DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Aplicação da Lei Geral de Proteção de Dados com Utilização de Modelos de Anonimização de Dados em Ambiente de Nuvem Pública

Juliano Rodrigues Ferreira

Brasília, 17 de abril de 2023

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

Aplicação da Lei Geral de Proteção de Dados com Utilização de Modelos de Anonimização de Dados em Ambiente de Nuvem Pública

Juliano Rodrigues Ferreira

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Professora Edna Dias Canedo, Ph.D, CIC/FT/UnB _____
Orientadora

Professor Fábio Lúcio Lopes de Mendonça, Ph.D, FT/UnB _____
Co-Orientador

Professor Laerte Peotta de Melo, Ph.D, Banco do Brasil _____
Examinador Externo

Professor João José Costa Gondim, Ph.D, CIC/FT/UnB _____
Examinador interno

FICHA CATALOGRÁFICA

FERREIRA, JULIANO RODRIGUES
Aplicação da Lei Geral de Proteção de Dados com Utilização de Modelos de Anonimização de Dados em Ambiente de Nuvem Pública [Distrito Federal] 2023.
xvi, 43 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2023).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Proteção de Dados	2. Segurança da Informação
3. Repositórios Seguros	4. LGPD
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

FERREIRA, J. R. (2023). *Aplicação da Lei Geral de Proteção de Dados com Utilização de Modelos de Anonimização de Dados em Ambiente de Nuvem Pública*. Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 43 p.

CESSÃO DE DIREITOS

Juliano Rodrigues Ferreira
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Dedico esse trabalho à minha família, pelo constante apoio e, principalmente, incentivo para que fosse possível conciliar as atividades profissionais com este importante desafio e processo tão rico de aprendizagem.

AGRADECIMENTOS

Agradeço a todos aqueles que apoiaram este processo de aprendizagem. Agradeço à instituição em que exerço minhas atividades pelo apoio institucional e por ter viabilizado em parceria com a Universidade de Brasília a realização desse trabalho. À UnB por todo o apoio durante esse processo, ao corpo docente, em especial à Professora Edna Dias Canedo, pelo suporte, cobranças e incentivos essenciais para a conclusão deste trabalho. Agradeço, também, à secretaria do programa PPEE que sempre se colocou a disposição para apoiar em todas as questões, e agradeço a Deus pela oportunidade desse período de aprendizado.

RESUMO

Este estudo pretende avaliar e aplicar técnicas de proteção de dados, considerando as diretrizes indicadas na Lei Geral de Proteção de dados (LGPD). Através de procedimentos práticos e análise de resultados buscando uma proteção adequada dessas informações com técnicas de criptografia e anonimização de dados, garantindo, além de adesão à legislação, a manutenção de performance e transparência para o usuário final. Trata-se de um desafio adicional à aplicação desse modelo de proteção de dados considerando o ambiente de nuvem pública e suas características específicas de acesso, armazenamento, gerenciamento de chaves criptográficas, manipulação de informações e de desempenho de tal ambiente.

ABSTRACT

This study aims to evaluate and apply data protection technology, considering the guidelines indicated in the General Personal Data Protection Law (LGPD). Through practical procedures and analysis of results seeking adequate protection of this information with encryption techniques and anonymization of data, ensuring, in addition to adherence with legislation, the maintenance of performance and transparency for the end user. It is an additional challenge to apply this data protection model considering the public cloud environment and its specific characteristics of access, storage, cryptographic key management, information manipulation, and performance of that environment.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A necessidade de prover e implementar mecanismos sólidos de segurança da informação sempre foi uma questão estratégica para empresas públicas e privadas [1, 2], constituindo-se como aspecto de grande relevância para as áreas de Tecnologia da Informação e Comunicação (TIC) destas organizações [3],[4].

Com o advento da Lei n º 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais (LGPD)[5], a necessidade de implementar mecanismos de segurança e privacidade ganhou ainda mais notoriedade, visto que veio estabelecer normas legais mais rígidas no tocante à segurança da informação e dos dados dos usuários manipulados [6], conduzindo a um caminho de melhores práticas, sob o manto desse novo aparato legislativo [7],[5].

A LGPD apresenta como princípio fundamental a necessidade de proteção de dados pessoais. Neste sentido, mecanismos de segurança da informação devem ser adotados tanto em sua manipulação e disponibilização, quanto em seu armazenamento [7], [8].

Para garantir a adesão às premissas da LGPD é importante uma análise adequada das informações armazenadas para reduzir os riscos envolvidos, tanto no aspecto de utilização do dado como no problema de reidentificação [9]. E nesse contexto destaca-se importância e a necessidade de uma gestão adequada dessas informações realizando uma governança de dados [9], com o objetivo de identificar adequadamente quais informações devem ser protegidas, qual técnica de anonimização aplicar e como avaliar a manutenção da capacidade de utilização dessa informação.

Com isso, recursos como tokenização de dados, criptografia, anonimização, gerenciamento de chaves utilizadas para criptografia de dados, e controle e registro de eventos de acessos de usuários passam a constituir importantes meios de proteção dessas informações [10],[11]. Estas práticas passam a ser exigidas das organizações, independentemente de onde esses dados estejam armazenados, seja em ambientes de data center, big data, container ou em nuvem [7, 5, 12]. Considerando a utilização desses recursos, é importante mencionar que a proteção de dados pessoais envolve todos aqueles que têm interesse na informação, que são os controladores do dado, processadores (usuários da informação) do dado e o indivíduo a quem o dado se refere [13].

Outra abordagem considerada para a proteção de dados é o uso da técnica de pseudonimização de dados [13], uma técnica estruturada na criptografia de dados. Os dados pseudonimizados podem ser considerados um sub conjuntos dos dados pessoais. Com a criptografia dos dados identificadores de informação pessoal (subconjunto de dados que identifica um indivíduo), e mantendo os dados auxiliares armazenados separadamente, é possível atribuir somente os dados a um indivíduo com informações adicionais ou com a chave criptográfica. Porém, nesse caso, é importante considerar que os dados pseudonimizados permitem, de alguma maneira, a reidentificação de um determinado indivíduo [13].

Neste sentido, dado o contexto da LGPD é preciso identificar e propor um modelo de proteção

de dados para viabilizar a proteção dessas informações em ambiente de nuvem, possibilitando a anonimização de dados de maneira transparente e sem redução de performance para o usuário [10, 6], garantindo o cumprimento das premissas impostas pela LGPD [5]. Quando tais dados se encontram fora do ambiente de TIC da organização, como, por exemplo, em nuvem pública, a concretização dessa proteção passa a se constituir como um desafio adicional.

1.1 PROBLEMA DE PESQUISA

O gerenciamento de informações em ambiente de TIC envolve, com frequência, a manipulação de informações pessoais. Frente aos desafios reforçados pela LGPD, uma questão importante é a garantia de que o armazenamento e o trânsito de informações ocorram de maneira segura e aderente à legislação [14]. Para tanto, é preciso definir e validar as técnicas necessárias para essa finalidade. Esta questão só poderá ser considerada adequadamente tratada com a aplicação correta de técnicas de criptografia, anonimização e pseudonimização dessas informações, este problema deve ser superado considerando a manutenção de performance e usabilidade do ambiente de TIC.

O aspecto da proteção de dados se torna ainda mais relevante ao serem considerados os impactos causados às empresas, até mesmo à Administração Pública com a possibilidade de vazamento de dados que deveriam permanecer protegidos. Existem dois aspectos importantes relacionados à proteção de dados, o primeiro é em relação à segurança informação, ao risco relacionado ao ataque à infraestrutura de TIC, que pode resultar em perda de dados, alteração dessas informações ou coleta indevida de informações [15].

O segundo aspecto é em relação à conformidade com a legislação e a normatização de melhores práticas. A proteção de dados pessoais passa a ser requisito obrigatório em muitos países em razão das legislações vigentes relacionadas à proteção de dados, como, por exemplo, no Brasil, após a LGPD entrar em vigor [5]. Desta maneira, os problemas decorrentes de não proteção de dados e de garantia da sua privacidade são de prejuízo institucional, considerando o ativo que é a informação armazenada, e, em muitos casos, o prejuízo decorrente de sanções relacionadas às legislações vigentes, como multas pecuniárias.

1.2 JUSTIFICATIVA

A necessidade de se estudar maneiras de anonimização de dados para atender premissas da Lei Geral de Proteção de Dados Pessoais é de grande importância para as organizações públicas e privadas. A documentação e a implementação de modelos seguros de anonimização de dados pessoais são resultados importantes para garantir tanto a segurança e proteção dos dados armazenados, como a possibilidade de acesso eficiente à essas informações, quando se trata de acessos legítimos [10].

A aplicação de técnicas de anonimização de dados com ênfase na LGPD está associada à área de concentração de "Segurança Cibernética", inserida na linha de pesquisa "Segurança e Inteligência Cibernética", que apresenta como uma de suas áreas de interesse a Lei Geral de Proteção de Dados e suas implicações na área cibernética.

Com o objetivo de atender as melhores práticas [16] relacionadas à proteção de dados, o processo de anonimização de dados fica caracterizado como um recurso importante de segurança e proteção dessas informações. A própria LGPD [5] recomenda a utilização dessa técnica de anonimização de informações para o tratamento de dados, principalmente, aqueles que contenham dados pessoais.

1.3 OBJETIVOS

1.3.1 Objetivo geral

O objetivo geral deste trabalho é propor um modelo de anonimização de dados pessoais com ênfase nas premissas definidas na Lei Geral de Proteção de Dados Pessoais (LGPD) para o contexto de ambiente de nuvem pública. O intuito é que esse modelo de anonimização, considerando um conjunto de dados, realize a anonimização dessas informações utilizando técnicas combinadas de anonimização como a substituição, generalização, adição de ruído, cifração e agregação. O objetivo é realizar avaliação e análise da execução dessas técnicas e seu resultado final no conjunto de dados selecionados.

1.3.2 Objetivos específicos

Para atingir o objetivo geral, definiu-se os seguintes objetivos específicos:

- Realizar uma revisão de literatura em relação às técnicas de anonimização e pseudonimização de dados existentes na literatura;
- Selecionar as técnicas de anonimização e pseudonimização de dados a serem utilizadas no contexto dessa pesquisa;
- Selecionar um subconjuntos de dados em uma base de dados hipotética;
- Realizar a aplicação das técnicas de anonimização e pseudonimização de dados selecionadas na base de dados;
- Propor um modelo de proteção de dados, em conformidade com a LGPD, que permita realizar a anonimização de dados em contexto de nuvem pública;
- Realizar a aplicação do modelo proposto em uma nuvem pública;
- Analisar a performance da solução proposta e realizar ajustes, caso necessário.

1.4 RESULTADOS ESPERADOS

Como resultado deste trabalho espera-se verificar a aplicação de técnicas de anonimização adequadas, considerando o tipo de dado pessoal identificado. Entre os resultados almejados estão:

- uma proposta de classificação de dados pessoais considerando aspectos de governança de dados, possibilitando a seleção de técnicas de anonimização adequadas;
- a anonimização de um subconjunto de dados selecionados, considerando o armazenamento dessas informações em um ambiente de nuvem pública. Sendo possível verificar a eficiência do resultado final da anonimização dessas informações;
- a apresentação de um modelo de implementação de proteção de dados pessoais para as informações a serem armazenadas em nuvem, com a identificação dos riscos envolvidos no processo;
- o modelo de implementação de proteção de dados que implique na possibilidade de utilização do dado pelo usuário, permitindo uma camada de proteção de dados pessoais, porém, com a possibilidade da utilização adequada dessas informações.

1.5 METODOLOGIA DE PESQUISA

A metodologia proposta para desenvolver este trabalho e atingir os objetivos gerais e específicos será um estudo de caso, estruturado em um protocolo de pesquisa, na análise e na produção do caso e, por fim, em seu relato, com as conclusões do trabalho. O protocolo de pesquisa incluirá a fundamentação conceitual da pesquisa, perguntas norteadoras e proposições de estudo.

Uma etapa imprescindível para o desenvolvimento do presente trabalho foi o levantamento dos dados considerados passíveis de proteção por apresentarem informações pessoais. Diante disso, definiu-se qual técnica de anonimização de dados seria melhor a ser aplicada. Esta etapa foi realizada através de seleção de amostra de dados e com utilização de soluções de TIC que auxiliem no mapeamento das informações armazenadas no ambiente. Destaca-se a importância de realizar um comparativo com as melhores práticas e buscar parâmetros de outras organizações em relação ao volume de dados protegidos quando comparado ao volume de dados totais, esse aspecto apresenta-se como relevante para entender o volume de dados passíveis de anonimização em infraestrutura de nuvem pública.

A aplicação de modelos e técnicas de anonimização de dados no ambiente de TI estruturado em nuvem pública é um dos intuitos desta pesquisa, pois esta etapa da metodologia possibilita a compreensão da qualidade da anonimização de dados, da pseudonimização, situação em que é possível a reversão da informação previamente anonimizada, e também do desempenho de acesso e manipulação das informações quando submetidas a essas técnicas de proteção de dados.

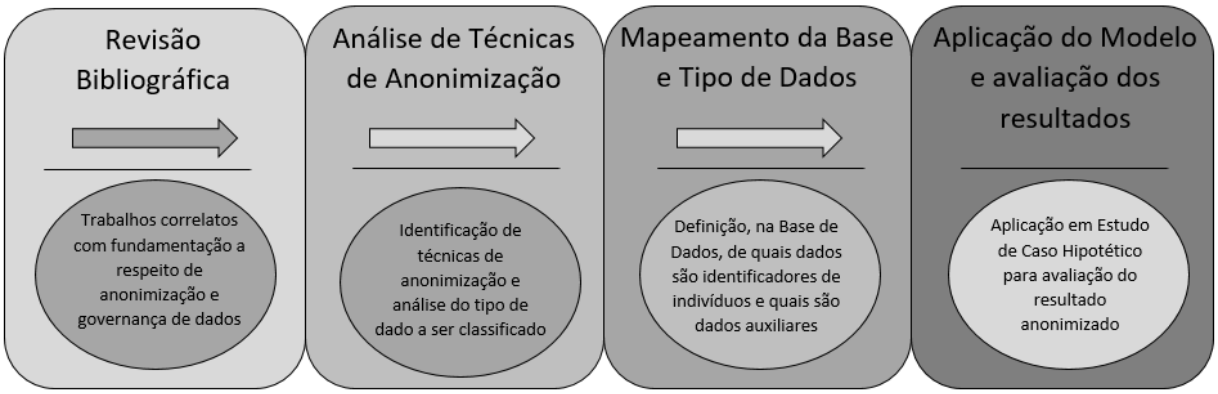

Figura 1.1: Metodologia de pesquisa adotada

Assim, como método de pesquisa, conforme apresentado na Figura 1.1, apresenta-se uma revisão bibliográfica a respeito dos princípios da proteção de dados, técnicas de anonimização, riscos envolvidos e governança de dados.

