Universidade Federal
do Rio de Janeiro
Escola Politécnica

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 2023

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 POLITÉCNICA 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 xxxx

Rio de Janeiro

Fevereiro de 2023

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. Excetuam-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizaçõees 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 violaçõo 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

DEDICATÓRIA

TODO

AGRADECIMENTO

TODO

RESUMO

Inserir o resumo do seu trabalho aqui. O objetivo é apresentar ao pretenso 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

1 Introdução **1**

1.1 Tema 1

1.2 Delimitação 1

1.3 Justificativa 2

1.4 Objetivos 2

1.5 Metodologia 3

1.6 Descrição 3

2 Fundamentação **4**

2.1 Privacidade 4

2.1.1 Privacidade e o Indivíduo 4

2.1.2 A Sensibilização sobre privacidade 5

2.1.3 A *General Data Protection Regulation* 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 Registros, Atributos e Tipos 9

2.2.3 Tipos 9

2.2.4 Consulta 10

2.2.5 OLTP e OLAP 10

2.2.6 *Data Warehouses* 11

2.2.7 Ferramentas e Processos de ETL 11

2.3 Anonimização e Reidentificação 12

2.3.1 Anonimização 12

2.3.2 Reidentificação, Identificadores e Quase Identificadores 13

2.4 Técnicas de Anonimização 14
2.4.1 Resposta Aleatorizada 15
2.4.2 Privacidade Diferencial 16
2.4.3 Análise de Risco de Reidentificação 17
2.4.4 k-anonimato 17
2.4.5 l-diversidade 18
2.4.6 (n,t)-proximidade 18
2.4.7 δ-presença 19

3 Metodologia **20**
3.1 Softwares de Anonimização 20
3.1.1 Anonymizer 20
3.1.2 Data::Anonymization 21
3.1.3 ARX - Open Source Data Anonymization Software 22
3.1.4 Presidio 23
3.2 Extração de Tipos Semânticos 25
3.2.1 Busca em Dicionário 25
3.2.2 Expressões Regulares 26
3.2.3 SATO 26

4 Implementação **27**
4.1 Requisitos e Limitações 27
4.2 Tecnologias Utilizadas 28
4.3 Componentes 29
4.3.1 Analisador de Banco de Dados 29
4.3.2 Gerador de Dados 31

5 Validação **33**
5.1 Dados de Teste 33
5.2 Resultados 33

6 Conclusão **34**

Bibliografia **35**

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, além de Direito Digital e Privacidade.

Assim, este projeto implementa uma nova ferramenta de anonimização e dessensibilização de código aberto.

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 per-

mitam compatibilizar um banco de dados relacional contendo informações sensíveis a uso por terceiros para o propósito do desenvolvimento de aplicações. As implementações destas técnicas, quando existentes, serão estudadas a partir de softwares de código aberto.

1.3 Justificativa

A sociedade atual existe num momento em que a conectividade é amplamente difundida e contínua aonde um indíviduo mantém-se constantemente conectado a outros individuos e instituições através de serviços digitais.

Esses serviços digitais são acessíveis via internet e acumulam um grande volume de dados sobre seus usuários. Ações indevidas habilitadas por estes dados podem trazer prejuizos para os indivíduos e para a sociedade como um todo.

O risco trazido pela possibilidade de um abuso desses dados levou a criação de diversas regulações ao redor do mundo. A adaptação de instituições a essas regulações protege a sociedade mas gera uma complexidade adicional em projetos de engenharia de software, que passam a estar sob exigências mais rígidas sobre o acesso a esses dados.

Assim, surge a necessidade de ferramentas que auxiliem no cumprimento dessas regulações, protejam a privacidade dos usuários e desonerem o processo de desenvolvimento de software.

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 utiliza-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 vai houve validação, etc. Em resumo, você de 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 se dedica a apresentar a interseção do assuntos da engenharia e da privacidade de dados, trazendo um panorama das definições e técnicas existentes. Além disso, o capítulo também discute ferramentas existentes, limitadas a aquelas cuja licença garante a abertura do código-fonte, com o intuito de prospectar as funcionalidades fornecidas por esses programas e possíveis cenários de aplicação.

