UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

DANILO FIGUEIREDO DE OLIVEIRA

Proteção de privacidade de dados em ambiente de *big data analytics*: um estudo da realidade brasileira

São Paulo
2023

DANILO FIGUEIREDO DE OLIVEIRA

Proteção de privacidade de dados em ambiente de *big data analytics*: um estudo da realidade brasileira

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 08 de dezembro de 2022. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Edmir Parada Vasques Prado

São Paulo

2023

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Figueiredo de Oliveira, Danilo
Proteção de privacidade de dados em ambiente de big data analytics: um estudo da realidade brasileira / Danilo Figueiredo de Oliveira; orientador, Edmir Parada Vasques Prado. -- São Paulo, 2023.
136 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2023.
Versão corrigida

1. Privacidade de dados. 2. Problemas de privacidade. 3. Privacidade em big data analytics. 4. Big data analytics. 5. Proteção de dados. I. Prado, Edmir Parada Vasques, orient. II. Título.

Dissertação de autoria de Danilo Figueiredo de Oliveira, sob o título **"Proteção de privacidade de dados em ambiente de *big data analytics*: um estudo da realidade brasileira"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 08 de dezembro de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Edmir Parada Vasques Prado
Universidade de São Paulo
Presidente

Profa. Dra. Gisele da Silva Craveiro
Universidade de São Paulo

Prof. Dr. Gilberto Perez
Universidade Presbiteriana Mackenzie

Aos meus pais que sempre priorizaram a minha educação. Que sempre me deram apoio, conselhos e amor. Inspiraram a minha formação como cidadão e a fundamentação dos valores que carrego comigo. Não há palavras que façam jus à admiração, gratidão e amor que sinto por eles, mas que esta singela dedicatória seja nota desses sentimentos. Ao meu orientador Edmir, também meu professor em três disciplinas na graduação, pelos ensinamentos e orientações, pelo compromisso com a excelência, e pela dedicação à docência e à pesquisa.

Agradecimentos

Primeiramente, a Universidade de São Paulo, em especial, a Escola de Artes Ciências e Humanidades, por ter propiciado um ambiente de excelência de ensino e pesquisa desde a minha graduação, no qual adquiri competências que me fizeram evoluir como cidadão e profissional, e fortaleci a valorização e paixão pela ciência e tecnologia.

A todos os painelistas que se dispuseram a participar desta pesquisa por meio da técnica Delphi, que, em meio a tantos compromissos e responsabilidades cotidianas, gentilmente dedicaram tempo para contribuir com seus conhecimentos a esta pesquisa.

"Better be despised for too anxious apprehensions than ruined by too confident security."

(Edmund Burke)

Resumo

OLIVEIRA, Danilo Figueiredo. **"Proteção de privacidade de dados em ambiente de *big data analytics*: um estudo da realidade brasileira"**. 2023. 136 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2023.

Privacidade é reconhecida internacionalmente como um direito humano fundamental e tem sido um tema cada vez mais importante ao passo que a humanidade aumenta o uso de produtos e serviços digitais e, consequentemente, a geração e utilização de dados pessoais. A quantidade de dados gerados a cada momento é enorme e exige tecnologias específicas para que sejam coletados, armazenados e disponibilizados para tomada de decisão, o que foi denominado como sistemas de *big data analytics* (BDA). Nesse cenário, diversos problemas de privacidade de dados podem ocorrer, e as organizações ainda sofrem para mitigar os riscos de violação de privacidade. Governos e organizações internacionais têm criado regulamentos para evitar e punir abusos por parte das organizações aos indivíduos, mas não têm sido suficientes para evitar grandes vazamentos de dados, entre outros problemas de privacidade de dados. Assim, o objetivo deste estudo foi analisar os problemas de privacidade de dados no contexto de BDA, bem como as suas causas, e identificar as principais ações e práticas que podem ser adotadas para evitar, minimizar ou resolver esses problemas identificados a partir de uma revisão sistemática da literatura. Para tanto, adotou-se como parâmetro as avaliações de especialistas coletadas por meio da técnica Delphi, considerando a eficiência, eficácia e factibilidade das soluções propostas pela literatura. Como resultado da aplicação da técnica Delphi, verificou-se concordância forte ou muito forte em 10 das 14 dimensões avaliadas, o que implica em confiabilidade alta ou muito alta no ranqueamento dos nove problemas em relação a cinco dos sete conjuntos de causas, e no ranqueamento dos 10 conjuntos de solução em relação a cinco conjuntos de causas. Portanto, conclui-se que, dado os sete conjuntos de causas, foram identificados os principais problemas causados por esses conjuntos e as melhores soluções para mitigar esses conjuntos de causas. De forma geral, dentre os nove problemas de privacidade de dados em BDA "roubo ou acesso não autorizado a dados" foi o principal, segundo os especialistas, seguido por "fraudes e outros crimes". Os principais conjuntos de causas desses nove problemas foram "ataques e vulnerabilidade de segurança" e "revelação ou inferência de dados não autorizados". Os principais conjunto de soluções para os sete conjunto de causas foram "governança de dados" e "políticas internas de proteção de privacidade".

Palavras-chaves: Privacidade de dados. Problemas de privacidade. Privacidade em *big data analytics*. *Big data analytics*. Proteção de dados.

Abstract

OLIVEIRA, Danilo Figueiredo. **Data privacy protection in big data analytics environment: a study of the Brazilian reality**. 2023. 136 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2023.

Privacy is internationally recognized as a fundamental human right and has become an increasingly important issue as humanity increases the use of digital products and services and, consequently, the generation and use of personal data. The amount of data generated at each moment is huge and requires specific technologies to be collected, stored and made available for decision-making, which has been called big data analytics (BDA) systems. In this scenario, several data privacy problems can occur, and organizations still struggle to mitigate the risks of privacy breaches. Governments and international organizations have created regulations to prevent and punish abuse by organizations against individuals, but it has not been enough to prevent large data leaks, among other data privacy problems. Thus, the objective of this study was to analyze data privacy problems in the context of BDA, as well as their causes, and to identify the main actions and practices that can be adopted to avoid, minimize or solve these problems identified from a systematic literature review. For this purpose, the expert assessments collected using the Delphi technique were adopted as parameters, considering the efficiency, effectiveness and feasibility of the solutions proposed by the literature. As a result of applying the Delphi technique, there was strong or very strong agreement in 10 of the 14 dimensions evaluated, which implies high or very high reliability in the ranking of the nine problems in relation to five of the seven sets of causes, and in the ranking of the 10 sets of solution in relation to five sets of causes. Therefore, it is concluded that, given the seven sets of causes, the main problems caused by these sets and the best solutions to mitigate these sets of causes were identified. Overall, among the nine data privacy issues in BDA, "theft or unauthorized access to data" was the main one, according to the experts, followed by "fraud and other crimes against victims", and "unfeasibility to maintain anonymous". The main sets of causes of these nine problems were "attacks and security vulnerability", and "inference or disclosure of unauthorized data". The two main sets of solutions for the seven sets of causes were "data governance", and "internal privacy protection policies".

Keywords: Data privacy. Privacy issues. Privacy in big data analytics. Big data analytics. Data protection.

Lista de figuras

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

AD	Análise de dados
AES	Advanced Encryption Standard
ANPD	Autoridade Nacional de Proteção de Dados
BDA	Big Data Analytics
BI	Business intelligence
BPM	Business process management
CD	Ciência de dados
CE	Critério de exclusão
CI	Critério de inclusão
CIA	Confidentiality, integrity and availability
DMBOK	Data Management Body of Knowledge
DNC	Do Not Collect
DNT	Do Not Track
DPO	Data Privacy Officer
DW	Data warehouse
ED	Engenharia de dados
ECHR	European Convention on Human Rights
ETL	Extraction, transformation and load
EUA	Estados Unidos da América
GD	Governança de dados
HDFS	Hadoop Distributed File System
IA	Inteligência artificial

IaaS	Infrastructure as a Service
LGPD	Lei Geral de Proteção de Dados Pessoais
NDA	Non disclosure agreement
NSA	National Security Agency
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
OLAP	Online analytical processing
ONU	Organização das Nações Unidas
PaaS	Platform as a Service
PII	Personally Identifiable Information
R1	Primeira rodada
R2	Segunda rodada
RSL	Revisão sistemática da literatura
SaaS	Software as a Service
SGBD	Sistema gerenciador de banco de dados
SI	Sistemas de informação
SQL	Structured Query Language
SPI	Sensitive Personal Information
SSL	Secured Socket Layer
VPN	Virtual Private Network
W	Coeficiente de concordância de Kendall

Sumário

1 Introdução

Atualmente a humanidade vive uma era de transformações tecnológicas rápidas e constantes, principalmente quando se refere à análise e ao tratamento de dados e informações, o que alguns autores como Wu *et al.* (2014) e Kitchin (2014) classificam como uma verdadeira revolução. De fato, a sociedade moderna vem adotando inovações de forma acelerada desde o século 19. Como apontam Müller *et al.* (2016), entre 1800 e 1840 uma inovação levava em torno de 66 anos para ter o uso massificado. Esse tempo de adoção caiu para 50 anos entre 1840 a 1880, 27 anos entre 1880 a 1920, e atualmente pode levar menos de dois anos para atingir a casa da centena de milhões de usuários, como o jogo *Candy Crush*, lançado em 2012 (DREISCHMEIER; CLOSE; TRICHET, 2015).

As consequências dos avanços tecnológicos podem ser positivas ou negativas, e isso depende mais de como se aplica a tecnologia do que a sua mera existência. Por exemplo, câmeras podem ser usadas como ferramenta de segurança ou de repressão autoritária, ou seja, é percebível que a mesma tecnologia pode ser utilizada para fins distintos (DOYLE; LIPPERT; LYON, 2013).

A popularização da internet e dos computadores pessoais têm grande relevância nas transformações tecnológicas recentes, pois atualmente uma parcela considerável das interações humanas acontecem e são registradas por meios informatizados (NORRIS; SOLOWAY, 2009). Isso significa bilhões de pessoas produzindo novos dados em intervalos curtos de tempo, em grande volume e de uma variedade enorme de fontes. Isso é conhecido como os três Vs (velocidade, variedade e volume) que definem o *big data* (KITCHIN, 2014).

Desafios técnicos como armazenamento e processamento de grandes volumes de dados, e dilemas éticos na utilização desses dados, surgem devido às interações dos usuários com sistemas informatizados, que estão constantemente gerando novos dados. Enquanto os desafios técnicos citados anteriormente têm sido resolvidos na academia e na indústria (KITCHIN, 2014), os dilemas éticos continuam com as mesmas preocupações citadas por Conger, Loch e Helft (1995), como transgressões à privacidade, precisão dos dados pessoais (que inclui completude e corretude dos dados em posse de terceiros), propriedade (de dados, ideias, processos, *hardware*, código etc.), e controle de acesso.

Das diversas ferramentas de análise, armazenamento e transformação de dados que surgiram, grande parte delas são utilizadas como componentes de sistemas de tomada de decisões táticas e estratégicas nas organizações, muitas vezes conhecidos como sistemas de *big data analytics* (BDA). As arquiteturas desses sistemas são bastante variadas, mas a sua finalidade geralmente é tirar o máximo de proveito dos dados disponíveis para essa organização (SHAYTURA *et al.*, 2016).

Porém, ocasionalmente, os interesses das organizações podem conflitar com os interesses de privacidade dos usuários, e a partir desse conflito surgem dilemas éticos, como, por exemplo, uma empresa que faz recomendações de publicações que afetam as emoções de seus usuários deliberadamente. Governos de várias regiões do mundo estão regulamentando o uso dos dados pessoais devido às diversas preocupações com privacidade. Embora os próprios governos adotem práticas eticamente questionáveis (NEMITZ, 2018; TERZI; SINANC; SAGIROGLU, 2015).

A classificação de uso correto ou ético dos dados é difusa, mas pressupõe-se que o uso de dados pessoais que sofrem com desvio de finalidade se caracteriza como violação de um princípio de boa-fé (BRASIL, 2018). Ademais, há vários parâmetros a serem considerados para compreender e avaliar os riscos à privacidade de dados (BARKER *et al.*, 2009).

Novos papéis têm surgido na indústria quando se trata de BDA, tais como engenheiro de dados, arquiteto de dados, analista de governança de dados, que se somam a papéis mais antigos como analista de inteligência de negócios. Isso é reflexo da crescente atenção que as organizações têm dado ao uso de seus dados para impulsionar seus negócios. Novas áreas, que exigem novos papéis especializados, são criadas para esse fim. Nessas áreas, usualmente dados de diversas fontes internas e externas são reunidas em um ou alguns repositórios centralizados, nos quais esses dados são transformados e disponibilizados às partes interessadas. Por essa natureza, esses ambientes do contexto de BDA costumam estar no centro da discussão de dados nas organizações, o que inclui proteção à privacidade.

Como relata Wang (2018), apesar dos esforços de várias organizações e profissionais, muitas áreas do conhecimento ainda carecem de estratégias adequadas de proteção de privacidade. Tenha-se a área da saúde como exemplo, na qual dados clínicos poderiam ser utilizados para fins discriminatórios por companhias de seguro e ainda não há consenso sobre as boas práticas de gestão de privacidade. Isto porque alguns autores como Abouelmehdi *et al.* (2017) consideram que a segurança e a privacidade em *big data* são as maiores barreiras para os pesquisadores de dados na área da saúde.

Diante deste contexto, o estudo dos problemas de privacidade se mostra relevante. Considerando a realidade brasileira, que possui baixa competitividade digital, se comparado a países desenvolvidos (IMD World Digital, 2020), o problema se intensifica mais ainda. Corrobora essa afirmação os recentes vazamentos de dados pessoais envolvendo significativa parcela da sociedade brasileira (G1, 2021; CNN Brasil, 2021a; CNN Brasil, 2021b; Olhar Digital, 2021).

1.1 Justificativa

Privacidade é reconhecida como um direito humano fundamental por empresas e instituições. Como exemplo tem-se a Convenção para a Protecção dos Direitos do Homem e das Liberdades Fundamentais (ECHR - *European Convention on Human Rights*) e a Carta dos Direitos Fundamentais da União Europeia (BROEDERS *et al.*, 2017). Por essa razão, diversos governos estão criando regulamentações de uso de dados pessoais, proteção de dados e privacidade. Além disso, a privacidade tem ganhado destaque devido ao uso massivo de dados pessoais por instituições e empresas, que muitas vezes levantam questionamentos a respeito dos limites éticos dessa utilização.

A popularização de algoritmos de aprendizado de máquina e ferramentas de análise de dados levaram as empresas a investirem em coleta, armazenamento e tratamento de dados, inclusive considerando os dados como ativos na avaliação do valor de mercado das empresas. Ou seja, aumentou a percepção de valor dos dados, que tiveram sua utilização aumentada, e isso inclui dados pessoais (WU *et al.*, 2014).

Apesar da validade da discussão ética, para as organizações privadas existe preocupações muito mais pragmáticas, como conformidade legal e risco operacional. No Brasil e na Europa, por exemplo, empresas já podem sofrer reveses legais por mau gerenciamento de privacidade. Vazamento de dados também pode impactar negativamente a imagem de uma companhia, gerando problemas para a própria operação da empresa (SERRADO *et al.*, 2020).

A pesquisa de Hashem *et al.* (2015) aponta que muitas ameaças e problemas de privacidade e confidencialidade existem em BDA, mesmo em plataformas de computação em nuvem. E indicam ao final do estudo que alguns dos maiores desafios e problemas a

serem endereçados pela academia e pela indústria são: proteção de dados, privacidade, questões legais e regulatórias e acesso a dados.

Müller *et al.* (2016) citam algumas práticas de organizações e governos em seus sistemas de BDA que, além de trazerem incerteza a respeito da validade e confiabilidade dos dados, podem levantar preocupações éticas quando os indivíduos não estão cientes que suas pegadas digitais estão sendo analisadas ou mesmo apenas armazenadas, ainda que o uso seja exclusivo para fins de pesquisas científicas.

Assim, se torna perceptível a crescente importância de debater e conciliar o uso eficiente de dados pessoais para fins comerciais com as questões éticas e legais que os envolvem. Segundo Abouelmehdi *et al.* (2017), as preocupações com privacidade e segurança de dados são bastante debatidas e representam o maior risco conhecido por qualquer pessoa familiarizada com BDA.

Por outro lado, há situações em que o compartilhamento de dados é necessário. Schadt *et al.* (2010) citam que dados de saúde coletados de grandes populações podem exigir compartilhamento aberto à sociedade para que modelos preditivos de doenças sejam criados. Esse é um exemplo claro de uma situação na qual a sociedade se beneficiaria da publicidade de dados que são considerados sensíveis de acordo com regulações como a LGPD (BRASIL, 2018). Porém o direito individual de privacidade pode ser violado se não houver a devida anonimização desses dados, podendo até mesmo ser uma transgressão do juramento ético de profissionais da área de saúde ("Juramento de Hipócrates"), ou dos direitos humanos, de acordo com a ECHR.

No estudo de Stahl e Wright (2018), na revisão de 809 publicações que discutiam ética em sistemas de informação (SI), foram encontrados 177 trabalhos que tratavam do tema proteção de dados e privacidade, o que fazia desse assunto o problema mais proeminente.

Apesar disso, como levantado por Singh *et al.* (2018) em uma revisão de 58 publicações científicas revisadas por pares de 2007 a 2016, a proteção de dados é uma fonte de grande preocupação para pesquisadores de todo o mundo. No entanto, o autor conclui que não há pesquisas suficientes na literatura a respeito de como resolver alguns dos principais problemas de confidencialidade e privacidade em BDA, comparações de diferentes soluções parciais propostas, e porque ou como elas podem ser aplicadas.

Como relata Singh *et al.* (2018), se não houver investimento de recursos, tempo e esforço pelas empresas para proteger dados sensíveis, podem surgir problemas sérios de

confidencialidade. Além disso, a fim de obter melhor controle de risco de privacidade em BDA, Wang (2018) conclui que são necessárias mais análises e pesquisas sobre os detalhes específicos da estrutura de gerenciamento ativo de privacidade.

Porém, Ying e Grandison (2017) concluíram a partir do seu modelo de avaliação de riscos de privacidade, que é muito difícil haver privacidade quando se trata de BDA, pois é inviável lidar com seus riscos de maneira efetiva. E diz que espera que a comunidade científica se envolva mais nessa discussão. Joshi e Kadhiwala (2017) também defenderam futuras pesquisas para soluções de problemas de segurança e privacidade em BDA.

Dado esse cenário, acredita-se que, dado um pesquisa bibliográfica das causas e soluções dos problemas de privacidade no contexto de BDA, explorar as melhores práticas para proteger a privacidade do ponto de vista legal, gerencial e técnico é importante para a evolução do conhecimento nessa área e contribuir com um dos mais sérios temas da atualidade em SI.

1.2 Objetivos

Este estudo pretende analisar os problemas de privacidade de dados no contexto de BDA, bem como as suas causas, e identificar as principais ações e práticas que podem ser adotadas para evitar, minimizar ou resolver esses problemas identificados a partir da revisão da literatura. Isso, tendo como parâmetro as avaliações de especialistas coletadas por meio da técnica Delphi, considerando a eficiência, eficácia e factibilidade das soluções propostas pela literatura.

Os objetivos específicos são:

- Identificar os problemas, causas e as soluções de privacidade de dados no contexto de BDA com base na literatura.
- Analisar no contexto brasileiro e com base na opinião de especialistas: os problemas de privacidade de dados; as causas desses problemas; e as soluções de proteção da privacidade de dados.
- Analisar, com base na opinião de especialistas, as relações entre causas e problemas, e entre causas e soluções.

1.3 Escopo da pesquisa

Por delimitação de escopo, esta pesquisa não abrangerá:

- Tópicos de segurança de dados que não tenham impacto direto em privacidade de dados.
- Pesquisas que discutam privacidade de dados no contexto de internet das coisas ou *blockchain* também não serão consideradas no mapeamento da literatura, pois esses tópicos têm desafios específicos que diferem bastante de outras aplicações de *big data*.
- Os aspectos políticos, econômicos, sociais, ambientais e legais serão considerados se, somente se, no mesmo contexto houver relação direta com tecnologia.

1.4 Estrutura da dissertação

As próximas seções deste trabalho estão divididas em capítulos e apêndices, conforme detalhado abaixo.

No Capítulo 2 são apresentados os conceitos fundamentais de privacidade de dados no contexto de BDA. O que inclui a identificação de características de privacidade, além de como identifica-las e classifica-las, e limitações do escopo de *big data*.

No Capítulo 3 é apresentada uma revisão sistemática da literatura (RSL) a fim de identificar o estado da arte dentre os trabalhos correlatos. Ou seja, serão revisados trabalhos, de acordo com alguns critérios detalhados nesse capítulo para identificar-se o que já foi estudado em torno de problemas de privacidade em sistemas de *big data analytics*, suas causas e possíveis soluções.

Em seguida, no Capítulo 4, é apresentado o modelo de referência da pesquisa, com o contexto da pesquisa e as questões norteadoras da pesquisa.

No Capítulo 5, detalha-se o método de pesquisa empregado neste trabalho. São apresentadas e discutidas as questões de pesquisa e o protocolo da RSL. Além dos resultados da condução dessa revisão, ou seja, como os trabalhos se enquadraram ou não nos critérios de exclusão e inclusão, e quais de fato foram considerados para o desenvolvimento desta pesquisa. Ademais, é fundamentada a técnica Delphi para coleta e tratamento de dados em campo.

No Capítulo 6 expõe a esquematização do painel, montagem do grupo de painelistas, a preparação e realização das duas rodadas do Delphi, e o resultado da aplicação do método Delphi. Enquanto que no Capítulo 7 são apresentados e analisados os resultados da pesquisa.

No Capítulo 8 apresenta as considerações finais, bem como as contribuições e limitações desta pesquisa, e sugestão de trabalhos futuros.

No Apêndice A alguns detalhes do método de busca, como as *strings* de busca e os filtros, são apresentados. O Apêndice B apresenta a relação dos artigos analisados na RSL. Ademais, o Apêndice C apresenta informações adicionais desses artigos. Por fim, o Apêndice D expõe o questionário aplicado aos painelistas.

2 Conceitos básicos da pesquisa

Neste capítulo são apresentados conceitos fundamentais para o melhor entendimento deste estudo, bem como uma revisão da literatura acerca dos principais construtos. Este capítulo aborda privacidade de dados e BDA, sendo que cada um desses tem uma seção na qual definem-se os construtos, elementos básicos e características desses assuntos.

2.1 Privacidade de dados

Segundo o dicionário Michaelis (2016), a palavra privacidade é definida por: vida privada; intimidade. No entanto é uma definição insuficiente para se entender esse construto, que tem definições muito amplas e difusas na literatura (STUTZMAN; HARTZOG, 2012).

A ideia de respeito à vida privada é muito antiga, como pode-se verificar no Juramento de Hipócrates, que em seu artigo 8 declara: "sobre aquilo que vir ou ouvir respeitante à vida dos doentes, no exercício da minha profissão ou fora dela, e que não convenha que seja divulgado, guardarei silêncio como um segredo religioso". No entanto, a definição de privacidade não é muito clara na literatura, presume-se com frequência que é um conceito globalmente uniforme, porém nem sempre isso é verdadeiro (BARKER *et al.*, 2009).

Segundo Solove (2002), os conceitos que formam a ideia de privacidade são: direito de ser deixado em paz, acesso limitado ao seu *ego*, sigilo, controle sobre as próprias informações pessoais, personalidade (que se constitui em individualidade, dignidade e autonomia) e intimidade.

Hartzog (2018) argumenta que há muita discordância na definição de privacidade, pela dificuldade em se definir o construto, mas destaca a concepção de privacidade em SI como sendo o controle do usuário sobre as configurações dos sistemas, por ser amplamente adotada por acadêmicos, executivos, legisladores, reguladores e juízes. No entanto, segundo o autor, a interface dos sistemas podem ser construídas para dificultar a navegação dos usuários por essas configurações, deixando opções escondidas, sobrecarregando os usuários com opções demais, ou maquiando reais intenções com textos ambíguos.

Isso afeta a autonomia dos usuários sobre os seus dados, que tomam decisões baseadas na confiança que eles têm no provedor do sistema, como quando usuários aceitam

as políticas de privacidade mesmo sem lê-las, o que explica dois dos três pilares que formam a privacidade, segundo a concepção do autor: autonomia e confiança. O terceiro pilar é obscuridade, que é a limitação do acesso aos dados por terceiros, como, por exemplo, restringir a visibilidade das informações de um perfil em uma rede social a apenas amigos, ou apagar frequentemente o histórico de navegação na internet. É notável que dados obscuros não significa que são inacessíveis, mas são difíceis de se obter a ponto que diminui-se os riscos de privacidade. Em redes sociais, como usuários podem sofrer más consequências por causa de suas informações públicas, a obscuridade pode tornar a informação suficientemente difícil de ser obtida ou compreendida, propiciando ao usuário mais autonomia, melhor socialização e relativa liberdade ao abuso de poder por terceiros (HARTZOG, 2018).

Violação ou vazamento de dados, por sua vez, pode ser definida como qualquer incidente que envolva perda ou exposição de registros pessoais digitalmente. Registros pessoais significam informações sobre uma pessoa que não podem ser obtidas facilmente por outros meios públicos; e essas informações só são conhecidas por um indivíduo ou por uma organização nos termos de um acordo de confidencialidade (KHAN; HOQUE, 2016).

A seguir são apresentados as questões de privacidade que se relacionam aos objetivos desta pesquisa. Ou seja, os aspectos legais e regulatórios da privacidade de dados, o gerenciamento de privacidade de dados nas organizações, e a diferenciação entre segurança de dados, proteção de dados e privacidade de dados.

2.1.1 Leis e regulamentos

Pelo menos desde 1971 privacidade é um tema presente em alguma legislação pelo mundo, tendo surgido nos EUA por meio do *Fair Credit Reporting Act* (Lei de Justo Relatório de Crédito, em uma tradução livre), e desde então outras se sucederam, como relatam Wall, Lowry e Barlow (2016). Ainda segundo esses autores, apesar de ter surgido em outro contexto, atualmente o tema privacidade de dados está intimamente ligado a SI, principalmente no âmbito da *internet*.

Como relata Greenleaf (2012), a União Europeia frequentemente é pioneira nas regulações e acordos relativos a privacidade de dados, mas outros organismos internacionais contribuíram para essa área ao longo do tempo, como a OCDE (Organização para a

Cooperação e Desenvolvimento Econômico) e a ONU (Organização das Nações Unidas) por meio do seu Conselho de Direitos Humanos. Um dos principais acordos internacionais foi a Convenção 108, que foi a convenção para a proteção de indivíduos com relação ao processamento automático de dados pessoais, aprovada pelo Conselho da Europa em 28 de janeiro de 1981. Reconhecidamente essa convenção serviu de referência a dezenas de regulações fora da Europa (GREENLEAF, 2012). Pela data de aprovação da Convenção 108, em abril de 2006 o Conselho da Europa definiu o dia 28 de janeiro como o Dia da Proteção de Dados (KIERKEGAARD *et al.*, 2011).