Destarte, ao revisar a teoria a respeito da classificação por tipo de dados das informações a serem protegidas e anonimizadas, apresenta-se um estudo de caso que demonstra a aplicação de técnicas de anonimização de dados em uma estrutura de dados hipotética, evidenciando o mapeamento das informações e resultado final da base de dados anonimizada, além de considerar no estudo de caso quando é adequado o uso da pseudonimização de dados e de demonstrar que, nessas situações, se trata de um processo de criptografia de informações.

1.6 PUBLICAÇÕES RESULTANTES DA PESQUISA

1. Ferreira, Juliano Rodrigues; Ribeiro, Cileno de Magalhães; Pincovscy, João Alberto; Canedo, Edna Dias; Mendonça, Fábio Lúcio Lopes. Mitigação dos Riscos à Privacidade através da Anonimização de Dados. RISTI (PORTO), v. E49, p. 573-585, 2022.

1.7 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está organizado por mais cinco capítulos relacionados ao conteúdo da pesquisa.

O capítulo dois, “Embasamento Teórico”, abordará os conceitos gerais que norteiam este trabalho, bem como os trabalhos correlatos.

O capítulo três, “Metodologia”, descreverá a metodologia de pesquisa apresentada, além dos passos necessários para melhor entendimento e as respostas ao problema de pesquisa.

O capítulo quatro, “Framework”, descreverá a configuração do estudo de caso abordado, quais os aspectos foram considerados e quais premissas foram utilizadas para o estudo de caso.

O capítulo cinco, “Análise de dados e resultados”, será referente a discussão, análise de dados e apresentação dos resultados obtidos no estudo, onde foi realizada uma avaliação considerando o problema de pesquisa, os resultados esperados e a hipótese considerada.

Por último, o capítulo seis, “Conclusão”, apresentará a conclusão para o trabalho, retomando as ideias iniciais e avaliando os resultados, sempre considerando os objetivos gerais e específicos anteriormente delimitados.

2 EMBASAMENTO TEÓRICO

Para atender o objetivo de uma proteção de dados pessoais eficiente, alguns conceitos precisam ser reforçados. A proteção de dados é o resultado final da aplicação de três importantes disciplinas: a governança de dados, a política de privacidade de dados e a aplicação de uma adequada política da informação [3][17].

A governança de dados envolve aspectos relacionados ao planejamento do fluxo de informações dentro da organização. Trata-se de uma abordagem que busca definir princípios a serem utilizados durante todo o ciclo de vida do dado dentro da organização[18]. A governança de dados é estruturada nos pilares da governança, gestão de riscos e conformidade (compliance)[7] e é a combinação desses conceitos que possibilita a definição de processos que apoia a estratégia de negócios, fazendo-a acontecer[4].

As iniciativas relacionadas à governança englobam o desenvolvimento de políticas e procedimentos e, ainda, a definição de responsabilidades e diretrizes que orientam as pessoas e os processos da organização[19]. A gestão de riscos é bastante associada à questão da segurança da informação, referindo-se à implementação de mecanismos de gerenciamento, tratamento e auditoria das informações. Ela também implica na necessidade de conhecer e gerenciar a utilização dos dados acessados [12] [20].

O aspecto de conformidade, ou compliance, refere-se aos aspectos legais, à necessidade de adesão à legislação, e não apenas com normativos legais, mas, também, com normas internas da organização e melhores práticas do mercado [14]. No contexto da governança de dados as ações de aprimoramento da segurança desses dados e da qualidade de dados são pontos essenciais. É a disciplina de governança de dados que busca agregar qualidade aos dados para produção de informações considerando os requisitos de confiabilidade e disponibilidade dessas informações[18][14].

Em relação ao tratamento de dados na organização, os dados pessoais precisam de uma atenção específica, por serem considerados dados que precisam de uma proteção adicional e um tratamento que considere as legislações que normatizam o tema, em especial a Lei Geral de Proteção de Dados Pessoais (LGPD), Lei n. 13.709, de 14 de agosto de 2018 [5]. Porém, a questão da privacidade de dados também é reforçada pela Lei n. 12527, de 18 de novembro de 2011, também conhecida como Lei de Acesso à Informação (LAI) [21] e por normas e padrões que indicam melhores práticas, como a norma ISO/IEC 27001, de outubro de 2005 [22], referente à tecnologia da informação e aos sistemas de gestão de segurança da informação, e a norma ISO/IEC 27701, de agosto de 2019 [23], referente aos requisitos necessários para um sistema de gestão de privacidade.

A norma ISO/IEC 27701 [23] foi planejada para apoiar as organização em relação ao tratamento de informações pessoais, gerenciando e reduzindo os riscos associados a esse tratamento

de informações. Por estabelecer mecanismos de controle e privacidade das informações dentro de uma organização e por ser uma regulamentação internacional que orienta a estruturação de um sistema de controle e gerenciamento de informações, trata-se se uma importante ferramenta para apoiar as organizações a se adequarem a LGPD.

A LGPD e a Norma NBR/ISO/IEC 27701 estão relacionadas uma com a outra. As duas abordagens, tanto da lei quanto a norma, reforçam diretrizes importantes de proteção de dados e gestão de risco[23]. A LGPD dedica um capítulo inteiro ao tratamento de dados pessoais ("Capítulo II - Do Tratamento de Dados Pessoais"), e normatiza as situações em que tais dados poderão ser objeto de tratamento pelas instituições. Em relação a essas situações em que é permitido o tratamento de dados pessoais, cabe destacar as seguintes possibilidades expressas na legislação[7][5]:

- **mediante o fornecimento de consentimento pelo titular**: este consentimento deve ser expresso pelo titular, considerando as informações fornecidas a respeito da finalidade de utilização das informações;
- **cumprimento de obrigação legal**: Nesse ponto entende-se a necessidade de uma previsão legal que obrigue o controlador dessas informações a utilizá-las para o cumprimento desta obrigação;
- **administração pública**: para o tratamento e uso compartilhado de dados necessários à execução de políticas públicas previstas em leis e regulamentos ou respaldadas em contratos, convênios ou instrumentos congêneres.

Dessa maneira a LGPD atua no sentido de restringir as situações em que é possível o tratamento de dados pessoais, com o objetivo de proteger o titular dessas informações, impedindo o de uso indevido destes dados sem a devida autorização. Em razão dessas restrições apontadas por esse normativo, ela LGPD, verifica-se que a regra geral é pelo não tratamento e utilização de dados pessoais pelas instituições de direito público ou privado. No entanto, um ponto apresentado por esta norma brasileira, no artigo 12, da seção II ("Do Tratamento de Dados Pessoais Sensíveis"), co capítulo II, descreve a seguinte ressalva em relação ao tratamento de dados pessoais[5]:

- **Capítulo II, Seção II, Art 12**: "Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei, salvo quando o processo de anonimização ao qual foram submetidos for revertido, utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido".

A própria legislação brasileira busca esclarecer o que pode ser considerado um "esforço razoável"no contexto desta lei no parágrafo 1, do artigo 12:

- **Esforço Razoável (Art 12, Parágrafo 1)**: "A determinação do que seja razoável deve levar em consideração fatores objetivos, tais como custo e tempo necessários para reverter o processo de anonimização, de acordo com as tecnologias disponíveis, e a utilização exclusiva de meios próprios".

Considerando essa diretriz prevista na legislação, é possível verificar a importância do processo de anonimização de dados [6] com o objetivo de adequação como a previsão normativa dessa legislação.

Em seu capítulo VII, Da segurança e das boas práticas, a LGPD apresenta na seção II, Das boas práticas e da Governança, o conceito de privacidade de dados, indicando requisitos mínimos de implementação de um programa de governança em privacidade. Com isso, a política de privacidade de dados, se tornou um importante elemento que decorre da aplicação de governança de dados e da busca por compliance com a legislação e melhores práticas. Uma importante referência para implementação da política de privacidade de dados é, conforme citado anteriormente, a norma ISO/IEC 27701 [23] que especifica requisitos e preconiza diretrizes para o estabelecimento, implementação, manutenção e melhoria contínua de uma sistema de gestão de privacidade da informação (SGPI).

A norma ISO/IEC 27701 é estruturada relatando requisitos relacionado ao SGPI que envolve requisitos para o planejamento, suporte, operação, monitoramento de desempenho e melhoria contínua dos processos relacionados a gestão de privacidade e segurança da informação. E, ainda, considerando a ISO/IEC 27701, a LGPD e a ISO/IEC 27001, é possível verificar a uma área comum entre as normas e a legislação, em relação a importância de implementação de uma política de segurança da informação sólida dentro das instituições.

A política de segurança da informação envolve a o tratamento efetivo desses dados, sua operação e manutenção. Com isso, um importante conceito merece destaque, trata-se do conceito de zero trust [24] (confiança mínima), que define que as ações dos usuários devem ocorrer com a concessão do menor privilégio possível para execução de determinada atividade.

O modelo de aplicação zero trust [24] é interessante por se tratar de uma abordagem que começa protegendo os dados, considerando que o acesso a esses dados é o objetivo final de uma tentativa de um ataque cibernético a sistemas de informação, esse conceito propõe que o modelo de segurança se inicie pela proteção do dado, e para realizar essa proteção as instituições precisam ser capazes de identificar seus dados, aonde estão armazenados, quais os privilégios de acesso, realizar o monitoramento desses ações e ser capazes de impedir acessos indevidos.

A arquitetura zero trust (ZTA) é uma proposta de um conjuntos de diretrizes para a implementação de um modelo de segurança estruturados em requisitos fundamentais de segurança[25], como a segmentação e micro-segmentação de rede, autenticação e controle de acesso, a criptografia e automatização e orquestração da segurança[26]. Porém, mesmo com a aplicação de políticas de segurança e monitoramento dos dados armazenados, quando forem dados pessoais estarão sujeitos ao que prescreve a LGPD. Os mecanismos de segurança da informação e governança de dados são requisitos adicionais que auxiliam a aderência com a legislação[20], no entanto, conforme descrito anteriormente, a anonimização desses dados seria uma alternativa para o cumprimento dos requisitos da legislação [27][18].

2.1 PRINCIPAIS TÉCNICAS DE ANONIMIZAÇÃO DE DADOS

Retomando a importância da aplicação de técnicas de anonimização de dados como instrumento para a adequada proteção de dados pessoais, verifica-se que essa alternativa é um instrumento preconizado tanto pela LGPD [8] quanto pela *General Data Protection Regulation* (GDPR) [28]. Porém, é importante ressaltar que existe um conjunto amplo de possibilidades de aplicação de diferentes técnicas de anonimização e entender o funcionamento das principais técnicas é de grande relevância para sua adequada utilização em cada situação e em cada tipo de informação a ser protegida [13].

Entre as principais técnicas de anonimização, a substituição, mascaramento, criptografia e tokenização ganham destaques por serem as mais utilizadas [13]:

- Mascaramento ou Supressão: A supressão ou mascaramento de dados é uma forma extrema de anonimato. Substitui as informações por algum valor fixo de texto pré-definido (ou em alguns casos, uma tarja preta).
- Generalização: Nessa técnica os dados são substituídos por valores de categorias mais amplas. Por exemplo: o valor 19 do campo “Idade” pode ser substituído por < "20", o valor "23"por "20« Idade < 30", ou seja, trata-se de organização em grupos de valores para evitar a associação imediata com o valor de um dado pessoal apresentado.
- Criptografia: É técnica mais amplamente conhecida de proteção de informações. Trata-se de encriptação de valores, utilizando um algoritmo criptográfico que realizado a cifração de informações utilizando uma valor constante nesse processo denominado chave criptográfica. Esse processo é considerado um tipo de pseudoanonimização [29][30], pois apesar dos dados estarem protegidos ao fim do processo, essa proteção pode ser revertida utilizando uma chave criptográfica que pode ser a mesma do processo de cifração, no caso da criptografia simétrica, ou diferente, no caso da criptografia assimétrica [31] [32]. A qualidade e complexidade do processo de criptografia para a proteção de dados é diretamente relacionado com o algoritmo de criptografia selecionado e com a complexidade e tamanho da chave criptográfica criada para essa finalidade.
- Tokenização: trata-se de uma abordagem não matemática para proteger dados em repouso que substitui dados sensíveis por substitutos não sensíveis, referidos como fichas. As fichas não têm qualquer significado ou valor extrínseco ou explorável. A Tokenização não altera o tipo ou comprimento dos dados, o que significa que pode ser processada por sistemas herdados, tais como bases de dados que podem ser sensíveis ao comprimento e tipo de dados[33].

A utilização de técnicas adequadas de anonimização de dados, de fato, auxilia no objetivo final de proteção dessas informações, porém a possibilidade de falhas nesse processo existe, possibilitado a reidentificação dessas informações o que pode tornar ineficaz o processo de proteção dessas informações [16],[13].

Dados armazenados em ambiente de nuvem pública podem adicionar riscos a esse processo, pois potencializa a possibilidade de associação de informações anonimizadas com informações de outras fontes de informação possibilitando a reidentificação dessas informações e, com isso, realizando a reversão do processo de anonimização desses dados[6],[34].

2.2 PRINCÍPIOS E ETAPAS PARA IMPLEMENTAÇÃO DA LGPD

Existem princípios relativos a LGPD que são balizadores para a governança de dados e implementação da proteção de dados. No Artigo 6º da LGPD são apresentados esses princípios e a respectiva definição, conforme segue [5],[19]:

"Art. 6º As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios:

I - finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;

II - adequação: compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;

III - necessidade: limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;

IV - livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;

V - qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;

VI - transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;

VII - segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;

VIII - prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;

IX - não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;

X - responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados

pessoais e, inclusive, da eficácia dessas medidas. "

Nesse estudo, considerando a governança e proteção de dados é importante destacar os princípios, previsto na LGPD, da qualidade dos dados, segurança e prevenção. Esses são princípios estruturante para a proteção de dados efetiva, pois indicam diretrizes para a governança de dados no caso do princípio da qualidade de dados, diretrizes para a segurança da informação e, também, direcionamentos para proteção de dados, através dos princípios da segurança e prevenção [18].

2.3 A ANONIMIZAÇÃO E PSEUDONIMIZAÇÃO DE DADOS E A LGPD

A Lei geral de proteção de dados reforça a importância de técnicas de anonimização e pseudonimização de dados para a efetiva proteção de dados e cumprimento da legislação. No Art. 12º da LGPD é expresso que “Os dados anonimizados não serão considerados dados pessoais para os fins desta Lei, salvo quando o processo de anonimização ao qual foram submetidos for revertido, utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido"[5].