O capítulo 4 propõe um software a ser construído e define as suas especificações, mostrando as decisões técnicas tomadas, a arquitetura e explicando seus diferentes componentes.

O capítulo 5 explica como foi realizado o processo de validação do software construído e demonstra seus casos de uso.

O capítulo 6 apresenta a conclusão deste trabalho.

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 aceitaveis 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 individuos 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 forncida por este dispositivo legal é aplicável a individuos 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 inviduos pertencentes a União Europeia por instituições não pertencentes para o oferencimento 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 diretivas 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 individuos 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 requere-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 responsaveis 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 Registros, Atributos e Tipos

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 estruturada dentro de um conjunto de dados. Alguns destes campos podem permitir a associação entre registros em diferentes bancos de dados ao possuírem valores únicos, implicando em risco de privacidade.

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

2.2.3 Tipos

Atributos podem especificar tipos de dados. Tipos definem como a informação é armazenada e quais processamentos podem ser realizados.

De acordo com Zhang[8] e Donahue[9], os tipos de dados para uma coluna de banco de dados podem ser atômicos ou semânticos.

2.2.3.1 Tipos Atômicos

Tipos atômicos são aqueles que definem o formato de armazenamento do dado no banco de dados: Inteiros, cadeias de caracteres (string), variáveis booleanas são exemplos de tipos atômicos. Esses tipos não dependem da existência e do contexto do dado para sua definição e em alguns casos podem ser obtidos através dos metadados do registro.

2.2.3.2 Tipos Semânticos

Tipos semânticos dependem do contexto e do conteúdo dos dados, podendo ser extraidos a partir do conteúdo com técnicas como procura em dicionário, expressões regulares, aprendizado profundo[10] e reconhecimento de entidades nomeadas.

No exemplo anterior, a coluna nome pode ser representada em um banco de dados relacional pelo tipo atômico cadeia de caracteres (string) e seu tipo semântico pode ser inferido como nome completo ou pessoa.

2.2.4 Consulta

Uma consulta é uma função matemática aplicável ao banco de dados que permite obter 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 escrever consultas. Estes são popularmente conhecidos como bancos de dados SQL.

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

2.2.5 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).

Inicalmente, os mesmos bancos de dados eram utilizados para ambos os propósitos. A tendência em separar os sitemas OLTP dos sistemas OLAP deu origem as *Data Warehouse*s.[11]

2.2.6 *Data Warehouse*s

De acordo com Ali El-Sappagh et al. [12], 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.

2.2.7 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. [13] 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 proximo passo do processo de ETL é a transformação. Este passo tem o objetivo de conformar o conjunto de dados de entrada as 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 contém dados pessoais como CPF, nome e telefone. Estes permitem atribuir um registro a um individuo 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 caracteristicas 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[14] sobre o conjunto de dados através de substuições, agregações e supressões de registros e seus atributos com o intuito final de gerar um novo conjunto de dados cuja possibilidade de divulgação deve mitigar o risco para os individuos e instituições envolvidos.

2.3.2 Reidentificação, Identificadores e Quase Identificadores

Intuitivamente, os atributos que levam a associação direta de um individuo 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.

Suponha que uma farmácia 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 de um consultório médico, considerando um agente terceiro que possui acesso a ambos os bancos de dados, da farmácia e do consultório médico, é possível realizar a junção dos dados dos dois registros correlacionando as informações clínicas e de consumo de medicamentos dos clientes.

Atributos que podem ser utilizados para identificação exclusiva de um indivíduo são chamados Identificadores. Já atributos que podem ser utilizados para identificar um indivíduo quando cruzados com outros dados são chamados Quase Identificadores.

Em 2008 Arvind Narayanan e Vitaly Shmatikov demonstraram uma metologia robusta[15] 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[16] 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 implicitas, possivelmente violando a privacidade dos individuos 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 associaveis a um único individuo.