Mais recentemente foram aprovadas diversas leis que tratam de proteção de dados. Entrou em vigor em 2020 a lei nº 13.709, de 14 de agosto de 2018, também conhecida como Lei Geral de Proteção de Dados Pessoais (LGPD) (BRASIL, 2018). Na União Europeia foi implantado em maio de 2018 o Regulamento Geral sobre a Proteção de Dados (também conhecido como GDPR, sigla de *General Data Protection Regulation*). Segundo a redação dessas leis, organizações que falharem na proteção de dados podem receber multas milionárias, dentre outras penalidades. Porém, nesses textos regulatórios há pouca ou nenhuma definição teórica dos temas que são abordados neste capítulo, como privacidade e proteção de dados (BRASIL, 2018; European Parliament and Council of European Union, 2016).

Apesar de haver poucas definições no texto da GDPR, o termo "processamento de dados pessoais" é definido, e abrange uma gama diversa de termos específicos, conforme detalhado no quadro 1.

Quadro 1 – Ações comparadas ao "processamento de dados pessoais" do GDPR

Ação	**Exemplos de processamento de dados pessoais do GDPR**
Operar	Adaptação; alteração; recuperação; consulta; uso; combinação
Armazenar	Organização; estruturação; armazenamento
Reter	Oposto a: apagamento, exclusão; destruição
Colecionar	Coleta; gravação
Compartilhar	Transmissão; disseminação; disponibilização. Oposto a: restrição; bloqueio
Alterar	Adaptação; alteração; uso; combinação (por terceiro não autorizado)
Vazamento	Recuperação; consulta (por terceiro não autorizado)

Fonte: adaptado de Colesky, Hoepman e Hillen (2016)

Comumente o conjunto legal nos Estados Unidos da América (EUA) usa o termo "privacidade", enquanto na Europa e no Brasil é frequente o termo "proteção de dados"

(BRASIL, 2018; COLESKY; HOEPMAN; HILLEN, 2016). Embora ressalve que esses termos não são intercambiáveis, os pesquisadores Colesky, Hoepman e Hillen (2016) combinaram as ideias presentes em cada construto para melhorar a compreensão, formando o termo "*privacy protection*", proteção da privacidade, em uma tradução livre e como é referenciada doravante.

A LGPD também utiliza algumas definições, como em seu quinto artigo dados pessoais e dados pessoais sensíveis (ou simplesmente dados sensíveis). Sendo que dado pessoal é qualquer dado que possa levar à identificação de uma pessoa física, e dado pessoal sensível é considerado aquele que pode levar a uso discriminatório e prejudicial ao indivíduo, como informações de etnia, religião, opinião política, biométricos etc. (BRASIL, 2018). Esses conceitos são conhecidos pelas siglas PII (*Personally Identifiable Information*) e SPI (*Sensitive Personal Information*) (SCHWARTZ; SOLOVE, 2011).

Além dos dados pessoais identificáveis, também existe o conceito de quase-identificador, que é frequentemente citado na literatura. Esse conceito define os atributos que não vinculam diretamente uma pessoa, mas podem servir para re-identificar um indivíduo quando os valores de vários atributos são combinados (GRUSCHKA *et al.*, 2018).

Outras definições da LGPD do Brasil (2018) são: controlador, operador e encarregado. O controlador é a pessoa (natural ou jurídica) que tem o poder de tomar as decisões, como demandar o tratamento dos dados, e responde legalmente às infrações, mesmo por danos causado pelo operador (responsabilidade solidária). O GDPR define esse papel como "responsável", embora na regulação europeia apenas pessoas jurídicas possam ter essa classificação (European Parliament and Council of European Union, 2016).

O operador é um terceiro que realiza o tratamento dos dados quando solicitado por um controlador. Essa figura é equivalente ao "subcontratante" do GDPR, e também pode responder legalmente pelas infrações caso haja danos provocados pelo descumprimento das orientações do controlador. Por fim, o encarregado, que no GDPR é chamado de "DPO" (*Data Privacy Officer*), é a pessoa indicada pelo controlador para atuar como canal de comunicação entre o controlador, os titulares dos dados e o regulador. No Brasil, a ANPD (Autoridade Nacional de Proteção de Dados) exerce o papel de reguladora (BRASIL, 2018; European Parliament and Council of European Union, 2016).

2.1.2 Privacidade de dados e sistemas de informação

Sistemas que detectam, processam, disponibilizam e comunicam dados, foram classificados como onipresentes por Schaub, Konings e Weber (2015), pois estão presentes em diversas situações das nossas vidas e a tendência é estarem cada vez mais. Os autores dizem ainda, que isso gera inúmeras implicações de privacidade, pois os sistemas podem, potencialmente, reunir e trocar informações extremamente abrangentes dos usuários com pessoas ou empresas em qualquer lugar do mundo.

Segundo o guia DMBOK, os principais riscos que estão associados com segurança de informação estão ligados a conformidade legal, reputação, e consequente perda financeira (Dama International, 2017). As mesmas implicações podem ocorrer quando se trata de proteção de privacidade. Segundo Ahmadian *et al.* (2018), garantir a proteção e privacidade dos dados se tornou um grande problema para as empresas que utilizam dados pessoais de seus clientes em seus serviços informatizados. Segundo Colesky, Hoepman e Hillen (2016), a privacidade pode ser tratada não só como um atributo de segurança, mas como também um atributo de qualidade.

Violações de privacidade não resultam somente de processos organizacionais ruins, pois podem acontecer mesmo quando a organização está em conformidade com regulações e boas práticas. Portanto, pode-se supor que há fatores importantes na proteção de privacidade que vão além do devido cumprimento processual de regras, leis e das boas práticas. Além disso, até recentemente, as principais pesquisas sobre o tema focavam na prática dos funcionários individualmente, e não no contexto da organização como um todo (WALL; LOWRY; BARLOW, 2016).

Como explicitado pela LGPD de Brasil (2018), os dados coletados de terceiros devem ser utilizados apenas para o propósito definido entre as partes, no entanto organizações privadas e governos, ao redor do mundo, têm coletado dados para seus sistemas de *big data* sem um propósito pré-estabelecido, esperando que em algum momento possam extrair valor desses dados, conforme expõem Constantiou e Kallinikos (2015), Müller *et al.* (2016), o que pode ser caracterizado como uma violação de privacidade.

Segundo Cavoukian (2012), o projeto de sistemas pode e deve sofrer forte influência da privacidade. A partir dessa ideia surgiu o conceito de *privacy by design*, que é uma abordagem de engenharia de *software* no qual exige-se que a privacidade seja levada

em consideração ao longo de todo o processo de engenharia. Por consequência, pode-se inferir que a própria forma de concepção dos sistemas de proteção de privacidade pode influenciar a comunicação interna de uma organização e vice versa. Pois de acordo com Conway (1968), o sistema projetado por uma organização é como uma cópia da própria estrutura de comunicação dessa organização, isso é conhecido como Lei de Conway, em referência ao seu autor. Essa hipótese foi evidenciada em diversos estudos, incluindo a pesquisa de MacCormack, Baldwin e Rusnak (2012). Ou seja, por dedução, é possível supor que, se a arquitetura de sistemas está correlacionada com a estrutura de comunicação da organização, diferentes esquemas de comunicação podem favorecer ou prejudicar a devida proteção de privacidade de dados.

Nos estudos de Barker *et al.* (2009), foi proposta uma taxonomia formal de privacidade que é composta por quatro dimensões:

- Finalidade: refere-se ao uso pretendido dos dados, ou seja, com qual propósito as informações pessoais são divulgadas ou disponibilizadas. Isso é medido com uma variável ordinal que varia de "propósito único" a "qualquer propósito".
- Visibilidade: se trata de quem tem acesso a o quê. A medida varia de "ninguém" até "todos" ou "mundo", passando por "somente o proprietário" e "terceiros", por exemplo.
- Granularidade: descreve o nível de detalhamento das informações que, além de "nenhuma", varia de "existencial" a "específico".
- Retenção: é o período de armazenamento dos dados. Pode variar de fração de segundo a, virtualmente, "sempre".

Ainda segundo Barker *et al.* (2009), a privacidade está garantida se todas essas dimensões forem devidamente comunicadas e aceitas pelos proprietários dos dados e o acordo continuar sendo rigidamente cumprido. Ou seja, a informação pode ser específica, pública e retida para sempre sem prejuízo à privacidade de dados, desde que isso seja conscientemente autorizado pela pessoa proprietária dos dados.

2.1.3 Privacidade de dados e segurança da informação

Segurança da informação é frequentemente subdividida em três pilares: confidencialidade, integridade e disponibilidade (CHEN; ZHAO, 2012). A confidencialidade existe

quando o acesso às informações está restrito apenas a quem é necessário, integridade se trata de ter as informações incorruptíveis e, por fim, as informações devem estar disponíveis a quem deve acessá-las.

A LGPD de Brasil (2018) fundamenta a disciplina da proteção de dados de forma bastante coerente com o que se define como privacidade na literatura, no entanto há autores que identificam diferenças entre os conceitos, como Barker *et al.* (2009). Ou seja, como evidenciado por Kokott e Sobotta (2013), não existe consenso sobre a utilização desse construto, que ora é tratado como subconjunto de privacidade de dados, ora é tratado quase como sinônimo de segurança de dados, na qual privacidade de dados está contida. No entanto, é notável que a expressão "proteção de dados" é comumente utilizada no âmbito do direito (BRASIL, 2018; European Parliament and Council of European Union, 2016; KOKOTT; SOBOTTA, 2013).

Como abordado anteriormente, os dados são classificados como seguros, quando há confidencialidade, integridade e disponibilidade, culminando na sigla CIA (com A de *availability*). Analogamente, Joshi e Kadhiwala (2017) mencionam a medida PAIN, acrônimo de privacidade, autenticação, integridade, e não repúdio (irretratabilidade à autoria de uma ação específica), que serve como *benchmark* de privacidade. Nesse modelo, percebe-se claramente a distinção (e interseção) entre segurança de informação e privacidade de dados.

Evidentemente, existem diversos impactos negativos nas organizações quando há falhas de segurança, como quebra de confiança na organização, conflitos de direitos de propriedade dos dados, aumento nos custos, atrasos e problemas de relacionamento com os clientes (RANJAN; FOROPON, 2021).

E as ameaças à segurança são multiplicadas pelo volume, velocidade e variedade do *big data*, isso é observado mesmo com os avanços da computação em nuvem nessa área (HASHEM *et al.*, 2015). Portanto a segurança dos dados na nuvem deve ser avaliada rotineiramente a fim de garantir que os níveis de serviços acordados estejam sendo atendidos.

Uma preocupação adicional são as agências de inteligência e segurança governamentais que podem utilizar dados de cidadãos em benefício próprio sem a devida permissão. Portanto as políticas de privacidade devem ser amplas o suficiente para incluir até mesmo situações como essas, nas quais os governos podem pressionar provedores de computação em nuvem a liberarem o acesso aos dados dos clientes (HASHEM *et al.*, 2015).

Colesky, Hoepman e Hillen (2016) argumentam que proteção da privacidade é um atributo de qualidade, assim como segurança. Ou seja, se a segurança é considerada como parâmetro para definir se um sistema tem boa qualidade, assim deveria ser com proteção de privacidade.

Além disso, segundo Ranjan e Foropon (2021), os dados utilizados em processos de inteligência competitiva devem autenticar as fontes de dados, pois o impacto negativo para o negócio pode ser muito grande em caso de dados não íntegros, inautênticos ou de baixa qualidade. E como se verifica nos estudos de Chen, H.L.Chiang e C. Storey (2018), grande parte do *big data* vem de fontes externas. Ademais, a devida preparação prévia para situações de risco permite respostas rápidas a vulnerabilidades, o que pode até mesmo se tornar uma vantagem competitiva (BOSE, 2008; DAS, 2010; ROSS; MCGOWAN; STYGER, 2013). Como integridade é tida um dos três pilares de segurança da informação, esses aspectos dos dados em BDA se tornam relevantes para a segurança dos sistemas.

Segundo Ranjan e Foropon (2021), as organizações sentem dificuldade em alterar suas estratégias de dados internos para capturarem dados que podem gerar inteligência competitiva devido à sensibilidade dos dados, à privacidade e aos desafios de compartilhamento desses dados. No entanto, os autores recomendam que as organizações promovam as devidas políticas de segurança de dados e proteção de privacidade a fim de se sentirem seguras e aptas a capturar, armazenar e extrair valor, mesmo de dados sensíveis.

Foi evidenciado ao longo desta subseção que questões de segurança de dados também estão associadas a impactos na privacidade de dados, e que o mau gerenciamento de segurança e privacidade pode gerar consequências gravíssimas, desde danos à imagem da organização até desvantagem competitiva.

2.2 Big data analytics

A análise metodológica de dados tem evoluído ao longo do tempo, ao passo que o acesso a informações cresce em qualidade e quantidade (DAVENPORT, 2013). Mundialmente a competição no mercado se tornou mais acirrada e, muitas vezes, global, então vantagens competitivas providas por dados se torna cada vez mais necessário para o sucesso organizacional (MCAFEE; BRYNJOLFSSON, 2012; RANJAN; FOROPON, 2021). Nesse

cenário, *analytics* ganhou muita relevância nos últimos anos e permanece em tendência de alta (GOOGLE, 2020c). Logo, esse conceito é detalhado nesta seção.

Analytics, de forma simples, pode ser entendida como a análise de dados e estatísticas realizada de forma sistemática por meios computacionais, como definido por Oxford University (2020). E é comumente relacionado ao *big data*, pois como dizem Gandomi e Haider (2015), o *big data* por si só tem pouca ou nenhuma utilidade. Isso é, seu potencial só é aproveitado se é utilizado em tomada de decisão.

No entanto, não existe consenso na literatura sobre a definição de *big data*, diversos autores consideram diferentes parâmetros para defini-la, embora haja definições que se tornaram mais usuais. Aborda-se ao longo desta seção algumas dessas diferentes definições.

Big data se tornou uma *buzzword*, ou seja, algo que está em moda. Alguns autores inclusive dão ao *big data* o status de fenômeno cultural, acadêmico e tecnológico (BOYD; CRAWFORD, 2012). Uma implicação disso foi diversos fornecedores usarem o termo para seu próprio benefício comercial em contextos nos quais não seria devidamente aplicado, causando confusão a respeito do tema (MITHAS *et al.*, 2013). Mas, por outro lado, mesmo na academia a definição e o escopo são bastante diversos. Porém é resumível como o uso organizacional de grandes quantidades de dados para apoiar processos de tomada de decisão (RANJAN; FOROPON, 2021).

Em uma das subseções, são apresentadas algumas das diferentes tecnologias que estão envolvidas nesses ambientes, pois isso ajuda na definição do escopo do restante do trabalho de pesquisa, além de permitir melhor compreensão da prática e materialização dos conceitos de tratamento de dados, *big data* e *analytics*.

As tecnologias são empregadas em diferentes arquiteturas, nas quais se comunicam com diversos outros sistemas e diferentes pessoas. Em uma das subseções, são abordadas as tecnologias e arquiteturas dos ambientes nos quais esses sistemas de *big data* estão contidos.

Existe na indústria e na literatura científica boas práticas de gerenciamento e governança de dados, é reservada uma subseção para tratar a definição de governança de dados e dos principais métodos e ferramentas de gerenciamento de dados.

Por fim, explana-se como a indústria tem abordado essas questões e perspectivas futuras para o mercado de dados, *big data* e *analytics*.

2.2.1 Conceito de *big data*

Uma definição comum e bastante aceita de *big data* são os 3 Vs, como citam Chen, H.L.Chiang e C. Storey (2018). Diversos outros autores, consultorias, como a Gartner, e outras empresas de tecnologia, como a IBM, definem de forma similar (GANDOMI; HAIDER, 2015). Os 3 Vs são: volume, variedade e velocidade. Ou seja, como é explicado em outro estudo bastante citado, *big data* tem como contexto um grande volume de dados, vindos de diversas fontes diferentes e em intervalos de tempo extremamente curtos (MCAFEE; BRYNJOLFSSON, 2012).

Porém, surge o questionamento: a partir de qual volume, por exemplo, os dados são considerados "*big*"? Uma pesquisa conduzida pela IBM em 2012 indicou que mais da metade dos 1144 entrevistados consideraram mais de um *terabyte*, outros consideram a casa dos *petabytes*. Da mesma forma, a partir de quantas fontes se tem variedade suficiente para se classificar como *big data*? Os dados precisam estar em tempo real para serem considerados velozes ou basta que sejam gerados no intervalo de minutos? Há quem leva em consideração até mesmo o formato dos arquivos (vídeos, arquivos estruturados, arquivos não estruturados etc.) e o setor (varejo, telecomunicações etc.) para estabelecer um limiar. Ou seja, é impraticável cravar parâmetros precisos para a definição de *big data* (GANDOMI; HAIDER, 2015).

Nessa mesma linha, uma definição criativa é mencionada por Madden (2012): *big data* pode significar que os dados são grandes demais, rápidos demais e variados demais para ferramentas comuns processarem.

Outros autores adicionam outros Vs à definição, como Hashem *et al.* (2015) e Fiorinia *et al.* (2018) que mencionam "valor" (ou seja, os dados devem agregar valor ao negócio), e Müller *et al.* (2016) que defende "veracidade" como parte da definição (isto é, os dados devem vir de fontes confiáveis e ter corretude). No estudo de Ranjan e Foropon (2021), são mencionados autores que vão além e adicionam até 7 Vs: variabilidade, veracidade, visualização, valor, validade, vulnerabilidade e volatilidade.

Além da definição do *big data* em si, os autores Terzi, Sinanc e Sagiroglu (2015) classificam também as características do *big data*, separadas em dez categorias, conforme a figura 1.

Figura 1 – Características do *big data*

Fonte: adaptado de Terzi, Sinanc e Sagiroglu (2015)

2.2.2 Conceito de *analytics*

Schadt *et al.* (2010) afirmam que o ritmo de crescimento das dificuldades de armazenar e analisar dados é exponencial, o que nos leva às novas tendências de processamento de *big data*, como a ciência de dados e o *fast analytics*, que é gerar informação para tomada de decisão com dados recém gerados rapidamente, que cada vez mais fazem parte do processo de inteligência de negócios.

Análise de dados não é algo novo, existe há tempos, mesmo antes da era digital. E já nos anos 50 do século 20, computadores digitais lidavam com dados de forma que nenhum humano conseguiria lidar tão rapidamente. Porém, a já mencionada popularização de computadores pessoais e o rápido crescimento do poder de processamento e armazenamento permitiu que novas ferramentas fossem criadas especificamente para esse fim. A isso pode-se dar o título de *Analytics 1.0*, ou mesmo de "era pré-*big data*", popularizada como *business intelligence*, ou simplesmente BI (DAVENPORT, 2013).

O *Analytics 2.0*, já envolvia ferramentas de *big data*, pois os 3 Vs já estavam presentes. Os produtos desenvolvidos com base em inteligência artificial, como sistemas de recomendação, se popularizaram, e o maior desafio se tornou coletar mais dados para alimentar os algoritmos e as análises. A estratégia comum era ingerir o máximo de dados possíveis e depois verificar como seriam utilizados, em vez de estruturar os dados a priori

como na era anterior. A evolução ao Analytics 3.0 está sendo a exponenciação da era anterior, que agora exige enriquecimento dos dados, novas técnicas de gerenciamento de dados, *analytics* embarcado na operação, e não somente nos sistemas analíticos, times de dados multidisciplinares e escalabilidade do *analytics* (DAVENPORT, 2013).

Frequentemente a definição de analytics está associada a *insights*, que é o termo em inglês que Santos (2018) definiu como "uma súbita percepção da solução de um problema ou dificuldade". Como na publicação de Cooper (2012), que define *analytics* como um processo de desenvolvimento de *insights* acionáveis por meio da definição de problemas e aplicação de modelos estatísticos e análises de dados reais ou simulados. Enquanto que, segundo Jagadish *et al.* (2014), BDA pode ser entendido como um subprocesso da extração de *insights* do *big data*.

Segundo Gandomi e Haider (2015), o *big data* é indissociável do que se define como *analytics*, pois o valor dessa grande quantidade de dados só pode ser bem aproveitado se as organizações tenham processos eficientes de geração de *insights*. Na pesquisa de Jagadish *et al.* (2014), esse processo é separado em estágios, que por sua vez são divididos em dois grupos: gerenciamento de dados e *analytics*. Gerenciamento de dados envolve processos e tecnologias para adquirir, armazenar, preparar e disponibilizar os dados para a análise. Enquanto *analytics* se refere às técnicas usadas para extrair inteligência dos dados. Logo, BDA pode ser visto como um subprocesso da extração de *insights* do *big data*.

Segundo Chen, H.L.Chiang e C. Storey (2018), *analytics* se refere principalmente a técnicas fundamentadas em mineração de dados e análises estatísticas, sendo que a maioria dessas técnicas depende de tecnologias como sistemas gerenciadores de banco de dados (SGBD), armazém de dados, popularmente conhecido como *data warehouse* ou simplesmente DW, ETL (sigla de *extraction, transformation and load*, em português, extração, transformação e carga), OLAP (acrônimo de *online analytical processing*, em português, processamento analítico em tempo real) e BPM (sigla de *business process management*, em português, gerenciamento de processos de negócio).

Quando se trata de modelagem estatística, a literatura tradicionalmente distingue dois grandes tipos de abordagens: a modelagem explicativa, que visa testar estatisticamente hipóteses baseadas em dados empíricos, e modelos preditivos, que visam fazer previsões sobre eventos futuros ou desconhecidos com base em dados históricos. Os estudos de BDA podem seguir uma abordagem explicativa ou preditiva, mas em ambos os casos os modelos

devem ser atualizados continuamente para refletir as mudanças nos comportamentos dos objetos representados pelos dados (MÜLLER *et al.*, 2016).

A potencialização de *analytics* a partir do *big data* se dá, por exemplo, devido à amostra de dados viabilizada pelo *big data*. Pois permite atingir pelo menos dois grandes objetivos gerais de entendimento de dados que não seriam possíveis em situações comuns. O primeiro é exploração de estruturas ocultas em diferentes subpopulações de dados, que tradicionalmente são tratadas como *outliers* em amostras menores. O segundo é a identificação e extração de características comuns em subpopulações mesmo quando há grandes variações individuais (FAN; HAN; LIU, 2014).

Alguns autores, como Dev Mishra e Beer Singh (2017), defendem que as organizações devem gerenciar os riscos e impactos que *analytics* pode causar na privacidade e segurança dos dados antes mesmo de iniciarem as análises. Portanto, é fundamental que a proteção de privacidade seja levada em consideração ao tratarmos de *analytics*.

Como *big data* define-se meramente pelas características dos dados (em termos de volume, variedade e velocidade, por exemplo), é evidente que se torna indissociável do uso prático desses dados, neste trabalho sintetizados pelo termo *analytics*. Portanto, BDA tornou-se parte do escopo da pesquisa.

2.2.3 Arquitetura e tecnologias de BDA

Segundo Bose (2008) e Ranjan e Foropon (2021), o sucesso do BDA e sua precisão dependem muito das ferramentas e técnicas usadas para analisar os dados. Além disso, parte das causas ou das soluções dos problemas de privacidade de dados podem estar atrelados ao desenho de arquitetura ou às tecnologias adotadas. Portanto, nesta subseção serão discutidas as principais tecnologias e algumas das diferentes arquiteturas de BDA.

As características do *big data* geram alguns desafios bastante complexos, como acúmulo de ruído devido a alta dimensionalidade dos dados (muitas variáveis, pois cada variável é uma dimensão), alto custo computacional, instabilidade algorítmica, e a dificuldade na agregação de dados de fontes múltiplas que usam tecnologias distintas entre si (FAN; HAN; LIU, 2014).

Por esses motivos, dentre outros, o desenho arquitetural é uma atividade que exige bastante técnica, principalmente à medida que a velocidade e o volume dos dados aumentam.

Por exemplo, se houver suspeita de uma transação fraudulenta com cartão de crédito, idealmente a transação é negada de imediato, porém uma análise completa do histórico de um usuário específico provavelmente não será viável em tempo real. Uma alternativa possível é ter resultados parciais com antecedência. Ou seja, o desafio fundamental é obter resultados práticos a partir de ambiente analíticos complexos (JAGADISH *et al.*, 2014).

Os SGBDs tradicionais ficam aquém das necessidades de processamento de muitos dados tempestivamente. Embora muitos sistemas comerciais como Greenplum, Netezza, Teradata ou Vertica relatam ser capazes de lidar com bancos de dados com vários *petabytes*, sistemas de código aberto como MySQL e PostgreSQL ficam muito atrás em termos de escalabilidade. E o custo dessas ferramentas comerciais são proibitivos para muitas empresas (MADDEN, 2012).

Dentre os problemas enfrentados pelos bancos de dados relacionais tradicionais de código aberto, está a importação de dados, que tende a ser lenta, limitando sua capacidade de lidar com dados em tempo real. Outro problema é a falta de suporte para estatísticas e modelagem, pois geralmente não se paralelizam efetivamente com grandes quantidades de dados (MADDEN, 2012).

Por outro lado, o chamado *data lake* tem o propósito de ser um repositório único de dados, estruturados ou não, armazenados em seu formato bruto ou processado, a baixo custo, utilizando computação distribuída (FARRUGIA; CLAXTON; THOMPSON, 2016; CHEN; CHEN; HUANG, 2018).

Outra etapa importante do trabalho de engenharia de dados é a modelagem de banco de dados, Jagadish *et al.* (2014) classifica a atividade como "arte" e ressalta que é executada no contexto empresarial por profissionais altamente pagos. Ainda segundo o autor, atualmente, há duas tendências importantes na arquitetura de dados: o uso de computação em nuvem e softwares de uso livre e código aberto.

A computação em nuvem se caracteriza por softwares, infraestruturas e plataformas confiáveis fornecidos pela *internet* em centros de dados remotos. Como essas tecnologias são ofertadas como serviço, surgem os termos IaaS, PaaS e SaaS, respectivamente infraestrutura, plataforma e software "*as a service*" ("como um serviço", em português). A computação em nuvem tem crescido rapidamente, e está cada vez mais presente do setor de SI e negócios, pois permitem a realização de tarefas complexas de forma escalável, abrangente e rápida. O custo costuma ser muito menor do que manter servidores locais, o que tem

permitido a implantação de diversos modelos de negócios e experimentos científicos que eram inviáveis sem o serviço de nuvem (HASHEM *et al.*, 2015).

Em contraponto aos bancos de dados relacionais, é muito comum a utilização de bancos de dados NoSQL como mecanismo de armazenamento e recuperação de dados que não são modelados em tabelas tradicionais de linhas e colunas. O próprio nome NoSQL se dá pelo fato de esses bancos de dados não oferecerem de forma nativa consultas SQL (*Structured Query Language*). A grande vantagem do NoSQL está nos mecanismos para armazenar e recuperar grandes volumes de dados distribuídos, que executam muito rapidamente, além de viabilizarem a utilização de dados semiestruturados. Há diferentes tipos de bancos de dados NoSQL, como chave-valor, orientado a colunas, orientado a documentos, orientado a grafos etc.. Portanto, esse tipo de tecnologia é bastante útil à arquiteturas de *big data* (HASHEM *et al.*, 2015).

