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 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 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

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. 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

Opcional.

AGRADECIMENTO

Sempre haverá. Se não estiver inspirado, aqui está uma sugestão: dedico este trabalho ao povo brasileiro que contribuiu de forma significativa à minha formação e estada nesta Universidade. Este projeto é uma pequena forma de retribuir o investimento e confiança em mim depositados.

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.

ABSTRACT

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

Key-words: word, word, word.

SIGLAS

UFRJ - Universidade Federal do Rio de Janeiro

WYSIWYG - *What you see is what you get*

Sumário

Lista de Figuras

Lista de Tabelas

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 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 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 Teórica

O presente capítulo propõe-se a introduzir o leitor ao entendimento acadêmico existente sobre o conceito de privacidade, especificamente no que contempla a área de Tecnologia da Informação com foco nos pontos de interesse ao presente trabalho, principalmente com relação a anonimização e desanonimização de informações pessoais. Visando uma maior familiarização do leitor ao domínio deste trabalho, também serão introduzidos conceitos de armazenamento de dados.

2.1 Privacidade

2.1.1 A Sensibilização sobre privacidade

De acordo com [1], 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.

A ampla disponibilidade de dados privados na Internet trouxe o tema às pautas dos debates políticos mundiais, culminando na implementação de regulamentações para definir qual é o tratamento adequado a estas informações, de modo a impedir o abuso no seu uso e garantir o direito a privacidade de cada indivíduo.

2.1.2 O entendimento multifacetado de privacidade

Warren e Brandeis - "Direito à reserva de informações pessoais e da própria vida pessoal"

2.1.3 A General Data Protection Regulation europeia

2.1.4 A Lei Geral de Protecao de Dados brasileira

2.2 Bases de Dados

2.2.1 Tipos, Atributos e Registros

De maneira geral, um registro é a unidade de armazenamento de dados. 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 digital, 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 [2]. Tipos impõem restrições nas operações que podem ser realizadas nos atributos, podem dar sentido semântico [3], e definem o formato de armazenamento da informação, garantindo sua consistência.

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

2.2.2 ETL

2.3 Anonimização e Reidentificação

2.3.1 Anonimização

Por muitas vezes atributos em registros contém dados pessoais como CPF, nome, endereço e telefone. É fácil perceber que estes permitem atribuir um registro a um individuo especifico 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[4] 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 não implica em um risco para os individuos e instituições envolvidos.

2.3.2 Reidentificação

É trivial perceber que 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.

Em 2008 Arvind Narayanan e Vitaly Shmatikov demonstraram uma metologia robusta[5] que permitiu quebrar o anonimato de um conjunto de dados da Netflix. Os pesquisadores utilizaram as preferencias individuais, recomendações e registro de transações do conjunto anonimizado e dados públicos disponíveis no 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[6] em 2004 ao recuperar informações genéticas de pacientes a partir dos padrões de visitação e dados disponíveis no ambiente informacional da saúde.

2.4 Modelos de Privacidade

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.

Paul Ohm[7] 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.

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[8] 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[9] 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[10], 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[11] 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

[12]
TODO

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] 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.

[2] CARDELLI, L., WEGNER, P., "On understanding types, data abstraction, and polymorphism", *ACM Computing Surveys (CSUR)*, v. 17, n. 4, pp. 471–523, 1985.

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

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

[5] 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.

[6] 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.

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

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

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

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

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

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