Dessa maneira é possível verificar o potencial de utilização de técnicas de anonimização para minimizar o risco de não aderência com a legislação, tendo em vista que a própria lei considera que os dados pessoais após passarem pelo processo de anonimização deixam de ser dados pessoais para fins da legislação. Trata-se de uma maneira de atendimento aos requisitos da legislação de maneira mais imediata. Porém, a utilização de técnicas de anonimização de dados pessoais impacta na possibilidade de utilização dessas informações, em razão de parte da informação ser anonimizada sem possibilidade de visualização e entendimento do seu conteúdo [15],[14]. Para minimizar o prejuízo na visibilidade da informação é importante o adequado mapeamento dos dados para que a anonimização ocorra no mínimo necessário para proteção de dados pessoais [13].

Outra possibilidade prevista na Legislação [5] é a possibilidade de aplicação de técnicas de pseudonimização para a proteção de dados, nos casos em que a informação protegida só poderá ter seu processo de anonimização revertido com o auxílio de informação adicional, como a chave criptográfica por exemplo [29].

Em seu artigo 13º, a LGPD, indica a seguinte definição: "Na realização de estudos em saúde pública, os órgãos de pesquisa poderão ter acesso a bases de dados pessoais, que serão tratados exclusivamente dentro do órgão e estritamente para a finalidade de realização de estudos e pesquisas e mantidos em ambiente controlado e seguro, conforme práticas de segurança previstas em regulamento específico e que incluam, sempre que possível, a anonimização ou pseudonimização dos dados, bem como considerem os devidos padrões éticos relacionados a estudos e pesquisas.", e, ainda, segue no parágrafo 4º desse mesmo artigo com a seguinte definição: "Para os efeitos deste artigo, a pseudonimização é o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida

separadamente pelo controlador em ambiente controlado e seguro"[5],[17].

Dessa maneira a própria legislação indica uma alternativa de proteção de dados, com o objetivo de evitar a possibilidade de associação da informação a um indivíduo sem a utilização de uma informação adicional. Nesse cenário, a informação adicional citada pela legislação trata-se da chave criptográfica a ser utilizada para a reversão do processo de criptografia utilizado[30]. A LGPD ainda reforça a necessidade de que essa informação adicional seja mantida separadamente pelo controlador em ambiente seguro [20].

2.4 A CRIPTOGRAFIA E GERENCIAMENTO DE CHAVES CRIPTOGRÁFICAS

Conforme citado anteriormente, a Criptografia como técnica de pseudonimização é de grande importância em razão do potencial de proteção de dados e a possibilidade de utilização da informação, de acordo com a política de privacidade, em razão da necessidade da instituição [27].

A Criptografia, além de ser um importante instrumento de proteção de dados e apoio a aderência com a LGPD, acrescenta serviços importantes de segurança da informação ao dado manipulado [32]. Através da utilização dessa técnica serviços criptográficos importantes são associados, conforme podem ser descritos na sequência de acordo com a ISO/IEC 17799:2005[35]:

- **Confidencialidade**: Trata-se da propriedade que limita o acesso a informação tão somente às entidades legítimas, ou seja, àquelas autorizadas pelo proprietário da informação;
- **Integridade**: Se apresenta como a propriedade que garante que a informação manipulada mantenha todas as características originais estabelecidas pelo proprietário da informação, incluindo controle de mudanças e garantia do seu ciclo de vida;
- **Autenticidade**:É a propriedade que garante que a informação é proveniente, de fato, da fonte anunciada e que não foi alvo de mutações ao longo de um processo;
- **Irretratabilidade**: É referente a propriedade que garante a impossibilidade de negar a autoria em relação a uma transação anteriormente feita.

Esses serviços citados reforçam a eficiência de utilização dessa técnica para proteção de dados pessoais[32]. Porém, para utilização dessa técnica é importante, conforme preconizado pela legislação, o adequado gerenciamento das chaves criptográficas utilizadas. O gerenciamento de chaves criptográficas envolve o armazenamento adequado, a proteção dessa informação, o inventário de chaves, o controle do ciclo de vida dessa chave criptográfica, bem como o gerenciamento de utilização adequada dessa informação[31].

A criptografia (cifração) será a técnica de anonimização e proteção de dados utilizada nesse estudo[30]. O algoritmo criptográfico tem papel determinante na qualidade da proteção de dados realizada. Considerando o provedor de nuvem pública que será utilizado nesse estudo, o algoritmo

utilizado é o AES 256 bits ou AES 128 bits. O AES é uma algoritmo simétrico comum, é um padrão de criptografia avançada[32].

O AES, abreviação para Advanced Encryption Standard, é uma cifra de bloco simétrico, utilizando a mesma chave para cifração e decifração, sendo que nesse modelo de algoritmo a chave criptográfica utilizada pode ser de tamanho variado, 128, 192 ou 256 bits[36]. O tamanho da chave influencia na complexidade de reversão do texto cifrado sem a utilização das chaves utilizadas no processo.

No que se refere ao gerenciamento de chaves criptográficas, o controle do ciclo de vida dessas informações é de grande importância para a manutenção da proteção de dados. Segundo o National Institute of Standards and Technology (NIST) [37], que é uma agência governamental não regulatória da administração de tecnologia do Departamento de Comércio dos Estados Unidos, referência em padronizações associadas a Cibersegurança, o ciclo de vida de uma chave criptográfica é iniciado com a sua geração e somente finalizado com seu descarte e adequada eliminação. Passando durante esse processo pelos estados [32]:

- **geração**: criação da chave, sendo que ainda não está pronta para uso;
- **pré-ativação**: aguarda o período de utilização ou a emissão de um certificado;
- **ativada**: disponível para uso;
- **suspensa**: o uso da chave está temporariamente suspenso. Neste estado não pode mais realizar operações de cifra ou assinatura, mas pode realizar a recuperação de dados ou verificação de assinaturas;
- **inativada**: não pode ser mais utilizada para cifra ou assinatura digital, sendo mantida para o processamento de dados cifrados ou assinados antes da inativação;
- **comprometida**: a chave tem a sua segurança afetada e não pode mais ser usada em operações criptográficas.;
- **destruída**: é indica nesse estado que uma chave não é mais necessária. A destruição da chave é o estágio final e pode ser atingido devido ao fim do ciclo de uso dela ou do comprometimento de sua segurança.

O gerenciamento de chaves criptográficas envolve, ainda, a realização de cópias de segurança [3], em razão da chave criptográfica ser o único instrumento que associado ao algoritmo criptográfico utilizado pode reverter a anonimização das informações, a ausência de cópias de segurança aumenta significativamente os riscos associados ao processo, com consequências possíveis de perda total de acesso a informação [38].

2.5 A INFRAESTRUTURA DE NUVEM PÚBLICA E PROTEÇÃO DE DADOS

A Infraestrutura em nuvem é o conceito utilizado para descrever uma infraestrutura virtual de Tecnologia da Informação que pode ser acessada através diretamente da Internet ou de uma rede configurada para essa finalidade. Trata-se de um modelo de entrega de funcionalidades e infraestrutura de TI como serviço, por exemplo. [39]. A Infraestrutura em nuvem inclui servidores (capacidade de processamento), armazenamento, serviços, aplicações, rede e segurança[40]. É uma alternativa a infraestrutura de TI física e local, como um datacenter da instituição.

A possibilidade de instituições optar pela utilização de infraestrutura em nuvem ocorre em razão dos benefícios associados a esse modelo de disponibilização de serviços de TIC. A principal vantagem inicial desse modelo está relacionada a rentabilidade desse modelo, que decorre da redução de custos operacionais e da possibilidade de utilização de recursos computacionais de acordo com a real demanda por serviços de TIC [39]. Outro aspecto está associado a agilidade de implementação de serviços conforme a demanda, esse modelo permite a configuração de serviços de maneira mais ágil acarretando maior flexibilidade as organizações usuárias desse modelo de contratação [41].

Porém os aspectos mais relevantes do modelo de infraestrutura em nuvem está na possibilidade de ampliação da segurança dos sistemas suportados, escalabilidade e confiabilidade [42]. A segurança baseada em nuvem minimiza os riscos de ataques de negação de serviço em razão dos mecanismos de proteção associados, e, amplia, a alta disponibilidade e suporte dos serviços oferecidos [41]. Em relação a escalabilidade esse modelo permite a ampliação da capacidade computacional de maneira mais ágil, sem a necessidade de aquisição local de recursos. E no caso da confiabilidade, esta decorre do fato que a computação em nuvem tem o potencial de ser altamente distribuída, com diferentes datacenter físicos que possibilitam uma maior capacidade de redundância dos serviços suportados.

Esse modelo ganha ainda mais importância no contexto de proteção de dados, em razão da virtualização do armazenamento de dados, assim como ocorre com o processamento e serviços de TIC [42]. O Armazenamento em nuvem exige cuidados adicionais para garantir a aderência com as legislações de proteção de dados. Ocorre que nesse modelo o dado é armazenado externamente a infraestrutura física do controlador do dado, sendo armazenado em uma solução de armazenamento virtual.

Com isso é importante considerar que o detentor do dado em última instância é a provedora do serviço de armazenamento em nuvem, por isso o aspecto de proteção de dados se torna tão sensível nesse cenário [43]. A proteção de dados em ambiente de nuvem ocorre por camadas, sendo a primeira camada o controle de acesso a essas informações e o monitoramento dos recursos e acessos, porém além disso é necessário, principalmente para dados pessoais ou dados sensíveis, a proteção de dados com recursos de anonimização ou pseudonimização dos dados pessoais para evitar que essa informação seja compartilhada de maneira indevida [44].

Nesse contexto com a realização da proteção de dados com técnicas de criptografia, os dados

podem ser armazenados em nuvem e as chaves utilizadas para cifração dessa informação não são armazenadas em nuvem, são armazenadas de maneira local e segura, com isso mesmo no caso de um vazamento ou ataque malicioso às informações armazenadas, não será possível a visualização da informação em texto claro. E, assim, como a LGPD preconiza que os dados protegidos e que não possam ser reassociados com a utilização de esforços razoáveis estão aderentes com a legislação, sendo considerado esforços razoáveis aquelas ações que considerem fatores objetivos como o custo e tempo necessário para a reversão do processo [5],[44].

O fornecimento desse modelo de infraestrutura em nuvem pode ser oferecido por diferentes fornecedores de solução [45], trata-se de Infraestrutura como Serviço, conhecida como IaaS (Infrastructure as a Service), que é a o serviço oferecido de computação em nuvem provisionada através da internet. Segundo o Gartner, importante consultoria global de soluções de Tecnologia da Informação, os principais provedores de IaaS possuem a características de serem globais e entre eles estão a Google cloud, Amazon web services e Microsoft [40], [45].

Esses provedores apresentam a característica comum de descentralização dos recursos e infraestrutura de TIC, com datacenter em localidades diversas e muitas vezes com a informação sendo armazenada de maneira distribuída dificultando ainda mais a visibilidade do local de armazenamento físico da informação [42], [45].

2.6 TRABALHOS CORRELATOS

Existem estudos relacionados a proteção de dados pessoais, sendo a grande referência a própria Lei nº 13.709/2018 [5], Lei Geral de Proteção de Dados Pessoais, que pode ser considerada um guia, visto que indica princípios, conceitos, requisitos, penalidades e, inclusive, boas práticas de governança dessas informações. Existem pesquisas relacionadas as premissas definidas na LGPD.

Pinheiro [7] descreveu os conceitos relacionados à proteção de dados e realizou um comparativo com a General Data Protection Regulation (GDPR). Nesse estudo a autora ainda descreveu os aspectos legais da nova norma e os desdobramentos potenciais para os aspectos relacionados a tecnologia da informação, citando a necessidade de proteção de dados e a possibilidade de aderência à legislação em caso de aplicação de técnicas de anonimização de dados.

Blum et al. [17] coordenaram um estudo com especialistas e professores em proteção de dados pessoais e privacidade, em que abordaram as competências e atividades relacionados ao Data Protection Officer, DPO, facilitando o entendimento das características desse relevante papel na correta aplicação da LGPD. Os Autores ressaltaram, ainda, a possibilidade do encarregado da proteção de dados na instituição ser uma atribuição compartilhada, em um comitê ou unidade criada para essa finalidade.

Carvalho et al. [15] apresentaram um conjunto com as melhores práticas de governança de dados. Os autores abordaram quatro desafios da anonimização de dados. Os desafios apresen-

tados foram, primeiramente a a informação ainda pode ser considerada pessoal mesmo quando o nome não está associado diretamente aos demais dados, isso decorre do fato que a associação pode ocorrer através de outros dados. Um segundo desafio é que a possibilidade de associação em bancos de dados com grandes volumes é alta quando as técnicas de anonimização utilizadas são mais simples, como o mascaramento por exemplo. Um terceiro desafio é referente a classificação do tipo de dado a ser anonimizado, a correta classificação desse dado como sendo uma dado pessoal que pode ser utilizado como identificador é importante para evitar a reidentificação dessa informação. E, ainda, como um quarto desafio tem a preocupação a respeito da dificuldade de determinar se um dado parcialmente anonimizado possa ser reidentificado por um critério adicional que possa surgir posteriormente por uma mudança técnica de tecnologia ou por uma decisão legal. Dessa maneira, esses desafios se referem à possibilidade de associação da informação anonimizada com seu significado real em um contexto de big data. Nesse cenário, considerando o maior volume de informações (big data), aumenta a possibilidade de uma reidentificação do dado.

Canedo et al. [14] apresentaram um processo para apoiar na implementação da Lei Geral de Proteção de dados. O processo proposto pelos autores consiste de 14 etapas: 1) Estudar a LGPD e a POSIC da instituição; 2) Aplicação de Questionários; 3) Indicação do Data Protection Officer (DPO); 4) Mapear o Fluxo e Processamento de Informações; 5) Analisar e Implementar Políticas de segurança internas ou externas; 6) Mapeamento dos Riscos (o objetivo é identificar ameaças que podem afetar dados pessoais); 7) Formular Data Protection Impact Assessment (DPIA); 8) Aprovar DPIA (Data Protection Impact Assessment); 9) Formular Data Protection Policy; 10) Implementar e validar Data Protection Policy; 11) Analise do impacto da Implementação da Data Protection Policy; 12) Treinamentos; 13) Nova Concepção de Dados; e 14) ICT (Information and Communication Techonology) Governança de Dados (Data Protection).

Starchon e Pikulik [13] abordaram aspectos importantes das técnicas de anonimização, considerando princípios da privacidade de dados, e fazendo uma exposição das principais técnicas de anonimização. Destacando as seguintes técnicas: Mascaramento ou supressão; a Generalização; a própria técnica de criptografia; e Tokenização. Os autores apresentaram, ainda, a conceituação dessas técnicas para auxiliar na decisão de quando se aplicar cada tipo de procedimento de anonimização.

Carvalho et al. [46] apresentaram um estudo que reforçou a importância da anonimização de dados pessoais. Os autores ressaltaram a existência de limites na utilização dessas técnicas e apresentaram os riscos envolvidos quando a anonimização é realizada no contexto de dados massivos, pois trata-se de um cenário de maior dificuldade de se garantir a privacidade de informações pessoais pela possibilidade de reassociação dessas informações em um ambiente de Big Data.