Ademais, uma parte fundamental do processo de *analytics* é a visualização de dados, que é um campo que emprega canais de representação de conjunto de dados. Ou seja, transforma diferente tipos de dados em representações visuais que facilitam a compreensão dos dados por humanos (CHEN; GUO; WANG, 2015).

Como explicam Keim, Qu e Ma (2013), em quase todas as áreas do conhecimento o volume de dados e a rapidez em que ficam defasados se tornou um desafio muito complexo. Em meio a esse contexto, as ferramentas de visualização de dados devem ser eficazes em apresentar informações essenciais de grandes quantidades de dados e também em conduzir análises complexas, e isso se traduz em oportunidades de pesquisa para a comunidade de computação gráfica e visualização de dados.

Segundo Chen, Guo e Wang (2015), existem três grandes áreas de visualização de dados: visualização científica, visualização de informações e *visual analytics*. Sendo que a visualização científica ilustra estruturas e evoluções de propriedades físicas ou químicas no domínio espacial. A visualização de informações se concentra na representação de dados abstratos, não estruturados e de alta dimensão, incluindo dados de negócios, de redes sociais etc.. Por fim, *visual analytics* se trata da integração iterativa, interativa e dinâmica da inteligência humana e a inteligência das máquinas.

Existem diversas ferramentas de visualização de dados que realizam tarefas como filtragem e agregação de formas diferentes, podem usar ou não processamento em memória, utilizar *cache* ou acessar o banco de dados em cada acesso ao relatório, entre outras características técnicas e gerenciais, como custo e manutenibilidade. Cada caso pode ter

técnicas e ferramentas adequadas ao seu propósito, mas quando se trata de *analytics*, o principal desafio está em atender às especificidades de volume, variedade e velocidade dos dados (WU; BATTLE; MADDEN, 2014).

2.2.4 Governança de dados

Além do DMBOK (Dama International, 2017), nesta subseção também serão apresentados trabalhos da academia que discutem o tema governança de dados.

Governança de dados se trata do gerenciamento de dados de maneira ampla, como disponibilidade, relevância, usabilidade, integridade, qualidade, interoperabilidade, referência, segurança, entre outros (Dama International, 2017). Obviamente, a proteção de privacidade passa por uma governança adequada, como proteção contra o uso indevido, categorização, controle de acesso e etc.. O DMBOK, especificamente, divide a governança de dados em subtópicos conforme a figura 2.

Figura 2 – Escopo da governança de dados segundo o DMBOK

Fonte: DAMA International, 2017

Quando se fala em segurança de dados, é natural que se pense primeiro em confidencialidade dos dados, mas outro aspecto fundamental da segurança de *big data* é a integridade, que significa que os dados podem ser modificados apenas por partes au-

torizadas ou pelo proprietário dos dados para evitar o uso indevido (HASHEM *et al.*, 2015). Notoriamente, a boa governança de dados viabiliza maior eficiência nos aspectos de confidencialidade e integridade (Dama International, 2017).

Metadados são importantes por diversos fatores, como maior controle de acesso, viabilizar melhores práticas de qualidade, auxiliar na descoberta de dados, entre outros (Dama International, 2017). Porém, é um grande desafio gerenciar metadados em *big data*, e é frequentemente um trabalho manual e pouco eficiente (JAGADISH *et al.*, 2014).

Em seu estudo, Jagadish *et al.* (2014) explanam como humanos lidam bem com dados heterogêneos, tanto que a própria linguagem natural tem vários nuances e detalhes de grande complexidade. Porém, em contraste, máquinas lidam bem com dados homogêneos, portanto análise de dados frequentemente envolve estruturação em tabelas ou outros formatos estruturados. E esse é um dos grandes desafios na geração de metadados de forma automática.

Como é evidenciado por Hashem *et al.* (2015), o processamento de dados se limitava a conjuntos de dados limpos de fontes bem conhecidas, no entanto, com o surgimento do *big data*, os dados se originam de muitas fontes diferentes, dentre as quais muitas são pouco confiáveis. Isso gera um sério problema de qualidade dos dados, até mesmo tornando-os impróprios para uso. Portanto, pelo tamanho do desafio e pela importância da consistência dos dados, um dos tópicos mais importantes da governança de dados é o gerenciamento de qualidade (Dama International, 2017; HASHEM *et al.*, 2015).

Outra parte fundamental da governança de dados diz respeito a políticas de armazenamento e processamento de dados. E, além, também tem papel importante na definição de políticas de controle de acesso e definição de donos dos dados (*data owners*), sempre visando o equilíbrio entre a exposição ao risco e a criação de valor na organização (Dama International, 2017; HASHEM *et al.*, 2015).

Por fim, embora não seja um dos tópicos gerais do DMBOK, a criação de uma cultura de dados também pode estar inserida neste contexto de governança, pois como apresentado por Ranjan e Foropon (2021), é comum que se pense em projetos de dados como projetos de TI em vez de projeto de negócio, porém é notório que a extração de valor e *insights* depende fundamentalmente da participação ativa das áreas de negócio nos projetos de dados.

2.2.5 *Big data analytics* nas organizações

Naturalmente, a resolução de problemas relacionados à privacidade de dados é tomada em um contexto organizacional, portanto é importante que compreenda-se o papel do BDA nas organizações, como tratado nesta subseção, ao discutir as causas, problemas e soluções de privacidade de dados em BDA.

A abrangência do uso de *big data* é enorme, havendo exemplos bem sucedidos na área da saúde, planejamento urbano, transporte, meio ambiente, energia, educação, ciências sociais computacionais, economia e finanças, segurança e defesa, e assim por diante (JAGADISH *et al.*, 2014).

Notoriamente, quanto mais as empresas se caracterizam como orientadas a dados, melhor desempenham em medidas objetivas de resultados financeiros e operacionais. Como apresenta (MCAFEE; BRYNJOLFSSON, 2012), foi feito um ranqueamento de empresas por utilização de dados para tomada de decisão, das que mais usam, para as que menos usam. Essas empresas foram divididas por setores. A partir desse ranqueamento, foi observado que a tomada de decisão baseada em dados aumentou a produtividade das empresas em média em 5% e a lucratividade em 6%. Esse resultado foi corroborado pela contabilização de contribuições de trabalho, capital, serviços adquiridos e investimento em TI tradicional.

Além do desenvolvimento de vantagens competitivas, como apresentado anteriormente, uma revisão de literatura relacionou a teoria da administração com *big data*, pois o uso dos dados permite previsão, monitoramento, avaliação e adaptação de processos, entre outras aplicações que impactam na estratégia da organização (FIORINIA *et al.*, 2018). Ou seja, é recomendado que a orientação a dados esteja presente nas definições estratégicas das organizações.

Porém, como evidenciam (RANJAN; FOROPON, 2021), apesar do número crescente de empresas que lançam iniciativas de *big data*, nota-se que as empresas têm muitas limitações ao tentar converter o potencial de tais tecnologias em valor para os negócios. Os autores supracitados apresentam uma pesquisa com 175 especialistas de nível estratégico sobre abordagens de *big data*, que concluiu que organizações de diversos tamanhos, estruturas e setores têm grandes dificuldades em orquestrar a análise de *big data*, sugerindo que é um desafio intrínseco do contexto.

De fato, muitos dos algoritmos usados em BDA foram projetados para aplicações práticas, como pontuação de risco de crédito ou recomendação de produtos para clientes individuais, e isso gera até mesmo críticas por parte de pesquisadores sobre seus processos (MÜLLER *et al.*, 2016). Porém tem surgido uma vasta literatura que visa endereçar dificuldades em gerenciamento de dados, inclusive a partir da criação de artefatos de gestão de processos, construção de algoritmos e arquitetura de dados (RANJAN; FOROPON, 2021). Afinal, como concluem Ranjan e Foropon (2021), atualmente é imperativo que as organizações adotem *big data* para descobrir novos e importantes *insights* gerenciais.

2.3 Definição de problema, causa e solução

Segundo o dicionário de Oxford University (2020), um problema é definido por "um assunto ou situação considerada indesejável ou prejudicial e que precisa ser tratada e superada". E, como definido anteriormente, privacidade diz respeito à vida privada de indivíduos. Portanto, neste trabalho é tida como *problema* uma situação considerada indesejável ou prejudicial aos indivíduos que tenha sido causada por violação de privacidade no contexto de BDA.

Segundo o dicionário de Michaelis (2016), uma causa é definida por "aquilo que provoca o início ou determina a origem de algo; agente, origem, princípio". Ou, "razão pela qual se faz algo ou se provoca um acontecimento". Portanto, neste trabalho *causa* é aquilo que é origem de um problema de privacidade de dados, desde que esteja no contexto de BDA.

Segundo o dicionário de Oxford University (2020), uma solução é definida por "um meio de resolver um problema ou lidar com uma situação difícil". Portanto, neste trabalho, *solução* é uma forma de resolver ou lidar, por meio do contexto de BDA, com um ou mais problemas ou uma ou mais causas.

3 Problemas, causas e soluções de privacidade de dados no contexto de BDA

3.1 RSL sobre problemas, causas e soluções de privacidade de dados

No processo de RSL foram utilizadas outras pesquisas publicadas, que passaram por revisão de pares, para a extração de informações pertinentes ao objetivo deste trabalho. Essas pesquisas foram selecionadas a partir de critérios específicos de inclusão e exclusão.

Esta RSL seguiu o método proposto por Kitchenham (2004).

3.1.1 Protocolo da RSL

Esta seção detalha a RSL realizada nas bibliotecas *online* ACM Digital Library, IEE Xplore, Scopus e Web of Science.

A RSL tem como objetivo identificar problemas de privacidade de dados, suas causas e soluções, no contexto de BDA. O objetivo dessa revisão é obter esses desafios relatados na literatura científica recente para que sirva de insumo ao restante do desenvolvimento da pesquisa.

Algumas questões de pesquisa devem ser respondidas para que se tenha o devido embasamento para este estudo. Essas questões estão apresentadas a seguir.

1. Quais são os problemas encontrados no tratamento da privacidade de dados no contexto de BDA?
2. Quais são as causas identificadas ou sugeridas pelos problemas no devido tratamento da privacidade de dados no contexto de BDA?
3. Quais são as soluções identificadas ou sugeridas para resolver ou mitigar as causas dos problemas de privacidade de dados no contexto de BDA?

Nota-se que as perguntas são semelhantes, mas a primeira tem foco nos “problemas” e a segunda e terceira nas “causas” e “soluções’ desses problemas, respectivamente. Espera-se que existam mais relatos de problemas que causas apontadas, pois nem sempre essas causas são conhecidas ou mesmo objeto ou produto de estudo. Da mesma forma, nem todas as causas terão soluções. Então essa diferenciação se faz necessária.

Espera-se que o resultado dessa revisão provenha material suficiente para compilar, contabilizar, ordenar e categorizar as dificuldades em tratar a privacidade de dados no

contexto pretendido. E compilar, contabilizar, ordenar, categorizar, comparar e sugerir causas e possíveis soluções dos problemas de privacidade de dados no contexto de BDA.

A palavra "privacidade" é bastante elementar, a ponto que há poucos termos que possam substituí-la, mas houve alguns pouquíssimos estudos que utilizaram o termo "anonimização", mas também tratava de privacidade. Por outro lado, é possível que haja estudos que citam privacidade no resumo, sem que privacidade seja objeto de estudo. Portanto, decidiu-se por tornar obrigatório que houvesse a palavra "privacidade" no título ou nas palavras-chaves, com exceção de haver o radical relacionado a "anonimização", pois se essa palavra constasse no título ou nas palavras-chaves, bastaria que houvesse "privacidade" no resumo.

O termo BDA pode ser bastante abrangente e não é consensualmente definido, mas está associado a análise de dados, e não a sistemas transacionais, por exemplo. Então, para que fosse possível abranger o máximo de pesquisas possíveis nesse tema, optou-se por utilizar os termos "*business intelligence*" e "*relational database*", por exemplo, mesmo que não haja uma associação direta entre esses termos e BDA.

Outros sinônimos e palavras relacionadas também apareceram em outras bibliografias ou são termos ocasionalmente utilizados na indústria junto a *big data* ou *analytics*. Também foi identificado o mesmo problema de haver estudos que citam um desses termos ou palavras no resumo, mas não se aprofundam no assunto, portanto também se decidiu por exigir que esses termos e palavras se encontrassem no título ou dentre as palavras-chaves.

Por fim, exigiu-se a presença de palavras e termos associados com "problema" ou "dificuldade". As palavras foram encontradas com o auxílio do sistema de recomendação de sinônimos da ferramenta de busca do Google (2020a) e da ferramenta de tradução do Google (2020b). Também se restringiu a busca por trabalhos que têm algum desses termos no título ou dentre as palavras-chaves, para que se limitasse a pesquisas que deram a devida relevância a isso.

Todos os motores de busca utilizados têm a característica de reconhecer automaticamente plural. Ou seja, ao utilizar uma palavra no singular, o motor de busca também faz a pesquisa pela sua palavra equivalente no plural. Além disso, todos os critérios de inclusão (CI) puderam ser incorporados na própria *string* de busca.

Para uma pesquisa ser selecionada para a RSL, foi obrigatório atender a todos os nove CI, que foram organizados da seguinte forma:

- **CI1**: conter no título ou em palavras-chaves referência a "privacidade".
- **CI2**: conter no título ou em palavras-chaves referência a BDA ou correlatos.
- **CI3**: conter no título ou em palavras-chaves referência a "problemas" ou correlatos.
- **CI4**: a fonte do estudo deve ser conferência ou periódico.
- **CI5**: o tipo do documento deve ser artigo de periódico, artigo de conferência, revisão ou capítulo de livro (esse último apenas no caso de a fonte ser um periódico).
- **CI6**: a publicação deve estar em inglês.
- **CI7**: a publicação deve estar no estágio final de publicação.
- **CI8**: a área de estudo da publicação deve ser computação, mesmo que computação seja apenas uma de duas ou mais áreas.
- **CI9**: ter sido publicado a partir de 2016. Nesse ano foi aprovada a GDPR na UE, o que impactou a discussão sobre esse tema.

Se qualquer pesquisa atender a qualquer critério de exclusão (CE), ela é desconsiderada. Os CEs são apresentados a seguir:

- **CE1**: documento duplicado nas bases de dados pesquisadas.
- **CE2**: não permitir acesso e não ser encontrado por outras fontes (como em resposta a solicitação de acesso ao próprio autor).
- **CE3**: não abordar problemas, causas de problemas ou soluções de problemas de privacidade de dados.
- **CE4**: abordar o tema "privacidade", mas não ter como objeto de estudo a privacidade no contexto de BDA.
- **CE5**: o foco da pesquisa estar em tecnologias de "internet das coisas" ou "*blockchain*" em vez de ter foco em BDA.
- **CE6**: foco na privacidade de dados em modelos de inteligência artificial.
- **CE7**: apesar de abordar privacidade e BDA, não ter o objetivo de estudar problemas de privacidade de dados, ou suas causas ou soluções, no contexto de BDA.

É importante destacar que não foram encontradas publicações brasileiras em português que atendessem aos critérios de seleção, mesmo ao traduzir as expressões utilizadas no filtro. Portanto, foram consideradas apenas publicações em inglês. Além disso, ao ampliar a área de estudo para outras além de computação, como por exemplo ciência

da informação, os resultados foram numerosos em excesso e as publicações se tornaram inadequadas em sua maioria. Ou seja, o filtro perdeu qualidade.

A estratégia de extração e síntese de dados é baseada na abordagem sugerida por Keshav (2007). Ou seja, a leitura passa por até três etapas.

Na primeira parte da primeira etapa, lê-se apenas o título, palavras-chaves e resumo. Em seguida lê-se a introdução e os títulos das seções e subseções. E, por fim, a conclusão é lida.

A segunda etapa é a leitura de diagramas, ilustrações, e, eventualmente, quadros e tabelas. A terceira etapa é a leitura cuidadosa de todo o artigo.

Em qualquer parte de qualquer etapa a leitura pode ser interrompida caso seja percebido que o estudo não é útil para a revisão.

Os dados extraídos do texto foram registrados em planilhas e em documentos de texto. Ao fim da tabulação dos dados de todos os artigos, aqueles elencados para compor o texto da revisão foram sintetizados. Ou seja, as ideias relevantes são inseridas ao longo do texto desta pesquisa.

3.1.2 Condução da RSL

Esta seção apresenta o método de seleção dos artigos que serão aproveitados na RSL.

As buscas ocorreram em fevereiro de 2021, e foram encontrados 745 resultados, considerando as *strings* de busca e os filtros, que representam todos os CI. Essa data de corte foi o marco do início desta RSL, por isso artigos posteriores não fizeram parte da RSL. Os CE foram feitos manualmente, por etapas, conforme detalhado na tabela 1. A segunda coluna da tabela 1 apresenta a quantidade de resultados descartados por CE, a terceira coluna indica a quantidade de publicações restantes após a aplicação de cada CE.

Estão apresentadas no Apêndice B as 61 pesquisas que foram selecionadas para a RSL após a aplicação dos CE e CI.

Tabela 1 – Quantidade de estudos desconsiderados por CE

Etapa	Eliminados	Restante
CE1	76	669
CE2	42	627
CE3	75	552
CE4	89	463
CE5	80	383
CE6	25	358
CE7	297	61

Fonte: Danilo Figueiredo de Oliveira, 2023

3.2 Apresentação e análise dos resultados

Nesta seção é apresentada a literatura a respeito dos problemas de privacidade de dados em BDA, assim como as causas desse tipo de problema e as soluções propostas, que podem ser meramente mitigação de risco ou soluções parciais. Afinal, como alguns autores demonstram, a classificação de o que é um risco ou problema de privacidade pode depender do contexto, e soluções definitivas podem ser impossíveis (YING; GRANDISON, 2017).

Muitos autores segregam os processos de *big data* por etapas ou por papéis das partes envolvidas, como Choudhary e Garg (2019), Singh *et al.* (2018) que classificam quatro diferentes papéis envolvidos: (i) provedor de dados, (ii) coletor de dados, (iii) minerador de dados e, finalmente, (iv) tomadores de decisão. Cada um tem desafios diferentes para a proteção da privacidade. No entanto, esta seção é divida simplesmente por problemas e causas e soluções, independentemente das etapas ou dos envolvidos.

As formas tradicionais de proteção da privacidade individual não são eficazes no contexto de *big data*, e os indivíduos têm pouco controle sobre o uso e as análises que serão feitas a partir desses dados (GHANI; HAMID; UDZIR, 2016). Adicionalmente, Singh *et al.* (2018) argumenta que os métodos de controle de acesso tradicionais também são insuficientes, pois existem muitas demandas diferentes na utilização do *big data*, e a granularidade e as especificidades dessas demandas podem inviabilizar ou, pelo menos, dificultar muito o controle de acesso.

Segundo Ying e Grandison (2017), muitos creem que a proteção de privacidade em *big data* é possível meramente por meio da aplicação de algoritmos de criptografia ou mascaramento, mas essa suposição é equivocada, pois somente isso não é suficiente.

Segundo os autores, se não houver uma análise crítica sobre os atuais sistemas e processos, poderia haver problemas de privacidade de dados com o potencial de inviabilizar campos de pesquisa inteiros.

Para o processamento de dados sensíveis existem pelo menos as técnicas baseadas em distorção de dados, criptografia de dados, e publicação restrita. As perturbações comumente usadas são: randomização, permutação, condensação e privacidade diferencial. No método tradicional, também existem métodos de criptografia e ofuscação que se concentram em problemas de publicação e mineração de dados (HU *et al.*, 2019).

Ao desenvolver um modelo matemático de estimação de risco de privacidade, Ying e Grandison (2017) sugeriram que não existe um único limiar a partir do qual uma coleção de dados se torna não segura. Embora o modelo forneça esse valor, os autores deixam claro que esse limiar muito provavelmente é contextual, ou seja, variará com base no contexto da coleta de dados.

Ainda de acordo com a pesquisa de Ying e Grandison (2017), embora o modelo tivesse como premissa que os administradores dos dados sejam conscientes e tomem medidas para proteger a privacidade, verificou-se que é probabilisticamente impossível proteger a privacidade quando se trata de *big data*.

Segundo Canbay, Vural e Sagiroglu (2019), ao mudar do domínio de dados tradicionais para o domínio de BDA, novas ameaças surgem com novas oportunidades e as ameaças existentes tornam-se mais complexas e arriscadas. E, embora o problema de privacidade seja NP-Difícil, novos estudos são necessários para superar esse problema no domínio de BDA.

De qualquer forma, é importante que se entenda como se dão os problemas de privacidade de dados. De acordo com quatro modelos estatísticos diferentes, Alashoor, Han e Joseph (2017) concluíram que a medida que os usuários têm consciência dos possíveis problemas de privacidade e de vulnerabilidades, suas preocupações com privacidade aumentam. Além disso, enquanto a familiaridade com *big data* tem correlação negativa com essas preocupações, o conhecimento a respeito das implicações do *big data* tem correlação positiva com as preocupações de privacidade. Apesar dessa associação, o estudo não avaliou causalidade.

Contudo, existe um fenômeno citado por Eastin *et al.* (2016) que é chamado de "paradoxo da privacidade". A hipótese desse fenômeno é que muitos usuários aceitam alguma perda de privacidade para fazer negócios digitais, apesar de expressarem altos

níveis de preocupação com a privacidade de suas informações. Os autores concluíram que o paradoxo é verificável empiricamente, ou seja, se mostrou verdadeiro no estudo.

3.2.1 Problemas de privacidade de dados em BDA

Nesta pesquisa "problema" é definido como uma situação considerada indesejável ou prejudicial às partes envolvidas da organização e que precisa ser tratada e superada. Essa situação deve ter sido em decorrência de violação de privacidade.

Naturalmente, algumas aplicações tradicionais de BDA, como astronomia, não armazenam ou divulgam informações pessoais, portanto não costumam ter problemas de privacidade significativos. Mas quando se trata de redes sociais, varejo, governos, por exemplo, a privacidade é uma questão muito importante. Além disso, seus usuários frequentemente têm ciência e preocupações a respeito de privacidade. Um em cada três consumidores reconheceu que oferece informações falsas (como cartões de crédito, senhas, detalhes de endereço, etc.) para obter privacidade (PATEL *et al.*, 2017).

Segundo Khanan *et al.* (2019), os principais desafios de privacidade e segurança em BDA são os seguintes:

- Mineração e análise de dados escalável preservando a privacidade.
- Riscos na terceirização e uso de ferramentas de terceiros.
- Cultura de aprendizagem organizacional e competências.
- Falta de infraestrutura para garantir a segurança dos dados, em especial dados de diferentes fontes.

Ao passo que os autores Shamsi e Khojaye (2018) classificam categorizam as possíveis violações de privacidade em BDA em quatro tipos:

- Rastreamento pelo governo.
- Coleta de informações por prestadores de serviços.
- Ataques de re-identificação.
- Violações de dados.

Segundo Abouelmehdi *et al.* (2016), não é possível afirmar que existem mecanismos de segurança de informação infalíveis. Em primeiro lugar, porque as vulnerabilidades não são todas conhecidas e, em segundo lugar, porque os sistemas e tecnologias estão evoluindo

rapidamente. No entanto, existem diversas técnicas para mitigar os riscos de violação de segurança e privacidade.

Apesar de convencionalmente haver o pressuposto que a análise de dados e a proteção da privacidade se contradizem, os autores Wieringa *et al.* (2021) defendem que essa é uma presunção muito restrita, porque as empresas podem implementar uma ampla gama de métodos que atendem a diferentes graus de privacidade, ao mesmo tempo que lhes permite lidar com todas as responsabilidades de análise de dados.

Conforme BDA se torna cada vez mais parte de todos os setores da economia, tornam-se possíveis novas soluções para uma infinidade de desafios, como tratamento médico mais eficaz, melhoria da segurança alimentar e prevenção do tráfico humano. No entanto, também traz consigo novos desafios sociais (DABAB *et al.*, 2018). Esses problemas sociais são agrupados pelo autor em cinco categorias gerais: privacidade e segurança, reutilização de dados, precisão de dados, acesso e arquivamento, e preservação de dados.

Privacidade e segurança referem-se principalmente à geração inicial de dados sobre indivíduos, incluindo a geração secundária por meio da associação com outros conjuntos de dados existentes. A reutilização de dados, por outro lado, está preocupada com o reaproveitamento dos dados de seus destinatários pretendidos e processos para outros usos. Podem surgir problemas de precisão de dados quando várias fontes de dados com controles e processos de verificação diferentes influenciam a qualidade geral dos dados e o grau em que os dados estão corretos. O acesso aos dados diz respeito aos indivíduos e organizações que têm acesso a quaisquer dados que façam parte do processo de BDA. E a preservação dos dados se refere à catalogação histórica dos dados uma vez decorrida a sua utilização inicial. Em todos esses casos, os indivíduos que geram dados têm pouco ou nenhum controle sobre esses dados (DABAB *et al.*, 2018).

Em sua pesquisa, Wang (2018) destaca algumas situações problemáticas em cada etapa do ciclo de vida do *big data*. Na etapa de coleta de dados, no caso de dados pessoais serem coletados por um serviço terceirizado não confiável, esses dados podem ser vazados para pessoas ou organizações mal-intencionadas. Na etapa de integração e armazenamento, a falta de criptografia e serviços terceirizados não confiáveis são citados. Na etapa de análise de dados, são citados ataques de classificação e clusterização, entre outros. E, por fim, na etapa de interpretação dos dados, é possível haver vazamento de dados a partir dos próprios metadados.

Além disso, quando não há respeito ao consentimento do uso de dados dos usuários, há um risco enorme de problemas de privacidade, e a principal causa desse risco é a falta de regulações. Mas mesmo quando há normas, também é necessário haver, por parte das organizações, disciplina em seguir o acordo feito com os usuários, pois não só os dados podem ser utilizados para fins não acordados, como novos dados pessoais e sensíveis podem ser inferidos a partir da integração de dados de outras fontes (WANG, 2018).

Porém, como explicam Singh *et al.* (2018), o tomador de decisão também tem responsabilidades, e deve respeitar os cinco aspectos da informação: autoridade, precisão, objetividade, atualidade e cobertura. Isso está em conformidade com o que Terzi, Sinanc e Sagiroglu (2015) concluem, pois, segundo eles, o acordo entre a empresa e o indivíduo deve ser determinado por políticas bem definidas.

Ademais, Terzi, Sinanc e Sagiroglu (2015) dissertam que os dados pessoais devem ser desidentificados e movidos por canais de comunicação seguros. Mas mesmo nesses casos, modelos de inteligência artificial podem ter implicações antiéticas se houver discriminação indevida. Além disso, mesmo dados anonimizados podem ser re-identificados indevidamente.

