessoais e da própria vida pessoal” [3], pode ser vista como uma ferramenta fundamental para garantia da manutenção da dignidade do indivíduo, tendo assim caráter de direito fundamental.

### **2.1.2 A Sensibilização sobre privacidade**

De acordo com [4], a Web 2.0 e o consequente crescimento da cultura participativa aonde os próprios usuários consumidores são também geradores de conteúdo implicou numa percepção de risco relacionado a privacidade cada vez maior.

O contexto tecnológico existente permite que instituições públicas e privadas façam uso profuso de dados pessoais para desempenhar suas atividades. Dentro deste contexto, as atividades realizadas por indivíduos são frequentemente rastreadas e utilizadas como insumo; Em alguns casos, sem seu devido conhecimento ou consentimento.

A percepção sobre a ampla disponibilidade de dados na Internet trouxe o tema às pautas dos debates políticos mundiais, culminando na implementação de regulamentações para padronizar o tratamento adequado a estas informações, de modo a impedir o abuso no seu uso e garantir o direito a privacidade individual.

### **2.1.3 A *General Data Protection Regulation* europeia**

A GDPR é o conjunto de regras europeias definidas para o tratamento de dados, criadas com o objetivo de garantir que o processamento de dados respeite os direitos fundamentais de cada indivíduo.

Seu texto no diário oficial da Europa afirma que o avanço tecnológico acelerado e a globalização impuseram novos desafios para a proteção de dados pessoais[5] e visa provisionar a consistência necessária para remover barreiras no fluxo de dados entre diferentes países da União Européia.

Para isto, o texto da GDPR instaura que a aplicação de suas regras deve ser homogêneas entre os países e abre margem para que regras locais sejam definidas de acordo com a necessidade dos agentes públicos, como por exemplo, para questões de segurança nacional.

A lei define o detalhamento dos direitos e deveres de quem se utiliza do processamento de dados pessoais, assim como mecanismos para fiscalização do respeito as regras e sanções aplicáveis nos países membros. A proteção fornecida por este dispositivo legal é aplicável a indivíduos de qualquer nacionalidade ou residência.

A GDPR gera a necessidade de adaptação também fora da Europa, pois estabelece que o processamento de dados oriundos de indivíduos pertencentes a União Europeia por instituições não pertencentes para o oferecimento de bens e serviços ou rastreamento também estará sujeito a este conjunto de regras.

É estabelecido que os provedores de serviços devem coletar o consentimento dos usuários sobre o uso de seus dados para diferentes escopos, informando o usuário sobre estes de maneira clara, permitindo a seleção parcial de escopos, sem a utilização de *dark patterns*, como caixas de seleção pré-marcadas ou impedimento da utilização do serviço. O consentimento deve ser fornecido livremente pelo usuário, no sentido em que, não se deve obrigar o usuário a fornecer consenso para um escopo se a ação que o usuário deseja realizar pertence ou depende de um outro escopo. Informações textuais sobre a utilização dos dados coletados devem estar escritas de maneira clara, sem ambiguidades e facilmente acessíveis.

Além disso, o texto da lei também especifica que o processamento de dados deve ser utilizado somente quando o objetivo de uma atividade não puder ser realizado por outros meios e requer que nestes casos, a instituição implemente mecanismos

para prevenção de acessos não autorizados aos dados e seus instrumentos de processamento, além de realizar o expurgo destes dados com a periodicidade adequada. Para que o processamento ocorra de maneira legal, o consentimento do usuário deve ter sido previamente concedido à instituição e a instituição deve ser capaz de demonstrar este consentimento.

#### **2.1.4 A Lei Geral de Proteção de Dados brasileira**

No Brasil, a lei número 13.709/2018[6] é a responsável pelas diretrizes para tratamento de dados pessoais. O texto possui múltiplos tópicos similares a lei europeia. No momento da escrita deste trabalho, a lei começará a vigorar em maio de 2021.