Paul Ohm[17] 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 Resposta Aleatorizada

A resposta aleatorizada é uma técnica proposta por Warner[18] 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. Em outras palavras, torna possível identificar a proporção de indivíduos com determinada caracteristica em um grupo de estudo, como por exemplo consumo de alcool e drogas ilícitas[19], sem permitir atribuição da caracteristica a um único indivíduo e permite obter dados sobre comportamentos socialmente inaceitaveis sem expor a identidade do participante em uma pesquisa.

Para um atributo binário, esta técnica pode ser exemplificada a partir do seguinte algoritmo, como mostrado por Warner: Um pesquisador solicita que um volutário pesquisado jogue uma moeda sem que o pesquisador veja seu resultado. Caso o resultado deste arremesso seja cara, o volutário joga novamente a moeda e responde 0 ou 1 de acordo com o resultado da moeda. Caso o arremesso inicial seja coroa, o voluntário responde a resposta verdadeira. Desta forma, não é possível saber se o voluntário pertence ou não a um grupo porém é possível extrair informações relativas ao grupo a partir da agregação dos resultados de vários voluentários.

A aplicação computacional desta técnica foi demonstrada por Agrawal e Srikant[20] 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.

Uma outra implementação foi demonstrada por Erlingsson et al[21], coletando estatísticas de uso do navegador Google Chrome com garantias de privacidade dos usuários.

Nota-se que este tipo de algoritmo pode ser implementado diretamente no ponto de coleta de dados, permitindo que somente dados cuja privacidade já está garantida sem a necessidade de processamento adicional sejam armazenados.

2.4.2 Privacidade Diferencial

De acordo com Dwork[22], a privacidade diferencial é uma definição e não um algoritmo específico. Isso implica na existência de uma classe de algoritmos diferencialmente privados satisfazendo uma condição determinada.

Um algoritmo deste tipo deve respeitar a definição matemática de privacidade diferencial[23] conforme demonstrada a seguir.

Seja X um banco de dados contendo n registros, cada um pertencente a um único indivíduo. Seja $\mathbb{D}$ o conjunto formado por todos os possíveis bancos de dados. Seja $\mathbb{Q}$ o conjunto formado por todas as possíveis consultas. Existe um algortimo M, referido como mecanismo, que tem como argumentos um banco de dados X e opcionalmente uma consulta q, que é uma função aplicada ao conjunto de dados.

Seja também a diferença simétrica entre dois bancos de dados X e X' denotadas por $X \ominus X'$, equivalente ao conjunto de registros que aparece em um banco ou outro mas não em ambos (análoga a operação lógica XOR). Considera-se neste caso. Finalmente, seja ε o fator de privacidade. Então:

Definição 1 *[23] Um mecanismo satisfaz privacidade ε-diferencial se para todo par de bancos de dados X,X' e para todo subconjunto $S \subseteq Range(M)$ e toda consulta $q \in \mathbb{Q}$(caso aplicável)*

$$\frac{\Pr\left[M\left(X, q\right) \in S\right]}{\Pr\left[M\left(X', q\right) \in S\right]} = e^{\varepsilon |X \ominus X'|}$$

Em outras palavras, um mecanismo é dito diferencialmente privado se a presença ou ausência de um registro no conjunto de entrada causa uma alteração mínima nas probabilidades dos resultados deste mecanismo.

Com relação ao fator ε, é possível encontrar seu valor variando entre 0.01 e 7 na literatura, ocasionalmente definido de forma arbitrária[24].

Outra vantagem desta classe de algoritmos é que pela própria definição, não é necessário saber quais informações adicionais um atacante possui para construir um modelo de segurança dos dados. Ou seja, há uma garantia de pior caso: Mesmo que

um atacante possua acesso a todas os registros menos um único, não será possível inferir a informação que falta.

A definição também implica no fato de que pode-se compor um algoritmo com uma garantia de privacidade específica a partir da associação de algoritmos com garantia de privacidade menor e conhecida[25].

Desta forma, a privacidade diferencial fornece um arcabouço para avaliação quantitativa da garantia de privacidade provida por um algoritmo e pode ser usada como uma ferramenta de projeto[26].