Segundo Ghani, Hamid e Udzir (2016), mesmo quando os dados são devidamente coletados, posteriormente pode haver problemas de privacidade. Por exemplo, quando há novos dados pessoais gerados a partir da análise dos dados fornecidos a priori. Por isso, os autores defendem que os usuários devem ser novamente consultados nas situações que fogem do que foi previamente autorizado por eles.

É importante destacar que os dados têm diferentes níveis de criticidade, como elencaram Rama Devi e Rajesh Babu (2019), os dados podem ser classificados como:

- Identificadores explícitos, como nomes e número de identidade.
- Semi-identificadores, que podem ser associados a outros dados para reconhecer uma pessoa.
- Identificadores sensíveis, que são dados que podem levar à hostilidade ao indivíduo, como doenças, renda, etc..
- Identificadores não confidenciais, que são características que não causam problemas, independentemente de serem descobertos.

Conforme supracitado, quando se trata de BDA, frequentemente as organizações podem criar, a partir de seus dados, informações que nunca foram explicitamente solicitadas aos usuários. Três circunstâncias em que essa violação de privacidade pode ocorrer são

enumeradas por Ghani, Hamid e Udzir (2016). Primeiramente, quando novas fontes de dados, em conjunto com os dados já coletados, passam a possibilitar a inferência de dados pessoais não permitidas a priori pelos usuários. Segundo, quando dados não pessoais, como os hábitos de compras de um indivíduo, são utilizadas para gerar dados pessoais. E, por fim, quando dados confidenciais são armazenados e processados em um local não protegidos, facilitando vazamentos.

Identificadores sensíveis, porém, podem ser difíceis de serem identificados, pois não há consenso sobre qual tipo de dado pode ter viés discriminatório. Portanto, Gambs (2019) sugere que haja pesquisas sociológicas a fim de estudar quais dados seriam discriminatório, portanto sensíveis.

Dados de saúde são amplamente considerados como dados sensíveis, porém tipicamente carecem da devida proteção, conforme evidenciado por Khan e Hoque (2016). Segundo os autores, dados médicos são altamente privados, pois as pessoas, no geral, não querem que terceiros saibam sobre suas condições médicas ou psicológicas. Vazamento desse tipo de dados pode a cobertura de seguro ou o emprego de indivíduos, por exemplo.

Além disso, há uma tendência crescente de invasão de registros médicos, pois a exploração de dados médicos é negócio lucrativo. Nos EUA, um número de previdência social roubado pode ser vendido por 25 centavos de dólar no mercado clandestino, o número de cartão de crédito por um dólar, e um registro médico pode variar de dez a mil dólares (KHAN; HOQUE, 2016).

No entanto, tipicamente os hospitais têm baixa segurança, por isso é relativamente fácil para os *hackers* obterem uma grande quantidade de dados médicos. O setor governamental, como os sistemas de departamentos de saúde pública que contêm dados relacionados à saúde, também é um alvo cada vez maior de *hackers* por duas razões principais. Primeiro, eles são geralmente mais vulneráveis, pois são sistemas mais antigos que executam *softwares* mais antigos e menos seguros. Em segundo lugar, eles são ricos em dados como informações de identificação pessoal, saúde e informações financeiras. De acordo com o Quinto Estudo de Referência Anual de Privacidade e Segurança de Dados de Saúde de 2015, que cobriu 90 organizações de saúde nos EUA, mais de 90% dos prestadores de serviços de saúde tiveram uma violação de dados e 40% tiveram mais de cinco violações de dados nos últimos dois anos (KHAN; HOQUE, 2016).

Embora não tenha sido maioria nas publicações analisadas, alguns trabalhos apontam problemas mais técnicos. Como os problemas de privacidade em armazenamento de

dados não relacionais que são alertados por Dev Mishra e Beer Singh (2017), por conta de limitações de arquitetura. No entanto, Joshi e Kadhiwala (2017) mencionam um método que visa garantir a segurança e privacidade dos dados em bancos não relacionais, e esse método é divido em quatro fases: pré-tratamento, identificação de atributos sensíveis, fragmentação de dados sensíveis e fase de reconstrução de dados.

Rama Devi e Rajesh Babu (2019) destacam que dados não estruturados, como mensagens instantâneas, apresentação de *slides*, áudios, imagens, etc., correspondem, em média, a mais de 90% dos dados de uma organização. O que gera um grande desafio de tratamento e proteção de dados.

A infraestrutura também deve ser avaliada na proteção de privacidade, pois o risco de vazamento de segurança e privacidade de *big data* é parcialmente cruzado com o risco de segurança de rede (JIANG; SHI; ZHOU, 2019).

Frequentemente vazamento de dados aparece na literatura com destaque dentre os problema de privacidade, como na pesquisa de Joshi e Kadhiwala (2017). Singh *et al.* (2018) defendem que vazamento de dados é um dos maiores problemas para as empresas, e argumentam que já causou muitos problemas. E, segundo Tiwari *et al.* (2019), o ativo mais importante de uma organização que trabalha com SI são os dados e, portanto, os dados precisam ser protegidos contra criminosos cibernéticos.

Mas outros problemas também são lembrados. Especificamente na coleta de metadados, diversas informações podem ser utilizadas para identificação pessoal. E a coleta desses metadados por governos pode causar preocupações sobre os direitos humanos e as liberdades fundamentais (AGARWAL; GUPTA; SHARMA, 2019).

Mas as preocupações de privacidade de dados em BDA são variadas. As preocupações listadas por Dev Mishra e Beer Singh (2017) são essas:

- O anonimato pode se tornar impossível.
- As análises dos dados não são completamente precisas.
- Inexistência de proteção legal para as pessoas envolvidas.
- Inteligência em segurança e auditoria de conformidade.
- Ações antiéticas baseadas em interpretações.
- Violações de privacidade e incidentes de fraude.
- Discriminação.
- O mascaramento de dados pode ser revertido e revelaria informações pessoais.

- A segurança da informação é um problema com a escala do *big data*.
- *Big data* existirá para sempre, portanto sempre será um desafio.
- Preocupações em *e-discovery* (*electronic discovery*).
- Patentes e direitos autorais podem se tornar irrelevantes.

Diferentes áreas podem ter diferentes problemas, causas e soluções, como no setor financeiro, no qual Singh *et al.* (2018) citam autenticação e autorização como dois dos primeiros desafios que exigem de atenção imediata das organizações. Enquanto na área da saúde a coleta de dados é um ponto crítico (GHANI; HAMID; UDZIR, 2016).

3.2.2 Causas e soluções de problemas de privacidade em BDA

Na literatura analisada, tanto as causas quanto as soluções encontradas aos problemas de privacidade de dados em BDA passam por aspectos organizacionais e técnicos, e permeia diversos níveis verticais e horizontais na estrutura das organizações, pois, como afirma Ali Khan, Sudhakar Reddy e Manoj Kumar (2019), as relações interpessoais nas organizações têm papel chave no ambiente de dados.

A filtragem, controle e gerenciamento de indicadores de risco de vazamento de dados já são suficientes para reduzir em grande medida os riscos de violação de privacidade de dados, segundo Jiang, Shi e Zhou (2019). Portanto, é importante a definição desses indicadores. Na área da saúde, Ghani, Hamid e Udzir (2016) definiram quatro grupos de indicadores para avaliar segurança e privacidade de dados em *big data*. São esses:

- Segurança de *big data* e vazamento de privacidade causado pela fase de coleta de dados.
- Segurança de *big data* e vazamento de privacidade causado pela fase de transmissão de dados.
- Violações de segurança e privacidade de *big data* causadas pela fase de armazenamento de dados.
- Violações de segurança e privacidade de *big data* causadas pela fase de uso e compartilhamento de dados.

Como nem toda quebra de confidencialidade leva a violação de privacidade, pois nem todos os dados são PII, Ying e Grandison (2017) propuseram o particionamento

dos dados em três categorias principais: dados sensíveis, quase-identificadores e dados benignos (que não identificam indivíduos e não são confidenciais). A partir disso, pesquisadores e profissionais de privacidade aplicariam anonimização, pseudonimização ou outros mecanismos de desidentificação, para criar versões de dados que preservem a privacidade.

Como defendem Wang (2018), para resolver problemas de privacidade de *big data*, é necessária uma estrutura de gerenciamento de privacidade abrangente, que inclui sistemas de monitoramento, avaliação e gerenciamento de risco ativos, gerenciamento e regulação de responsabilidades, e fornecimento de suporte técnico para gerenciamento de privacidade de *big data*. Pois, como argumentam Wang (2018), um suporte técnico ruim ou inexistente nas aplicações de *big data* pode resultar em vazamento de dados durante a coleta, análise, processamento, armazenamento, conversão ou exclusão de dados, seja maliciosamente ou não.

No entanto, não surpreendentemente, as principais soluções propostas nos trabalhos analisados visam resolver o problema mais frequentemente mencionado: confidencialidade. A seguir expõe-se as pesquisas que sugerem abordagens como criptografia e desidentificação em uma subseção apartada das demais soluções.

3.2.3 Criptografia e desidentificação

A técnica mais comum de proteção de privacidade e acesso a dados é proteger o sistema de gerenciamento de dados, em vez de proteger os dados em si. No entanto, essa abordagem pode gerar brechas de segurança. A criptografia "em pouso" e "em trânsito", ou seja, respectivamente no armazenamento e na transmissão, ao encapsular os dados sensíveis no ambiente da computação em nuvem, podem melhorar a proteção de dados. Além do devido gerenciamento de chaves de acesso ao *big data*, independentemente do provedor de nuvem (HASHEM *et al.*, 2015). No entanto, Alabdullah, Beloff e White (2018), Gambs (2019) alertam para o alto custo computacional que criptografia pode gerar.

Também com foco em computação em nuvem, Maohong, Aihua e Hui (2018) propõem um mecanismo de proteção de privacidade baseado em criptografia mista e separada. Em primeiro lugar, o conjunto de dados criptografados é disperso e armazenado em vários servidores na nuvem, e outro servidor na nuvem é configurado para calcular os dados relevantes de acordo com os requisitos do usuário. O resultado final da criptografia

é sintetizado, enviado ao usuário para o ambiente *on premise*, e descriptografado pelo usuário.

Em seu estudo, Abouelmehdi *et al.* (2017) também defendem o uso da criptografia de certos dados armazenados para a preservação da privacidade. Mas os autores vão além, e sugerem quatro pilares de proteção: autenticação. criptografia, mascaramento de dados e controle de acesso. De forma análoga, Varshney *et al.* (2020) defendem que a melhor técnica para proteger os dados é a criptografia funcional, mas defende também o uso da anonimização, que, apesar de permitir a de-anonimização, é um processo valioso e viável para minimizar o risco de vazamento de dados e brechas de privacidade.

Duas abordagens são sugeridas por Vatsalan, Karapiperis e Gkoulalas-Divanis (2019) para proteger a privacidade de dados, técnicas baseadas em criptografia, que são computacionalmente custosas, e técnicas baseadas em geração de ruídos, que são computacionalmente eficientes, mas os resultados perdem precisão. Portanto, os autores sugerem uma abordagem híbrida. Similarmente, a abordagem sugerida por Mehrotra *et al.* (2020) se baseia no uso de criptografia somente para dados sensíveis e confidenciais.

Alguns autores citam a criptografia homomórfica, que permitiria o processamento arbitrário de dados criptografados, embora até o momento não haja criptografia totalmente homomórfica (BONDEL *et al.*, 2020; Ramya Devi; Vijaya Chamundeeswari, 2020). Apesar disso, Bondel *et al.* (2020) concluem que a de-identificação fornece a abordagem mais promissora.

Para proteger a privacidade dos usuários, Terzi, Sinanc e Sagiroglu (2015) mencionam uma arquitetura capaz de anonimizar dados sensíveis em *logs* a partir da criptografia por chave simétrica AES (*Advanced Encryption Standard*), e posterior armazenamento no HDFS (*Hadoop Distributed File System*). Nesse caso, a qualidade do anonimato é medida por métricas baseadas em k-anonimato.

O Hadoop é uma estrutura de software de código aberto para armazenamento de dados e execução de aplicativos em *clusters* de *hardware* comum. Esta estrutura é baseada em um sistema de arquivos distribuído (HDFS) e em um modelo de programação paralela, conhecida como MapReduce. Quando o Hadoop foi criado, o problema de segurança não era uma prioridade enfrentada pela distribuição e processamento paralelo de dados, portanto a diversidade e o aumento do volume de dados trocados fazem com que os problemas de segurança não sejam realmente resolvidos (ABOUELMEHDI *et al.*, 2016).

Na área da saúde, Singh *et al.* (2018) argumenta que as informações não devem ser divulgadas ou recuperadas facilmente, nem serem vazadas para hackers, obviamente, pois são informações confidenciais que podem causar prejuízo financeiro, dentre outros. Para evitar isso, o autor sugere técnicas utilizadas que incluem anonimização e mascaramento dos registros por meio da técnica MapReduce.

Em sua pesquisa, Saxena (2017) descreve o sistema central de dados de saúde do governo de Omã. Relativo à privacidade e confidencialidade do portal do governo local, chamado e-Oman, há a menção clara de que todas as sessões e dados são encriptados e todas as comunicações efetuadas através deste portal estão protegidas de intrusos e *hackers*. A política de privacidade do portal prevê que o portal coleta informações pessoais dos indivíduos, é informado que todas as informações coletadas por meio deste portal são retidas ou encaminhadas para agências ou departamentos governamentais apropriados para uso posterior, mas que essas informações não devem ser vendidas ou transferidas a terceiros sem o consentimento de um indivíduo. No entanto, o autor afirma que mesmo se a informação for anonimizada, é possível que haja de-anonimização ou re-identificação quando uma infinidade de bancos de dados são agregados, e há chances de que os armazenamentos de dados possam ser acessados ilegalmente por estranhos, seja por atividade maliciosa ou descuido. No caso de vazamento de privacidade de um sistema como esse, o impacto seria enorme e envolveria pessoas de toda a nação.

Existem várias métricas formais para medir o grau de anonimato de um conjunto de dados. Segundo a GDPR, sem fornecer uma definição precisa ou concreta de anonimato, considera-se um conjunto de dados anônimo quando a re-identificação só é possível com grande esforço ou meios improváveis (European Parliament and Council of European Union, 2016; GRUSCHKA *et al.*, 2018).

Para o processamento de dados pessoais, o GDPR define uma série de requisitos legais, organizacionais e técnicos e propõe diferentes métodos. Como solicitar ao titular dos dados o seu consentimento para o processamento de dados pessoais, e minimização de dados, ou seja, limita-se a coleta, armazenamento e uso de dados pessoais aos dados que são relevantes, adequados, e necessários para cumprir a finalidade para a qual os dados são processados. Além disso, a técnica pseudonimização é explicitamente mencionada como uma medida de minimização de dados. Em dados pseudonimizados, os PIIs são substituídos por outros identificadores gerados aleatoriamente. Outra técnica de minimização de dados

é a restrição da duração do armazenamento de dados ao período necessário somente (European Parliament and Council of European Union, 2016; GRUSCHKA *et al.*, 2018).

No entanto, várias técnicas mais elaboradas são mencionadas na literatura, sendo as abordagens mais comuns k-anonimato, l-diversidade, t-proximidade, e privacidade diferencial (GRUSCHKA *et al.*, 2018). Similarmente, as técnicas e abordagens citadas por Abouelmehdi *et al.* (2017) são desidentificação baseada em k-anonimato, l-diversidade e t-proximidade, além de um modelo de computação em nuvem híbrida, no qual os dados não confidenciais ficam na nuvem pública e dados confidenciais são armazenados *on premise*, e anonimização baseada em identidade.

Em sua pesquisa, Canbay, Vural e Sagiroglu (2019) detalha as técnicas de anonimização supracitadas:

- K-anonimato: um modelo que fornece um registro indistinguível de pelo menos k-1 outros registros e fornece uma solução para a violação de divulgação de identidade.
- L-diversidade: garante que cada classe de equivalência inclua no mínimo L dados sensíveis diferentes. Essa técnica fornece uma solução contra violações de divulgação de atributos.
- T-proximidade : garante que a distribuição de dados sensíveis em uma classe de equivalência não pode exceder um valor T relativo à distribuição de dados sensíveis em toda a tabela, e é o modelo que fornece uma solução para a divulgação do atributo.
- Delta-presença: torna difícil a descoberta de informações de associação de um indivíduo para o invasor e oferece uma solução para a divulgação de atributos confidenciais.

Outras técnicas complementares às supracitadas são:

- M-invariância: é uma técnica de proteção de privacidade que avalia a variância dos elementos de um conjunto de dados ao publicar diversas versões desse conjunto, pois a variância dos elementos do conjunto pode revelar indivíduos.
- P-sensibilidade: um modelo de proteção de privacidade que considera a diversidade de consultas e informações semânticas ao tornar anônimos os locais dos usuários.

Ademais, como meios de aumentar a proteção de privacidade, Singh *et al.* (2018) sugerem paralelizar, com a técnica MapReduce, a anonimização de pequenas parcelas de

dados locais, separar dados sensíveis e não sensíveis em diferentes nuvens, e a aplicação de criptografia de identidade multinível em nível de arquivos e de blocos.

3.2.4 Demais abordagens de solução

Alguns autores definem princípios que norteiam a proteção de privacidade, como Alwabel (2020) elenca:

- Aviso, abertura e transparência
- Escolha, consentimento e controle
- Definição de escopo e minimização
- Acesso e precisão
- Salvaguardas de segurança
- Conformidade
- Propósito
- Limitação de uso e retenção
- Responsabilização

No entanto, Alwabel (2020) conclui que os princípios não são suficientes, pois é necessário traduzir os princípios em prática. Alguns autores sugerem inclusive a participação da sociedade civil de forma geral para educar e trazer ao debate a privacidade de dados (FULLER, 2019).

Em sua pesquisa, Aloysius *et al.* (2018) concluíram que a coleta de grandes quantidades de informações dos clientes tiveram impactos negativos em suas percepções da imagem da loja em todos os cenários. Na mesma pesquisa, identificaram que a digitalização móvel foi percebida favoravelmente pelos clientes em termos de facilidade de uso, mas, por outro lado, a combinação de digitalização móvel e pagamento móvel aumentou as preocupações dos clientes com privacidade, tendo efeito negativo nas intenções dos clientes de usar o aplicativo.

Durante as fases inciais do fluxo de processamento dos dados, as informações pessoais confidenciais devem ser excluídas para garantir que as informações mais sensíveis não tenham chance de vazar durante o processo, segundo Yu e Tsai (2016).

Por outro lado, ao publicar informações nas fases finais do processamento de dados, alguns meios de proteger a privacidade são: redução da granularidade, desassociação de

quase-identificadores, agrupamento de conjuntos de dados, embaralhamento de informações entre os grupos, geração de ruídos nos dados, substituição de alguns dados, mascaramento de dados, distorção probabilística e geração de dados sintéticos (SINGH *et al.*, 2018). A diminuição da granularidade do dados também é mencionada por Dev Mishra e Beer Singh (2017) como um meio de reduzir os riscos.

Na fase de aquisição de dados, Venkatraman e Venkatraman (2019) sugere a aplicação do método de proveniência de dados, usado para determinar a origem dos dados no ambiente analítico. Pela variedade de fontes de dados do BDA, os autores propõem também a adaptação da tecnologia de proveniência de dados para detectar anomalias na fase de aquisição de dados. No entanto, a coleta de metadados deve respeitar a conformidade de privacidade.

Sendo o acesso aos dados compartilhados uma das maiores causas dos problemas de privacidade, a autenticação bidirecional de um usuário é uma das possíveis soluções de preservação da privacidade, e na fase de transmissão de dados, pode-se autenticar o usuário e criptografar os dados antes da transmissão (SINGH *et al.*, 2018). Além disso, atribuir um dono ao dado pode dar mais segurança ao processo de gestão de acesso (FARRUGIA; CLAXTON; THOMPSON, 2016).

Uma das técnicas populares de processamento de dados é clusterização, isso por conta da sua capacidade de analisar dados desconhecidos. A ideia básica dessa técnica é dividir os dados sem rótulos em grupos diferentes. No entanto, o maior problema com os algoritmos de clusterização existentes é sua dependência de formatação de dados, o que o impede de ser utilizado em dados não estruturado do BDA (ALABDULLAH; BELOFF; WHITE, 2018).

Alguns métodos de proteção de privacidade mencionados por Gruschka *et al.* (2018) e Shozi e Mtsweni (2017) são:

- Supressão: remover os valores de um atributo completamente ou substituí-los por um valor fictício (normalmente um asterisco). Essa operação geralmente é executada em identificadores explícitos.
- Generalização: substitui-se valores específicos por valores mais gerais ou mais abstratos dentro da taxonomia do atributo, por exemplo, data de nascimento é generalizada para idade em anos, ou idade em anos generalizada para um intervalo de anos. Essa operação geralmente é realizada em quase-identificadores.

- Agregação: aumenta-se a cardinalidade de conjuntos, por exemplo, agrega-se todas as cidades pertencentes a uma região sob o nome da região ou da principal cidade do conjunto. Também é comumente realizada em quase-identificadores.
- Permutação: particiona-se os dados em grupos e embaralha-se os valores confidenciais em cada grupo. Como consequência, a relação entre quase-identificadores e dados confidenciais é eliminada.
- Perturbação: substituição dos valores de forma que a ligação com os dados originais seja removida, mas mantendo as propriedades estatísticas semelhantes. Um método típico para perturbação é adicionar ruído.
- Ruído aleatório: adiciona-se ruído aos dados pode reduzir as possibilidades de re-identificação. Geralmente aplicado a dados numéricos.
- Dados sintéticos: substituição dos valores originais por valores simulados de distribuições de probabilidade. Geralmente, também aplicável a dados numéricos.

No entanto, mesmo com a aplicação dessas técnicas, a privacidade de um indivíduo não pode ser considerada segura (SHOZI; MTSWENI, 2017). Mas é possível avaliar diferentes abordagens de de-identificação e anonimização para encontrar o melhor equilíbrio entre privacidade e utilidade dos dados (GRUSCHKA *et al.*, 2018).

Além das técnicas de controle de acesso e a criptografia, amplamente investigadas pela academia, Bertino (2016) também apresenta outras de abordagens de confidencialidade:

- Mesclar as diversas políticas de controle de acesso.
- Administrar de forma automatizada as autorizações e permissões de acesso ao *big data*.
- Aplicação de políticas heterogêneas de controle de acesso a dados de multimídia.
- Aplicar políticas de controle de acesso às ferramentas de armazenamento de dados.

Em relação a análise e processamento de dados privados, Bertino (2016) apresenta os seguintes métodos:

- Técnicas para controlar o que é extraído e verificar se os dados são usados para o fim pretendido.
- Suporte à privacidade pessoal e populacional em modelos de IA (inteligência artificial).
- Boa usabilidade das políticas de privacidade de dados.
- Cuidado com a qualidade dos dados, devido a suas implicações em privacidade.

- Modelagem de risco.
- Definição e gestão de proprietários dos dados.
- Estrutura do ciclo de vida dos dados.

Adicionalmente, Bertino (2016) enfatiza outras duas preocupações:

- Proteção de dados contra ameaças internas: monitoramento do comportamento dos funcionários e usuários internos. No entanto, isso pode envolver outros problemas de privacidade e, portanto, requer uma troca cuidadosa entre riscos de segurança e privacidade individual.
- Engenharia de *software* com consciência de privacidade: desde a concepção do *software*, devem ser identificadas as partes do código que lidam com dados confidenciais, além de cuidar da capacidade dos aplicativos em trabalhar com dados anônimos e gestão de permissões. Além de limpeza de memória para excluir permanentemente os dados confidenciais, ferramentas de criação de perfis de uso e log de auditoria para detecção de anomalias no uso de dados.

Os autores Hemlata e Gulia (2018) definem quatro abordagens para lidar com problemas de privacidade, conforme segue:

- Aviso e consentimento.
  - Os indivíduos devem saber para onde irão suas informações pessoais.
  - As pessoas devem ter conhecimento das interpretações e inferências que podem ser tiradas de seus dados pessoais usando técnicas de mineração de *big data* e *analytics*.
- Acesso e participação.
  - Apresentar aos usuários como seus dados são utilizados.
  - Descrever todos os direitos de acesso fornecidos aos usuários.
  - Permitir ao indivíduo corrigir seus próprios dados quando estão errados.
- Possibilidade de DNT (*Do Not Track*, em tradução livre, "não rastreie") e DNC (*Do Not Collect*, em tradução livre, "não colete")
- Desidentificação e re-identificação.

Em sua pesquisa, Liu (2019) foca nos problemas e requisitos técnicos de proteção de privacidade nos diferentes estágios do ambiente de BDA. Jiang, Shi e Zhou (2019) também

dividem os problemas e riscos de privacidade de acordo com as etapas do *big data*, que enumeram da seguintes maneira:

- Fase de coleta de dados.
- Fase de transmissão de dados.
- Fase de armazenamento de dados.
- Fase de uso e compartilhamento de dados.

Na fase de coleta de dados, os principais problemas apontados pelos autores são: função de posicionamento de dispositivos vestíveis, a falta de conhecimento do paciente sobre seus direitos de privacidade e comportamento malicioso de terceiros. A fim de evitar violação de privacidade, é recomendado o uso de tecnologias de proteção de privacidade, como criptografia, privacidade local diferencial, privacidade de gráfico social e privacidade de rastreamento de localização, além da obtenção do consentimento do usuário, considerando as características culturais do país (JIANG; SHI; ZHOU, 2019).

Na fase de transmissão, são apontados como principais riscos a falta de um padrão de protocolo de transferência de dados unificado, vulnerabilidades do mecanismo de serviço e ataques de hackers. Os autores recomendam a padronização de um protocolo de comunicação, o uso de redes privadas virtuais (conhecida também como VPN, *virtual private network*), ou o uso de protocolo de comunicação SSL (sigla para *secured socket layer*) na fase de transmissão de dados (JIANG; SHI; ZHOU, 2019).

Na terceira fase, armazenamento de dados, os principais riscos são pessoas internas roubando informações, vulnerabilidade virtual e vulnerabilidade de *firewall*. E os autores sugerem o estabelecimento de um sistema de gestão de autorização, definição consistente de responsabilidade e implementação de regulamentos de confidencialidade. Além de prover ferramentas para a proteção de privacidade, punição severa para roubo ilegal de dados e atualização constante dos *softwares* e *hardwares* de segurança (JIANG; SHI; ZHOU, 2019).

Por fim, na fase de uso e compartilhamento de informações, são apontados como principais riscos a despadronização das plataforma de informações hospitalares, comportamento desonesto do adquirente de dados, falta de leis e regulações especiais, falta de auditoria de segurança e falta de confiabilidade do certificado digital. E em resposta a esses problemas, são sugeridas a padronização da interação da plataforma hospitalar e do prontuário eletrônico, o fortalecimento das leis e regulações (isso especificamente na China)

e mecanismos de supervisão sobre a aplicação de *big data* no contexto da saúde. Além disso, também é sugerido um sistema confiável de gerenciamento de identidade digital para que o acesso seja gerenciável, controlável e rastreável, e por fim, o estabelecimento de uma cultura de proteção à privacidade e repúdio ao vazamento e roubo de dados (JIANG; SHI; ZHOU, 2019).