Gunawan e Mambo [47] exploraram a anonimização de dados através do método de substituição. Os resultados experimentais mostraram que o método proposto reduziu com sucesso a probabilidade de sucesso de ataque em um banco de dados anonimizado minimizando a perda de informações. Outro aspecto que foi abordado pelos autores foi sobre a possibilidade de acompanhar valores associados ao banco de dados para avaliar tendências dos usuários. Exemplo

utilizado nesse artigo para ilustrar é o de consulta de web sites, quando é verificado o conjunto de valores acessados é possível verificar a tendência de uso do usuário, com isso essa informação passa a ser passível de ser tratada como informação pessoal e devendo ser protegida de maneira adequada.

Questões relacionadas as estratégias para planejamento de projetos de privacidade considerando a GDPR foram tratados por Saltarella et al. [27], em estudo que tratou da importância de aplicação de medidas de segurança da informação não apenas para proteção dos ativos digitas como para o cumprimento de requisitos previstos em legislações vigentes.

Em relação a aplicação de técnicas de privacidade e proteção de dados, como a técnica de anonimização de dados, Prasser et al. [48] apresentou um estudo de caso aplicando, de fato, essas técnicas utilizado uma ferramenta open source denominada ARX Data Anonymization. Nesse trabalho o autor relatou o funcionamento de técnicas e modelos de transformação de dados, como a generalização, supressão e agregação, e ainda o funcionamento conjunto desses modelos de proteção de dados.

Considerando os trabalhos correlatos apresentados, foi possível mapear as diferentes abordagens apresentadas. Foram identificadas abordagens relacionadas as legislações vigentes, em relação a diretrizes de privacidade de dados, também foi abordada uma proposta para implementação da proteção de dados e, ainda, foram descritas técnicas de anonimização de dados. A Tabela 2.1 auxilia o entendimento ao mapear os principais temas abordados nos trabalhos correlatos, com a indicação do que pretende ser a contribuição desse trabalho, bem como o principal diferencial dessa proposta.

Tabela 2.1: Tabela comparativa trabalhos correlatos. (Fonte: Autor)

	Aspectos da Legislação	**Privacidade e Proteção de Dados**	**Técnicas de Anonimização**	**Modelo de Anonimização em Nuvem**
Pinheiro [7]	Sim	Sim		
Blum et al. [17]	Sim	Sim		
Carvalho et al. [15]	Sim	Sim	Sim	
Canedo et al. [14]	Sim	Sim	Sim	
Starchon e Pikulik [13]		Sim	Sim	
Carvalho et al. [46]	Sim	Sim	Sim	
Gunawan e Mambo [47]			Sim	
Saltarella et al. [27]	Sim	Sim		
Prasser et al. [48]			Sim	
Proposta desse Trabalho	**Sim**	**Sim**	**Sim**	**Sim**

2.7 SÍNTESE DO CAPÍTULO

Nesse capítulo foram tratados conceitos que irão direcionar e fundamentar esse estudo. Considerando os referenciais teóricos que abordam o tema e citando estudos correlatos relevantes para a discussão proposta. A partir dessa revisão conceitual será tratado no capítulo seguinte a descrição da metodologia que se pretende utilizar para atingir os objetivos propostos para esse estudo.

3 METODOLOGIA

A Metodologia utilizada busca atender aos objetivos desse estudo, implica na realização da revisão bibliográfica e aplicação de um Estudo de Caso que possa apoiar a compreensão dos aspectos relacionados a efetiva proteção dos dados pessoais [7].

Foi realizado no Capítulo 2 uma revisão bibliográfica com o objetivo de identificar as principais técnicas de anonimização utilizadas para proteção de dados. As fontes de dados pesquisadas para fundamentar esse estudo foram as bases digitais Computer Science Bibliography (DBLP) e a base do Institute of Electrical And Eletronics Engineers (IEEE) Xplore.

As pesquisas com o objetivo de fundamentação dos aspectos relacionados à governança e proteção de dados tiverem um intervalo de seleção considerando os últimos 5 anos com o objetivo de análise de artigos mais recentes a respeito dessa temática. Em relação ao estudo de segurança da informação e técnicas de criptografia o intervalo de pesquisa de artigos publicados foi ampliado para 10 anos, considerando a possibilidade de estudos já terem sido realizados por se tratar de um tema já abordado com mais frequência.

Considerando a seleção dos artigos analisados, foram priorizados nesse estudo os artigos associados a validação das técnicas de anonimização com a finalidade de proteção de dados. Para validação dos conceitos apresentados nessa pesquisa será elaborado um protocolo de pesquisa para aplicação de Estudo de Caso. Para isso é necessário o estudo das técnicas de anonimização que podem ser utilizadas na aplicação desse estudo de caso [6].

3.1 SELEÇÃO DA BASE DE DADOS

Considerando conceitos já abordados na revisão bibliográfica, na sequência da metodologia proposta deve ser realizada a seleção da base de dados de amostra que deve ser utilizada no Estudo de Caso. Para fins desse estudo será utilizada uma amostra de dados reais disponibilizados abertamente pelo Portal Brasileiro de Dados abertos (dados.gov.br) [49], em que são franqueados a quem tiver interesse informações sobre diversos serviços públicos. Com o interesse de adotar uma amostra que apresente informações consideradas dados pessoais, foi selecionada uma base de dados com uma amostra de solicitações de usuários para o serviço da administração pública denominado TaxiGOV, que trata-se de um serviço de transporte de servidores e colaboradores da Administração Pública Federal em deslocamentos a trabalho com o uso de serviço de táxis [50]. Dessa maneira, foi selecionada uma base de dados, representada pela amostra apresentada na Tabela 3.1, que lista as corridas realizadas pelo serviço Taxigov.

Tabela 3.1: Amostra com dados de utilização do serviço Taxigov . (Fonte: Autor)

Motivo	**Origem**	**Destino**	**Nome passageiro**
4 - Outros	GAMA	Palácio da Alvorada	DOUGLAS BORGES DE SOUSA
4 - Outros	SIA	Palácio da Alvorada	ANTONIANNI ARAUJO DE SOUSA
1 - Reuniao	ASA NORTE	Ministerio da Economia	ELIZETE MEIRELES DA COSTA
1 - Reuniao	BRASILIA	Ministerio do Desenvolvimento Regional	JULIANA GRANDE POUSA
1 - Reuniao	BRASILIA	B Hotel Brasilia, Asa Norte	ANAMARIA DANDREA CORBOA
1 - Reuniao	LAGO SUL	Edificio Porto Real, QMSW 4 Lt. 6	GILMAR ANTONIO ALVES SOUTO
1 - Reuniao	ASA SUL	Departamento de Infraestrutura	ABDSANDRYK CUNHA SOUZA
1 - Reuniao	ASA NORTE	Ministerio da Ciencia e Tecnologia	ABEL BARBOSA NETO SOUTO
1 - Reuniao	ASA NORTE	B Hotel Brasilia, Asa Norte	ABEL BARBOSA NETO SOUTO
1 - Reuniao	BRASILIA	Superquadra Norte 411, Asa Norte	ABEL BARBOSA NETO SOUTO
1 - Reuniao	ASA SUL	Ministerio da Saude - Bloco G	ABEL DA SILVA MUNIZ
1 - Reuniao	ASA NORTE	SGO Q 1 Ae	ABIGAIR APARECIDA SANTOS
1 - Reuniao	SUDOESTE	Anexo do Ministerio da Saude	ABIGAIR APARECIDA SANTOS
1 - Reuniao	BRASILIA	B Hotel Brasilia, Asa Norte	ABILIO AUGUSTO MAIA PINTO
1 - Reuniao	ASA NORTE	B Hotel Brasilia, Asa Norte	ABILIO AUGUSTO MAIA PINTO
1 - Reuniao	ASA SUL	Ministerio da Cidadania - GM/MC	ABIMAEL RIBEIRO DA SILVA
3 - Visita	ASA NORTE	SGO Q 1 Ae	ABNER DA SILVA SOUZA
1 - Reuniao	ASA NORTE	Venancio Shopping, Setor Comercial Sul	ABNER DA SILVA SOUZA
1 - Reuniao	Z. CIVICO	Instituto Nacional do Seguro Social	ABNER DA SILVA SOUZA
1 - Reuniao	BRASILIA	SQN 405 Bl. D, Asa Norte	ABRAAO BILLY VILA FLOR
1 - Reuniao	ASA SUL	Imprensa Nacional, Industrias Graficas	JULIANA GRANDE POUSA

É possível perceber, ainda, observando a Tabela 3.1, que essa amostra apresenta em texto claro informações como motivo da utilização do serviço, bairro de origem da corrida, local de destino e ainda o nome do passageiro. Em Starchon and Pikulik [13] é descrito a ação de mapeamento das informações para que os dados sejam separados em três categorias distintas: o conjunto completo de dados, os dados de identificação pessoal e os dados que não identificam um indivíduo. Considerando a Tabela 3.1, e realizando uma avaliação considerando proposta apresentada por Starchon and Pikulik [13], é possível identificar o campo "Nome passageiro"como um dado de identificação pessoal, afinal é um campo que pode identificar diretamente um indivíduo.

Tabela 3.2: Amostra com dados de passageiros e valores do serviço Taxigov . (Fonte: Autor)

Nome passageiro	Cpf solicitante	Ano mês	Km total	Vl corrida	Qt corrida
DOUGLAS BORGES DE SOUSA	***.425.428-**	202010	20.2	59.99	1
ANTONIANNI ARAUJO DE SOUSA	***.987.231-**	202001	22.9	66.41	1
ELIZETE MEIRELES DA COSTA	***.378.031-**	201912	74.7	222.14	19
JULIANA GRANDE POUSA FIDELIS	***.771.511-**	201911	2.0	5.8	1
ANAMARIA DANDREA CORBOA	***.0.1.14.-**	202107	23.18	67.22	1
GILMAR ANTONIO ALVES DE SOUTO	***.039.451-**	201911	81.4	236.06	6
ABDSANDRYK CUNHA DE SOUZA	***.930.011-**	201911	3.5	10.15	1
ABEL BARBOSA NETO SOUTO	***.356.801-**	201911	10.2	41.47	7
ABEL BARBOSA NETO SOUTO	***.356.801-**	201912	29.7	109.04	18
ABEL BARBOSA NETO SOUTO	***.356.801-**	202001	1.7	5.8	1
ABEL DA SILVA MUNIZ	***.165.457-**	202003	3.56	14.74	2
ABIGAIR APARECIDA DOS SANTOS	***.369.406-**	201912	5.8	20.59	2
ABIGAIR APARECIDA DOS SANTOS	***.369.406-**	202002	2.8	17.4	3
ABILIO AUGUSTO MAIA PINTO	***.138.495-**	201912	5.2	15.08	2
ABIMAEL RIBEIRO DA SILVA	***.260.201-**	202102	74.2	220.37	4
ABNER DA SILVA SOUZA	***.562.732-**	202007	23.6	70.09	1
ABRAAO BILLY VILA FLOR DORIA	***.729.935-**	202204	14.7	46.45	1
ABRAAO BILLY VILA FLOR DORIA	***.729.935-**	202206	12.53	44.5	3
ABRAAO VILA FLOR DORIA	***.729.935-**	201912	6.3	18.27	1
ABSAI DE SOUSA CAMARGO	***.935.861-**	202008	12.5	37.13	2
ABSAI DE SOUSA CAMARGO	***.935.861-**	202204	44.28	139.93	4
ACAUA BROCHADO	***.448.808-**	202003	2.6	7.54	1
ADA BENTO DOS SANTOS	***.613.241-**	201911	7.5	21.75	2
ADA BENTO DOS SANTOS	***.613.241-**	202001	5.0	14.5	2
ADA REGINA NOGUEIRA VIANA	***.108.501-**	202002	107.7	312.33	4
ADA REGINA NOGUEIRA VIANA	***.108.501-**	202102	8.7	25.84	2
ADAILSON LOPES TEIXEIRA	***.686.041-**	201911	90.4	262.16	25
ADAILSON LOPES TEIXEIRA	***.686.041-**	201912	138.41	402.26	36
ADEMIR BARROS DE CARVALHO	***.713.721-**	202207	20.95	74.35	2
ADEMIR LAPA	***.372.629-**	201911	9.8	34.22	5
ADEMIR LAPA	***.372.629-**	201912	3.6	11.6	2
ADEMIR LAPA	***.372.629-**	202002	26.4	76.56	4
ADENILDA MARIA DE ARAUJO	***.792.611-**	202107	36.59	115.62	3
ADENILDA MARIA DE ARAUJO	***.792.611-**	202109	10.42	32.94	1
ADENILTON SOUZA	***.421.226-**	202102	46.7	150.59	22
ADENILTON SOUZA	***.421.226-**	202103	91.1	759.42	44
ADENILTON SOUZA	***.421.226-**	202202	22.3	73.03	10
ADENISIO ALVARO DE OLIVEIRA	***.131.334-**	201911	3.2	9.28	1
ADENISIO ALVARO DE OLIVEIRA	***.131.334-**	202206	3.98	14.57	2
ADHEMAR RANCIARO NETO	***.365.928-**	201912	3.9	11.89	2
ADI BALBINOT JUNIOR	***.692.621-**	202202	2.1	6.63	1
ADI BALBINOT JUNIOR	***.692.621-**	202206	1.99	7.1	1
ADIEL PEREIRA ALCANTARA	***.727.161-**	202003	18.0	52.2	1

Analisando a tabela 3.2, verifica-se que os nomes dos passageiros são apresentados em texto

claro. No caso do dado CPF do solicitante, este já é apresentado na amostra selecionada com a técnica de anonimização de mascaramento para evitar a associação imediata entre nome e cpf do solicitante.