2.4.3 Análise de Risco de Reidentificação

A análise de risco de reidentificação é o processamento realizado sobre um conjunto de dados para inferir quais atributos admitem um risco de privacidade e podem ser utilizados para recuperar informações sobre um indivíduo neste conjunto.

Em contraste a algoritmos de privacidade diferencial, que podem ser aplicados diretamente na coleta de dados em alguns casos, os critérios de privacidade utilizados para análise de risco geralmente consideram um conjunto de dados pré-existente.

Assim, algoritmos que almejam atingir esses critérios geralmente operam sobre a agregação de valores, sendo especialmente úteis para o lançamento de dados com privacidade protegida como dados médicos e resultados de pesquisa.

2.4.4 k-anonimato

k-anonimato é um critério de privacidade para um conjunto de dados definido por Sweeney[27].

Um conjunto de dados satisfaz k-anonimato para $k > 1$ se para cada combinação de quase indentificadores, existirem ao menos k registros[28].

É possível encontrar a implementação deste critério de privacidade como a generalização de um atributo específico, por exemplo, substituindo o endereço completo pelo bairro ou a idade pela faixa etária.

Nestes casos, observa-se que a adequação ao critério de privacidade compromete o tipo semântico e possívelmente o tipo atômico do dado.

2.4.5 l-diversidade

A l-diversidade[29] é um critério de privacidade que visa mitigar dois problemas encontrados no k-anonimato: a possibilidade de um atacante descobrir o valor de atributos que não possuem uma diversidade significativa e a possibilidade de um atacante possuir conhecimento agregável externo ao conjunto de dados atacado.

Para satisfazer l-diversidade, todos as combinações de quase identificadores devem possuir cada uma l valores diversos para seus atributos sensíveis.

Como uma extensão do k-anonimato, l-diversidade também pode implicar na generalização de atributos e na disrupção dos tipos.

A l-diversidade possui duas vulnerabilidades[30] que podem ser exploradas por um atacante para obter informações privadas: Quando a distribuição geral possui uma distorção(*skewness*) ou quando a diversidade se dá através de informações semanticamente equivalentes, por exemplo, sinônimos da mesma palavra como valor de atributo.

2.4.6 (n,t)-proximidade

A (n,t)-proximidade[30] é um critério de privacidade criado com o intuito de resolver as falhas da l-diversidade e do k-anonimato.

Sejam classes de equivalência neste contexto, quaisquer grupos de registros que incluam uma combinação única em si de multiplos quase identificadores. Para satisfazer este critério de privacidade, a distribuição de qualquer classe de equivalência deve ser próxima a distribuição de uma classe de equivalência suficientemente grande.

Quantitativamente, a distância entre as distribuições não deve ultrapassar o limiar t enquanto cada classe de equivalência deve conter ao menos n registros.

2.4.7 δ-presença

A δ-presença[31] é uma métrica de privacidade criada para cenários nos quias o conhecimento da presença de um indivíduo em um conjunto de registros implica em um risco de privacidade.

O parâmetro δ representa a máxima certeza que se pode ter sobre a presença de um individuo no conjunto de dados anonimizado. Convenientemente, este parâmetro também representa o risco na divulgação do conjunto de dados.

Os conjuntos de dados anonimizados por esta técnica apresentam uma distorção menor que aqueles processados com objetivo de k-anonimato para as mesmas garantias de privacidade.

Capítulo 3

Metodologia

Este capítulo apresenta o resultado da prospecção das ferramentas de código aberto existentes com o intuito de entender as capacidades fornecidas, determinar os requisitos necessários a uma nova solução além de investigar possíveis componentes a serem reutilizados na solução final.

3.1 Softwares de Anonimização

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. Foi considerada a primeira linha dos arquivos README dos repositórios como nome do projeto.