Em sua pesquisa, Agarwal, Gupta e Sharma (2019) também separam o fluxo do *big data* em quatro momentos, cada um com seus requisitos de privacidade. Esses momentos são: coleta, armazenamento, uso da computação em nuvem e processamento.

As soluções para as questões de privacidade de *big data* nos momentos citados por Agarwal, Gupta e Sharma (2019) são:

- Uso de fonte confiável de rede.
- Criptografia de dados.
- Anonimização de dados pessoais e confidenciais.
- Confidencialidade de dados e monitoramento de acesso.
- Monitoramento e vigilância dos dados.

Guerriero (2017) defende que haja uma camada intermediária (*middleware*) que aplica regras de controle de acesso aos dados de entrada no contexto de dados por *streaming*, e que cada fluxo tenha uma classificação de acesso específico.

Especialmente com foco em *data lakes*, Chen, Chen e Huang (2018) propõe uma arquitetura específica para a preservação de privacidade no compartilhamento de dados entre pares não confiáveis. Além de um protocolo de compartilhamento de dados e uma política de cobrança pelo *download* de dados.

Dados não-estruturados costumam ser mais difíceis de serem protegidos, pois geralmente as técnicas de proteção de privacidade exige que se saiba os quase-identificadores, dados sensíveis, etc (MEHTA; RAO, 2016). Nesse caso, os autores recomendam que a análise de dados tem que ser feita somente após os dados estarem garantidamente anonimizados ou criptografados.

Os autores Jadon e Mishra (2019) reforçam a ideia que tecnologias de armazenamento de dados não-estruturados, como NoSQL, não fornecem robustez aos dados, portanto é ainda mais importante que se tenha atenção à autorização e autenticação de usuários. Segundo Vonitsanos *et al.* (2020), os bancos de dados NoSQL têm um papel valioso no BDA, uma vez que são adequados para armazenar ou recuperar dados em alta

velocidade em nós distribuídos, no entanto a proteção de dados e o controle de acesso são alguns dos desafios ao se utilizar essa tecnologia.

Khan e Hoque (2016) descrevem em sua pesquisa um sistemas centralizado de saúde de Bangladesh e seus riscos de vazamento de dados. Os autores discutem a lucratividade que esses dados potencialmente trariam a *hackers* e a falta de um responsável quando se trata de vazamento de dados dos repositórios do governo.

O autor propõe, como chave de acessos aos dados, o uso dos celulares pelos pacientes desse sistema de saúde, que funcionaria da seguinte forma: primeiramente um código é gerado para cada registro do paciente usando os dados de identificação do paciente. Em seguida, todos os dados capazes de identificar pacientes individuais são removidos do prontuário. A chave seria gerada a partir de três atributos: número do celular, nome e sexo de um paciente (KHAN; HOQUE, 2016).

No entanto, o autor menciona as dificuldades que esse modelo enfrentaria. Em Bangladesh, muitas pessoas não sabem sua data de nascimento, principalmente idosos de baixa escolaridade. Outros problemas são relacionados o uso do número do celular, pois muitas pessoas usam vários números ou por vezes mudam de número, por diversas razões (KHAN; HOQUE, 2016).

Tendo como local de estudo a China, três setores são citados por Jiao (2021) em relação a problemas com segurança e privacidade: desenvolvimento de negócios e consumo; área da saúde; e governo. Como contramedida aos problemas dessas áreas, o autor cita três ações legislativas: legislação de proteção de privacidade em *big data*; refinamento das legislações já existentes; e restrição de uso e transmissão de análises em *big data*.

Segundo Strang e Sun (2020), embora a maioria dos países tenha legislação para proteger os pacientes contra o uso impróprio de seus dados, os provedores do domínio da saúde tendem a simplesmente evitar o registro de certos atributos de identificação. Outro problema é que até mesmo o HIPAA nos Estados Unidos permite que um hospital ignore as regras se tiver um motivo justificável. Nesses casos, a criptografia pode ser uma solução para esse problema, mas o autor sugere que haja *softwares* e *hardwares* aprimorados para tornar mais rápido e acessível os dados criptografados no setor de saúde.

Ainda no setor de saúde, à medida que mais instalações médicas mudam para registros médicos eletrônicos, prevê-se que o número de violações aumentará com o tempo se outros controles e regulamentos não forem promulgados e se a interação humano-

computador dos controles de segurança e privacidade não forem posteriores analisados (SCHMEELK, 2019).

Como alertam Martin *et al.* (2020), algumas organizações, como empresas do setor varejistas, podem deliberadamente optar por não cumprir os regulamentos e apenas esperar evitar que as irregularidades não sejam detectadas, o que pode criar desigualdade competitiva no setor. Além disso, os varejistas podem não ter incentivos a realizar práticas de minimização de dados, que exigiriam que os dados do consumidor sejam eliminados de seus sistemas ao longo do tempo. E até mesmo o direito dos consumidores de serem esquecidos podem ser desrespeitados, dado que os dados do consumidor já estão profundamente incorporados nos vários sistemas e processos da empresa.

Os autores Jurkiewicz (2018) concluem que o autopoliciamento por empresas que utilizam BDA não funciona em muitos casos, bem como os códigos de ética da indústria e de profissionais. Portanto, segurem que o governo aplique penalidades e faça cumprir as regulações de proteção de privacidade.

Nos Estados Unidos, as políticas de privacidade de BDA podem ser implementadas em dois níveis: federal e estadual. Os tratamentos da política de BDA dos EUA geralmente mostram uma falta de uma política federal coerente que abranja diversos setores, pois a maioria das políticas focam em setores específicos, como saúde, educação e instituições financeiras. As políticas de dados estão contidas na legislação de privacidade mais ampla, como HIPAA, FERPA e GLBA/FCRA para esses setores, respectivamente (DABAB *et al.*, 2018).

Como explica Chang, Ji e Arami (2019), a ocorrência frequente de vazamento de privacidade pode representar um problema de segurança pública e de ameaça à segurança da vida pessoal. De acordo com as estatísticas da Delegacia de Polícia de Zhongguancun de Pequim, o número de casos de fraude de telecomunicações em 2012 foi responsável por 32% de todos os casos de crimes.

No contexto supracitado, Dabab *et al.* (2018) defende que seja criada uma agência de fiscalização com a capacidade de impor penalidades a corporações, organizações e setores que não protegem adequadamente os dados, além de formação de políticas específicas para proteção de privacidade e financiamento governamental de pesquisas de proteção de privacidade em BDA. Complementarmente, Scotti (2017) defende que as políticas e regulamentações, em todos os níveis de governo, não devem incorporar soluções tecnológicas específicas, mas sim ser declaradas em termos dos resultados pretendidos,

além de concentrar-se mais nas aplicações reais do BDA e menos na coleta e análise dos dados.

No entanto, a mera ocorrência de crimes cibernéticos e violações de segurança de dados em seus sistemas pode trazer consequências ruins para as organizações. A perda de confidencialidade, integridade e disponibilidade de dados, resultante de furto de dados, vazamento de dados, ciberespionagem e sabotagem de dados, ameaçam o capital relacional, estrutural e humano da organização. Assim, potencialmente gera risco de reputação, dano à imagem da marca, falta de competitividade e inovação, perda do valor do conhecimento para a tomada de decisão, e dano aos ativos de infraestrutura (La Torre; DUMAY; REA, 2018; TIWARI *et al.*, 2019).

O contexto organizacional, como estratégias, cultura, estrutura e políticas da empresa, influencia as práticas e cultura de segurança organizacional, planejamento e política de segurança e privacidade, e estratégias de mitigação de risco. Salleh e Janczewski (2016) atribui às organizações várias questões de segurança e privacidade, pois para que as organizações obtenham o benefício pretendido de BDA e protejam os dados de violações de segurança, as organizações são obrigadas a fazer alterações e melhorias em termos de seus processos de negócios, além de fazer uma mudança incremental em seu modelo de negócios. E para fazer as mudanças na cultura e a conscientização de um esforço bem-sucedido, o papel da alta administração é vital na promoção da cultura de segurança e privacidade, e no fornecimento de suporte e recursos necessários, pois, do contrário, a falta de suporte da alta administração pode impedir ou dificultar esses esforços.

Em países democráticos, regulamentos são importantes para a própria manutenção da democracia, como exemplificado por Gambs (2019), o escândalo da Facebook e Cambridge Analytica demonstra que privacidade também é um ingrediente essencial para a democracia. Além disso, Scotti (2017) alerta que as organizações governamentais podem usar dados coletados inadequadamente para atentar contra a liberdade e a discriminar pessoas.

Gambs (2019) menciona diversas técnicas de proteção de privacidade, como k-anonimidade, no qual um conjunto estruturados de dados pessoais é transformado de modo que qualquer indivíduo se diferencia de pelo menos outros k-1. O autor cita também sanitização de dados, adição de ruído, e supressão de dados, e destaca a privacidade diferencial, que garante que, para qualquer análise feita em um conjunto de dados que tem essa propriedade, adicionar ou remove um único indivíduo do conjunto não muda

significativamente a probabilidade de vazamento de identidade como resultado da análise. Por fim, Gambs (2019) também menciona o *membership attack*, ataque no qual o objetivo é descobrir se um indivíduo está presente ou não em um conjunto de dados. A técnica que visa proteger um conjunto de dados desse tipe de ataque é o *delta-presence*, porém ainda é uma área de estudo muito incipiente.

A privacidade diferencial é uma técnica robusta por causa de sua prova matemática rigorosa e pode resistir a várias formas de ataque com o máximo de conhecimento prévio do invasor. O DP obtém proteção de privacidade adicionando ruído aleatório que satisfaz uma distribuição específica no conjunto de dados. Sendo que o mecanismos de adição de ruído comumente usados incluem mecanismo de Laplace, mecanismo exponencial, e mecanismo gaussiano. Importante ressaltar que a privacidade diferencial é uma definição e não um algoritmo (HU *et al.*, 2019; ALABDULLAH; BELOFF; WHITE, 2018). E, segundo Purandhar e Saravana Kumar (2019), é bastante útil para proteger dados de saúde.

3.2.5 Síntese dos problemas, causas e soluções

Com base nos resultados da RSL, foram identificados problemas, causas e soluções, sumarizados abaixo.

- Problemas:

  P1: Ameaça à vida ou à liberdade.

  P2: Assédio moral ou discriminação.

  P3: Constrangimento ou dano reputacional.

  P4: Desvantagens em negociações.

  P5: Fraudes e outros crimes.

  P6: Inviabilidade de manter-se anônimo.

  P7: Re-identificação de dados anonimizados.

  P8: Roubo ou acesso não autorizado a dados.

  P9: Vigilância ilegal.

As causas e soluções dos problemas de privacidade de dados foram limitadas ao escopo de atuação da organização, ignorando, por exemplo, diferença entre leis entre países, pois entende-se que as organizações não têm poder de decisão sobre esse aspecto.

- Causas:

  C1: Ataques e vulnerabilidade de segurança.

  C2: Deficiência da gestão de BDA.

  C3: Desafios técnicos de BDA.

  C4: Empoderamento e comunicação com o usuário.

  C5: Gestão de acesso inadequada.

  C6: Deficiência de gestão organizacional.

  C7: Revelação ou inferência de dados não autorizados.

- Soluções:

  S1: Anonimização (irreversível).

  S2: De-identificação por ruído e perturbação.

  S3: De-identificação por generalização.

  S4: De-identificação por pseudoanonimização, supressão e mascaramento de dados.

  S5: Governança de dados.

  S6: Políticas internas de proteção de privacidade.

  S7: Controle pelo usuário de seus dados.

  S8: Criptografia.

  S9: Controle de acesso.

  S10: Sanitização de dados.

As causas e as soluções foram agrupadas em conjuntos para melhor entendimento e organização, conforme quadros 2 e 3.

Quadro 2 – Causas específicas por grupo de causas

Conjunto	Causas específicas
Ataques e vulnerabilidade de segurança	Alto valor de mercado de dados médicos Ataques externos por *hackers* ou *malwares* Baixa preocupação com segurança em setores específicos Chaves de criptografia fracas Uso de ferramentas de terceiros
Deficiência da gestão de BDA	Falta de propósito de uso dos dados Falta de transparência do uso dos dados Re-propósito do uso dos dados Retenção indevida de dados
Desafios técnicos de BDA	Complexidade ao tratar de dados não estruturados Complexidade técnica da anonimização Falta de mensuração de privacidade Má qualidade dos dados Metadados com PII Ordenação dos dados publicados Protocolo de transmissão de dados inadequado Publicação indevida ou inadequada de dados Quantidade de dados Uso de chave natural PII como chave de negócio Variedade de fontes
Empoderamento e comunicação com o usuário	Desinformação, desconhecimento ou despreocupação dos usuários Falta de controle do usuário Não consentimento de uso pelo usuário Usuário controla inadequadamente seus dados
Gestão de acesso inadequada	Acesso excessivamente granular Acesso ilegal ou não autorizado Acesso inadequado por terceiros Autorização excessiva Falta de controle de acesso sobre os dados da organização
Deficiência de gestão organizacional	Responsabilização inadequada Comportamento malicioso de funcionários Comportamento malicioso de terceiros Cultura fraca em privacidade Falta de apoio da alta gestão Falta de capacitação técnica Falta de regulamentos internos
Revelação ou inferência de dados não autorizados	Inferência de dados anonimizados Inferência de dados mais granulares Inferência de dados novos Inferência de dados por cruzamento com outras bases Inferência de PIIs Re-identificação indevida de dados

Fonte: Danilo Figueiredo de Oliveira, 2023

Quadro 3 – Técnicas e métodos por grupo de soluções

Conjunto	**Técnicas e métodos**
Anonimização	K-anonimização L-diversidade T-proximidade M-invariância P-sensibilidade Delta-presença Privacidade diferencial
De-identificação por ruído e perturbação	Permutação Dado aleatório Desassociação entre quase-identificadores e identificadores Ruído aleatório Substituição por dados sintéticos Substituição por distribuição probabilística Troca de dados Condensação Embaralhamento de dados
De-identificação por generalização	Por domínio Sub-árvore Particionamento único Multi-particionamento Bucketização *Recoding* *Top-coding* Agregação Clusterização Sintetização
De-identificação por pseudoanonimização, supressão e mascaramento de dados	Supressão de registro Supressão de valor Supressão de célula Pseudoanonimização Mascaramento de dados
Governança de dados	Alertas de privacidade Arquitetura de nuvem híbrida Classificação dos dados Linhagem de dados (*data provenance*) Garantia de qualidade de dados Verificação de qualidade de dados
Políticas internas de proteção de privacidade	Auditoria Apoio da alta gestão Capacitação técnica Conscientização Cultura organizacional *Non disclosure agreement* (NDA) Responsabilização (*accountability*)
Controle pelo usuário de seus dados	Consentimento explícito Transparência Definição de propósito de uso dos dados Comunicação e notificação Política de privacidade
Criptografia	AES Criptografia homomórfica De-criptografia verificável Computação multipartidária segura Criptografia funcional
Controle de acesso	Definição de papéis Definição de propósito de acesso Definição de obrigações Granularidade de acesso Autenticação VPN Controle de acesso baseado em risco Limitação de acesso à nuvem Restrição de publicações
Sanitização de dados	Exclusão de SPI Expurgo de dados Sanitização de dados pré-publicação Supressão de PII Minimização de dados

Fonte: Danilo Figueiredo de Oliveira, 2023

4 Modelo de referência da pesquisa

Por meio deste capítulo é apresentado o modelo de referência da pesquisa. Este modelo é baseado na literatura pesquisada e visa servir com um guia para a pesquisa empírica deste trabalho. Inicialmente é apresentado o contexto da pesquisa e em seguida as questões que norteiam a pesquisa.

4.1 Contexto da pesquisa

Nesta pesquisa *big data* é apresentado como um componente de SI em organizações, e BDA se refere especificamente ao processamento, análise e apresentação dos dados do *big data*, excluindo-se dessa definição as etapas de extração, ingestão e armazenamento de dados.

Com base no objetivo de pesquisa, segurança de dados não faz parte do escopo da pesquisa, a não ser pelos métodos e técnicas que atendem aos propósitos tanto de proteção de privacidade quanto de segurança. A RSL mostrou os desafios relacionados à proteção de privacidade, que podem ser categorizados em aspectos técnicos que podem ser endereçados a partir de instrumentos tecnológicos, aspectos organizacionais, que envolvem políticas internas das organizações, e aspectos legais, que dizem respeito à interação da organização com órgãos governamentais.

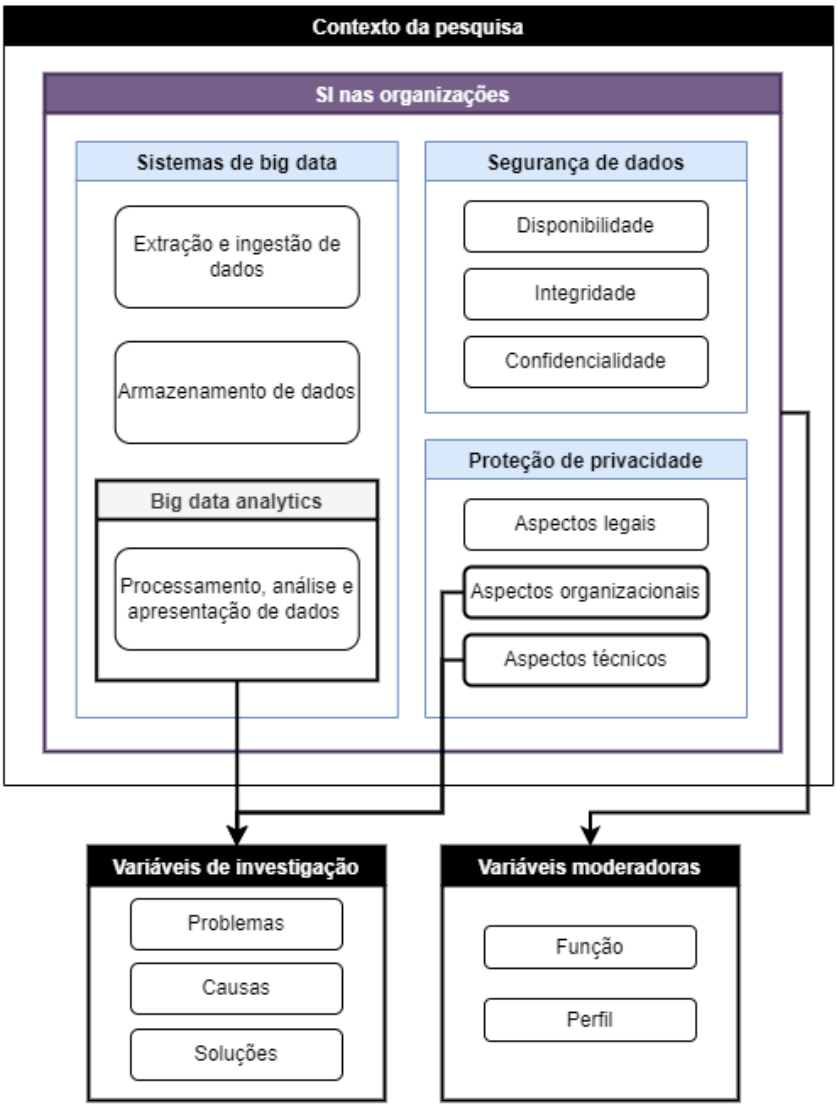
Além disso, a pesquisa tem quatro variáveis moderadoras, que representam a divisão dos painelistas em quatro grupos dependendo de suas responsabilidades técnicas e organizacionais. Os grupos são gestores e especialistas, e generalistas e não-generalistas.

Também não houve segregação de organizações por ramo ou porte. Ou seja, para esta pesquisa basta que a organização esteja inserida no contexto de BDA. Além disso, quaisquer problemas, causas e soluções que surgirem na RSL serão consideradas, independentemente de estarem no escopo operacional, tático ou estratégico da organização.

Considerando o *framework* PESTEL, acrônico em inglês para Político, Econômico, Social, Tecnológico, Ambiental, e Legal, como explicado por Issa, Chang e Issa (2010), esta pesquisa aborda apenas o aspecto tecnológico, sendo os demais aspectos considerados apenas nos contextos em que houver relação direta com tecnologia.

A figura 3 apresenta o contexto da pesquisa e as variáveis de investigação.

Figura 3 – Contexto da pesquisa



Fonte: Danilo Figueiredo de Oliveira, 2023

4.2 Questões norteadoras da pesquisa

A seguir são elencadas as questões norteadoras da RSL e da aplicação da técnica Delphi.

- Questões respondidas por meio de análise de conteúdo da RSL, conforme método proposto por Bardin (1977):
  Q1: Quais são os problemas encontrados no tratamento da privacidade de dados no contexto de BDA?

Q2: Quais são as causas identificadas ou sugeridas pelos problemas no devido tratamento da privacidade de dados no contexto de BDA?
Q3: Quais são as soluções identificadas ou sugeridas para resolver ou mitigar os problemas de privacidade de dados no contexto de BDA?

- Questões respondidas por meio da aplicação da técnica Delphi:
Q4: Quais os principais problemas relacionados à proteção da privacidade no contexto de BDA nas organizações brasileiras?
Q5: Quais as principais causas desses problemas no contexto brasileiro?
Q6: Quais as principais soluções para essas causas no contexto brasileiro?

4.3 Variáveis de pesquisa

Esta seção descreve as variáveis de investigação e variáveis moderadoras utilizadas nesta pesquisa.

4.3.1 Variáveis de investigação

A pesquisa possui três variáveis de investigação. Estas são variáveis relacionadas diretamente às questões de pesquisa.

- **Problemas**: representa uma situação considerada indesejável ou prejudicial a um indivíduo e gerada por problemas de privacidade de dados no contexto de BDA. Trata-se de uma situação social, e não tecnológica. Trata-se de uma variável do tipo nominal, com dez categorias, sendo as nove primeiras aquelas identificadas na RSL e a última a categoria denominada "Outros", que permite a ampliação do estudo com base na coleta de dados junto aos especialistas.
- **Causas**: representa aquilo que faz com que exista ou aconteça um problema de privacidade de dados no contexto de BDA (origem ou motivo do problema). Trata-se de um problema tecnológico ou um problema organizacional dentro do contexto de tecnologia de uma organização. É uma variável nominal, com oito categorias no total, sendo sete identificadas na RSL e a categoria "Outras", que permite a ampliação do estudo com base na coleta de dados junto aos especialistas.

- **Soluções**: são técnicas, métodos e práticas que podem resolver ou tornar mais branda uma ou mais causas. É uma variável nominal, com onze categorias, sendo dez identificadas na RSL e a categoria "Outras", que permite a ampliação do estudo com base na coleta de dados junto aos especialistas.

4.3.2 Variáveis moderadoras

A pesquisa possui duas variáveis moderadoras. A utilização dessas variáveis permite enriquecer a análise dos dados sob um ponto de vista específico a partir de comparação de diferentes perfis e funções dos painelistas, similarmente como utilizada por Ayabe (2018).

- **Função**: representa o tipo de função exercida pelo respondente da pesquisa. Trata-se de uma variável do tipo nominal com duas categorias: gestor ou técnico.
- **Perfil**: trata-se do perfil profissional do respondente da pesquisa em relação ao conhecimento e atividade profissional exercida. É uma variável do tipo nominal com duas categorias: generalistas (painelistas que indicaram conhecimento em mais de uma especialidade) e não-generalistas (painelistas que indicaram conhecimento em apenas uma especialidade).

5 Método de pesquisa

Neste capítulo estão apresentados os procedimentos metodológicos desta pesquisa. São abordadas as definições de tipo de pesquisa e as fases da pesquisa. Em seguida, está detalhada a coleta de dados, e, por fim, a análise e tratamento desses dados.

5.1 Tipo de pesquisa

Nesta pesquisa é utilizada a abordagem quantitativa, que se centra na objetividade, pois tem enfoque menor na interpretação do objeto ou contexto, e permite a formulação de hipóteses e quadro teórico com mais rigor que pesquisas qualitativas (GERHARDT; SILVEIRA, 2009). Ou seja, como evidenciado por Baptista e Campos (2017), a abordagem quantitativa tem como característica a neutralidade do pesquisador, pois a análise sobre os dados coletados tem menos subjetividade.

Esta é uma pesquisa de natureza aplicada, pois objetivamente visa gerar conhecimento para aplicações práticas e dirigido à solução de problemas específicos (GERHARDT; SILVEIRA, 2009).

Segundo Gil (2002), as pesquisas podem ser classificadas em três grandes grupos, com base em seus objetivos gerais: exploratória, descritiva e explicativa. Esta é uma pesquisa exploratória, pois, de acordo com o objetivo geral definido, visa proporcionar maior familiaridade com o problema, tornando-o mais explícito e constituindo hipóteses.

Este trabalho envolve pesquisa de campo, pois há coleta de dados junto a pessoas especialistas nos temas de interesse. Neste trabalho, a coleta de dados é realizada por meio da técnica Delphi.

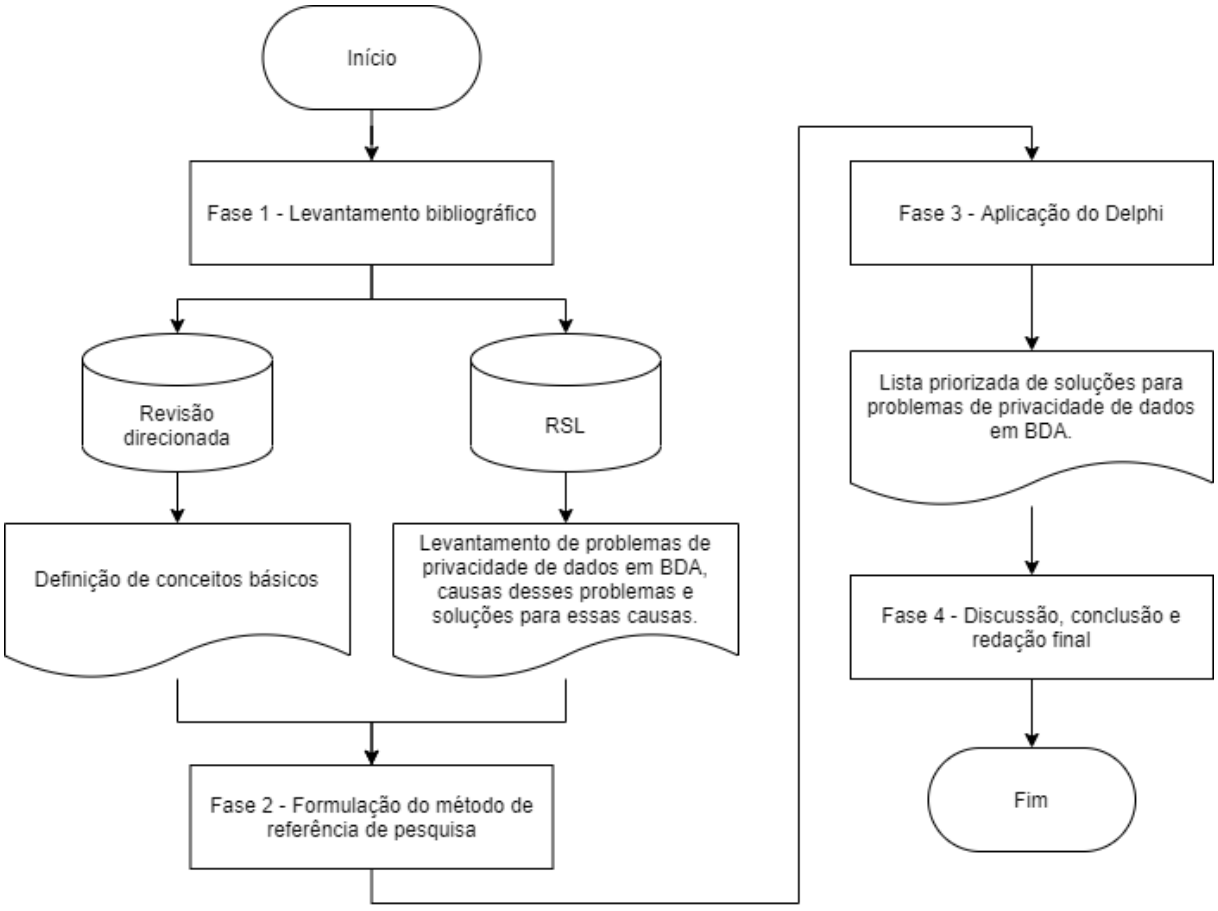
5.2 Fases de pesquisa

As fases de pesquisa, conforme representadas na figura 4, foram divididas em quatro.