A lei considera como dados pessoais informações relacionadas a pessoa natural identificada ou identificável e tratamento como qualquer operação realizada com dados pessoais. Além disso, definem-se agentes de tratamento de dados - Controlador: aquele a quem compete as decisões referentes ao tratamentos de dados pessoais; Operador: Aquele que realiza o tratamento de dados pessoais em nome do controlador. Ambas entidades podendo ser pessoa natural ou jurídica, de direito público ou privado. Titular é a pessoa natural cujos dados pessoais são objeto de tratamento.

Assim como a GDPR, a lei brasileira é destinada a proteção de indivíduos no território nacional independentemente da origem e localização da instituição que processa os dados.

Com relação as obrigações a serem cumpridas pelos agentes de tratamento de dados, a lei brasileira institui a necessidade de coleta de consentimento revogável e fornecimento de acesso às informações de tratamento dos dados, incluindo a finalidade, forma, duração e identificações. Também requiere-se que o controlador informe previamente o titular dos dados sobre mudanças na sua finalidade de uso.

A lei nacional cria a Agência Nacional de Proteção de Dados e o Conselho Nacional de Proteção de Dados Pessoais e Privacidade, órgãos governamentais responsáveis pela fiscalização e aplicação das sanções definidas no texto da lei.

Um ponto do texto de específico interesse neste trabalho é a anonimização, definida como a utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio das quais um dado perde a possibilidade de associação direta ou indireta, a um indivíduo. Dados anonimizados não são considerados como dados pessoais se o processo de anonimização não for factivelmente reversível, considerando-se o custo e o tempo necessários.

## **2.2 Bases de Dados**

Mostram-se a seguir, conceitos relacionados a armazenamento eletrônico de dados que serão explorados ao longo do trabalho.

### **2.2.1 Dados Estruturados e Não-Estruturados**

Quando se preenche um formulário em uma folha de papel, as informações inseridas estarão espacialmente confinadas, de acordo com o contexto previamente existente, no caso o nome dos campos impressos na folha.

Esse confinamento espacial dos dados em cada campo implica na existencia de uma estrutura, aonde é possível localizar uma informação específica de acordo com o nome do campo.

Neste exemplo, o nome do campo é um metadado, ou seja, uma informação sobre o dado específico. O mesmo raciocínio se aplica a dados preenchidos em um formulário eletrônico.

Todavia, quando se escreve uma carta, realiza-se um telefonema, há um espalhamento das informações apresentadas e não será possível associar diretamente uma informação a um rótulo como no caso do formulário. Em outras palavras, o meio com que a informação é apresentado não implica em uma estrutura que possa ser utilizada para recuperar, sem processamento adicional, uma parte dessa informação.

De acordo com Plejic et Al., dados estruturados sempre são apresentados juntos aos metadados[7].

## 2.2.2 Tipos, Atributos e Registros

De maneira geral, um registro é uma unidade de dados estruturados. Por exemplo, num consultório médico, existem diversas fichas cadastrais dos pacientes, contendo nome, números de documentos, informações de contato e histórico de consultas. Essas fichas, quando implementadas em um meio eletrônico, seja através de um software que acessa um banco de dados ou mesmo de uma planilha, podem ser armazenadas sob a forma de linhas em uma tabela. Cada linha desta tal tabela hipotética é um registro e cada uma das colunas nesta linha é um atributo do registro.

Tabela 2.1: Exemplos de registros para ficha cadastral de pacientes em uma clínica

| Nome        | Idade | Logradouro       | Telefone  | Data da Última Consulta |
|-------------|-------|------------------|-----------|-------------------------|
| João Silva  | 32    | Rua do Trigo, 1  | 9999-9999 | 02/01/2020              |
| Maria Souza | 30    | Rua das Neves, 2 | 9999-5555 | 02/01/2020              |

Os atributos, ou campos, dos registros armazenam a menor unidade de informação dentro de um conjunto de dados. Alguns destes campos permitem a associação entre diferentes registros ao possuírem valores únicos. Suponha que uma loja armazena o número de telefone de seus clientes, assim como seu histórico de compras.

Se o número de telefone é o mesmo que os clientes utilizaram para a ficha cadastral do consultório médico mostrada anteriormente, considerando também um agente que possui acesso a ambos os dados, da loja e do consultório médico, é possível realizar a junção dos dados dos dois registros. Assim, correlacionando as informações clínicas e de consumo dos clientes.