3.1.1 Anonymizer

Anonymizer[32] foi desenvolvida na linguagem de programação Ruby pela companhia européia 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[33] é uma solução criada pela ToughtWorks 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[34], 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 extender 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[35] explicam que o seu objetivo é produzir um software livre com alto grau de automação que fornece uma gama variada de tecnicas de anonimização. Várias destas técnicas estão descritas em publicações especificas 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 típicamente 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 númericos 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[36] 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 instituicionais.

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 amazenamento 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 lingua inglesa são suportados.[37] Dentre as modificações necessárias

para utilização da ferramenta no contexto nacional está a utilização de um modelo adequado com suporte a lingua portuguesa. Este modelo está disponível no site da biblioteca[38], 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 entre diferentes serviços através de *circuit-breaking*, tentativas, balanceamento de carga, dentre outras técnicas. O uso de containers diminui a ocorrencia de problemas no gerenciamento de dependências[39] uma vez que estas e suas respectivas configurações estarão inclusas nas imagens utilizadas para montagem, reduzindo o esforço de implementação.

3.2 Extração de Tipos Semânticos

A extração de tipos semânticos é o processo que possibilita extrair o significado de cada coluna em um conjunto de registros. Nesta sessão, são mostradas técnicas e softwares que permitem realizar esta tarefa.

3.2.1 Busca em Dicionário

A busca em dicionário para extração de tipo semântico consiste em contar as ocorrências de palavras de vários conjuntos em uma coluna.

Cada conjunto é formado pelos possíveis valores de cada tipo semântico a ser detectado por essa técnica. Após a comparação, observa-se qual dos conjuntos possui a maior quantidade de palavras na coluna trabalhada.

Também é possível estabelecer um limiar mínimo aceitável na contagem de palavras do conjunto na coluna no qual para valores menores que este, o tipo semântico

da coluna é considerado não detectado, sendo necessário um processamento mais adequado ao conjunto de dados da coluna.

3.2.2 Expressões Regulares

Uma expressão regular é uma cadeia de caracteres que representa um padrão textual[40], podendo ser usada para comparação, busca e verificação de cadeias de caracteres em geral.

É possível utilizar expressões regulares a partir de bibliotecas inclusas em diversas linguagens de programação como Python, C#, Perl, dentre outras.

Tipos semânticos que seguem um padrão único, como por exemplo um número de CPF, podem ser detectados a partir de expressões regulares.

3.2.3 SATO

SATO[8] é um modelo de aprendizado de máquina que realiza predição de tipos semanticos.

A abordagem apresentada por esta solução, além de permitir a predição a partir de uma única coluna, também permite utilizar colunas vizinhas como contexto local ou outras tabelas como contexto global, melhorando a acurácia da predição.

A implementação do SATO foi criada a partir do Sherlock[10], que utiliza aprendizado profundo para detecção de tipos semanticos, e a expande para incluir a influência do contexto.

Capítulo 4

Implementação

Este capítulo dedica-se a aprofundar a discussão sobre a solução criada, suas premissas, restrições e motivações. O capitulo se encerra com o detalhamento da solução e de seus componentes.

4.1 Requisitos e Limitações

A principal premissa a ser levada em consideração para implementação é que a estrutura dos dados deve ser conservada. Isso significa que as tabelas, as colunas, seus tipos e suas restrições relacionais, como chaves primárias e estrangeiras, deverão manter a compatibilidade com o banco de dados de entrada.

Isso reduz a gama de técnicas de anonimização passiveis de utilização neste trabalho, pois algumas destas dependem de operações que alteram os tipos semânticos dos dados. Por exemplo, a já apresentada anteriormente, substituição de uma idade específica (27 anos) por uma faixa etária (20-30 anos) para atingir k-anonimização.

Para colunas cujos valores são numéricos, este trabalho abre mão da fidelidade estatística por preferência a simplicidade de implementação. Colunas numéricas serão anonimizadas a partir da substituição por uma distribuição uniforme.

Deste modo, a ferramenta de anonimização projetada ao longo deste trabalho visa prioritariamente cumprir o caso de uso em que um terceiro deve ter acesso à um banco de dados análogo, preenchido com dados gerados, com mesmo sentido

semantico. Mantém-se um subconjunto da utilidade original enquanto os dados são dessensibilizados, assim como se é esperado de um sistema de anonimização.