Tabela 3.3: Amostra com dados de passageiros do serviço Taxigov e data de nascimento . (Fonte: Autor)

Nome passageiro	**Cpf solicitante**	**Data Nascimento**	**Sexo**
DOUGLAS BORGES DE SOUSA	757.425.428-32	05/07/1961	M
ANTONIANNI ARAUJO DE SOUSA	842.987.231-27	20/03/1967	M
ELIZETE MEIRELES DA COSTA	547.378.031-14	11/09/1971	F
JULIANA GRANDE POUSA FIDELIS	946.771.511-21	23/11/1978	F
ANAMARIA DANDREA CORBOA	521.107.112-14	25/06/1968	F
GILMAR ANTONIO ALVES DE SOUTO	722.039.451-91	30/05/1974	M
ABDSANDRYK CUNHA DE SOUZA	788.930.011-15	18/04/1959	M
ABEL BARBOSA NETO SOUTO	312.356.801-34	11/03/1957	M
ADAILSON LOPES TEIXEIRA	454.686.041-11	18/01/1971	M
ABEL DA SILVA MUNIZ	121.165.457-21	20/08/1974	M
ABIGAIR APARECIDA DOS SANTOS	437.369.406-12	24/02/1972	F
ADENILDA MARIA DE ARAUJO	212.792.611-13	04/03/1975	F
ABILIO AUGUSTO MAIA PINTO	978.138.495-51	27/06/1978	M
ABIMAEL RIBEIRO DA SILVA	795.260.201-89	05/09/1967	M
ABNER DA SILVA SOUZA	621.562.732-13	13/04/1973	M
ABRAAO BILLY VILA FLOR DORIA	649.729.935-44	14/08/1965	M
ADI BALBINOT JUNIOR	978.692.621-44	10/11/1969	M
ABSAI DE SOUSA CAMARGO	247.935.861-15	17/04/1958	M
ADEMIR LAPA	746.372.629-44	14/05/1968	M
ACAUA BROCHADO	324.448.808-56	10/01/1970	M

O Cenário apresentado na tabela 3.3 é de inclusão de dados cadastrais dos usuários com informações pessoais como data de nascimento e CPF completo do indíviduo. Trata-se de um cenário mais sensível considerando a necessidade adicional de proteção dessas informações.

3.2 SELEÇÃO DE FERRAMENTA DE ANONIMIZAÇÃO

Conforme descrito no capítulo 2 desse estudo, o processo de anonimização é um processo que envolve técnicas computacionais e matemáticas para atingir o objetivo de proteção dos dados. Dessa maneira para a execução desses processos de anonimização nas amostras selecionadas para validação do estudo, é importante a seleção e utilização de uma ferramenta que auxilie nesse processo.

A ferramenta de anonimização selecionada e utilizada nesse trabalho é denominada ARX Anonymization Tool [51]. Trata-se de uma ferramenta open source que permite a transformação estruturada de dados pessoais utilizando métodos de anonimização disponibilizados pela solução [52]. A ferramenta ARX suporta a utilização de diferentes técnicas de transformação ou anonimização de dados, tais como, embaralhamento, generalização, mascaramento, categorização,

tokenização e micro agregação [51] [52].

A ferramenta possui uma interface gráfica simples e intuitiva o que facilita a sua utilização nesse estudo, permitindo a aplicação, de maneira simples, das técnicas de anonimização previstas nesse estudo para as amostras selecionadas.

3.3 SELEÇÃO DE AMBIENTE DE NUVEM PÚBLICA

Após a seleção das bases de dados para o estudo, e para dar sequência na aplicação da metodologia proposta, deve ser selecionado e configurado um ambiente de nuvem pública para armazenamento dessas bases de dados. O ambiente selecionado para esse estudo é a nuvem Microsoft, denominada Microsoft Azure [53].

Nesse ambiente de nuvem pública é possível realizar o armazenamento das bases de dados após a realização dos processos de proteção de dados que serão descritos no detalhamento do estudo de caso. Em um contexto de nuvem pública é importante considerar que passa a existir uma gestão compartilhada da segurança da informação, para a seleção desse ambiente para a aplicação do estudo de caso será considerado uma arquitetura específica para disponibilizar área de armazenamento dessas informações através de autenticação e controle de acesso.

Porém, mesmo com essa camada de segurança proporcionada pelo controle de acesso, resta ainda a questão de que o armazenamento de dados ocorrerá em ambiente externo, ou seja, não local. Em última instância, essas informações ainda poderiam ser acessadas pelos administradores do ambiente de colaboração que administram a infraestrutura utilizada em nuvem pública. Com isso, é importante buscar implementar alternativas de proteção de dados, conforme as mencionadas nesse trabalho, como a implementação de técnicas de anonimização dos dados sensíveis e pessoais.

3.4 *FRAMEWORK* PARA ESTUDO DE CASO

O estudo de caso será estruturado com a definição e seleção de uma base de dados hipotética que contenha uma amostra de dados considerados informações de identificação pessoal ou dados auxiliares, e por essa razão, passíveis de aplicação de mecanismos de proteção de dados conforme preconizado pela LGPD [5], [10].

Para realização desse protocolo de pesquisa um conceito importante precisará ser utilizado, trata-se da governança de dados, é através do mapeamento das informações armazenadas em base de dados, e identificando os tipos de dados armazenados que será possível a identificação e seleção da técnica mais adequada de anonimização para cada tipo de dado mapeado [6]. Dessa maneira, a aplicação de conceitos de governança de dados é uma etapa importante da realização do estudo de caso.

Com a identificação e mapeamento dos dados a serem protegidos, e seleção da técnica de anonimização a ser utilizada em cada cenário, será aplicado nesse estudo de caso a anonimização de uma amostra de dados para que seja possível a comparação da base de dados selecionada antes do processo de proteção de dados e o cenário após o processo realizado.

E é através da análise dos resultados obtidos que se espera responder ao problema de pesquisa apresentado, que trata da questão de como buscar garantir o armazenamento seguro de dados de identificação pessoal, implementando técnicas de proteção dessas informações, de maneira aderente com a legislação [5].

Como etapa final da realização do estudo de caso, é importante descrever os mecanismos de proteção de uma base de dados [6], com dados pessoais armazenados, quando esse armazenamento ocorre em nuvem pública, ou seja, quando as informações armazenadas não estão, exclusivamente, na infraestrutura de tecnologia da instituição.

Nesse cenário é ainda mais relevante a adequada proteção de dados, pois os mecanismos de segurança da infraestrutura não estão sendo realizados e geridos pela instituição encarregada desses dados, e sim, sendo uma gestão de segurança delegada para a provedora de serviços de nuvem pública [11].

Para a aplicação da proteção de dados nesse estudo de caso, duas abordagens associadas a anonimização de dados devem ser consideradas. O primeiro aspecto é relacionado a aplicação de técnicas de anonimização apenas nas informações mapeadas e identificadas como dados pessoais. Com isso deve ser considerado que a anonimização trata-se de um processo irreversível em que o dado pessoal não pode mais ser associado a um determinado indivíduo e, por esse motivo, essa informação perde, de maneira significativa seu potencial de utilização [54].

Uma segunda abordagem que deve ser apresentada é a proposta de anonimização de dados que considera a proteção das informações, mas, também, a possibilidade de utilização das informações em momento posterior, e necessário é o caso de aplicação da técnica de criptografia das informações. Nesse contexto, uma informação adicional faz parte do processo para que ocorra a cifração, trata-se da chave criptográfica, nesse caso com a utilização de uma chave criptográfica de decifração o processo pode ser revertido, com essa possibilidade de reversão esse processo é denominado pseudoanonimização de dados. No entanto, esse processo pode ser considerado uma técnica de anonimização de dados se somados alguns critérios que dificultem o processo de reversão dessa informação [55], [54].

Segundo Doneda e Machado [54], quando esse processo de cifração dos arquivos ocorre com um armazenamento adequado da chave de cifração em local diverso, esse processo pode ser considerado uma anonimização de dados dado a dificuldade de reversão do processo. Nesse estudo de caso o modelo proposto envolve a possibilidade de anonimização parcial da base de dados e, também, a alternativa de cifração dos arquivos para o posterior armazenamento em nuvem e garantido que a chave de cifração seja armazenada em local diverso.

4 FRAMEWORK

O framework apresentado nesse capítulo considera dois modelos para a realização da proteção de dados de dados pessoais, o primeiro considerando a anonimização desses dados e um segundo modelo considerando a criptografia do conjunto completo de dados.

4.1 MODELO PARA ANONIMIZAÇÃO DOS DADOS

Esse modelo é estruturado considerando 4 (quatro) etapas para sua implementação, conforme descrito abaixo.

Etapa 1: Realização da classificação dos dados apresentados.

A identificação e classificação do tipo de dados presentes na amostra utilizada para esse estudo foi realizada considerando a os campos de informação presentes nas tabelas apresentadas no capítulo anterior, os campos de informação presentes nas tabelas 3.1, 3.2 e 3.3 seguem listados abaixo.

Tabela 4.1: Campos apresentados nas amostras selecionadas para o estudo . (Fonte: Autor)

Tabela 3.1	motivo, origem, destino, Nome passageiro
Tabela 3.2	Nome passageiro, cpf, ano mês, km total, vl corrida, qt corrida
Tabela 3.3	Nome passageiro, cpf solicitante, data nascimento, sexo

Considerando o conceito de dado pessoal como aquele que remete diretamente ou indiretamente a um individuo, conforme apresentado no capítulo 2 desse estudo, e ainda citando a lei geral de proteção de dados [5] que de maneira expressa cita que "o dado pessoal é aquele que possibilita a identificação, direta ou indireta, da pessoa natural", é possível verificar analisando os campos indicados na tabela 4.1 que os campos "nome passageiro", "cpf solicitante"e "data nascimento"são informações que devem ser tratadas como dados pessoais. Sendo os demais campos tratados como dados auxiliares, que são informações que se não associadas a outras não possuem o potencial de associação direta a um individuo.

Etapa 2: Seleção de técnica de anonimização a ser utilizada.

Após o mapeamento, identificação e classificação, realizado na etapa inicial, dos tipos de dados apresentados na amostra, deve ser selecionada o tipo de técnica de anonimização mais adequada a ser utilizada nesse processo.

As principais técnicas de anonimização foram apresentadas no capítulo 2, seção 2.1, são elas: Mascaramento e supressão, generalização, criptografia e tokenização.

O processo de anonimização de dados, conforme descrito no capítulo 2 desse estudo, deve

considerar o tipo de informação a ser protegida e a possibilidade de se manter a possibilidade de utilização da informação [13]. Dessa maneira é necessário uma avaliação dos campos identificados na etapa de classificação do tipo de informação para verificação do tipo de dado a ser protegido.

Com isso, deve ser verificada as informações apresentadas em cada campo. No caso do campo "nome passageiro"o conteúdo informado é o nome completo do indivíduo, nessa situação é necessário evitar a visualização da informação completa, uma técnica de anonimização a ser utilizada nessa atuação é a técnicas de embaralhamento ou "scrambling"das informações [13], dessa maneira o nome completo não pode ser utilizado para associação direta ou indireta a um indivíduo. Utilizando o primeiro registro da amostra da tabela 3.1 para exemplificar teríamos: "DOUGLAS BORGES DE SOUSA"para "UAGBOS EGDUSAO EL ALGUA", o que deixaria a informação anonimizada.

Para o campo "cpf solicitante"a informação apresentada é o número cpf (cadastro de pessoa física) completo o que é um identificador único do individuo. Dessa maneira é necessário anonimizar, e nesse caso a técnica de anonimização adequada seria o Mascaramento ou "Masking"[13]. Nessa técnica parte da informação é preservada para que a informação ainda possa ser parcialmente utilizada. Utilizando o primeiro registro da amostra da tabela 3.3 para exemplificar teríamos: "757.425.428-32", para "***.425.428-**", assim a informação ficaria anonimizada, porém ainda teria uma utilidade parcial para realização de consultas, filtros ou validação de informações, pois parte do dado continuaria íntegro.

No caso do terceiro campo identificado na etapa anterior como dado pessoal, que é o campo "data nascimento"a informação apresentada é data de nascimento completa do individuo. Nesse caso a associação não é imediatada ao indivíduo, porém o potencial de identificação indireta é muito elevado, por isso é uma informação que deve ser tratada como dado pessoal. A técnica de anonimização adequada para utilizar para esse tipo de informação seria a generalização, citada na seção 2.1 desse trabalho. É a técnica que permite a proteção da informações porém preserva parte da sua utilidade ao permitir que a informação seja tratada como um intervalo de valores [13]. Para melhor compreensão dessa técnica e utilizando o primeiro registro da amostra da tabela 3.3 para exemplificar teríamos: "05/07/1961"para "maior de 60 anos"ou ainda, outra possibilidade seria a de se estabelecer uma faixa de valores e verificar em qual faixa a informação se adequaria, ficando, por exemplo, "maior que 60 menor que 70", assim a informação com a data precisa de nascimento estaria protegida, mas a informação ainda teria uma utilidade parcial para realização de relatórios, filtros ou consultas.

Etapa 3: Configuração e aplicação da ferramenta de apoio à anonimização dos dados.

Conforme descrito na seção 3.2, a ferramenta selecionada para auxiliar na execução do processo computacional de anonimização de dados é a ARX anonymization tool. A solução é de fácil "download"e instalação [51], a ferramente não exige licenciamento para utilização, é uma ferramenta open source e compatível com sistema operacional windows.

A solução ARX permite a abertura dos arquivos com as bases de dados selecionadas para a

amostra, e permite a configuração da técnica de anonimização a ser utilizada conforme o caso. A definição das técnicas de anonimização a serem utilizadas em cada situação foram definidas e descritas na etapa anterior desse framework. A etapa de configuração da ferramenta envolve o carregamento do arquivo com a amostra para a ferramenta em formato .cvs (comma separated values) e seleção da técnica de transformação dos dados a ser utilizada, com isso a ferramenta executa o processo e gera uma nova tabela com os dados anonimizados.

Etapa 4: Execução da anonimização dos dados

A etapa de execução do processo é decorrência imediata da etapa anterior de configuração da ferramenta. É nessa uma última etapa do framework proposto que a operação é de fato realizada e processada pela ferramenta, com a geração de uma nova base de dados com os dados anonimizados de maneira irreversível. Dessa maneira, uma vez realizado o processo não é possível retornar ao conteúdo da base de dados anterior ao processo. Com isso a nova base gerada pode ser disponibilizada para acesso pois estará devidamente protegida e aderente a legislação.

4.2 MODELO PARA PROTEÇÃO DE ARQUIVOS

Conforme citado na metodologia apresentada, uma segunda abordagem de proteção de dados pode ser utilizada: a proteção do arquivo completo através da criptografia desse arquivo. Com isso o framework para proteção de dados utilizando criptografia é estruturado considerando as etapas descritas nesta seção: avaliação e classificação do conjunto de dados presentes no arquivo (Etapa 1), seleção do tipo de criptografia a ser utilizada (Etapa 2), preparação da nuvem pública a ser utilizada (Etapa 3) e execução da cifração e armazenamento do arquivo (Etapa 4).

Etapa 1: Avaliação e classificação do conjunto de dados presentes no arquivo.

A avaliação do arquivo deve ocorrer considerando a natureza e sensibilidade das informações e se deve ser aplicada o modelo de proteção de dados mais rigoroso que é a criptografia de todo arquivo, inviabilizando a abertura do arquivo.

Com o objetivo de propor um framework que possibilite a proteção do arquivo completo, a amostra apresentada na tabela 3.1 pode ser considerada como uma amostra que apresenta dados pessoais e, também, um conjunto de dados sensíveis como os campos "motivo", "origem"e "destino"que de acordo com o contexto de utilização podem trazer prejuízos ao detentor dos dados em caso de divulgação. Dessa forma, pode ser escolhido o modelo de proteção do arquivo completo, que, apesar de aumentar significativamente o nível de proteção das informações, inviabiliza a possibilidade de utilização, ainda que parcial, das informações contidas no arquivo.