Além disso, cada atributo pertence a um tipo de dado. No exemplo da loja, o campo de idade é armazenado como um tipo número inteiro não negativo, assim como o nome é armazenado como uma cadeia de caracteres. A nível de registro, é possível definir um tipo a partir do conjuntos de valores que este pode assumir [8]. Tipos impõem restrições nas operações que podem ser realizadas nos atribu-

tos, podem dar sentido semântico [9], e definem o formato de armazenamento da informação, garantindo sua consistência.

A quantidade de atributos em um registro determina a sua dimensionalidade. Diz-se que o conjunto de dados é esparsa quando dois registros aleatórios estão sempre distantes no espaço multidimensional formado por seus atributos.

### 2.2.3 Consulta

Uma consulta é uma função aplicada ao banco de dados. Essa função tem por objetivo obter os valores de atributos para registros de acordo com os critérios nela definidos.

Softwares de bancos de dados que implementam o padrão ISO 9075 são comuns para bancos de dados relacionais e permitem utilizar a linguagem SQL para realização de consultas.

Exemplos desses softwares são Postgres, MySQL, SQLite e Microsoft SQL Server.

### 2.2.4 OLTP e OLAP

Uma transação em um banco de dados é um conjunto de leituras e escritas com baixa latência. As transações ações comuns para recuperação e persistência de dados em várias aplicações, como por exemplo, adicionar produtos ao carrinho de compras de uma loja virtual, consultar o saldo bancário, etc. O padrão de acesso quando alguns registros são criados ou atualizados conforme as interações do usuário chama-se *Online Transaction Processing*(OLTP).

Quando os bancos de dados são utilizados para fins analíticos, no entanto, a principal característica do padrão de acesso deixa de ser a manipulação de alguns registros. Nesses casos de uso, as consultas mais comuns envolvem a agregação e processamento de uma grande quantidade de registros enquanto a modificação destes registros ocorre com menor frequência e geralmente, em lotes. Este padrão de acesso se chama *Online Analytics Processing*(OLAP).

Inicialmente, os mesmos bancos de dados eram utilizados para ambos os propósitos. A tendência em separar os sistemas OLTP dos sistemas OLAP deu origem as *Data Warehouses*. [10]

### **2.2.5 *Data Warehouses***

De acordo com Ali El-Sappagh et al. [11], um *Data Warehouse* é um conjunto de tecnologias que permitem a uma instituição tomar decisões melhores e mais rápidas, diferindo de bancos de dados tradicionais no sentido de serem variantes no tempo, orientados por assunto, não normalizados e otimizados para OLAP.

Sendo assim, o objetivo da *Data Warehouse* é fornecer o aparato tecnológico para a realização de análises sobre grandes volumes de dados agregados de diferentes fontes, sendo estas provenientes dos diferentes domínios de atuação da instituição.

### **2.2.6 Ferramentas e Processos de ETL**

Ferramentas de *Extract-Transform-Loading*(ETL) são soluções em software que permitem a integração de diferentes fontes de dados a uma *Data Warehouse*. De maneira geral, um processo de ETL realiza um mapeamento entre dois bancos de dados, possibilitando a aplicação de operações sobre o conjunto mapeado com o intuito de adequá-lo ao esquema do destino.

ETLs podem aceitar como entradas dados de diversas origens, como bancos de dados de provedores distintos, arquivos de texto ou planilhas. O passo inicial da ETL é a extração, sendo responsável pela captura dos dados. Este passo implementa as integrações para cumprir com os requisitos de operação específicos de cada origem, como por exemplo o protocolo de comunicação e o formato de armazenamento.

No momento da criação da *Data Warehouse*, o processo de extração é realizado em cima de um conjunto massivo de dados, realizando a extração inicial. Este processo é realizado uma única vez. [12] Após a extração inicial, os dados das fontes são obtidos a partir de Changed Data Capture. Isso implica que nas próximas extrações realizadas, apenas os dados novos ou alterados serão carregados. O software deve



determinar quais dados serão carregados. Isso pode ser feito a partir de técnicas como observar o log do banco de dados, auditar colunas específicas, observar a data de modificação de arquivos, dentre outras.