Ainda com intuito de simplificar a implementação, ao usar componentes cuja base é aprendizado de máquina, este trabalho se restringe a utilizar modelos prontos ao invés de treinar seus próprios modelos.

Com a intenção de criar uma solução extensível e reutilizável, optou-se por desenvolvimento dentro das premissas adequadas para software livre, de código aberto e disponibilizado na plataforma GitHub.

4.2 Tecnologias Utilizadas

O sistema implementado é capaz de gerar um banco de dados relacional anonimizado a partir de um banco de dados original, a partir da substituição dos dados reais por dados gerados, com mesmos tipos e estrutura.

A escolha de tecnologias para implementação tomou como critério a existência de bibliotecas para realizar as manipulações de dados previstas, facilidade de reuso, disponibilidade de documentação pública e também a familiaridade do autor.

Este trabalho utilizou Python como linguagem de programação principal tendo a biblioteca SQLAlchemy como provedor de acesso ao banco de dados para compatibilidade com multiplos Sistemas de Gerenciamento de Banco de Dados. No decorrer do processo de desenvolvimento, optou-se por restringir a implementação inicial existente em um único sistema de banco de dados relacional, especificamente Postgres.

Visando-se a expansibilidade futura, o código foi implementado de forma a definir pontos de extensão que permitem à um usuário interessado adicionar compatibilidade com outros SGBDs.

4.3 Componentes

Os componentes a seguir são os blocos lógicos do software e representam abstrações de alto nível das partes do sistema, visando facilitar o entendimento da solução como um todo.

Cada componente realiza operações encadeadas dentro de um fluxo único de transformação dos dados que se inicia com a leitura do banco de dados que se deseja disponibilizar de maneira a garantir a privacidade e termina com a população de um novo banco de dados. A seguir está a descrição ordenada de cada componente:

4.3.1 Analisador de Banco de Dados

O analisador de banco de dados é o componente responsável pelo inicio do fluxo de processamento da aplicação, obtendo as informações sobre o banco de dados e seus metadados a serem armazenadas em uma estrutura de dados comum que será utilizada posteriormente.

Essa estrutura de dados é serializada como um arquivo de texto no formato JSON e persistida em disco. Isso permite que um passo seguinte do fluxo de processamento possa reutilizar o resultado de uma execução prévia deste componente, removendo a necessidade de um novo acesso ao banco de dados de origem.

O analisador de banco de dados inclui o analisador de estrutura e o analisador de tipos semânticos.

4.3.1.1 Analisador de Estrutura

É o subcomponente do analisador de banco de dados responsável por capturar o desenho do banco de dados original, trazendo para o fluxo de dados suas colunas com os respectivos tipos atômicos e tabelas, além de amostras e estatísticas dos dados, a partir da realização das consultas necessarias para obtenção destas informações.

Este componente deve ser disponibilizado em uma implementação própria para cada dialeto de banco de dados a ser suportado.

Isso é uma restrição causada pelo fato de que cada dialeto de banco de dados possui o seus próprios detalhes de implementação para além da especificação SQL e nem todos os comandos são universalmente compatíveis.

4.3.1.2 Analisador de Tipos Semânticos

O analisador de tipos semânticos é o componente responsável por obter o significado de uma coluna dentro do contexto das informações presentes no banco de dados a partir de suas amostras e é aplicado em colunas textuais para permitir a futura geração de dados semanticamente coerentes com o banco de dados original.

Este processo pode ser feito a partir das tecnicas previamente apresentadas como busca em dicionário e expressão regulares. Porém, estas tecnicas apresentam limitações que as tornam inconvenientes para obtenção de alguns tipos semânticos importantes como endereço, etc.

Como uma tentativa de contornar este problema e melhorar a inferência de tipos semânticos, este componente utiliza uma solução externa criada a partir de técnicas de aprendizado de máquina.