- **Primeira fase**: esta fase envolve a coleta do repertório bibliográfico, no qual foram utilizadas pesquisas que definem conceitos básicos, obtidas por meio de revisão direcionada da literatura, e pesquisas que identificam problemas, causas de problemas e soluções de problemas de privacidade de dados em BDA, obtidas por meio de RSL.

- **Segunda fase**: nessa fase definiu-se o modelo de referência desta pesquisa para análise dos problemas, causas de problemas e soluções de problemas de privacidade de dados em BDA. A construção desse modelo partiu da RSL realizada na primeira fase.
- **Terceira fase**: Foi aplicado o método Delphi em um conjunto de profissionais especialistas em privacidade de dados e BDA.
- **Quarta fase**: Nesta fase serão analisados os resultados obtidos a partir da aplicação do método Delphi. Em seguida, serão elaboradas a discussão, conclusão e redação final da dissertação.

Figura 4 – Fases de pesquisa



Fonte: Danilo Figueiredo de Oliveira, 2023

5.3 Coleta, análise e tratamento dos dados

Este trabalho utilizou a técnica Delphi para coleta e tratamento de dados. Delphi é uma técnica de facilitação de grupo por um processo iterativo de várias rodadas de aplicação de questionário, projetado para transformar a opinião de especialistas em consenso do grupo. Muitos formatos podem ser empregados, mas, independente do formato, o uso apropriado dessa técnica exige um alto grau de precisão metodológica e rigor de pesquisa (HASSON; KEENEY; MCKENNA, 2000).

Após a análise estatística da opinião coletiva do grupo, os resultados da primeira rodada de aplicação do questionário auxiliam na formulação da segunda rodada. Isso ajuda a identificar itens que os painelistas podem ter deixado passar, ou seja, existe a oportunidade para os painelistas mudarem suas opiniões de acordo com as respostas do grupo da rodada anterior. Este processo é contínuo até que o consenso seja obtido ou até que o retorno de cada rodada diminua significativamente (HASSON; KEENEY; MCKENNA, 2000).

Skinner *et al.* (2015) descrevem alguns dos riscos e problemas que podem acontecer durante a aplicação da técnica Delphi como: critério ruins de seleção dos especialistas, desistência ou falta de comprometimento de alguns painelistas, o grupo pode sentir pressão para entrar em consenso ou pode haver complacência ao resultado da rodada anterior, levando ao falso consenso.

Delbecq, Gustafson e Van De Ven (1985) recomendam que o prazo de resposta dado aos painelistas de cada rodada seja de duas semanas, e estimam que a aplicação da técnica Delphi demanda no mínimo 45 dias de duração.

Segundo Powell (2003), os painelistas da pesquisa não são selecionados aleatoriamente, pois devem ser criteriosamente selecionados profissionais com experiência sólida no tema da pesquisa, pois o sucesso de um estudo Delphi depende da experiência combinada dos painelistas que compõem o grupo de especialistas. Porém, isso não garante que haja diversidade representativa.

Além da qualificações dos especialista, o segundo aspecto importante é a quantidade de painelistas, que pode variar de 15 a centenas de pessoas (POWELL, 2003). Mas, segundo Hsu e Sandford (2007), geralmente o grupo é composto por menos de 50 pessoas, e a maioria das vezes tem entre 15 e 20 painelistas. Porém há pouca evidência empírica sobre

o efeito do número de painelistas na confiabilidade ou validade dos processos de consenso (POWELL, 2003).

Conforme argumentado por Delbecq, Gustafson e Van De Ven (1985), grupos heterogêneos produzem uma proporção maior de soluções de alta qualidade e altamente aceitáveis do que grupos homogêneos. Essa ideia também é defendida por Rowe, Wright e Bolger (1991), que sugerem que os especialistas sejam provenientes de origens variadas, a fim de garantir uma ampla base de conhecimento. Segundo Powell (2003), também há evidências que indicam que a diversidade de membros do painel de especialistas resulta em melhor desempenho, pois pode permitir a consideração de diferentes perspectivas e uma gama mais ampla de alternativas.

De acordo com Hsu e Sandford (2007), Powell (2003), é importante que qualquer indivíduo do grupo não saiba a resposta de outro painelista, pois a anonimidade é um fator importante de sucesso dessa técnica. Isso, pois as ações de outras pessoas em situações de grupo podem inibir a criatividade e a possibilidade de resolver questões ambíguas e conflitantes, e é importante que as respostas não sejam enviesadas por um ou poucos indivíduos mais influentes. Afinal, a técnica tem o princípio de que o grupo tem desempenho melhor do que qualquer membro individualmente em casos em que há falta de consenso e conhecimento incompleto sobre o tema pesquisado (POWELL, 2003). No entanto, segundo Powell (2003), é importante notar que os resultados da técnica Delphi representam a opinião de especialistas, e não fato indiscutíveis.

O questionário da primeira rodada pode ser não-estruturado, e buscar respostas abertas (HSU; SANDFORD, 2007; ROWE; WRIGHT; BOLGER, 1991). Isso permite aos painelistas um espaço relativamente livre para desenvolver o tópico sob investigação (ROWE; WRIGHT; BOLGER, 1991). Nesse caso, uma análise qualitativa dos resultados é realizada e isso fornece a base ao segundo questionário e os subsequentes.

Porém, segundo Hsu e Sandford (2007), Powell (2003), também podem ser utilizadas perguntas semiestruturadas ou estruturadas no questionário, como em casos nos quais a literatura existente ou uma pesquisa auxiliar forma a base para a primeira rodada.

A segunda rodada e as subsequentes são mais específicas, com questionários que buscam a quantificação de descobertas anteriores (POWELL, 2003). Embora a possibilidade de mais de três rodadas seja oferecida, há uma necessidade de equilibrar tempo, custo e possível fadiga dos painelistas (HASSON; KEENEY; MCKENNA, 2000; ROWE; WRIGHT; BOLGER, 1991).

Segundo Powell (2003), não existe apenas um método consistente para relatar os resultados da técnica Delphi, diversas abordagens diferentes são usadas na literatura. Dentre essas abordagens, incluem-se representação gráfica, apresentação textual de resultados estatísticos, delineamento de tendências centrais, variância e *rankings*. Powell (2003) argumenta também que os dados da segunda rodada e subsequentes podem ser analisados usando técnicas de classificação. Além disso, é importante haver um meio de mostrar a dispersão das pontuações e a indicação da resposta do painelista em relação ao quadro geral.

Conforme sugerido por Linstone e Turoff (1976), a aplicação do Delphi pode ser definida em quatro fases amplas e distintas:

- Fase 1: exploração do assunto via questionário, no qual cada indivíduo contribui com informações consideradas pertinentes.
- Fase 2: compreensão de como o grupo vê a questão em termos de concordância, percepção de importância, viabilidade, etc..
- Fase 3: se houver discordância significativa, é realizada a exploração e avaliação dessa discordância para identificar as razões subjacentes às diferenças.
- Fase 4: avaliação final após todas as informações coletadas anteriormente terem sido analisadas e as avaliações enviadas aos painelistas.

Nesta pesquisa, o resultado da RSL foi utilizado como insumo para a elaboração do questionário da primeira rodada. O questionário foi estruturado, mas permitiu respostas abertas (opcionais). A partir da segunda rodada, as perguntas serão estruturadas e não permitirão respostas abertas. O prazo máximo de resposta para cada rodada é de duas semanas. O número mínimo de painelistas ao final da segundo rodada é 15.

O coeficiente de concordância W de Kendall (W) foi utilizado para medir a concordância das opiniões fornecidas pelos painelistas a partir de uma lista ordenada de respostas (SCHMIDT, 1997). Ou seja, o coeficiente foi utilizado para definir se haveria uma nova rodada ou se o painel seria finalizado. Segundo Schmidt (1997), o método de Kendall é preferível em relação a outros métodos por sua facilidade e simplicidade. O coeficiente de concordância é o valor de W resultante da fórmula a seguir:

$$W = \frac{12S^2}{m^2 n(n^2-1) - m\sum_{j=1}^{m} Tj}$$

em que:

- A variável S trata-se da soma dos desvios padrão de todos os elementos.
- A variável m trata-se da quantidade de painelistas no grupo.
- A variável n trata-se da quantidade de elementos avaliados no grupo.
- A variável $Tj = \sum_{i=1}^{g_j}(t_i^3 - t_i)$, na qual t_i é o número de postos empatados no i-ésimo agrupamento de empates e g_j é o número de grupos de empate no j-ésimo conjunto de ordenação.

A interpretação para o valor de W segue os valores dispostos no quadro 4.

Quadro 4 – Interpretação do coeficiente de concordância de Kendall (W)

W	**Interpretação**	**Confiança no *ranking***
Menor ou igual a 0,1	Concordância muito fraca	Nenhuma
Maior que 0,1 até 0,3	Concordância fraca	Baixa
Maior que 0,3 até 0,5	Concordância moderada	Média
Maior que 0,5 até 0,7	Concordância forte	Alta
Maior que 0,7 até 1	Concordância muito forte	Muito alta

Fonte: Schmidt (1997)

6 Aplicação da técnica Delphi

Neste capítulo são apresentados detalhes da aplicação da técnica Delphi

A aplicação ocorreu em duas rodadas. A primeira envolveu 17 painelistas, e a segunda envolveu 16 painelistas. Um dos painelistas da primeira rodada esteve incomunicável.

Doravante, são consideradas **dimensões** o ranqueamentos de problemas para cada causa (sete causas) e o ranqueamentos de soluções para cada causa, totalizando 14 dimensões.

6.1 Esquematização do painel

No próprio instrumento utilizado, foi informado aos painelistas que nenhum participante terá sua identidade revelada na condução do questionário ou a posteriori, seja na dissertação ou em qualquer produção acadêmica posterior ao questionário. A identificação se faz necessária única e exclusivamente para a organização do pesquisador durante aplicação do questionário.

E também foi explicado que a técnica Delphi consiste em aplicar um questionário a alguns painelistas e, em seguida, caso não cheguem a um consenso no grupo, pode-se aplicar mais rodadas de questionário com o resultado médio da rodada anterior apresentado aos painelistas. Assim, os painelistas podem revisar suas respostas com base na avaliação média dos demais. Esse processo é encerrado quando chega-se ao consenso, ou quando for verificado que não há mais ganhos significativos com novas rodadas.

O instrumento continha um *briefing* acerca das principais definições dos itens a seguir:

- BDA
- Privacidade
- Segurança da informação
- Quase-identificadores
- Problema
- Causa
- Solução
- K-anonimato

- L-diversidade
- T-proximidade
- M-invariância
- P-sensibilidade
- Delta-presença
- Exemplos de técnicas de ruído e perturbação
- Exemplos de técnicas de generalização
- Exemplos de técnicas de pseudoanonimização, supressão e mascaramento
- Criptografia homomórfica
- De-criptografia verificável

Ademais, havia a lista de causas e soluções agrupados em conjuntos, conforme os quadros 2 e 3.

O questionário, conforme apresentado no Apêndice D, inclui duas perguntas para identificação de perfil e de contato do painelista.

As outras duas perguntas do questionário foram criadas com base nos problemas, causas, e soluções. O objetivo foi viabilizar a associação de problemas com causas e causas com soluções. Para cada combinação, há uma matriz para preenchimento das notas.

O questionário foi testado com dois profissionais em dados, ambos especializados em engenharia de dados, antes de serem aplicados aos painelistas.

6.2 Montagem do grupo de painelistas

Os painelistas foram selecionados a partir da pesquisa de perfis de profissionais em dados a partir da rede social LinkedIn. Foram priorizados profissionais de primeiro grau de conexão com o autor desta pesquisa. No total, foram abordados 59 profissionais, do quais 17 participaram da primeira rodada e, desses, 16 participaram da segunda rodada.

Todos os painelistas tinham no mínimo cinco anos de experiência com dados e estavam em papéis de especialistas em suas respectivas carreiras em dados. Dos 16 painelistas, sete desempenham atividades técnicas e nove desempenham atividades de supervisão.

Os painelistas tinham domínio em pelo menos uma especialidade relacionada a dados: análise de dados (AD), ciência de dados (CD), engenharia de dados (ED), e governança de

dados (GD). No questionário foi dada a possibilidade de adicionar outras especialidades em um campo aberto. Um dos painelistas mencionou duas outras especialidades: privacidade de dados e segurança da informação.

As quatro especialidades consideradas emergiram da área de computação com foco em BDA. Uma pessoa, por exemplo, pode ser especialista em privacidade de dados, mas na perspectiva do direito, lidando com leis e outras regulações, ou ser especialista em segurança da informação, mas sem qualquer relação com BDA. Por esse motivo não foram consideradas as especialidades privacidade de dados e segurança da informação nesta pesquisa.

O quadro 5 apresenta os painelistas com suas respectivas funções profissionais, tempo de experiência em anos, e especialidades.

Quadro 5 – Painelistas por tempo de experiência e especialidade

Painelista	**Função**	**Experiência (anos)**	**AD**	**CD**	**ED**	**GD**	**Outros**
1	Gestão	>5<=10	x			x	
2	Gestão	>5<=10	x	x	x	x	
3	Técnica	>5<=10			x		
4	Técnica	>5<=10		x			
5	Técnica	>5<=10		x	x	x	
6	Gestão	>5<=10			x		
7	Gestão	>5<=10			x		
8	Gestão	>5<=10	x		x	x	
9	Gestão	>5<=10			x		
10	Técnica	>5<=10	x		x		
11	Gestão	>5<=10	x		x		
12	Técnica	>10	x				
13	Técnica	>5<=10	x		x	x	
14	Gestão	>5<=10				x	x
15	Gestão	>5<=10	x		x	x	
16	Técnica	>5<=10	x		x	x	

Fonte: Danilo Figueiredo de Oliveira, 2023

6.3 Preparação e realização da primeira rodada

Todos os painelistas receberam o questionário por e-mail, bem como orientações de preenchimento e o prazo de duas semanas para preenchimento. Porém, o prazo teve que ser estendido para que a quantidade mínima de respondentes fosse atingida. A décima

sétima resposta foi recebida após 68 dias. As respostas foram recebidas por diferentes canais, por e-mail, LinkedIn e Whatsapp.

Cada painelista preencheu duas matrizes, uma relacionando técnicas com causas e outro problemas com causas, conforme demonstrado no apêndice D.

Todas as respostas dos 17 painelistas foram consolidadas e o grau de concordância pôde ser avaliado. Após a primeira rodada, várias dimensões avaliadas tinham concordância apenas moderada, e uma delas concordância fraca. Portanto, foi realizada outra rodada a fim de atingir maior grau de consenso dentre os painelistas.

6.3.1 Resultado para causas e problemas

O quadro 6 apresenta o resultado da primeira rodada do Delphi para a matriz de causas e problemas. É possível notar que houve concordância moderada no ranqueamento dos problemas em cinco dos sete conjuntos de causas, além de uma concordância forte e outra fraca. Isso implica em confiabilidade média em cinco dimensões, além de uma alta e outra baixa, conforme a interpretação de Kendall apresentada no quadro 4. É importante notar que a confiabilidade no *ranking* depende do intervalo de W, conforme quadro 4.

Quadro 6 – W no ranqueamento de problemas em relação às causas na rodada 1 (R1)

Causas	**Coeficiente W**	**Concordância R1**
C1	0,481	Moderada
C2	0,326	Moderada
C3	0,533	Forte
C4	0,255	Fraca
C5	0,470	Moderada
C6	0,387	Moderada
C7	0,451	Moderada

Fonte: Danilo Figueiredo de Oliveira, 2023

Após a primeira interação, foi possível verificar que os painelistas obtiveram concordância forte apenas no ranqueamento dos problemas em relação ao C3 (desafios técnicos de BDA). Por essa razão se faz necessária a realização de uma segunda rodada.

6.3.2 Resultado para causas e soluções

O quadro 7 apresenta o resultado da primeira rodada do Delphi para a matriz de causas e soluções. Percebe-se que houve concordância moderada no ranqueamento dos conjuntos de soluções em três dos sete conjuntos de causas, além de concordância forte em outros três, e muito forte em um conjunto de causas.

O resultado da primeira rodada nesse contexto foi superior aos problemas por causas, o que indica maior concordância a priori no ranqueamento de soluções para as causas. Ou seja, os painelistas formaram quatro ranqueamentos de confiabilidade alta ou muito alta, e nenhum abaixo de confiabilidade média, mesmo sem *feedback* dos demais painelistas.

Os resultados indicam que se poderia parar a aplicação do Delphi na primeira rodada para causas e soluções. Porém, como foi necessária uma nova rodada para causas e problemas, estendeu-se a realização da segunda rodada para causas e soluções.

Quadro 7 – W no ranqueamento de soluções em relação às causas na R1

Causas	**Coeficiente W**	**Concordância R1**
C1	0,415	Moderada
C2	0,545	Forte
C3	0,472	Moderada
C4	0,624	Forte
C5	0,745	Muito forte
C6	0,613	Forte
C7	0,301	Moderada

Fonte: Danilo Figueiredo de Oliveira, 2023

6.4 Preparação e realização da segunda rodada

Na segunda rodada os questionários continham as médias das respostas dos painelistas ao lado da resposta de cada um deles, conforme ilustrado na figura 5. Isso permitiu que cada painelista comparasse suas respostas com as médias do grupo. Essa comparação permitiu que cada um pudesse ajustar, ou não, sua resposta em função da resposta do grupo.

Cada questionário da segunda rodada foi enviado aos painelistas por meio do mesmo canal pelo qual a resposta da primeira rodada foi enviada. O prazo de resposta foi de

Figura 5 – Matriz do questionário da segunda rodada

		Conjunto de técnicas (soluções)																					
		S1 - Anonimização		S2 - De-identificação por ruído e perturbação		S3 - De-identificação por generalização		S4 - De-identificação por pseudoanonimização, supressão e mascaramento		S5 - Governança de dados		S6 - Políticas internas de proteção de privacidade		S7 - Controle pelo usuário de seus dados		S8 - Criptografia		S9 - Controle de acesso		S10 - Sanitização de dados		S11 - Outros	
Conjunto de causas	C1 - Ataques e vulnerabilidade de segurança	2	3,9	2	3,4	2	2,9	2	3,4	5	3,3	5	3,7	3	2,5	5	4,6	5	4,1	1	2,7	3	2
	C2 - Deficiência da gestão de BDA	1	2,2	1	1,8	1	2,0	1	2,2	5	4,5	5	3,4	5	2,6	3	2,2	5	2,9	5	2,8	3	2
	C3 - Desafios técnicos de BDA	1	2,5	1	2,3	1	2,2	1	2,3	5	3,7	4	3,3	3	2,2	2	2,1	2	2,3	2	2,4	3	2
	C4 - Empoderamento e comunicação com o usuário	2	2,6	2	2,3	2	2,7	2	2,3	5	3,9	5	3,5	5	3,7	5	2,6	5	2,6	5	3,2	3	3,5
	C5 - Gestão de acesso inadequada	3	2,9	2	2,4	2	2,6	2	2,7	5	4,3	5	3,7	5	3,1	5	3,2	5	4,3	2	2,4	3	3,5
	C6 - Problemas de gestão organizacional	2	2,3	2	2,3	2	2,2	2	2,3	5	3,9	5	4,3	5	2,9	2	2,5	5	3,5	3	2,7	3	4
	C7 - Revelação ou inferência de dados não autorizados	5	4,4	5	3,8	5	3,8	5	3,8	5	3,3	5	3,3	5	2,7	5	4,1	5	3,2	5	3,7	5	5

Fonte: Danilo Figueiredo de Oliveira, 2023

duas semanas, porém também teve que ser estendido para que o máximo de respostas fossem obtidas, respeitando quantidade mínima de 15 respostas. A décima sexta resposta foi recebida 28 dias após o início da segunda rodada.

Por fim, as respostas foram consolidadas e o grau de concordância de cada dimensão foi avaliado. Na segunda rodada, 11 painelistas fizeram alterações e seis dimensões aumentaram o nível de concordância. Apenas uma dimensão permaneceu no nível fraco. Portanto, foi considerado o nível de concordância como satisfatório, e rodadas adicionais não foram necessárias.

6.4.1 Resultado para causas e problemas

Após a aplicação do questionário na segunda rodada do Delphi, a categoria de concordância aumentou em cinco dimensões, e se manteve igual nas outras duas, conforme verificado no quadro 8. Como resultado, em cinco ranqueamentos há confiabilidade alta ou muito alta, um tem confiabilidade média e outro confiabilidade baixa. Apesar de a concordância ter diminuído em duas dimensões (C3 e C4), ambas permaneceram na mesma

escala de concordância e confiabilidade no *ranking*. O *p-value*, que é o menor nível de significância para rejeitar a hipótese nula, é sempre abaixo de 0,01.

Quadro 8 – W no ranqueamento de problemas em relação às causas na rodada 2 (R2)

Causas	**W R1**	**W R2**	**Diferença**	**Conc. R1**	**Conc. R2**	***P-value***
C1	0,481	0,757	57%	Moderada	Muito forte	0,000
C2	0,326	0,520	59%	Moderada	Forte	0,000
C3	0,533	0,501	-6%	Forte	Forte	0,000
C4	0,255	0,208	-18%	Fraca	Fraca	0,004
C5	0,470	0,740	57%	Moderada	Muito forte	0,000
C6	0,387	0,489	26%	Moderada	Moderada	0,000
C7	0,451	0,654	45%	Moderada	Forte	0,000

Fonte: Danilo Figueiredo de Oliveira, 2023

6.4.2 Resultado para causas e soluções

Após a aplicação do questionário na segunda rodada do Delphi, a categoria de concordância aumentou em duas dimensões, e se manteve igual nas outras cinco, conforme verificado no quadro 9.

Como resultado, em três ranqueamentos há confiabilidade muito alta, em dois há confiabilidade alta, e nos outros dois ranqueamentos há confiabilidade média. Apenas o ranqueamento de C5 teve uma leve queda na concordância, de apenas 4%, mas permanecendo na categoria de concordância muito forte.

Quadro 9 – Concordância no ranqueamento de soluções em relação às causas na R2

Causas	**W R1**	**W R2**	**Diferença**	**Conc. R1**	**Conc. R2**	***P-value***
C1	0,415	0,430	3%	Moderada	Moderada	0,000
C2	0,545	0,563	3%	Forte	Forte	0,000
C3	0,472	0,721	53%	Moderada	Muito forte	0,000
C4	0,624	0,770	23%	Forte	Muito forte	0,000
C5	0,745	0,717	-4%	Muito forte	Muito forte	0,000
C6	0,613	0,671	9%	Forte	Forte	0,000
C7	0,301	0,347	15%	Moderada	Moderada	0,000

Fonte: Danilo Figueiredo de Oliveira, 2023

De forma geral, houve menos variação que as relações entre causas e problemas apresentadas no quadro 8.

7 Análise dos resultados

Neste capítulo são apresentados os resultados do Delphi e suas análises. As duas matrizes do questionário foram dispostas a fim de relacionar as causas com os problemas e as causas com as soluções. A primeira seção refere-se à análise de causas e problemas. As análises consideraram todo o grupo de painelistas, e subgrupos de acordo com as variáveis moderadoras desta pesquisa. Na seção seguinte foi realizado o mesmo procedimento para causas e soluções.

Conforme explicitado nos capítulos anteriores, todos esses itens têm como foco pessoas, e não organizações, pois se trata de um estudo sobre privacidade, o que pressupõe impacto à vida privada de indivíduos.

7.1 Análise de causas e problemas

Esta seção se refere à análise da associação entre causa e problemas, com subseções de análise por amostra total e variáveis de moderação.

7.1.1 Análise de todo o grupo de painelistas

A tabela 2 apresenta o ranqueamento, com empates, ao final da segunda rodada. O ranqueamento é feito de P1 a P9, do primeiro ao nono, sendo o primeiro o mais associado à causa e o nono o menos associado.

Tabela 2 – Ranqueamento dos problemas por conjunto de causas na segunda rodada

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
C1	9	5	3	7	2	4	8	1	6
C2	9	8	7	5	4	3	2	1	6
C3	9	5	7	5	4	2	1	3	8
C4	5	5	8	2	2	1	4	7	8
C5	9	6	7	8	1	4	3	2	5
C6	9	4	5	8	1	3	7	2	6
C7	6	8	5	9	4	3	2	1	6

Fonte: Danilo Figueiredo de Oliveira, 2023

A tabela 3 apresenta o total de pontos ao final da segunda rodada com os respectivos totais por conjunto de causas e problemas. Os pontos se referem às somas das notas de zero a cinco dadas pelos painelistas.

Tabela 3 – Total de pontos dos problemas e conjuntos de causas na segunda rodada

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	P7	**P8**	**P9**	**Total**
C1	57	*70*	*74*	*63*	*76*	*71*	58	*77*	*65*	**611**
C2	37	39	42	50	52	55	58	59	48	**440**
C3	37	40	39	40	47	53	54	50	38	**398**
C4	44	44	42	48	48	53	46	43	42	**410**
C5	43	56	55	48	*68*	62	*64*	*66*	57	**519**
C6	49	58	56	50	*65*	60	53	*64*	54	**509**
C7	61	60	62	51	*67*	*68*	*70*	*71*	61	**571**
Total	**328**	**367**	**370**	**350**	**423**	**422**	**403**	**430**	**365**	

Fonte: Danilo Figueiredo de Oliveira, 2023

Percebe-se que os painelistas atribuíram ao final da segunda rodada, de forma geral, maior nota a P8 (roubo ou acesso não autorizado a dados), mas próximo de P5 (Fraudes e outros crimes). Enquanto que o problema menos associado às causas foi P1 (ameaça à vida ou à liberdade).