O próximo passo do processo de ETL é a transformação. Este passo tem o objetivo de conformar o conjunto de dados de entrada às suas expectativas de uso, provendo um conjunto de dados consistente e sem ambiguidades a partir da limpeza e agregação da entrada. As regras utilizadas para transformação são armazenadas sob a forma de metadados.

A conclusão do processo de ETL se dá com o passo de carga. Essa etapa do processo é responsável por escrever os dados resultantes da transformação nos repositórios de dados que serão utilizados pelos usuários finais em seus respectivos mecanismos de armazenamento.

## **2.3 Anonimização e Reidentificação**

### **2.3.1 Anonimização**

Por muitas vezes atributos em registros contêm dados pessoais como CPF, nome, endereço e telefone. É fácil perceber que estes permitem atribuir um registro a um indivíduo específico diretamente.

Essa atribuição, entretanto, é por muitas vezes indesejada. Tomando como exemplo a divulgação do resultado de testes para medicamentos, é importante para a sociedade saber características que possam determinar o comportamento da substância pesquisada em certos grupos de indivíduos. Por outro lado, é fundamental para os próprios indivíduos que seus dados pessoais não estejam presentes nos resultados públicos destes testes.

Para garantir que estas duas condições sejam satisfeitas, é possível implementar um processo de anonimização sobre o conjunto de dados trabalhados. Este consiste em um processamento irreversível[13] sobre o conjunto de dados através de substituições, agregações e supressões de registros e seus atributos com o intuito final de

gerar um novo conjunto de dados cuja possibilidade de divulgação não implica em um risco para os indivíduos e instituições envolvidos.

### 2.3.2 Reidentificação

É trivial perceber que os atributos que levam a associação direta de um indivíduo com seus registros devem ser suprimidos do conjunto de dados. Esta intuição entretanto não leva em consideração a possibilidade de reidentificação de pessoas no conjunto anonimizado através do processamento de múltiplos registros de fontes diversas.

Em 2008 Arvind Narayanan e Vitaly Shmatikov demonstraram uma metodologia robusta[14] que permitiu quebrar o anonimato de um conjunto de dados da Netflix. Os pesquisadores utilizaram as preferências individuais, recomendações e registro de transações do conjunto anonimizado e dados públicos disponíveis no site *Internet Movie Database* (IMDB), revelando a identidade dos usuários, seu alinhamento político aparente, dentre outras informações pessoais.

Outro exemplo deste tipo de ataque foi demonstrado por Malin e Sweeney[15] em 2004 ao recuperar informações genéticas de pacientes a partir dos padrões de visitação e dados disponíveis no ambiente informacional da saúde.

## 2.4 Técnicas de Anonimização

A avaliação de riscos sobre dados sensíveis pode ser feita a partir da definição dos seguinte perfis: A instituição, responsável pela governança dos dados; O adversário, que deseja abusar do conjunto de dados divulgado para obter informações implícitas, possivelmente violando a privacidade dos indivíduos representados nos registros; O terceiro, interessados em informações analíticas e estruturais no conjunto de dados divulgado. A anonimização transforma um conjunto de dados inicial em um conjunto anonimizado, ou seja, livre de informações pessoais identificáveis.

Paul Ohm[16] diz que "Dados podem ser úteis ou perfeitamente anônimos mas nunca ambos." Neste sentido, uma aplicação de anonimização bem sucedida deve

impedir o adversário de recuperar informações a partir da associação deste com informações auxiliares de sua posse e ao mesmo tempo deve manter a utilidade dos dados para utilização de terceiros.

Neste escopo, algumas técnicas podem ser utilizadas para implementar a anonimização de registros, de acordo com as metas de compromisso entre utilidade e segurança das informações.

### **2.4.1 Aleatorização e adição de ruído**

A aleatorização é uma técnica proposta por Warner[17] com o objetivo de resolver o seguinte problema: estimar qual parcela de pessoas em uma população possui um atributo específico sem que seja possível obter a informação para uma amostra individual.