Etapa 2: Seleção do tipo de criptografia a ser utilizada.

A criptografia utilizada para a proteção dessas informações pode utilizar algoritmos de criptografia simétricos ou assimétricos, no entanto o gerenciamento de chaves no modelo de criptografia assimétrica é mais complexo e com desempenho prejudicado nesse modelo. A criptografia simé-

trica , aquela em que a chave utilizada para cifração é a mesma utilizada para decifração, é mais utilizada nesse modelo. A chave a ser utilizada deve ser complexa e armazenada em local distinto do arquivo protegido, pois a dificuldade de reversão do processo é o que caracteriza essa alternativa como uma opção efetiva de proteção de dados.

Um segundo passo a ser realizada na etapa de seleção do tipo de criptografia a ser utilizado é a seleção do algoritmo criptográfico. A opção nesse estudo é pela utilização da criptografia AES (Advanced Encryption Standard) de 256 bits, trata-se de um modelo amplamente validade e reconhecidamente seguro e estável. Para a escolha desse algoritmo é importante verificar que é oferecido e suportada pela solução de nuvem pública a ser utilizada.

Etapa 3: Preparação da nuvem pública a ser utilizada

A seleção da nuvem pública utilizada nesse estudo é importante para a definição de como se dará o armazenamento e, também, como se dará o processo de cifração do arquivo, e armazenamento da chave criptográfica utilizada no processo.

A nuvem pública utilizada nesse estudo é o Microsoft Azure. O Azure fornece diferentes modelos de criptografia para armazenamento dos arquivos. Conforme definido na etapa anterior, o algoritmo de criptografia utilizado é o AES, que é suportado e oferecido pela nuvem Azure de maneira nativa, realizando a cifração e decifração dos dados armazenados de maneira transparente ao usuário.

Outro ponto relevante de preparação da nuvem pública é a definição do gerenciamento das chaves de criptografia. Existem dois modelos de gerenciamento dessas chaves, o primeiro é o modelo de delegação do gerenciamento dessas chaves para o provedor de nuvem pública, ou, uma segunda alternativa em que o próprio cliente realize o gerenciamento dessas chaves criptográficas. Esse segundo modelo é o que iremos adotar nesse estudo em razão da necessidade de que as chaves utilizados no processo de cifração sejam armazenadas em local diverso dos dados armazenados, para, conforme descrito no capítulo 2, garantir que o processo de reversão de criptografia dessas informações seja mais complexo, garantindo, dessa forma, que esses dados possam ser tratados como dados anonimizados.

Etapa 4: Execução da cifração e armazenamento do arquivo.

A execução do processo de cifração dos arquivos proposto nesse modelo pode ser realizado diretamente pelo provedor como um serviço de criptografia. O provedor selecionado oferece um local de armazenamento que para a nuvem pública selecionada nesse estudo é denominado Azure Storage. Nesse serviço de armazenamento oferecido é possível habilitar o recurso de cifração dos dados a serem armazenados e, nível de serviço, durante o processo de transferência dos dados (upload) para o local de armazenamento, o Azure Storage.

Esse processo de configuração e habilitação desse serviço deve ser realizado considerado todas as etapas anteriores, principalmente a etapa de governança e classificação dos dados a serem armazenados. O processo de cifração previsto nessa etapa deve ser aplicado apenas aos dados a serem armazenados que precisem ser protegidos em razão dos custos computacionais envolvidos

no processo. Esse processo por ser realizado na etapa de armazenamento impacta na performance de armazenamento desses dados e, da mesma maneira, também reduz a velocidade de acesso a essas informações, devido a necessidade de realização do processo de decifração desses dados para disponibilização ao usuário final.

Conforme descrito na etapa anterior o gerenciamento das chaves criptográficas será realizado pelo cliente da solução, e não pelo provedor de nuvem pública. Dessa maneira, é possível realizar a configuração do gerenciamento de chaves na solução de duas maneiras. A primeira alternativa é de armazenamento em nuvem pública em local diverso da estrutura de armazenamento, esse local é um cofre de senha oferecido pela solução e denominado Azure Key Vault. E uma segunda alternativa que é a configuração no lado do cliente de um hardware específico para essa finalidade, denominado HSM - Hardware Security Model. Essa segunda alternativa eleva a complexidade da configuração, devido a necessidade de implantação de um hardware físico e implica, também, na integração do provedor de nuvem pública com o equipamento do lado do cliente. Dessa maneira, nesse estudo iremos utilizar o primeiro modelo, com gerenciamento de chaves criptográficas pelo cliente utilizando o serviço Azure Key Vault.

Conforme definido e descrito na etapa 1 desse framework, considerando a amostra selecionada, para a Tabela 3.3 diante da sensibilidade do conjunto completo dos dados com informações pessoais foi realizada a criptografia do arquivo inteiro para na sequência realizar a transferência (upload) para o ambiente de nuvem. Dessa maneira a proteção ocorre para o arquivo completo e não para cada dado separadamente.

A opção de utilização desse tipo de técnica de proteção do arquivo completo é interessante, também, quando existe a necessidade eventual de reutilização dessa informação quando necessário, é o caso de dados cadastrais. Caso a opção fosse pela anonimização de cada campo de informação, como a anonimização é um processo que não pode ser revertido, a utilização completa da informação estaria comprometida. No caso de utilização da Criptografia o processo pode ser revertido e a informação reutilizada.

5 ANÁLISE DE DADOS E RESULTADOS

Considerando a aplicação do modelo proposto de framework apresentado no capítulo anterior (framework) é importante analisar os dados e os resultados obtidos para a validação do modelo apresentado. O modelo descrito pretende confirmar as hipóteses apresentadas para a proteção de dados levando em conta dois cenários diferentes de acordo com a avaliação e a classificação dos dados representados na amostra selecionada, sendo um cenário com a aplicação de técnicas de anonimização conforme o dados a serem protegidos, e um segundo cenário considerando a necessidade de proteção do conjunto completo de dados com a criptografia do arquivo a ser armazenado.

5.1 RESULTADO DO PROCESSO DE ANONIMIZAÇÃO DE DADOS

A aplicação das etapas do framework descritas na seção 4.1 na amostra listada na Tabela 3.1 apresentou um resultado de uma base de dados anonimizada, conforme Tabela 5.1:

Tabela 5.1: Amostra com dados anonimizados do Taxigov. (Fonte: Autor)

Motivo	**Origem**	**Destino**	**Nome passageiro**
4 - Outros	GAMA	Palácio da Alvorada	DOUGLWT UORGOT DO TOUTW
4 - Outros	SIA	Palácio da Alvorada	WNTONPWNNP WRWUJO DO TOUTW
1 - Reuniao	ASA NORTE	Ministerio da Economia	OLPZOTO MOPROLOT DW COTTW
1 - Reuniao	BRASILIA	Ministerio do Desenvolvimento	JULPWNW GRWNDO POUTW
1 - Reuniao	BRASILIA	B Hotel Brasilia, Asa Norte	WNWMWRPW DWNDROW CORUOW
1 - Reuniao	LAGO SUL	Edificio Porto Real, QMSW 4 Lt. 6	GPLMWR WNTONPO WLVOT TOUTO
1 - Reuniao	ASA SUL	Departamento de Infraestrutura	WUDTWNDRYK CUNHW TOUZW
1 - Reuniao	ASA NORTE	Ministerio da Ciencia e Tecnologia	WUOL UWRUOTW NOTO TOUTO
1 - Reuniao	ASA NORTE	B Hotel Brasilia, Asa Norte	WUOL UWRUOTW NOTO TOUTO
1 - Reuniao	BRASILIA	Superquadra Norte 411, Asa Norte	WUOL UWRUOTW NOTO TOUTO
1 - Reuniao	ASA SUL	Ministerio da Saude - Bloco G	WUOL DW TPLVW MUNPZ
1 - Reuniao	ASA NORTE	SGO Q 1 Ae	WUPGWPR WPWROCPDW TWNTOT
1 - Reuniao	SUDOESTE	Anexo do Ministerio da Saude	WUPGWPR WPWROCPDW TWNTOT
1 - Reuniao	BRASILIA	B Hotel Brasilia, Asa Norte	WUPLPO WUGUTTO MWPW PPNTO
1 - Reuniao	ASA NORTE	B Hotel Brasilia, Asa Norte	WUPLPO WUGUTTO MWPW PPNTO
1 - Reuniao	ASA SUL	Ministerio da Cidadania - GM/MC	WUPMWOL RPUOPRO DW TPLVW
3 - Visita	ASA NORTE	SGO Q 1 Ae	WUNOR DW TPLVW TOUZW
1 - Reuniao	ASA NORTE	Venancio Shopping, Setor Comercial	WUNOR DW TPLVW TOUZW
1 - Reuniao	Z. CIVICO	Instituto Nacional do Seguro Social	WUNOR DW TPLVW TOUZW
1 - Reuniao	BRASILIA	SQN 405 Bl. D, Asa Norte	WURWWO UPLLY VPLW FLOR
1 - Reuniao	ASA SUL	Imprensa Nacional Ind Graficas	JULPWNW GRWNDO POUTW

Analisando o resultado apresentado na tabela 5.1 em que foi definido na etapa de classificação

do tipo de dado que o campo "Nome passageiro"deveria ser anonimizado utilizando a técnica de embaralhamento, é possível confirmar que o resultado após o processo de anonimização não permite mais a associação direta ou indireta das informações contidas nesse campo a um indivíduo. No entanto, ainda é possível a utilização das demais informações como os campos "motivo", "origem"e "destino"em texto claro, sendo essa uma vantagem do processo de anonimização aplicado apenas ao campo que contém dados pessoais.

Quando os dados apresentados na tabela 3.3 foram utilizados para a aplicação das etapas do framework descritas na seção 4.1 o resultado de uma base de dados anonimizada pode ser verificado na Tabela 5.2 que segue abaixo:

Tabela 5.2: Amostra com dados anonimizados de passageiros do serviço Taxigov . (Fonte: Autor)

Nome passageiro	**Cpf solicitante**	**Data Nascimento**	**Sexo**
DOBGLZT BORGWT DW TOBTZ	***.425.428-**	> 60	M
ZNTONPZNNP ZRZBJO DW TOBTZ	***.987.231-**	> 50 < 60	M
WLPZWTW MWPRWLWT DZ COTTZ	***.378.031-**	> 50 < 60	F
JBLPZNZ GRZNDW POBTZ FPDWLPT	***.771.511-**	> 40 < 50	F
ZNZMZRPZ DZNDRWZ CORBOZ	***.107.112-**	> 50 < 60	F
GPLMZR ZNTONPO ZLVWT DW TOBTO	***.039.451-**	> 40 < 50	M
ZBDTZNDRYK CBNHZ DW TOBZZ	***.930.011-**	50 < 60	M
ZBWL BZRBOTZ NWTO TOBTO	***.356.801-**	50 < 60	M
ZDZPLTON LOPWT TWPXWPRZ	***.686.041-**	> 40 < 50	M
ZBWL DZ TPLVZ MBNPZ	***.165.457-**	> 40 < 50	M
ZBPGZPR ZPZRWCPDZ DOT TZNTOT	***.369.406-**	> 50 < 60	F
ZDWNPLDZ MZRPZ DW ZRZBJO	***.792.611-**	> 40 < 50	F
ZBPLPO ZBGBTTO MZPZ PPNTO	***.138.495-**	> 40 < 50	M
ZBPMZWL RPBWPRO DZ TPLVZ	***.260.201-**	> 60	M
ZBNWR DZ TPLVZ TOBZZ	***.562.732-**	> 50 < 60	M
ZBRZZO BPLLY VPLZ FLOR DORPZ	***.729.935-**	> 60	M
ZDP BZLBPNOT JBNPOR	***.692.621-**	> 60	M
ZBTZP DW TOBTZ CZMZRGO	***.935.861-**	> 60	M
ZDWMPR LZPZ	***.372.629-**	> 60	M
ZCZBZ BROCHZDO	***.448.808-**	> 50 < 60	M

Analisando o resultado apresentado na tabela 5.2 podemos verificar a anonimização dos campos "Nome passageiro", "cpf solicitante"e data de nascimento. Para cada campo a informação contida foi anonimizada utilizando uma diferente técnica de anonimização, sempre com o objetivo de proteger a informação e evitar que essa seja associada a algum indivíduo, mas procurando manter o potencial de utilização da informação.

No caso do campo "cpf solicitante"a utilização da técnica de mascaramento permite ainda a visualização de parte do cpf, o que pode fazer com que a informação ainda tenha utilização para consultas, validações ou filtro dessas informações.

Para o campo "data nascimento' a técnica aplicada foi a de generalização, com isso não é possível verificar a data de nascimento exata do indivíduo mas a informação ainda é útil para

consultas, filtros e geração de relatórios com informação de fazia etária dos usuários do serviço taxigov.

5.2 RESULTADO DO PROCESSO DE CRIPTOGRAFIA DE ARQUIVOS

As etapas apresentadas no framework descritas na seção 4.2 têm o objetivo de realizar a criptografia do arquivo completo, e dessa maneira protegendo o conjunto de dados de maneira completa. O processo de cifração utilizado é o fornecido de maneira nativa pelo provedor de nuvem pública. O resultado esperado nesse modelo de proteção de dados é a geração de um arquivo cifrado utilizando o arquivo original com os dados em claro.

Nesse estudo de caso os arquivos originais são disponibilizados em formato padrão csv (comma-separated values), o arquivo pode ser aberto e manipulado pelo usuário em texto claro, sendo que a proteção dos dados utilizando a técnica de cifração ocorre no momento de realização da transferência do arquivo (upload) para a estrutura de armazenamento em nuvem, nesse processo a chave criptográfica é fornecida pela aplicação cliente e armazenada na infraestrutura local, sendo que a transferência do arquivo é finalizada já com os dados devidamente cifrados. O processo realizado é representado na Figura 5.1.