Os valores em itálico são maiores ou iguais ao primeiro quartil (63), com base nesses valores verifica-se que P5 e P8 são problemas associados a mais causas, e têm mais causas associadas a eles. Enquanto que P1 e P4 são problemas menos complexos em relação a quantidade de causas fortemente associadas.

Conforme tabela 4, resultado semelhante é obtido ao analisar as associações que estão no primeiro quartil.

Tabela 4 – Associações entre problemas e causas que pertencem ao primeiro quartil

	Concord.	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**Tot.**
C1	**Muito forte**	0	1	1	1	1	1	0	1	1	**7**
C7	**Forte**	0	0	0	0	1	1	1	1	0	**4**
C5	**Muito forte**	0	0	0	0	1	0	1	1	0	**3**
C6	Moderada	0	0	0	0	1	0	0	1	0	2
C2	Forte	0	0	0	0	0	0	0	0	0	0
C4	Fraca	0	0	0	0	0	0	0	0	0	0
C3	Forte	0	0	0	0	0	0	0	0	0	0
	Total	0	1	1	1	**4**	2	2	**4**	1	

Fonte: Danilo Figueiredo de Oliveira, 2023

Na tabela 4 os valores acima de 63 estão no primeiro quartil. Em seguida, foi aferida a mediana das somas das pontuações no quartil superior. Em seguida, na tabela 4, as causas foram divididas em dois grupos, conforme a mediada da frequência em que elas tiveram resultados no primeiro quartil.

Além disso, nota-se que o conjunto de causas com maior pontuação é C1, e com menor pontuação é C3. Sendo assim, conclui-se que ataques e vulnerabilidade de segurança (C1) é o fator mais associado aos problemas listados, ao passo que desafios técnicos de BDA (C3) é menos associado dentre os conjuntos de causas. O ranqueamento completo de causas é apresentado no quadro 10, conforme informações da tabela 3.

Quadro 10 – *Ranking* dos conjunto de causas em relação aos problemas por pontuação total

Posição	Conjunto de causas	Pontos
1	Ataques e vulnerabilidade de segurança (C1)	611
2	Revelação ou inferência de dados não autorizados (C7)	571
3	Gestão de acesso inadequada (C5)	519
4	Deficiência de gestão organizacional (C6)	509
5	Deficiência da gestão de BDA (C2)	440
6	Empoderamento e comunicação com o usuário (C4)	410
7	Desafios técnicos de BDA (C3)	398

Fonte: Danilo Figueiredo de Oliveira, 2023

O quadro 11 apresenta o ranqueamento dos problemas por pontuação total, conforme informação da tabela 3.

Quadro 11 – Ranqueamento dos problemas por pontuação total

Posição	Problema	Pontos
1	Roubo ou acesso não autorizado a dados (P8)	430
2	Fraudes e outros crimes (P5)	423
3	Inviabilidade de manter-se anônimo (P6)	422
4	Re-identificação de dados anonimizados (P7)	403
5	Constrangimento ou dano reputacional (P3)	370
6	Assédio moral ou discriminação (P2)	367
7	Vigilância ilegal (P9)	365
8	Desvantagens em negociações (P4)	350
9	Ameaça à vida ou à liberdade (P1)	328

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C1:

O quadro 12 apresenta o ranqueamento de C1, conjunto de causas denominado "ataques e vulnerabilidade de segurança". Identifica-se que C1 está associado especialmente

a roubo ou acesso não autorizado a dados, com apenas um ponto a mais que o segundo colocado, que é "fraudes e outros crimes". Na última posição está ameaça à vida ou à liberdade, com um ponto a menos que o penúltimo, que é re-identificação de dados anonimizados.

Dentre outras, algumas das causas específicas que compõem C1 são: baixa preocupação com segurança em setores específicos, como pro exemplo saúde, ataques externos por *hackers* e *malwares*, e uso de ferramentas de terceiros. Pode-se inferir que esses incidentes maliciosos resultem em roubo ou acesso não autorizado a dados, além de fraudes e outros crimes. Outrossim, o problema de constrangimento ou dano reputacional também teve pontuação alta.

No entanto, apesar de darem maior importância a P8, os painelistas consideraram que C1 tem menor potencial em viabilizar a re-identificação de dados anonimizados. E, embora fraudes, dentre outros crimes, e constrangimento ou dano reputacional tenham tido notas mais altas, os painelistas ponderaram que há menor possibilidade de C1 ocasionar ameaça à vida ou à liberdade.

Quadro 12 – Ranqueamento dos problemas em relação a C1

Posição	**Problema**	**Pontos**
1	Roubo ou acesso não autorizado a dados (P8)	77
2	Fraudes e outros crimes (P5)	76
3	Constrangimento ou dano reputacional (P3)	74
4	Inviabilidade de manter-se anônimo (P6)	71
5	Assédio moral ou discriminação (P2)	70
6	Vigilância ilegal (P9)	65
7	Desvantagens em negociações (P4)	63
8	Re-identificação de dados anonimizados (P7)	58
9	Ameaça à vida ou à liberdade (P1)	57

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C2:

O quadro 13 apresenta o ranqueamento de C2, conjunto de causas denominado "deficiência da gestão de BDA". Novamente, roubo ou acesso não autorizado a dados é o principal problema associado, seguido por re-identificação de dados anonimizados, e inviabilidade de manter-se anônimo. Por último está, outra vez, ameaça à vida ou à liberdade.

É interessante perceber que P7 ganhou bastante relevância neste ranqueamento em comparação ao anterior, subindo de penúltimo para segundo. C2 inclui falta de propósito

de uso dos dados, falta de transparência do uso dos dados, re-propósito do uso dos dados, e retenção indevida de dados.

Assim, é presumível que quando há mudança de propósito de uso dos dados ou retenção indevida de dados, por exemplo, aumenta a chance de haver re-identificação de dados anonimizados.

Também é possível presumir que quando não há transparência no uso de dados ou não há definição de propósito de uso dos dados, por exemplo, a anonimidade dos indivíduos pode ser inviabilizada, o que pode explicar P6 na terceira posição do ranqueamento para C2.

Quadro 13 – Ranqueamento dos problemas em relação a C2

Posição	**Problema**	**Pontos**
1	Roubo ou acesso não autorizado a dados (P8)	59
2	Re-identificação de dados anonimizados (P7)	58
3	Inviabilidade de manter-se anônimo (P6)	55
4	Fraudes e outros crimes (P5)	52
5	Desvantagens em negociações (P4)	50
6	Vigilância ilegal (P9)	48
7	Constrangimento ou dano reputacional (P3)	42
8	Assédio moral ou discriminação (P2)	39
9	Ameaça à vida ou à liberdade (P1)	37

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C3:

O quadro 14 apresenta o ranqueamento de C3, conjunto de causas denominado "desafios técnicos de BDA". Nesse ranqueamento, P7 e P6 ocupam as duas primeiras posições, nessa ordem. E, novamente, ameaça à vida ou à liberdade aparece como último no ranqueamento.

C3 inclui, dentre outros, quantidade de dados, variedade de fontes, complexidade técnica da anonimização, dentre outras características inerentes, ou pelo menos muito comuns, a sistemas de BDA. Portanto, é factível entender que esse contexto facilite a re-identificação de dados anonimizados ou cause inviabilidade do indivíduo manter-se anônimo.

Esse entendimento é compatível com o que foi verificado na RSL apresentada no Capítulo 3.

Quadro 14 – Ranqueamento dos problemas em relação a C3

Posição	Problema	Pontos
1	Re-identificação de dados anonimizados (P7)	54
2	Inviabilidade de manter-se anônimo (P6)	53
3	Roubo ou acesso não autorizado a dados (P8)	50
4	Fraudes e outros crimes (P5)	47
5	Assédio moral ou discriminação (P2) Desvantagens em negociações (P4)	40
7	Constrangimento ou dano reputacional (P3)	39
8	Vigilância ilegal (P9)	38
9	Ameaça à vida ou à liberdade (P1)	37

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C4:

O quadro 15 apresenta o ranqueamento de C4, conjunto de causas denominado "empoderamento e comunicação com o usuário". No primeiro lugar está inviabilidade de manter-se anônimo. Na segunda posição há empate entre P4 e P5, respectivamente desvantagens em negociações, e Fraudes e outros crimes.

C4 inclui falta de controle do usuário, não consentimento de uso pelo usuário, controle inadequado de seus dados pelo usuário, e desinformação, desconhecimento ou despreocupação dos usuários. Isto é, as próprias ações dos usuários, seja por sua responsabilidade ou não, inviabilizam seu poder de manter-se anônimo, e pode prejudicá-los em negociações ou facilitar fraudes e outros crimes contra si.

Porém, conforme apresentado no quadro 8, esse ranqueamento tem confiabilidade baixa, por causa do W fraco.

Quadro 15 – Ranqueamento dos problemas em relação a C4

Posição	Problema	Pontos
1	Inviabilidade de manter-se anônimo (P6)	53
2	Desvantagens em negociações (P4) Fraudes e outros crimes (P5)	48
4	Re-identificação de dados anonimizados (P7)	46
5	Ameaça à vida ou à liberdade (P1) Assédio moral ou discriminação (P2)	44
7	Roubo ou acesso não autorizado a dados (P8)	43
8	Constrangimento ou dano reputacional (P3) Vigilância ilegal (P9)	42

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C5:

O quadro 16 apresenta o ranqueamento de C5, conjunto de causas denominado "gestão de acesso inadequada". O principal problema associado à C3 é Fraudes e outros crimes, seguido por roubo ou acesso não autorizado a dados. Enquanto ameaça à vida ou à liberdade teve a menor pontuação.

C5 inclui, dentre outros, acesso excessivamente granular, acesso inadequado por terceiros, e autorização excessiva. Assim como outros conjuntos de causas que tiveram fraudes e outros crimes como principais problemas associados, C5 viabiliza o uso malicioso dos dados. Ademais, esperar-se-ia que P8 (roubo e acesso não autorizado a dados) tivesse uma pontuação alta quando ao se tratar de gestão de acesso inadequada, e isso de fato ocorreu, embora tenha ficado na segunda posição.

Além disso, P7 (re-identificação de dados anonimizados) e P6 (inviabilidade de manter-se anônimo) tiveram pontuações relativamente altas.

Quadro 16 – Ranqueamento dos problemas em relação a C5

Posição	**Problema**	**Pontos**
1	Fraudes e outros crimes (P5)	68
2	Roubo ou acesso não autorizado a dados (P8)	66
3	Re-identificação de dados anonimizados (P7)	64
4	Inviabilidade de manter-se anônimo (P6)	62
5	Vigilância ilegal (P9)	57
6	Assédio moral ou discriminação (P2)	56
7	Constrangimento ou dano reputacional (P3)	55
8	Desvantagens em negociações (P4)	48
9	Ameaça à vida ou à liberdade (P1)	43

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C6:

O quadro 17 apresenta o ranqueamento de C6, conjunto de causas denominado "deficiência de gestão organizacional". Nesse ranqueamento, outra vez Fraudes e outros crimes desponta em primeiro, seguido por roubo ou acesso não autorizado a dados, com apenas um ponto a menos. Em último aparece novamente ameaça à vida ou à liberdade.

Dentre outros, C6 é composto por responsabilização inadequada (*accountability*), comportamento malicioso de funcionários e terceiros, falta de regulamentos internos, e cultura fraca em privacidade. É coerente que esses itens possam facilitar fraudes, roubos de dados, acesso não autorizado a dados, e outros crimes.

Ao passo que a concordância nesse ranqueamento foi moderada, a confiabilidade nesse *ranking* foi média.

Quadro 17 – Ranqueamento dos problemas em relação a C6

Posição	Problema	Pontos
1	Fraudes e outros crimes (P5)	65
2	Roubo ou acesso não autorizado a dados (P8)	64
3	Inviabilidade de manter-se anônimo (P6)	60
4	Assédio moral ou discriminação (P2)	58
5	Constrangimento ou dano reputacional (P3)	56
6	Vigilância ilegal (P9)	54
7	Re-identificação de dados anonimizados (P7)	53
8	Desvantagens em negociações (P4)	50
9	Ameaça à vida ou à liberdade (P1)	49

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C7:

O quadro 18 apresenta o ranqueamento de C7, conjunto de causas denominado "revelação ou inferência de dados não autorizados". Roubo ou acesso não autorizado a dados ficou na primeira posição, seguido por re-identificação de dados anonimizados, com um ponto a menos. Inviabilidade de manter-se anônimo, e Fraudes e outros crimes também tiveram pontuações relativamente altas e próximos ao primeiro colocado. Em último está desvantagens em negociações, consideravelmente distante do penúltimo colocado.

Dentre outras, algumas das causas específicas que compõem C7 estão inferência de dados anonimizados, dados mais granulares, e dados novos. De acordo com a avaliação dos painelistas, isso também potencialmente ocasiona fraudes, roubos de dados, acesso não autorizado a dados, e outros crimes. Como era possível pressupor, re-identificação de dados anonimizados teve muita relevância para C7, embora tenha ficado na segunda posição.

Quadro 18 – Ranqueamento dos problemas em relação a C7

Posição	Problema	Pontos
1	Roubo ou acesso não autorizado a dados (P8)	71
2	Re-identificação de dados anonimizados (P7)	70
3	Inviabilidade de manter-se anônimo (P6)	68
4	Fraudes e outros crimes (P5)	67
5	Constrangimento ou dano reputacional (P3)	62
6	Ameaça à vida ou à liberdade (P1) Vigilância ilegal (P9)	61
8	Assédio moral ou discriminação (P2)	60
9	Desvantagens em negociações (P4)	51

Fonte: Danilo Figueiredo de Oliveira, 2023

7.1.2 Análise por subgrupo de painelistas

Ao analisar o ranqueamento de causas associadas aos problemas por função e perfil, percebe-se que independente da categoria, o ranqueamento é muito semelhante. Sendo que a maior divergência foi verificada entre gestores e técnicos (67% de convergência). Portanto, é possível inferir que a função pode influenciar no ranqueamento dos especialistas na associação de problemas e causas.

Além disso, as quatro principais causas se mantiveram em todas as categorias, sendo que, dessas, as duas primeiras são de caráter técnico e as outras duas de gestão.

Essas quatro causas foram destacadas por terem a soma toda de pontos acima da mediana de todas as causas, conforme apresentado na tabela 3.

O quadro 19 apresenta a diferença absoluta e convergência dos ranqueamentos entre subgrupos.

Quadro 19 – Análise de convergência de *ranking* de problemas por subgrupos

		Função		**Perfil**		**Diferenças**					
	T[1]	**Gs**[2]	**Tc**[3]	**Gn**[4]	**Ng**[5]	**T-Gs**	**T-Tc**	**T-Gn**	**T-Ng**	**Gs-Tc**	**Gn-Ng**
C1	1	1	1	1	1	0	0	0	0	0	0
C7	2	2	2	2	2	0	0	0	0	0	0
C5	3	3	4	3	3	0	1	0	0	1	0
C6	4	4	3	4	4	0	1	0	0	1	0
C2	5	3	6	6	5	2	1	1	0	3	1
C4	6	7	5	5	7	1	1	1	1	2	2
C3	7	6	7	7	6	1	0	0	1	1	1
					Soma	**4**	**4**	**2**	**2**	**8**	**4**
					Convergência	**83%**	**83%**	**92%**	**92%**	**67%**	**83%**

[1] Total de painelistas.
[2] Gestores.
[3] Técnicos.
[4] Generalistas.
[5] Não-generalistas.

Fonte: Danilo Figueiredo de Oliveira, 2023

7.2 Análise de causas e soluções

Esta seção se refere à análise da associação entre causa e soluções, com subseções de análise por amostra total e variáveis de moderação.

7.2.1 Análise de todo o grupo de painelistas

A tabela 5 apresenta o ranqueamento completo ao final da segunda rodada, também podendo haver empates.

Tabela 5 – Ranqueamento dos soluções por conjunto de causas na segunda rodada

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
C1	3	7	9	6	5	4	8	1	2	9
C2	6	10	9	8	1	2	4	7	3	5
C3	3	8	10	6	1	2	9	6	4	4
C4	6	10	8	9	2	3	1	6	4	5
C5	6	10	9	8	2	3	4	4	1	7
C6	6	10	9	8	1	2	4	5	3	7
C7	2	8	9	5	3	7	10	1	3	5

Fonte: Danilo Figueiredo de Oliveira, 2023

A tabela 6 apresenta o total de pontos ao final da segunda rodada com os respectivos totais por conjunto de causas e conjunto de soluções.

Tabela 6 – Total de pontos dos conjuntos de soluções e causas na segunda rodada

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**Total**
C1	*64*	49	40	54	56	*60*	41	*74*	*66*	40	**544**
C2	38	24	26	28	*72*	*61*	46	33	51	45	**424**
C3	38	31	27	32	*60*	51	29	32	33	33	**366**
C4	32	25	28	26	*61*	*57*	*63*	32	46	42	**412**
C5	48	28	33	35	*71*	*61*	51	51	*74*	38	**490**
C6	41	30	31	34	*69*	*65*	44	42	54	39	**449**
C7	*69*	52	49	54	56	53	48	*73*	56	54	**564**
Total	**330**	**239**	**234**	**263**	**445**	**408**	**322**	**337**	**380**	**291**	

Fonte: Danilo Figueiredo de Oliveira, 2023

Percebe-se que os painelistas atribuíram ao final da segunda rodada, de forma geral, maior nota a S5 (governança de dados), seguido por S6 (políticas internas de proteção de privacidade). Enquanto que os conjuntos de soluções de menor nota foram os três relacionados a de-identificação. Respectivamente, de-identificação por pseudoanonimização, supressão e mascaramento de dados em oitavo, de-identificação por ruído e perturbação em nono, e de-identificação por generalização em décimo.

Os valores em negrito são maiores ou iguais à mediana (46), com base nesses valores verifica-se que S5 e S6 são soluções que atendem a mais causas, pois têm mais causas fortemente associadas a elas. Enquanto que S2, S3, S4 (todos tipos diferentes de

de-identificação) e S10 são soluções menos úteis para atender às causas por terem menos associações fortes com as causas.

Conforme exibido na tabela 7, o mesmo resultado é obtido ao analisar as associações que estão no primeiro quartil (acima de 56), sendo que o valor máximo de pontos possíveis a serem atribuídos é 80 (16 painelistas atribuindo nota cinco cada um). Em seguida, foi aferida a mediana das somas das pontuações no quartil superior. As soluções com total acima da mediada (2) foram as melhores soluções para atendem ao maior número de causas.

Tabela 7 – Associações entre soluções e causas que pertencem ao primeiro quartil

	Concord.	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**Tot.**
C7	Moderada	1	0	0	0	0	0	0	1	0	0	**2**
C1	**Moderada**	1	0	0	0	0	1	0	1	1	0	**4**
C5	**Muito forte**	0	0	0	0	1	1	0	0	1	0	**3**
C6	Forte	0	0	0	0	1	1	0	0	0	0	**2**
C2	Forte	0	0	0	0	1	1	0	0	0	0	2
C4	**Muito forte**	0	0	0	0	1	1	1	0	0	0	**3**
C3	Muito forte	0	0	0	0	1	0	0	0	0	0	1
	Total	**2**	**0**	**0**	**0**	**5**	**5**	**1**	**2**	**2**	**0**	

Fonte: Danilo Figueiredo de Oliveira, 2023

O quadro 20 apresenta o ranqueamento dos conjuntos de soluções por pontuação total.

Quadro 20 – Ranqueamento dos conjuntos de soluções por pontuação total

Posição	**Conjunto de soluções**	**Pontos**
1	Governança de dados (S5)	445
2	Políticas internas de proteção de privacidade (S6)	408
3	Controle de acesso (S9)	380
4	Criptografia (S8)	337
5	Anonimização (irreversível) (S1)	330
6	Controle pelo usuário de seus dados (S7)	322
7	Sanitização de dados (S10)	291
8	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	263
9	De-identificação por ruído e perturbação (S2)	239
10	De-identificação por generalização (S3)	234

Fonte: Danilo Figueiredo de Oliveira, 2023

O quadro 21 apresenta o ranqueamento dos conjunto de causas em relação aos conjuntos de soluções por pontuação total, conforme apresentado na tabela 6. Percebe-se

que C7 (revelação ou inferência de dados não autorizados) é ranqueado em primeiro como o mais associado às soluções, enquanto que C3 (desafios técnicos de BDA) tem a menor pontuação.

Exceto pela primeira e segunda posição, que estão invertidas, a ordem é idêntica ao ranqueamento de conjunto de causas em relação aos problemas apresentado no quadro 10 da seção anterior.

Quadro 21 – *Ranking* dos conjunto de causas em relação às soluções por pontuação total

Posição	**Conjunto de causas**	**Pontos**
1	Revelação ou inferência de dados não autorizados (C7)	564
2	Ataques e vulnerabilidade de segurança (C1)	544
3	Gestão de acesso inadequada (C5)	490
4	Deficiência de gestão organizacional (C6)	449
5	Deficiência da gestão de BDA (C2)	424
6	Empoderamento e comunicação com o usuário (C4)	412
7	Desafios técnicos de BDA (C3)	366

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C1:

O quadro 22 apresenta o ranqueamento de C1, conjunto de causas denominado "ataques e vulnerabilidade de segurança". Dentre os dez conjuntos de soluções, a criptografia ficou em primeiro lugar como meio de mitigar C1, relativamente distante do segundo colocado.

Na segunda e terceira posição ficaram, respectivamente, controle de acesso, e anonimização. De-identificação por generalização ficou empatado com sanitização de dados, ambos na última posição do ranqueamento, um ponto abaixo de controle pelo usuário de seus dados.

Portanto, os painelistas concordaram moderadamente (confiabilidade média no ranqueamento) que diferentes formas de criptografia que compõem S8, são as formas mais adequadas de evitar C1.

Em seguida está S9, que engloba definição de papéis, propósito e granularidade de acesso, autenticação, dentre outros que compõem esse conjunto de soluções, e técnicas de anonimização como k-anonimização e l-diversidade.

Esse entendimento é compatível com o que foi verificado na RSL apresentada no Capítulo 3.

Quadro 22 – Ranqueamento dos conjuntos de soluções em relação a C1

Posição	Conjunto de soluções	Pontos
1	Criptografia (S8)	74
2	Controle de acesso (S9)	66
3	Anonimização (irreversível) (S1)	64
4	Políticas internas de proteção de privacidade (S6)	60
5	Governança de dados (S5)	56
6	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	54
7	De-identificação por ruído e perturbação (S2)	49
8	Controle pelo usuário de seus dados (S7)	41
9	De-identificação por generalização (S3) Sanitização de dados (S10)	40

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C2:

O quadro 23 apresenta o ranqueamento de C2, conjunto de causas denominado "deficiência da gestão de BDA". O conjunto de técnicas governança de dados ocupa a primeira posição, consideravelmente acima do segundo colocado.Em seguida, políticas internas de proteção de privacidade ocupa o segundo lugar. É intuitivo que deficiência de gestão se resolva com governança, assim como políticas internas. Portanto a primeira e segunda posições são bastante coerentes.

Os três conjuntos de técnicas de de-identificação ocupam as últimas posições. Criptografia também obteve pontuação baixa nesse ranqueamento. O conjunto C2 inclui falta de propósito e de transparência de uso dos dados, re-propósito do uso dos dados, e retenção indevida de dados.

Portanto, se pressupõe que soluções muito técnicas como de-identificação, criptografia, e anonimização sejam menos associados na mitigação dessas causas mais relacionadas a gestão. E essa particularidade ocorreu no ranqueamento de C2 abaixo.

Análise de C3:

O quadro 24 apresenta o ranqueamento de C3, conjunto de causas denominado "desafios técnicos de BDA". Nesse ranqueamento, novamente governança de dados e políticas internas de proteção de privacidade ficaram, respectivamente, na primeira e na segunda posição. Nas últimas cinco posições estão os conjuntos de técnicas de de-identificação, criptografia, e controle pelo usuário de seus dados, que ficou em penúltimo.

Quadro 23 – Ranqueamento dos conjuntos de soluções em relação a C2

Posição	Conjunto de soluções	Pontos
1	Governança de dados (S5)	72
2	Políticas internas de proteção de privacidade (S6)	61
3	Controle de acesso (S9)	51
4	Controle pelo usuário de seus dados (S7)	46
5	Sanitização de dados (S10)	45
6	Anonimização (irreversível) (S1)	38
7	Criptografia (S8)	33
8	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	28
9	De-identificação por generalização (S3)	26
10	De-identificação por ruído e perturbação (S2)	24

Fonte: Danilo Figueiredo de Oliveira, 2023

Faz sentido que desafios técnicos inerentes ou comuns a sistemas de BDA tenham pouca relação com o controle pelo usuário de seus dados. Ao passo que alertas de privacidade, arquitetura de nuvem híbrida, classificação dos dados, linhagem de dados, dentre outras técnicas que compõem S5, podem ter bons resultados para lidar com má qualidade dos dados, metadados com PII, uso de chave natural PII como chave de negócio, dentre outras causas que compões C3.

Além disso, auditoria, apoio da alta gestão, capacitação técnica, conscientização, entre outros que compõem S6, aparentam importantes técnicas para o conjunto de desafios técnicos de BDA.

Quadro 24 – Ranqueamento dos conjuntos de soluções em relação a C3

Posição	Conjunto de soluções	Pontos
1	Governança de dados (S5)	60
2	Políticas internas de proteção de privacidade (S6)	51
3	Anonimização (irreversível) (S1)	38
4	Controle de acesso (S9) Sanitização de dados (S10)	33
6	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4) Criptografia (S8)	32
8	De-identificação por ruído e perturbação (S2)	31
9	Controle pelo usuário de seus dados (S7)	29
10	De-identificação por generalização (S3)	27

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C4:

O quadro 25 apresenta o ranqueamento de C4, conjunto de causas denominado "empoderamento e comunicação com o usuário". Como é presumível, para C4, controle pelo usuário de seus dados (S7) ficou na primeira posição, seguido por governança de dados e políticas internas de proteção de privacidade.

Compõem S7 os seguintes itens: consentimento explícito, transparência, definição de propósito de uso dos dados, comunicação e notificação, e política de privacidade. Evidentemente esses itens são relevantes para as causas que compõem C4, como desinformação, desconhecimento ou despreocupação dos usuários, falta de controle do usuário, não consentimento de uso pelo usuário.

Nas cinco últimas posições ficaram os três conjuntos de técnicas de de-identificação, criptografia, e anonimização. Assim como em C2, é presuntivo que soluções muito técnicas tenham menos efeito na mitigação dessas causas relacionadas à empoderamento e comunicação com o usuário.