A aplicação computacional desta técnica foi demonstrada por Agrawal e Srikant[18] no contexto da mineração de dados, propondo um procedimento para construir classificadores de árvore de decisão acurados a partir de um conjunto de dados cujos valores originais foram alterados. Essas alterações podem ser através da substituição de uma informação pela sua alocação dentro de um conjunto com risco menor ou através da adição de uma variável aleatória ao valor real.

Aproveita-se do fato de que cada atributo em um registro implica em um potencial risco de privacidade diferente. Um usuário pode não se sentir confortável para expor o valor exato de seu salário, mas não ter problemas em revelar a faixa salarial, por exemplo.

A aleatorização para um atributo binário pode ser exemplificada a partir do algoritmo mostrado por Dwork et al.[19]: Arremessa-se uma moeda. Caso coroa, então responde-se a informação original. Do contrário, realiza-se um novo arremesso e responde-se “sim” para coroa ou “não” para cara.

## **2.4.2 Privacidade Diferencial**

De acordo com Dwork[20], a privacidade diferencial é uma definição e não um algoritmo específico.

A autora argumenta que conjuntos de dados diferencialmente privados garantem que seja possível inferir conclusões a partir de um mesmo que um indivíduo opte por não permitir que seus dados sejam incluídos e permitem a um curador garantir que um indivíduo não seja afetado por permitir a utilização de seus dados.

Isso permite que um indivíduo participe de uma pesquisa sem que suas respostas individuais sejam reveladas.

TODO (EXPANDIR; possivelmente incluir a definição matemática)

## **2.4.3 Agregação**

TODO: (Possivelmente juntar as técnicas a seguir em um tópico único)

## **2.4.4 k-anonimização**

## **2.4.5 l-diversidade**

## **2.4.6 t-proximidade**

## **2.4.7 Avaliação**

Aspectos possíveis de avaliação: - eficiência - efetividade

## **2.4.8 Pseudonimização**

TODO:(VERIFICAR SE SERÁ PERTINENTE...)

## **2.4.9 Comparativo das técnicas de anonimização**

TODO

# Capítulo 3

## Metodologia

Este capítulo apresenta o resultado da prospecção das ferramentas de código aberto existentes com o intuito de entender as capacidades fornecidas e determinar os requisitos necessários a uma nova solução.

### 3.1 Descoberta de Ferramentas

Para descoberta de ferramentas de anonimização neste trabalho, foram utilizados projetos de código-aberto disponíveis em repositórios públicos no GitHub.

Especificamente, foram consideradas as métricas de stars e forks dos repositórios como medida de popularidade para seleção de 4 projetos relativos a anonimização de dados. Todos os projetos apresentavam novas modificações no código-fonte em Março de 2020, indicando que são mantidos pelas comunidade. Buscou-se inferir os casos de uso, funcionalidades, usabilidade e tecnologias utilizadas a partir da documentação dos projetos. Considerou-se a primeira linha dos arquivos README dos repositórios como nome do projeto.

#### 3.1.1 Anonymizer

Anonymizer[21] foi desenvolvida na linguagem de programação Ruby pela companhia europeia Divante e opera exclusivamente em bancos de dados MySQL. Segundo a documentação, sua funcionalidade mais importante é a formatação de dados. A ferramenta substitui os dados originais por dados gerados de acordo com o tipo.

Sua instalação é feita a partir de uma cópia do repositório de código-fonte para a máquina na qual se deseja executar.

O processo de anonimização é feito a partir de um arquivo de dump do banco de dados. Este arquivo pode estar na mesma máquina que executa o processo ou em uma máquina remota. O usuário deve criar um arquivo no formato JSON com os parâmetros do processo.

A ferramenta permite substituir os valores nas tabelas por valores fixos, valores em branco ou gerados, de acordo com a categoria. São suportadas as categorias `firstname`, `lastname`, `login`, `email`, `telephone`, `company`, `street`, `postcode`, `city`, `full_address`, `vat_id`, `ip`, `quote`, `website`, `iban`, `json`, `uniq_email`, `uniq_login`, `regon` (equivalente polonês ao CNPJ) e `pesel` (equivalente polonês ao CPF).