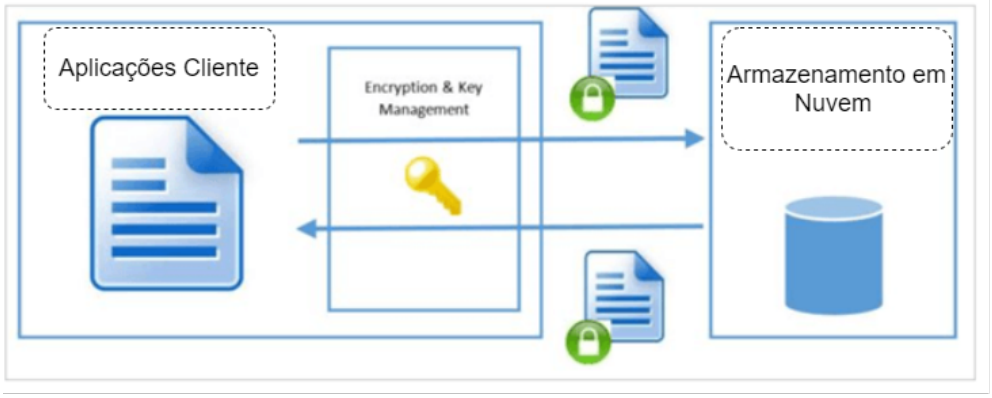

Figura 5.1: Processo de Proteção de dados com Criptografia - cliente

Considerando o processo representado na Figura 5.1, é possível perceber que o arquivo permanece em texto claro enquanto utilizado pela aplicação cliente, e o processo de anonimização utilização a criptografia do arquivo ocorre antes da ação de transferência do arquivo para o provedor de nuvem pública.

O processo de cifração do arquivo descrito na Figura 5.1 ocorre de maneira transparente para o usuário, pois é utilizado um serviço nativo de criptografia oferecido pelo próprio provedor de infraestutura de nuvem pública. Existe, no entanto, uma segunda maneira de execução do processo de cifração, em que ocorre a utilização da capacidade de processamento no ambiente e infraestrutura fornecida pelo provedor para execução desse processo, conforme representado na Figura 5.2.

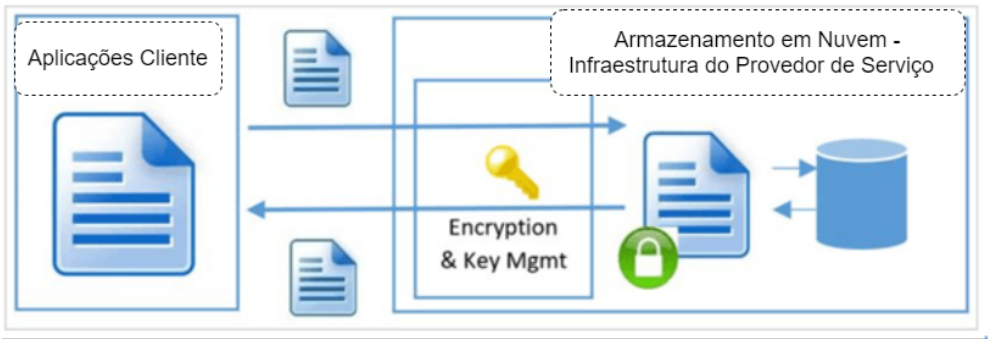

Figura 5.2: Processo de Proteção de dados com Criptografia - servidor

O segundo modelo de execução do processo pode ser uma decisão de projeto quando os arquivos e dados a serem criptografados exigirem uma capacidade maior de processamento para que ocorra com um desempenho adequado.

Neste estudo o resultado obtido foi através da execução do processo representado na figura 5.1 e considerando que a amostra utilizada nesse estudo foi de volume reduzido o desempenho do processo de cifração foi satisfatório.

O resultado de processo de cifração gera um arquivo que quando acessado na infraestrutura do provedor não permite o entendimento em razão do processo de alteração das informações através do processo matemático realizado pelo algoritmo criptográfico. A figura 5.3 demonstra uma representação do resultado obtido.

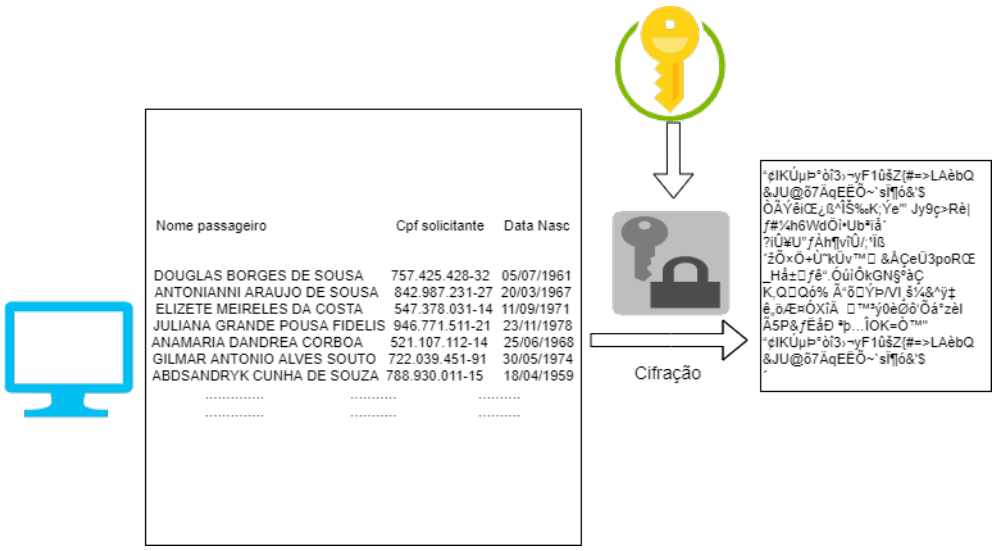

Figura 5.3: Representação gráfica do processo de cifração

É possível verificar que a amostra, inicialmente em texto claro, ao passar pelo processo de cifração tem todas as suas informações anonimizadas, diferente do processo de anonimização de cada campo e tipo de informação, nesse processo o resultado gerado inviabiliza a visualização de todas as informações contidas no arquivo.

A análise do resultado apresentado permite confirmar a validade desse modelo de proteção

dados, pois inviabiliza a a visualização das informações sem que o arquivo passe pelo processo de decifração das informações. E, conforme descrito do framework proposto, esse processo ocorre com a utilização da chave de cifração definida pelo cliente da solução. Dessa maneira, ainda que o modelo de criptografia seja executado pelo provedor de nuvem pública e, ainda, a definição do algoritmo criptográfico também fique a critério do provedor, a segurança de que o processo seja realizado de maneira que apenas o cliente tenha acesso ao texto claro desses arquivos é estruturada e garantida através do sigilo da chave criptográfica utilizada no processo.

Conforme descrito no framework apresentado o resultado desse processo foi obtido através da execução do processo criptográfico com a aplicação do algoritmo AES (Advanced Encryption Standard) que permite a utilização de uma chave criptográfica de até 256 bits o que eleva de maneira exponencial a dificuldade de decifração da informação criptografa sem o acesso ou conhecimento da chave de cifração utilizada. Com isso, o modelo de armazenamento dessas chaves se torna um ponto de importância significativo nesse processo.

na Figura 5.4 que segue abaixo, é apresentado o modelo de decifração dos dados cifrados previamente ao armazenamento. O Processo ocorre de maneira similar ao processo descrito na Figura 5.3 pois o processo, conforme detalhado anteriormente, ocorre com a utilização da mesma chave criptográfica, com objetivo de que o processo ocorra de maneira transparente ao usuário. Um aspecto importante observado nesse processo é que por se tratar de um processo de decifração no arquivo completo e não por campos de informações, após a conclusão do processo todos as informações ficam em texto claro o que torna o processo com características diversa do processo de anonimização conforme o tipo de dado a ser processado.

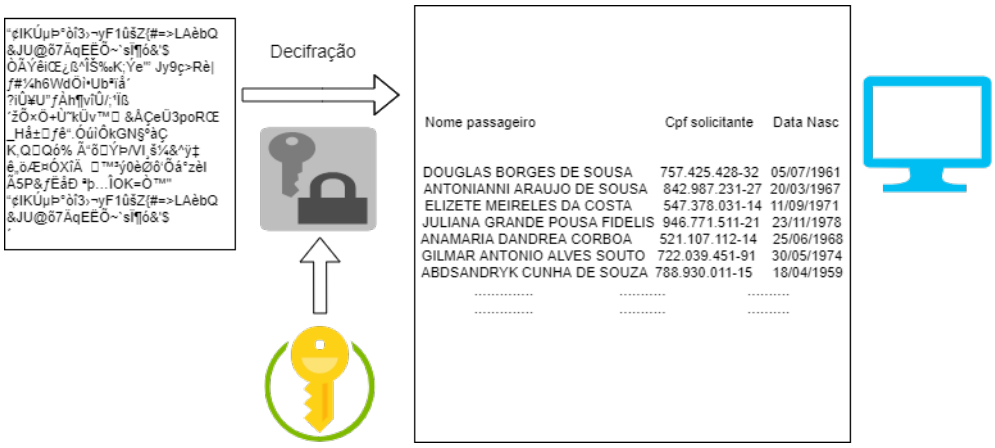

Figura 5.4: Representação gráfica do processo de decifração

Dessa maneira, a segurança e efetiva proteção dos dados armazenados ficam estruturadas na confiabilidade do algoritmo criptográfico utilizado e na proteção da chave de cifração utilizada no processo. Considerando esse aspecto, e conforme descrito no capítulo 2 desse estudo, o processo de cifração de informações pode ser considerado um processo de anonimização de dados caso o processo de reversão seja de grande complexidade [55] e a chave criptográfica utilizada no processo fique armazenada em local diverso do arquivo a ser protegido [54][55]. Nesse cenário deve ser utilizado um local de armazenamento de chaves criptográficas customizado para essa

finalidade de acesso restrito ao usuário cliente dessas chaves, sem possibilidade de administração do provedor de infraestrutura de nuvem pública.

O armazenamento das chaves criptográficas utilizadas nesse processo é realizado em uma estrutura definida para essa finalidade denominada Azure Key Vault em que as chaves são armazenadas e acessadas apenas pelo cliente da solução. A representação gráfica desse modelo pode ser verifica na Figura 5.5.

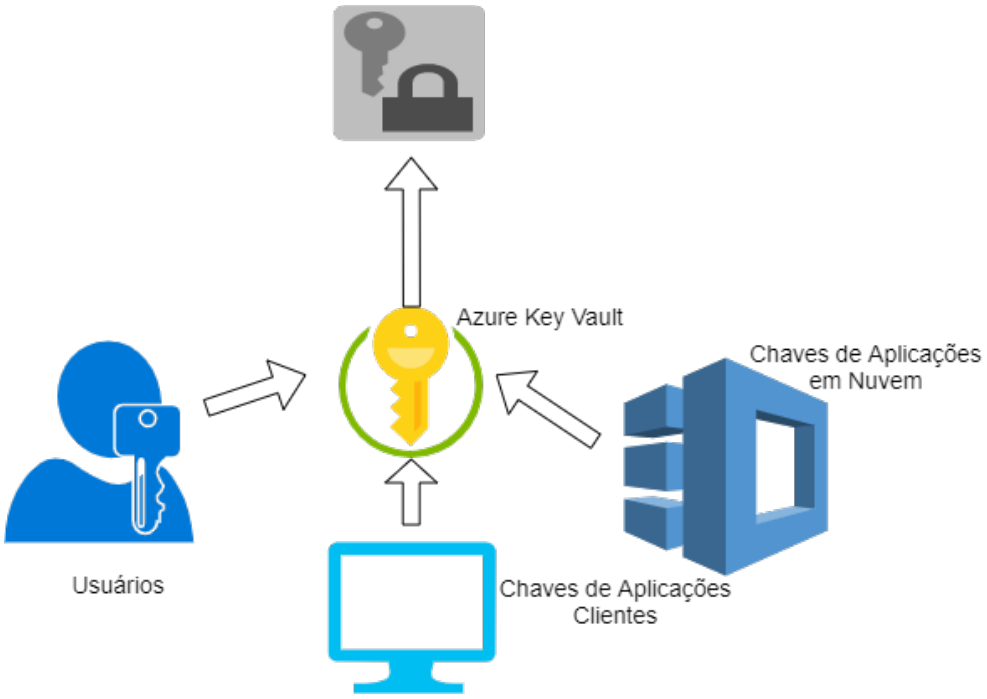

Figura 5.5: Representação gráfica do Azure Key Vault

Ao analisar o diagrama apresentado é importante reforçar o entendimento de que o fornecimento da chave criptográfica a ser utilizada no processo de cifração e a administração dessas chaves no Azure Key Vault são processos realizados pelo cliente da solução, através das aplicações, mesmo que em nuvem, e dos próprios usuários finais da solução.

6 CONCLUSÃO

Os principais objetivos deste estudo foram o de levantamento dos aspectos necessários a serem considerados em uma proposta de proteção de dados, com a apresentação de uma proposta de framework para a aplicação de processos de anonimização de dados, com o intuito de protegê-los, e, ainda, a consideração de um modelo de proteção de dados em nuvem pública.

A necessidade e a importância de proteção de dados se tornaram ainda mais relevante em razão dos aspectos legais decorrentes das recentes normatizações relacionadas à temática da governança de dados. Assim, verifica-se que os aspectos relacionados à governança e proteção apresentam características diversas passando de maneira transversal por diferentes disciplinas, processos e soluções de tecnologia da informação.

Neste estudo, foram abordados a importância de preparação do ambiente e o entendimento da importância da governança de dados para efetiva proteção das informações. Foi apresentado e proposto um framework para a realização efetiva de proteção de dados, considerando a necessidade de proteção adicional de dados pessoais. Em seguida foi realizada a aplicação do framework proposto considerando uma amostra para o estudo.

Foi possível verificar também a importância de combinação das técnicas de anonimização para a eficiente proteção de dados, foi demonstrado que diferentes níveis de proteção de dados podem ser obtidos considerando a sensibilidade da informação e a necessidade de utilização futura destas informações. A possibilidade de utilização de nuvem pública pode ser considerada inclusive para dados pessoais desde que utilizados instrumentos de proteção destas informações similares aos apresentados neste estudo.

Por fim, foi feita a análise dos resultados obtidos para validação do framework proposto, demonstrando a possibilidade de proteção de dados de maneira adequada quando considerado o mapeamento dos dados, a classificação das informações, a definição das técnicas de anonimização a serem utilizadas, e a validação da proteção dessas informações até mesmo para armazenamento em ambiente de nuvem pública.

6.1 TRABALHOS FUTUROS

O processo de proteção de dados envolve um conjunto diverso de etapas a serem aplicadas, envolvendo melhoria de processos e implementação de melhores práticas no ambiente de TI do cliente ou do provedor de infraestrutura de nuvem pública. Além da necessidade de evolução e maturidade dos processos de governança de dados nas instituições.

Aprofundar os mecanismos de proteção de dados em nuvem pública é interessante para validação de performance e eficiência do processo de anonimização de dados para grandes volumes

de informações e, ainda, seria interessante estudos que aprofundem a validação dos mecanismos de gerenciamento e proteção de grandes volumes de chaves criptográficas criadas pelo cliente da solução.

O próprio conceito de utilização de nuvem pública precisa ser mais explorado para que os ganhos relacionados à disponibilidade, à escalabilidade e, aos custos sejam estudados, porém sem que esses aspectos impliquem em uma redução de segurança de informações ou da proteção dos dados armazenados. Para isso, aspectos como melhores práticas, soluções tecnológicas e legislações relacionadas à proteção de dados precisam ser consideradas nestes estudos.