Este sistema externo, chamado SATO[8], é integrada à solução final a partir de chamadas HTTP. Para cada tabela do banco de origem, o analisador de tipos semânticos cria uma amostra no formato CSV contendo um subconjunto dos seus registros e a envia para a rota de classficação do SATO. Utilizou-se um modelo pré-treinado, incluido no repositório do SATO, chamado Type78.

Desta forma, é possível obter o tipo semântico de colunas que não poderiam ser processadas apenas com técnicas deterministicas.

A classificação semântica de cada coluna pode então ser utilizada em etapas seguintes do fluxo para geração de dados análogos.

4.3.1.3 Extrator Numérico

Para colunas de tipos numéricos, há um processamento que visa substituir o conjunto de dados originais com um conjunto cuja distribuição é análoga.

Para isso, coletam-se amostras do conjunto original. Essas amostras são ordenadas. Um número aleatório é sorteado de uma distribuição uniforme entre 0 e 1 e multiplicado pelo número de amostras.

A parte inteira deste número sorteado e escalado é utilizada como índice para acessar uma das amostras ordenadas do conjunto original.

Realiza-se então a interpolação linear dessa amostra e da amostra seguinte, com pesos proporcionais a parte fracionária do número índice. Essa operação gera um novo número que é armazenado em um vetor.

Esse processo é repetido até gerar um novo conjunto de números que caracteriza o conjunto numérico original. Quando ocorre a geração de dados, este novo conjunto passa pelo mesmo processo para ser expandido e gerar o conteúdo de uma coluna numérica anonimizada.

Um estudo aprofundado visando discutir valores ótimos para o tamanho do conjunto caracterizador e também das alterações nas funções estatísticas está fora do escopo deste trabalho e será deixada para eventuais continuações.

4.3.2 Gerador de Dados

O Gerador de dados é o componente responsável pela geração da carga de saída do fluxo de processamento.

Este componente requer como entrada o arquivo contendo as especificações geradas pelos componentes anteriores e utiliza como parâmetro para geração da estrutura e do conteúdo de um novo banco de dados.

O primeiro passo do fluxo de geração de dados é a criação das tabelas e suas relações, incluindo chaves primárias e extrangeiras.

Então, gera-se o conteúdo das tabelas. Para colunas que não possuem um elo com outra tabela, ou seja, que não são chaves estrangeiras, o dado é gerado utilizando um gerador selecionado a partir dos seus tipos atômico e semantico.

As colunas cujo dado é oriundo de outra tabela a partir de uma relação têm sua criação postergada até que a coluna de origem tenha dados gerados. Os dados gerados da origem são selecionados aleatóriamente para a coluna de destino.

Capítulo 5

Validação

5.1 Dados de Teste

Para validação do sistema implementado, utilizou-se uma versão para Postgres do banco de dados Northwind[41].

O modelo deste banco de dados foi inicialmente criado pela Microsoft para fins educacionais e representa a estrutura de uma loja, incluindo pedidos, clientes e funcionários. Pode-se considerar que os dados de um cliente, como por exemplo o endereço de seus pedidos, são PII e requerem anonimização.

O banco de dados foi implantado através de um container Docker.

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] ZHANG, D., SUHARA, Y., LI, J., *et al.*, "Sato: Contextual semantic type detection in tables", *arXiv preprint arXiv:1911.06311*, , 2019.

[9] DONAHUE, J., "On the semantics of "data type"", *SIAM Journal on Computing*, v. 8, n. 4, pp. 546–560, 1979.

[10] HULSEBOS, M., HU, K., BAKKER, M., *et al.*, "Sherlock: A deep learning approach to semantic data type detection". In: *Proceedings of the 25th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1500–1508, 2019.

[11] KLEPPMANN, M., *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. "O'Reilly Media, Inc.", 2017.

[12] 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.

[13] KIMBALL, R., ROSS, M., THORNTHWAITE, W., *et al.*, *The data warehouse lifecycle toolkit*. John Wiley & Sons, 2008.

[14] DIAS, F. M. C., "Multilingual automated text anonymization", *Instituto Superior Técnico of Lisboa*, , 2016.