Também é possível truncar todos os dados de uma tabela e executar comandos SQL arbitrários antes ou depois do processo.

### 3.1.2 Data::Anonymization

Data::Anonymization[22] é uma solução criada pela ThoughtWorks Inc usando a linguagem de programação Ruby. Sua instalação é feita através do gerenciador de pacotes do Ruby.

É possível utilizar a ferramenta em bancos de dados para os quais existam uma implementação de driver compatível com o Active Record[23], que é a implementação de mapeamento objeto relacional do framework Ruby on Rails. O repositório fornece exemplos de uso para SQLite, Postgres e MongoDB.

Seu funcionamento é similar ao de uma biblioteca. O usuário deve utilizar Linguagem Específica de Domínio fornecida pela ferramenta definindo as tabelas, suas relações e qual estratégia de anonimização deve ser aplicada a cada coluna.

O usuário pode definir se deseja que todos os campos sejam anonimizados e explicitar os campos que não devem ser anonimizados ou se deseja modificar apenas os

campos listados. Como exemplos de estratégias de anonimização é possível listar: geração de valores aleatórios, tanto para campos numéricos quanto textuais; sorteio de um valor a partir de uma lista; modelos formatados; resultados de uma consulta no banco de dados; deslocamento de valores de instantes de tempo e intervalos de tempo; geração de números de telefone, código postal e endereço.

Caso seja usado o modo em que todos os campos são anonimizados, a ferramenta aplicará estratégias de anonimização padrão dependendo do tipo de dado da coluna. Alguns registros podem ser ignorados a partir de verificações condicionais.

Também é possível estender a funcionalidade embutida e implementar um estratégia customizada utilizando a linguagem Ruby. Essa estratégia é então disponibilizada para uso na Linguagem Específica de Domínio.

Um detalhe de implementação importante é que a ferramenta altera os valores diretamente no banco de dados em que está conectada.

### 3.1.3 ARX - Open Source Data Anonymization Software

Os criadores do ARX[24] explicam que o seu objetivo é produzir um software livre com alto grau de automação que fornece uma gama variada de técnicas de anonimização. Várias destas técnicas estão descritas em publicações específicas e incluem algoritmos criados a partir de modelos estatísticos, teoria dos jogos, privacidade diferencial, dentre outras. O programa fornece uma interface gráfica para operação e sua entrada de dados é feita a partir de arquivos CSV.

O ARX utiliza um algoritmo de busca global para transformação de dados com *full-domain generalization* e supressão de registros. *Full-domain generalization* implica que todos os valores de um atributo são transformados ao mesmo nível de generalização em todos os registros.

Ou seja, um atributo (ou coluna) categórico de Gênero em um registro tipicamente possui os valores Masculino, Feminino e Outros. Esse atributo é generalizado a um nível hierarquico acima através da supressão. Ou seja, o valor armazenado no

registro é transformado em ”\*”. No caso de um atributo numérico de idade, a generalização transforma um valor específico em uma faixa etária. A extensão da faixa depende do nível de generalização aplicado. O resultado de todas as operações de transformação deve considerar a relação de custo-benefício entre anonimização e utilidade.

São implementadas também funções para agregação dos dados: Para valores numéricos é possível aplicar a média aritmética, média geométrica, mediana, moda, além de agregação em conjunto ou intervalo. Para valores categóricos, as operações de mediana, moda e agregação por conjunto estão disponíveis. O usuário também pode definir funções personalizadas indicando quais conjuntos de atributos deverão ser transformados em um valor comum.

O programa busca a solução ótima para a anonimização do conjunto de dados utilizando configurações definidas pelo usuário como o número máximo de registros que podem ser suprimidos, particionamentos do conjunto de dados a serem transformados utilizando técnicas diferentes, amostragem do conjunto de dados de entrada avaliadas contra diversos modelos de riscos de privacidade.