REFERÊNCIAS BIBLIOGRÁFICAS

1 YAYLA, A. A.; LEI, Y. Information security policies and value conflict in multinational companies. *Inf. Comput. Secur.*, v. 26, n. 2, p. 230–245, 2018. Disponível em: <https://doi.org/10.1108/ICS-08-2017-0061>.

2 FONTES, E. L. G. *Segurança da informação*. [S.l.]: Saraiva Educação SA, 2017.

3 SANTOS, E. E. dos; SOARES, T. M. M. K. Riscos, ameaças e vulnerabilidades: O impacto da segurança da informação nas organizações. *Revista Tecnológica da Fatec Americana*, v. 7, n. 02, p. 43–51, 2019.

4 GALEGALE, N. V.; FONTES, E. L. G.; GALEGALE, B. P. Uma contribuição para a segurança da informação: um estudo de casos múltiplos com organizações brasileiras. *Perspectivas em Ciência da Informação*, SciELO Brasil, v. 22, n. 3, p. 75–97, 2017.

5 REPúBLICA, P. da. *Lei Geral de Proteção de Dados Pessoais (LGPD)*. 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>.

6 RIBEIRO, R. C.; CANEDO, E. D. Using MCDA for selecting criteria of LGPD compliant personal data security. In: *DG.O*. [S.l.]: ACM, 2020. p. 175–184.

7 PINHEIRO, P. P. *Proteção de Dados Pessoais: Comentários à Lei n. 13.709/2018-LGPD*. [S.l.]: Saraiva Educação SA, 2020.

8 MACEDO, P. N. Brazilian general data protection law (lgpd). *Nartional Congress, accessed in Febreary 12, 2022*, 2018. Disponível em: <https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf>.

9 RHAHLA, M.; ALLEGUE, S.; ABDELLATIF, T. Guidelines for GDPR compliance in big data systems. *J. Inf. Secur. Appl.*, v. 61, p. 102896, 2021. Disponível em: <https://doi.org/10.1016/j.jisa.2021.102896>.

10 CARVALHO, A. P.; CARVALHO, F. P.; CANEDO, E. D.; CARVALHO, P. H. P. Big data, anonymisation and governance to personal data protection. In: *DG.O*. [S.l.]: ACM, 2020. p. 185–195.

11 CLOUD Security Alliance Big Data Working Group. Expanded top ten big data security and privacy challenges. 2013. Disponível em: <https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf>.

12 FERNANDES, M. A. de S.; OLIVEIRA, F. G. de; FERRAZ, F. S.; SILVA, D. A. da; CANEDO, E. D.; JR, R. T. de S. Impactos da lei de proteção de dados (lgpd) brasileira no uso da computação em nuvem. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informacao, n. E42, p. 374–385, 2021.

13 STARCHON, P.; PIKULÍK, T. GDPR principles in data protection encourage pseudonymization through most popular and full-personalized devices - mobile phones. In: SHAKSHUKI, E. M.; YASAR, A. (Ed.). *The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops, April 29 - May 2, 2019, Leuven, Belgium*. Elsevier, 2019. (Procedia Computer Science, v. 151), p. 303–312. Disponível em: <https://doi.org/10.1016/j.procs.2019.04.043>.

14 CANEDO, E. D.; CERQUEIRA, A. J.; GRAVINA, R. M.; RIBEIRO, V. C.; CAMÕES, R.; REIS, V. E. dos; MENDONÇA, F. L. L. de; JR., R. T. de S. Proposal of an implementation process for the brazilian general data protection law (LGPD). In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS 2021, Online Streaming, April 26-28, 2021, Volume 1*. SCITEPRESS, 2021. p. 19–30. Disponível em: <https://doi.org/10.5220/0010398200190030>.

15 CARVALHO, A. P.; CANEDO, E. D.; CARVALHO, F. P.; CARVALHO, P. H. P. Anonymisation and compliance to protection data: Impacts and challenges into big data. In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Proceedings of the 22nd International Conference on Enterprise Information Systems, ICEIS 2020, Prague, Czech Republic, May 5-7, 2020, Volume 1*. SCITEPRESS, 2020. p. 31–41. Disponível em: <https://doi.org/10.5220/0009411100310041>.

16 BRASHER, E. A. Addressing the failure of anonymization: Guidance from the european union's general data protection regulation. *Colum. Bus. L.*, v. 1, n. 209, 2018.

17 BLUM, R. O.; VAINZOF, R.; MORAES, H. F. *Data Protection Officer*. [S.l.]: Revista dos Tribunais, 2020.

18 BLOK, M. *Compliance e governança corporativa*. [S.l.]: Freitas Bastos, 2020.

19 LESSA, A. P. Proteção de dados pessoais: um plano viável de adequação da governança de dados à lgpd em empresas de pequeno porte. *Tecnologia em Gestão da Tecnologia da Informação-Unisul Virtual*, 2020.

20 LUCA Piras, Mohammed Ghazi Al-Obeidallah, Andrea Praitano, Aggeliki Tsohou, Haralambos Mouratidis, Beatriz Gallego-Nicasio Crespo, Jean Baptiste Bernard, Marco Fiorani, Emmanouil Magkos, Andres Castillo Sanz. DEFeND Architecture: A Privacy by Design Platform for GDPR Compliance. In International Conference on Trust and Privacy in Digital Business. *In International Conference on Trust and Privacy in Digital Business. AUSTRIA*, n. 78-93, 2019.

21 REPúBLICA, P. da. Lei geral de acesso a informação (lai). *Secretaria-Geral, Accessed in February 04, 2020*, 2011. <http://www.mpf.mp.br/atuacao-tematica/sci/normas-e-legislacao/legislacao/legislacao-em-ingles/law-12.527>.

22 ISO Central Secretary. *Information technology — Security techniques — Information security management systems — Requirements*. Geneva, CH, 2013. Disponível em: <https://www.iso.org/standard/54534.html>.

23 ISO Central Secretary. *Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines*. Geneva, CH, 2019. Disponível em: <https://www.iso.org/standard/71670.html>.

24 NEALE, C.; KENNEDY, I.; PRICE, B. A.; NUSEIBEH, B. The case for zero trust digital forensics. *CoRR*, abs/2202.02623, 2022. Disponível em: <https://arxiv.org/abs/2202.02623>.

25 SYED, N. F.; SHAH, S. W.; SHAGHAGHI, A.; ANWAR, A.; BAIG, Z. A.; DOSS, R. Zero trust architecture (ZTA): A comprehensive survey. *IEEE Access*, v. 10, p. 57143–57179, 2022. Disponível em: <https://doi.org/10.1109/ACCESS.2022.3174679>.

26 UEHARA, M. Zero trust security in the mist architecture. In: BAROLLI, L.; YIM, K.; ENOKIDO, T. (Ed.). *Complex, Intelligent and Software Intensive Systems - Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021), Asan, Korea, 1-3 July 2021*. Springer, 2021. (Lecture Notes in Networks and Systems, v. 278), p. 185–194. Disponível em: <https://doi.org/10.1007/978-3-030-79725-6_18>.

27 SALTARELLA, M.; DESOLDA, G.; LANZILOTTI, R. Privacy design strategies and the GDPR: A systematic literature review. In: MOALLEM, A. (Ed.). *HCI for Cybersecurity, Privacy and Trust - Third International Conference, HCI-CPT 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24-29, 2021, Proceedings*. Springer, 2021. (Lecture Notes in Computer Science, v. 12788), p. 241–257. Disponível em: <https://doi.org/10.1007/978-3-030-77392-2_16>.

28 REGULATION, G. D. P. *EU data protection rules*. 2018. Disponível em: <https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en>.

29 NEUMANN, G.; GRACE, P.; BURNS, D.; SURRIDGE, M. Pseudonymization risk analysis in distributed systems. *J. Internet Serv. Appl.*, v. 10, n. 1, p. 1:1–1:16, 2019. Disponível em: <https://doi.org/10.1186/s13174-018-0098-z>.

30 RAHMADIKA, S.; FIRDAUS, M.; LEE, Y.; RHEE, K. An investigation of pseudonymization techniques in decentralized transactions. *J. Internet Serv. Inf. Secur.*, v. 11, n. 4, p. 1–18, 2021. Disponível em: <https://doi.org/10.22667/JISIS.2021.11.30.001>.

31 Microdata anonymization techniques. In: TILBORG, H. C. A. van; JAJODIA, S. (Ed.). *Encyclopedia of Cryptography and Security, 2nd Ed*. Springer, 2011. p. 778. Disponível em: <https://doi.org/10.1007/978-1-4419-5906-5_1288>.

32 STRANSKY, C. *Challenges in using cryptography - End-user and developer perspectives*. Tese (Doutorado) — University of Hanover, Hannover, Germany, 2022. Disponível em: <https://www.repo.uni-hannover.de/handle/123456789/12204>.

33 MAJEED, A.; HWANG, S. O. A practical anonymization approach for imbalanced datasets. *IT Prof.*, v. 24, n. 1, p. 63–69, 2022. Disponível em: <https://doi.org/10.1109/MITP.2021.3132330>.

34 MAJEED, A.; LEE, S. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*, v. 9, p. 8512–8545, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2020.3045700>.

35 ISO Central Secretary. *ABNT NBR ISO/IEC 17799:2005*. Geneva, CH, 2005.

36 MENEZES, A.; STEBILA, D. The advanced encryption standard: 20 years later. *IEEE Secur. Priv.*, v. 19, n. 6, p. 98–102, 2021. Disponível em: <https://doi.org/10.1109/MSEC.2021.3107078>.

37 GREENBERG, C. S.; MASON, L. P.; SADJADI, S. O.; REYNOLDS, D. A. Two decades of speaker recognition evaluation at the national institute of standards and technology. *Comput. Speech Lang.*, v. 60, 2020. Disponível em: <https://doi.org/10.1016/j.csl.2019.101032>.

38 BRAEKEN, A. Public key versus symmetric key cryptography in client-server authentication protocols. *Int. J. Inf. Sec.*, v. 21, n. 1, p. 103–114, 2022. Disponível em: <https://doi.org/10.1007/s10207-021-00543-w>.

39 CALDERONI, L.; MAIO, D.; TULLINI, L. Benchmarking cloud providers on serverless iot back-end infrastructures. *IEEE Internet Things J.*, v. 9, n. 16, p. 15255–15269, 2022. Disponível em: <https://doi.org/10.1109/JIOT.2022.3147860>.

40 GOOGLE. O que é a google cloud. *Google, Último acesso em 31 de Julho de 2022*, 2022. <https://cloud.google.com/docs/get-started?hl=pt-br>.

41 ABURUKBA, R.; KADDOURA, Y.; HIBA, M. Cloud computing infrastructure security: Challenges and solutions. In: *International Symposium on Networks, Computers and Communications, ISNCC 2022, Shenzhen, China, July 19-22, 2022*. IEEE, 2022. p. 1–7. Disponível em: <https://doi.org/10.1109/ISNCC55209.2022.9851812>.

42 ASAD-UR-REHMAN; AGUIAR, R. L.; BARRACA, J. P. Fault-tolerance in the scope of cloud computing. *IEEE Access*, v. 10, p. 63422–63441, 2022. Disponível em: <https://doi.org/10.1109/ACCESS.2022.3182211>.

43 SAUBER, A. M.; ELKAFRAWY, P. M.; SHAWISH, A. F.; AMIN, M.; HAGAG, I. M. A new secure model for data protection over cloud computing. *Comput. Intell. Neurosci.*, v. 2021, p. 8113253:1–8113253:11, 2021. Disponível em: <https://doi.org/10.1155/2021/8113253>.

44 ALNAJRANI, H. M.; NORMAN, A. A. The effects of applying privacy by design to preserve privacy and personal data protection in mobile cloud computing: An exploratory study. *Symmetry*, v. 12, n. 12, p. 2039, 2020. Disponível em: <https://doi.org/10.3390/sym12122039>.

45 NADEEM, F. Evaluating and ranking cloud iaas, paas and saas models based on functional and non-functional key performance indicators. *IEEE Access*, v. 10, p. 63245–63257, 2022. Disponível em: <https://doi.org/10.1109/ACCESS.2022.3182688>.

46 CARVALHO, A. P.; CANEDO, E. D.; CARVALHO, F. P.; CARVALHO, P. H. P. Anonimisation, impacts and challenges into big data: A case studies. In: FILIPE, J.; SMIALEK, M.; BRODSKY, A.; HAMMOUDI, S. (Ed.). *Enterprise Information Systems - 22nd International Conference, ICEIS 2020, Virtual Event, May 5-7, 2020, Revised Selected Papers*. Springer, 2020. (Lecture Notes in Business Information Processing, v. 417), p. 3–23. Disponível em: <https://doi.org/10.1007/978-3-030-75418-1_1>.

47 GUNAWAN, D.; MAMBO, M. Data anonymization for hiding personal tendency in set-valued database publication. *Future Internet*, v. 11, n. 6, p. 138, 2019. Disponível em: <https://doi.org/10.3390/fi11060138>.

48 PRASSER, F.; EICHER, J.; SPENGLER, H.; BILD, R.; KUHN, K. A. Flexible data anonymization using ARX - current status and challenges ahead. *Softw. Pract. Exp.*, v. 50, n. 7, p. 1277–1304, 2020. Disponível em: <https://doi.org/10.1002/spe.2812>.

49 FEDERAL, G. Portal brasileiro de dados abertos. *GOV BR, Último acesso em 31 de Julho de 2022*, 2022. <https://dados.gov.br>.

50 FEDERAL, G. O que é o taxigov. *GOV BR, Último acesso em 31 de Julho de 2022*, 2022. <https://www.gov.br/economia/pt-br/assuntos/gestao/central-de-compras/taxigov>.

51 SOURCE open. *ARX Anonymization tool - overview*. 2023. Disponível em: <https://arx.deidentifier.org/overview/>.

52 PRASSER, F.; EICHER, J.; SPENGLER, H.; KUHN, K. A. Flexible data anonymization using arx — current status and challenges ahead. *J Software Pract Exper 50, 7 (2020)*, v. 7, p. 1277–1304, 2020.

53 GUNDU, S. R.; PANEM, C. A.; THIMMAPURAM, A. The dynamic computational model and the new era of cloud computation using microsoft azure. *SN Comput. Sci.*, v. 1, n. 5, p. 264, 2020. Disponível em: <https://doi.org/10.1007/s42979-020-00276-y>.

54 DONEDA DANILO; MACHADO, D. Proteção de dados pessoais e criptografia: tecnologias criptográficas entre anonimização e pseudonimização de dados. *Revista dos Tribunais Caderno Especial*, v. 998, n. 99-128, 2018.

55 SPINDLER, G.; SCHMECHEL, P. Personal data and encryption in the european general data protection regulation. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, v. 7, n. 1, 2016.

APÊNDICES