[15] 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.

[16] 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.

[17] OHM, P., "Broken promises of privacy: Responding to the surprising failure of anonymization", *UCLA l. Rev.*, v. 57, pp. 1701, 2009.

[18] 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.

[19] DÁVILA, O. L. S., TICERÁN, D. G., ROJAS, A. M. C., *et al.*, "Modelo de resposta aleatorizada: aplicação do modelo de Simmons", *Rev. Bras. Biom*, v. 28, n. 4, pp. 43–51, 2010.

[20] 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.

[21] ERLINGSSON, Ú., PIHUR, V., KOROLOVA, A., "Rappor: Randomized aggregatable privacy-preserving ordinal response". In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

[22] DWORK, C., "Differential privacy: A survey of results". In: *International conference on theory and applications of models of computation*, pp. 1–19, Springer, 2008.

[23] DWORK, C., MCSHERRY, F., NISSIM, K., *et al.*, "Differential privacy—a primer for the perplexed,"", *Joint UNECE/Eurostat work session on statistical data confidentiality*, v. 11, 2011.

[24] HSU, J., GABOARDI, M., HAEBERLEN, A., *et al.*, "Differential privacy: An economic method for choosing epsilon". In: *2014 IEEE 27th Computer Security Foundations Symposium*, pp. 398–410, IEEE, 2014.

[25] CUMMINGS, R., KREHBIEL, S., LAI, K. A., *et al.*, "Differential privacy for growing databases", *arXiv preprint arXiv:1803.06416*, , 2018.

[26] MCSHERRY, F., TALWAR, K., "Mechanism design via differential privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103, IEEE, 2007.

[27] SWEENEY, L., "k-anonymity: A model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, v. 10, n. 05, pp. 557–570, 2002.

[28] DOMINGO-FERRER, J., TORRA, V., "A critique of k-anonymity and some of its enhancements". In: *2008 Third International Conference on Availability, Reliability and Security*, pp. 990–993, IEEE, 2008.

[29] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J., *et al.*, "l-diversity: Privacy beyond k-anonymity", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, v. 1, n. 1, pp. 3–es, 2007.

[30] PRAKASH, M., SINGARAVEL, G., "A new model for privacy preserving sensitive Data Mining". In: *2012 Third International Conference on Computing,*

Communication and Networking Technologies (ICCCNT'12), pp. 1–8, IEEE, 2012.

[31] NERGIZ, M. E., ATZORI, M., CLIFTON, C., "Hiding the presence of individuals from shared databases". In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 665–676, 2007.

[32] DIVANTELTD, "Repositório Anonymizer", https://github.com/DivanteLtd/anonymizer, (Acesso em 16 de Maio de 2020).

[33] THOUGHTWORKS, "Repositório Data::Anonymization", github.com/sunitparekh/data-anonymization/, (Acesso em 16 de Maio de 2020).

[34] RUBY ON RAILS, *Active Record Basics*. (Acesso em 17 de Maio de 2020).

[35] PRASSER, F., EICHER, J., SPENGLER, H., *et al.*, "Flexible data anonymization using ARX—Current status and challenges ahead", *Software: Practice and Experience*, , 2020.

[36] MICROSOFT, "Repositório Microsoft Presidio", https://github.com/microsoft/presidio, (Acesso em 16 de Maio de 2020).

[37] IWANIR, E., "Fix surrounding context for unsupported languages", https://github.com/microsoft/presidio/issues/303, (Acesso em 15 de Junho de 2020).

[38] AI, E., "Available pretrained statistical models for Portuguese", https://spacy.io/models/pt, (Acesso em 15 de Junho de 2020).

[39] MERKEL, D., "Docker: lightweight linux containers for consistent development and deployment", *Linux journal*, v. 2014, n. 239, pp. 2, 2014.

[40] GOYVAERTS, J., LEVITHAN, S., *Regular expressions cookbook*. O'reilly, 2012.

[41] PTHOM, "northwind_psql", github.com/pthom/northwind_psql, (Acesso em 29 de Janeiro de 2023).