O ARX fornece uma solução robusta para anonimização, sendo a ferramenta mais flexível dentre as ferramentas avaliadas, ao custo de maior complexidade de operação. Entretanto, não necessariamente conserva os tipos de dados entre o conjunto original e o conjunto de saída (Uma idade é um tipo numérico, enquanto uma faixa etária é uma estrutura de dois valores numéricos).

### **3.1.4 Presidio**

Presidio é um software livre criado pela Microsoft[25] com o intuito de ajudar no gerenciamento de informações textuais sensíveis. Suas funcionalidades provém capacidades analíticas e anonimização para dados como nomes, lugares, números de documentos específicos dos Estados Unidos e Reino Unido, dados financeiros, carteiras de criptomoedas, dentre outros.



O programa é capaz de extrair entidades de dados textuais não estruturados a partir de listas, expressões regulares ou processamento de linguagem natural (NER). É possível expandir suas funcionalidades através da API ou alterando-se o código-fonte, permitindo que seu uso seja adaptado a diferentes necessidades institucionais.

Ao operar sobre um conjunto de dados textuais não-estruturados, a ferramenta detecta automaticamente informações pessoais (PII) e gera uma análise sobre os tipos de informações contendo uma pontuação que pode ser utilizada para inferir a qualidade do texto anonimizado. A saída deste processo de anonimização são os dados estruturados com as informações pessoais suprimidas.

No momento da escrita deste trabalho, a ferramenta também oferece em caráter experimental a possibilidade de detectar textos e aplicar anonimização por supressão em imagens a partir de reconhecimento óptico de caracteres.

A ferramenta possui suporte a API REST; aos mecanismos de armazenamento em nuvem Azure Blob Storage e S3; aos sistemas de fluxos Kafka e Azure Event Hub como entradas de dados. Sua saída pode ser enviada para bancos de dados MySQL, MSSQL e Postgres ou nos mecanismos de armazenamento e fluxos previamente listados.

É possível criar tarefas agendadas para realizar a ETL de anonimização dos dados periodicamente através de arquivos de configuração.

Sua implementação utiliza as linguagens de programação Go e Python. A arquitetura segue o padrão de microserviços em containers coordenados através de Kubernetes. Também utiliza-se Redis para cache.

As funcionalidades que dependem do processamento de linguagem natural utilizam a biblioteca SpaCy do Python. Porém, no momento de escrita, somente os modelos para língua inglesa são suportados.[26] Dentre as modificações necessárias para utilização da ferramenta no contexto nacional está a utilização de um modelo adequado com suporte a língua portuguesa. Este modelo está disponível no site da

biblioteca[27], assim como modelos para Alemão, Francês, Espanhol, dentre outros idiomas.

Os principais serviços realizam a Análise, Anonimização, Anonimização de Imagens e Reconhecimento Óptico de Caracteres. Além disso, também existem serviços únicos para API e agendamento, além do acesso de leitura e escrita as fontes de dados. Esta arquitetura fornece resiliência para as comunicações entre diferentes serviços através de *circuit-breaking*, tentativas, balanceamento de carga, dentre outras técnicas. O uso de containers diminui a ocorrência de problemas no gerenciamento de dependências[28] uma vez que estas e suas respectivas configurações estão inclusas nas imagens utilizadas para montagem, efetivamente facilitando a implementação.

## 3.2 Visão Geral

# Capítulo 4

## Implementação

### 4.1 Requisitos

### 4.2 Arquitetura Proposta

### 4.3 Desenvolvimento

# Capítulo 5

## Validação

### 5.1 Dados de Teste

### 5.2 Resultados

## Capítulo 6

## Conclusão

# Referências Bibliográficas

- [1] GOFFMAN, E., OTHERS, *The presentation of self in everyday life*. Harmondsworth London, 1978.
- [2] SÁNCHEZ ABRIL, P., LEVIN, A., DEL RIEGO, A., “Blurred boundaries: Social media privacy and the twenty-first-century employee”, *American Business Law Journal*, v. 49, n. 1, pp. 63–124, 2012.
- [3] WARREN, S. D., BRANDEIS, L. D., “The right to privacy”, *Harvard law review*, pp. 193–220, 1890.
- [4] RUIZ, E. E. S., “Comentários à Lei Geral de Proteção de Dados: Lei n. 13.709/2018, com alteração da Lei n. 13.853/2019”, 2020.
- [5] UNION, C. O. E., “Council regulation (EU) no 679/2016”, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>, 2016.
- [6] BRASIL, “LEI 13.709 DE AGOSTO DE 2018”, DOU, 2018.
- [7] PLEJIC, B., VUJNOVIC, B., PENCO, R., “Transforming unstructured data from scattered sources into knowledge”. In: *2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop*, pp. 924–927, 2008.
- [8] CARDELLI, L., WEGNER, P., “On understanding types, data abstraction, and polymorphism”, *ACM Computing Surveys (CSUR)*, v. 17, n. 4, pp. 471–523, 1985.
- [9] DONAHUE, J., “On the semantics of “data type””, *SIAM Journal on Computing*, v. 8, n. 4, pp. 546–560, 1979.
- [10] KLEPPMANN, M., *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. ”O’Reilly Media, Inc.”, 2017.

- [11] EL-SAPPAGH, S. H. A., HENDAWI, A. M. A., EL BASTAWISSY, A. H., “A proposed model for data warehouse ETL processes”, *Journal of King Saud University-Computer and Information Sciences*, v. 23, n. 2, pp. 91–104, 2011.
- [12] KIMBALL, R., ROSS, M., THORNTHWAITE, W., *et al.*, *The data warehouse lifecycle toolkit*. John Wiley & Sons, 2008.
- [13] DIAS, F. M. C., “Multilingual automated text anonymization”, *Instituto Superior Técnico of Lisboa*, , 2016.
- [14] NARAYANAN, A., SHMATIKOV, V., “Robust de-anonymization of large sparse datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE, 2008.
- [15] MALIN, B., SWEENEY, L., “How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems”, *Journal of biomedical informatics*, v. 37, n. 3, pp. 179–192, 2004.
- [16] OHM, P., “Broken promises of privacy: Responding to the surprising failure of anonymization”, *UCLA l. Rev.*, v. 57, pp. 1701, 2009.
- [17] WARNER, S. L., “Randomized response: A survey technique for eliminating evasive answer bias”, *Journal of the American Statistical Association*, v. 60, n. 309, pp. 63–69, 1965.
- [18] AGRAWAL, R., SRIKANT, R., “Privacy-preserving data mining”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 439–450, 2000.
- [19] DWORK, C., ROTH, A., OTHERS, “The algorithmic foundations of differential privacy”, *Foundations and Trends® in Theoretical Computer Science*, v. 9, n. 3–4, pp. 211–407, 2014.
- [20] DWORK, C., “Differential privacy: A survey of results”. In: *International conference on theory and applications of models of computation*, pp. 1–19, Springer, 2008.

- [21] DIVANTELTD, “Repositório Anonymizer”, <https://github.com/DivanteLtd/anonymizer>, (Acesso em 16 de Maio de 2020).
- [22] THOUGHTWORKS, “Repositório Data::Anonymization”, [github.com/sunitparekh/data-anonymization/](https://github.com/sunitparekh/data-anonymization/), (Acesso em 16 de Maio de 2020).
- [23] RUBY ON RAILS, *Active Record Basics*. (Acesso em 17 de Maio de 2020).
- [24] PRASSER, F., EICHER, J., SPENGLER, H., *et al.*, “Flexible data anonymization using ARX—Current status and challenges ahead”, *Software: Practice and Experience*, , 2020.
- [25] MICROSOFT, “Repositório Microsoft Presidio”, <https://github.com/microsoft/presidio>, (Acesso em 16 de Maio de 2020).
- [26] IWANIR, E., “Fix surrounding context for unsupported languages”, <https://github.com/microsoft/presidio/issues/303>, (Acesso em 15 de Junho de 2020).
- [27] AI, E., “Available pretrained statistical models for Portuguese”, <https://spacy.io/models/pt>, (Acesso em 15 de Junho de 2020).
- [28] MERKEL, D., “Docker: lightweight linux containers for consistent development and deployment”, *Linux journal*, v. 2014, n. 239, pp. 2, 2014.