Quadro 25 – Ranqueamento dos conjuntos de soluções em relação a C4

Posição	**Conjunto de soluções**	**Pontos**
1	Controle pelo usuário de seus dados (S7)	63
2	Governança de dados (S5)	61
3	Políticas internas de proteção de privacidade (S6)	57
4	Controle de acesso (S9)	46
5	Sanitização de dados (S10)	42
6	Anonimização (irreversível) (S1) Criptografia (S8)	32
8	De-identificação por generalização (S3)	28
9	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	26
10	De-identificação por ruído e perturbação (S2)	25

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C5:

O quadro 26 apresenta o ranqueamento de C5, conjunto de causas denominado "gestão de acesso inadequada". Nesse ranqueamento, presuntivamente a primeira posição é ocupada por controle de acesso, seguida por governança de dados e políticas internas de proteção de privacidade. As três últimas posições são ocupadas pelos conjuntos de técnicas de de-identificação.

Controle de acesso inclui definição de papéis, propósito e granularidade de acesso, autenticação, dentre outros, o que naturalmente mitiga as causas específicas que compõem

C5, como acesso excessivamente granular, ilegal ou não autorizado, autorização excessiva, dentre outros.

Embora a de-identificação possa ajudar a restringir o acesso aos dados reais indiretamente, essas técnicas têm pouco poder para influenciar a gestão de acesso inadequada.

Quadro 26 – Ranqueamento dos conjuntos de soluções em relação a C5

Posição	Conjunto de soluções	Pontos
1	Controle de acesso (S9)	74
2	Governança de dados (S5)	71
3	Políticas internas de proteção de privacidade (S6)	61
4	Controle pelo usuário de seus dados (S7) Criptografia (S8)	51
6	Anonimização (irreversível) (S1)	48
7	Sanitização de dados (S10)	38
8	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	35
9	De-identificação por generalização (S3)	33
10	De-identificação por ruído e perturbação (S2)	28

Fonte: Danilo Figueiredo de Oliveira, 2023

Análise de C6:

O quadro 27 apresenta o ranqueamento de C6, conjunto de causas denominado "deficiência de gestão organizacional". Novamente, governança de dados e políticas internas de proteção de privacidade ocupam a primeira e segunda posição no ranqueamento, e os conjuntos de técnicas de de-identificação ocupam as três últimas posições.

C6 inclui responsabilização inadequada, comportamento malicioso de funcionários e terceiros, falta de apoio da alta gestão, dentre outros. É notável que não são desafios técnicos, e dependem de soluções de gestão, logo o ranqueamento aparenta ser coerente. No entanto, pode causar surpresa que S6 tenha ficado abaixo de S5, pois parecem mais correlacionados.

Análise de C7:

O quadro 28 apresenta o ranqueamento de C7, conjunto de causas denominado "revelação ou inferência de dados não autorizados". Nas duas primeiras posições estão, respectivamente, criptografia e anonimização. Na última posição está controle pelo usuário de seus dados. É natural presumir também que a criptografia e a anonimização irreversível possam mitigar o risco de C7, como alguns dos itens que o compõem, inferência de dados anonimizados, dados mais granulares, e dados novos.

Quadro 27 – Ranqueamento dos conjuntos de soluções em relação a C6

Posição	Conjunto de soluções	Pontos
1	Governança de dados (S5)	69
2	Políticas internas de proteção de privacidade (S6)	65
3	Controle de acesso (S9)	54
4	Controle pelo usuário de seus dados (S7)	44
5	Criptografia (S8)	42
6	Anonimização (irreversível) (S1)	41
7	Sanitização de dados (S10)	39
8	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4)	34
9	De-identificação por generalização (S3)	31
10	De-identificação por ruído e perturbação (S2)	30

Fonte: Danilo Figueiredo de Oliveira, 2023

Ao passo que a concordância nesse ranqueamento foi moderada, a confiabilidade nesse *ranking* foi média.

Quadro 28 – Ranqueamento dos conjuntos de soluções em relação a C7

Posição	Conjunto de soluções	Pontos
1	Criptografia (S8)	73
2	Anonimização (irreversível) (S1)	69
3	Governança de dados (S5) Controle de acesso (S9)	56
5	De-identificação por pseudoanonimização, supressão e mascaramento de dados (S4) Sanitização de dados (S10)	54
7	Políticas internas de proteção de privacidade (S6)	53
8	De-identificação por ruído e perturbação (S2)	52
9	De-identificação por generalização (S3)	49
10	Controle pelo usuário de seus dados (S7)	48

Fonte: Danilo Figueiredo de Oliveira, 2023

7.2.2 Análise por subgrupo de painelistas

A análise do ranqueamento de causas associadas às soluções por segmento indicou 100% de convergência das principais causas em todas os segmentos. E considerando o ranqueamento total todas as comparações resultaram em mais de 90% de convergência. Ao contrário da associação de problemas e causas, o ranqueamento de soluções teve convergência alta mesmo entre gestores e técnicos.

Além disso, as quatro principais causas se mantiveram, independente da segmentação, iguais a dos problemas. A única diferença é que C1 e C7 inverteram a ordem. E, da mesma forma, das quatro principais causas, as duas primeiras são de caráter técnico e as outras duas de gestão.

Essas quatro causas foram destacadas por terem a soma toda de pontos acima da mediana de todas as causas, conforme apresentado na tabela 6.

O quadro 29 apresenta a diferença absoluta e convergência dos ranqueamentos entre subgrupos.

Quadro 29 – Análise de convergência de *ranking* de soluções por subgrupos

		Função		**Perfil**		**Diferenças**					
	T[1]	**Gs**[2]	**Tc**[3]	**Gn**[4]	**Ng**[5]	**T-Gs**	**T-Tc**	**T-Gn**	**T-Ng**	**Gs-Tc**	**Gn-Ng**
C7	1	1	1	1	1	0	0	0	0	0	0
C1	2	2	2	2	2	0	0	0	0	0	0
C5	3	3	3	3	3	0	0	0	0	0	0
C6	4	4	4	4	4	0	0	0	0	0	0
C2	5	5	5	5	5	0	0	0	0	0	0
C4	6	7	6	6	6	1	0	0	0	1	0
C3	7	6	7	7	7	1	0	0	0	1	0
					Soma	**2**	**0**	**0**	**0**	**2**	**0**
					Convergência	**92%**	**100%**	**100%**	**100%**	**92%**	**100%**

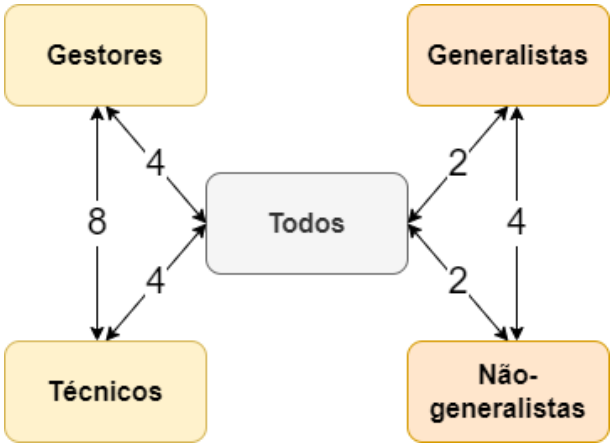
[1] Total de painelistas.
[2] Gestores.
[3] Técnicos.
[4] Generalistas.
[5] Não-generalistas.

Fonte: Danilo Figueiredo de Oliveira, 2023

7.3 Síntese dos resultados

Com a interpretação dos resultados, foi possível verificar que na maioria das dimensões os painelistas tiveram alto grau de concordância nos ranqueamentos. Além disso, independente da função ou do perfil dos painelistas, a convergência se mantém alta, exceto na comparação do ranqueamento de problemas e causas por função (gestor e técnico). A diferença entre técnicos e gestores também foi encontrada ao utilizar Delphi na pesquisa de Ayabe (2018). As figuras 6 e 7 ilustram essas comparações.

Figura 6 – Diferenças no ranqueamento de problemas por subgrupo



Fonte: Danilo Figueiredo de Oliveira, 2023

Figura 7 – Diferenças no ranqueamento de soluções por subgrupo

Fonte: Danilo Figueiredo de Oliveira, 2023

No ranqueamento geral, os problemas mais associado às causas são P5 (fraudes e outros crimes) e P8 (roubo ou acesso não autorizado a dados). Enquanto o menos associado é P1 (ameaça à vida ou à liberdade).

As causas mais associadas tanto aos problemas quanto às soluções são C1 (revelação ou inferência de dados não autorizados) e C7 (ataques e vulnerabilidade de seguranças). E as causas menos associadas tanto aos problemas quanto às soluções são C3 (desafios técnicos de BDA) e C4 (empoderamento e comunicação com o usuário).

Em relação às soluções, as mais associadas às causas são S5 (governança de dados) e S6 (políticas internas de proteção de privacidade). Por fim, as soluções menos associadas às causas são S2, S3, e S4, todas relacionadas à de-identificação, respectivamente por ruído e perturbação, por generalização, e por pseudoanonimização, supressão e mascaramento de dados.

O quadro 30 elenca os problemas e soluções mais associados a cada conjunto de causas.

Quadro 30 – Causas e seus problemas e soluções mais associados

Problemas associados	**Causas**	**Soluções associadas**
P8 - Roubo ou acesso não autorizado a dados	C1 - Ataques e vulnerabilidade de segurança.	S8 - Criptografia
P8 - Roubo ou acesso não autorizado a dados	C2 - Deficiência da gestão de BDA.	S5 - Governança de dados
P7 - Re-identificação de dados anonimizados	C3 - Desafios técnicos de BDA.	S5 - Governança de dados
P6 - Inviabilidade de manter-se anônimo	C4 - Empoderamento e comunicação com o usuário.	S7 - Controle pelo usuário de seus dados
P5 - Fraudes e outros crimes	C5 - Gestão de acesso inadequada.	S9 - Controle de acesso
P5 - Fraudes e outros crimes	C6 - Deficiência de gestão organizacional.	S5 - Governança de dados
P8 - Roubo ou acesso não autorizado a dados	C7 - Revelação ou inferência de dados não autorizados.	S8 - Criptografia

Fonte: Danilo Figueiredo de Oliveira, 2023

8 Considerações finais

O objetivo desta pesquisa foi analisar os problemas de privacidade de dados no contexto de BDA, bem como as suas causas, e identificar as principais ações e práticas que podem ser adotadas para evitar, minimizar ou resolver esses problemas identificados a partir da revisão da literatura.

O objetivo foi alcançado e questões de pesquisa foram respondidas de acordo com o resultado da RSL e da aplicação da técnica Delphi, por meio de pesquisa empírica de objetivo descritivo e natureza aplicada.

Após as buscas na literatura sobre o tema de problemas, causas e soluções de privacidade de dados em sistemas de BDA, descobriu-se nove problemas, 42 causas dos problemas de privacidade, e 68 técnicas de solução. Sendo que as causas foram agrupadas em sete conjuntos e as técnicas em 10 conjunto de soluções.

Foi verificado que na maioria das dimensões os painelistas tiveram alto grau de concordância nos ranqueamentos. Além disso, independente da função ou do perfil dos painelistas, a convergência se manteve alta, exceto na comparação do ranqueamento de problemas e causas por função (gestor e técnico).

Os problemas mais relatados em publicações na imprensa são também os de maior pontuação segundo os especialistas, como fraudes e outros crimes, e roubo ou acesso não autorizado a dados, que são frequentes nos noticiários. Enquanto que ameaça à vida ou à liberdade, desvantagens em negociações, e vigilância ilegal são menos frequentes no contexto brasileiro, mas são preocupações mais contumaz em países autoritários e não democráticos.

O ranqueamento das causas foi similar tanto para soluções quanto para problemas, exceto pelas duas primeiras posições que foram invertidas. A média da relação de causas e problemas foi maior que a média de causas e soluções, o que pode indicar que a primeira foi uma associação mais fácil para os especialistas.

As causas que estão envolvidas na gestão de segurança da informação (vulnerabilidades e ataques, revelação de dados não autorizados, e gestão de acesso) tiveram pontuações maiores, enquanto que desafios técnicos de BDA teve pontuação menor. Pode-se inferir que a complexidade causada pelos 3 Vs é menos relevante para as associações aos problemas e soluções que a gestão de segurança da informação.

E baseado na pontuação total de soluções, é possível inferir que as soluções organizacionais são mais relevantes na associação com as causas que soluções mais técnicas como de-identificação e sanitização de dados. Portanto, espera-se que, de forma geral, as causas são melhores resolvidas com boa governança de dados, políticas internas de proteção de dados, e controle de acesso.

8.1 Contribuições

A pesquisa teve contribuições práticas e teóricas. A contribuição teórica é a identificação das técnicas de solução, causas e problemas de privacidade de dados em sistemas de BDA. A RSL resultou em nove problemas, 42 causas dos problemas de privacidade, e 68 técnicas de solução ou mitigação dessas causas. Além disso, o agrupamento das causas em sete categorias e das técnicas em 10 categorias a partir da análise de conteúdo também contribui para o tema.

Ademais, a identificação de consenso ou falta de consenso para cada dimensão avaliada é outra contribuição teórica, pois pode indicar pontos que exigem maior discussão e estudo. Isso pode ser pela complexidade ou pela abrangência do assunto, que envolve uma quantidade grande técnicas. A falta de consenso no *ranking* de problemas para C4 (empoderamento e comunicação com o usuário) mostra a necessidade de novas pesquisas no tema.

A pesquisa contribuiu também para a prática gerencial e organizacional por meio da identificação e associação de causas com problemas e soluções. A partir dessa associação as organizações podem priorizar ações de proteção de privacidade de dados em sistemas de BDA a fim de otimizar esforços ao passo que se minimiza os riscos. Ainda, é possível verificar as principais causas associadas dado um problema, ou as principais soluções dada uma causa. Isso possibilita que ações mais específicas e corretas sejam tomadas a depender das peculiaridades de cada situação.

8.2 Limitações da pesquisa

Esta pesquisa foi realizada com 16 profissionais de dados brasileiros. A escolha desses especialistas partiu do primeiro grau de conexão com o autor da pesquisa. Portanto, essa amostra pode não ser generalizada para todo o país ou mundo.

Além disso, o agrupamento das 42 causas e 68 soluções em, respectivamente, sete e 10 grupos utilizando análise de conteúdo pode ser diferente a depender da técnica ou do autor.

Por fim, muitas das causas e soluções, devido ao avanço no desenvolvimento tecnológico e social, podem se tornar obsoletas ou incompletas com o passar do tempo.

8.3 Trabalhos futuros

Sugere-se pesquisas futuras para avaliar a relação detalhada entre causas e soluções. O agrupamento das causas e conjuntos pode ser aprimorado com o uso de técnicas como análise fatorial, entre outras. Além disso, a complexidade dos temas exige um profundo grau de especialização técnica. Portanto, pesquisas abordando técnicas específicas que, neste pesquisa, compuseram um conjunto de soluções, podem contribuir para o avanço do conhecimento acerca desses temas.

Outrossim, os painelistas deram notas que resultaram em dados quantitativos, porém não foram obtidas as opiniões qualitativas dos painelistas, como por exemplo, o racional no qual se basearam para dar as notas.

Esta pesquisa não abrangeu aspectos legais da privacidade de dados. No entanto, entender como dispositivos legais impactam na adoção de boas práticas de proteção de privacidade de dados auxilia priorização desse tema nas organizações.

Neste trabalho foram estudados sistemas de BDA, porém os dados de grandes organizações estão geralmente espalhados por diversos sistemas diferentes. Novas pesquisas podem dar foco nesses diferentes sistemas, que demandam diferentes níveis de acesso, desempenho computacional, disponibilidade, etc., o que culmina em diferentes abordagens de proteção de privacidade de dados.

Referências

ABOUELMEHDI, K.; BENI-HSSANE, A.; KHALOUFI, H.; SAADI, M. Big data emerging issues: Hadoop security and privacy. In: *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*. IEEE, 2016. p. 731–736. ISBN 978-1-5090-5146-5. Disponível em: ⟨http://ieeexplore.ieee.org/document/7905621/⟩. Citado 2 vezes nas páginas 50 e 57.

ABOUELMEHDI, K.; BENI-HSSANE, A.; KHALOUFI, H.; SAADI, M. Big data security and privacy in healthcare: A Review. *Procedia Computer Science*, Elsevier B.V., v. 113, p. 73–80, 2017. ISSN 18770509. Citado 4 vezes nas páginas 19, 21, 57 e 59.

AGARWAL, S.; GUPTA, M.; SHARMA, A. Big Data Privacy Issues Solutions. *Proceedings of the IEEE International Conference Image Information Processing*, v. 2019-Novem, p. 225–228, 2019. ISSN 2640074X. Citado 2 vezes nas páginas 54 e 65.

AHMADIAN, A. S.; STRÜBER, D.; RIEDIGER, V.; JÜRJENS, J. Supporting privacy impact assessment by model-based privacy analysis. *Proceedings of the ACM Symposium on Applied Computing*, p. 1467–1474, 2018. Citado na página 29.

ALABDULLAH, B.; BELOFF, N.; WHITE, M. Rise of Big Data - Issues and Challenges. *21st Saudi Computer Society National Computer Conference, NCC 2018*, IEEE, p. 0–5, 2018. Citado 3 vezes nas páginas 56, 61 e 69.

ALASHOOR, T.; HAN, S.; JOSEPH, R. C. Familiarity with big data, privacy concerns, and self-disclosure accuracy in social networking websites: An APCO model. *Communications of the Association for Information Systems*, v. 41, p. 62–96, 2017. ISSN 15293181. Citado na página 49.

Ali Khan, P. M.; Sudhakar Reddy, N.; Manoj Kumar, K. Data entities & its privacy with big data techniques in e-health systems. *International Journal of Engineering and Advanced Technology*, v. 9, n. 1, p. 232–235, 2019. ISSN 22498958. Citado na página 55.

ALOYSIUS, J. A.; HOEHLE, H.; GOODARZI, S.; VENKATESH, V. Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. *Annals of Operations Research*, Springer US, v. 270, n. 1-2, p. 25–51, 2018. ISSN 15729338. Citado na página 60.

ALWABEL, A. A. Privacy Issues in Big Data from Collection to Use. In: . Montreal: [s.n.], 2020. p. 382–391. Citado na página 60.

AYABE, F. *Fatores Críticos de Sucesso para Terceirização de Tecnologia da Informação no setor público brasileiro*. 103 p. Tese (Dissertação de mestrado), 2018. Disponível em: ⟨https://teses.usp.br/teses/disponiveis/100/100131/tde-16102018-102401/pt-br.php⟩. Citado 2 vezes nas páginas 76 e 107.

BAPTISTA, M. N.; CAMPOS, D. C. *Metodologias Pesquisa em Ciências - Análise Quantitativa e Qualitativa*. 2º. ed. [S.l.]: LTC, 2017. Citado na página 77.

BARDIN, L. *Análise de Conteúdo retirar*. [S.l.: s.n.], 1977. v. 22. 225 p. ISSN 1098-6596. ISBN 972-44-0020-4. Citado na página 74.

BARKER, K.; ASKARI, M.; BANERJEE, M.; GHAZINOUR, K.; MACKAS, B.; MAJEDI, M.; PUN, S.; WILLIAMS, A. A data privacy taxonomy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 5588 LNCS, p. 42–54, 2009. ISSN 03029743. Citado 4 vezes nas páginas 19, 25, 30 e 31.

BERTINO, E. Data Security and Privacy: Concepts, Approaches, and Research Directions. *Proceedings - International Computer Software and Applications Conference*, IEEE, v. 1, p. 400–407, 2016. ISSN 07303157. Citado 2 vezes nas páginas 62 e 63.

BONDEL, G.; GARRIDO, G. M.; BAUMER, K.; MATTHES, F. The use of de-identification methods for secure and privacy-enhancing big data analytics in cloud environments. *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, v. 2, n. Iceis, p. 338–344, 2020. Citado na página 57.

BOSE, R. Competitive intelligence process and tools for intelligence analysis. *Industrial Management and Data Systems*, v. 108, n. 4, p. 510–528, 2008. ISSN 02635577. Citado 2 vezes nas páginas 32 e 37.

BOYD, D.; CRAWFORD, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, v. 15, n. 5, p. 662–679, 2012. ISSN 1369118X. Citado na página 33.

BRASIL. *Lei Geral de Proteção de Dados Pessoais*. 2018. Disponível em: ⟨http://www.planalto.gov.br/ccivil{_}03/{_}ato2015-2018/2018/lei/l137⟩. Citado 6 vezes nas páginas 19, 21, 27, 28, 29 e 31.

BROEDERS, D.; SCHRIJVERS, E.; SLOOT, B. van der; BRAKEL, R. van; HOOG, J. de; Hirsch Ballin, E. Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data. *Computer Law and Security Review*, Elsevier Ltd, v. 33, n. 3, p. 309–323, 2017. ISSN 02673649. Citado na página 20.

CANBAY, Y.; VURAL, Y.; SAGIROGLU, S. Privacy Preserving Big Data Publishing. *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*, p. 24–29, 2019. Citado 2 vezes nas páginas 49 e 59.

CAVOUKIAN, A. Privacy by design [leading edge]. *IEEE Technology and Society Magazine*, IEEE, v. 31, n. 4, p. 18–19, 2012. ISSN 02780097. Citado na página 29.

CHANG, V.; JI, Z.; ARAMI, M. Privacy and ethical issues of big data in the airline industry. *COMPLEXIS 2019 - Proceedings of the 4th International Conference on Complexity, Future Information Systems and Risk*, p. 139–148, 2019. Citado na página 67.

CHEN, D.; ZHAO, H. Data security and privacy protection issues in cloud computing. *Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012*, v. 1, n. 973, p. 647–651, 2012. Citado na página 30.

CHEN, H.; H.L.CHIANG, R.; C. Storey, V. Business Intelligence and Analytics: From Big Data To Big Impact. *MIS Quarterly*, v. 36, n. 4, p. 1165–1188, 2018. ISSN 01406736. Citado 3 vezes nas páginas 32, 34 e 36.

CHEN, W.; GUO, F.; WANG, F.-y. A Survey of Traffic Data Visualization - Pegar referências citadas. *IEEE Transactions on Intelligent Transportation Systems*, v. 16, n. 6, p. 2970–2984, 2015. ISSN 1524-9050. Citado na página 39.

CHEN, Y. H.; CHEN, H. H.; HUANG, P. C. Enhancing the data privacy for public data lakes. *Proceedings of 4th IEEE International Conference on Applied System Innovation 2018, ICASI 2018*, IEEE, p. 1065–1068, 2018. Citado 2 vezes nas páginas 38 e 65.

CHOUDHARY, P.; GARG, K. An experimental technique on potential issues and prospective solution for preserving privacy in big data. *International Journal of Innovative Technology and Exploring Engineering*, v. 8, n. 8 Special Issue 3, p. 504–508, 2019. ISSN 22783075. Citado na página 48.

CNN Brasil. *50 milhões de senhas de e-mail de brasileiros podem ter sido vazadas*. 2021. Disponível em: ⟨https://www.cnnbrasil.com.br/business/2021/03/05/50-milhoes-de-senhas-de-e-mail-de-brasileiros-podem-ter-sido-vazadas⟩. Citado na página 20.

CNN Brasil. *Novo vazamento expõe dados telefônicos de mais de 100 milhões de brasileiros*. 2021. Disponível em: ⟨https://www.cnnbrasil.com.br/business/2021/02/10/novo-vazamento-expoe-dados-telefonicos-de-mais-de-100-milhoes-de-brasileiros⟩. Citado na página 20.

COLESKY, M.; HOEPMAN, J. H.; HILLEN, C. A Critical Analysis of Privacy Design Strategies. In: *Proceedings - 2016 IEEE Symposium on Security and Privacy Workshops, SPW 2016*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2016. p. 33–40. ISBN 9781509008247. Citado 4 vezes nas páginas 27, 28, 29 e 32.

CONGER, S.; LOCH, K. D.; HELFT, B. L. Ethics and information technology use: a factor analysis of attitudes to computer use. *Information Systems Journal*, v. 5, n. 3, p. 161–183, 1995. ISSN 13652575. Citado na página 18.

CONSTANTIOU, I. D.; KALLINIKOS, J. New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*, v. 30, n. 1, p. 44–57, 2015. ISSN 14664437. Citado na página 29.

CONWAY, M. E. How do committees invent. *Datamation*, v. 14, n. 4, p. 28–31, 1968. ISSN 00116963. Citado na página 30.

COOPER, A. What is "Analytics"? Definition and Essential Characteristics. *CETIS Analytics Series*, v. 1, n. 5, p. 1–10, 2012. ISSN 2051-9214. Disponível em: ⟨http://publications.cetis.ac.uk/2012/521⟩. Citado na página 36.

DABAB, M.; CRAVEN, R.; BARHAM, H.; GIBSON, E. Exploratory strategic roadmapping framework for big data privacy issues. In: *PICMET 2018 - Portland International Conference on Management of Engineering and Technology: Managing Technological Entrepreneurship: The Engine for Economic Growth, Proceedings*. [S.l.: s.n.], 2018. ISBN 9781890843373. Citado 2 vezes nas páginas 51 e 67.

Dama International. *Data Management Body of Knowledge (DMBOK)*. 2a edição. ed. Basking Ridge: Technics Publications, 2017. ISBN 9781634622349. Citado 3 vezes nas páginas 29, 40 e 41.

DAS, S. R. Business and market intelligence 2.0, part 2: The finance web: Internet information and markets. *IEEE Intelligent Systems*, IEEE, v. 25, n. 2, p. 74–78, 2010. ISSN 15411672. Citado na página 32.

DAVENPORT, T. H. Spotlight on Making Your Company Data-Friendly. *Harvard Business Review*, n. 5, p. 64–72, 2013. Citado 3 vezes nas páginas 32, 35 e 36.

DELBECQ, A. L.; GUSTAFSON, D. H.; Van De Ven, A. H. *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. 1985. Citado 2 vezes nas páginas 79 e 80.

Dev Mishra, A.; Beer Singh, Y. Big data analytics for security and privacy challenges. *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2016*, IEEE, p. 50–53, 2017. Citado 3 vezes nas páginas 37, 54 e 61.

DOYLE, A.; LIPPERT, R.; LYON, D. *Eyes everywhere: The global growth of camera surveillance*. [S.l.: s.n.], 2013. 1–392 p. ISBN 9780203141625. Citado na página 18.

DREISCHMEIER, R.; CLOSE, K.; TRICHET, P. *The digital imperative*. Boston, 2015. Citado na página 18.

EASTIN, M. S.; BRINSON, N. H.; DOOREY, A.; WILCOX, G. Living in a big data world: Predicting mobile commerce activity through privacy concerns. *Computers in Human Behavior*, Elsevier Ltd, v. 58, p. 214–220, 2016. ISSN 07475632. Citado na página 49.

European Parliament and Council of European Union. *General Data Protection Regulations*. 2016. Disponível em: ⟨https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679{&}from⟩. Citado 5 vezes nas páginas 27, 28, 31, 58 e 59.

FAN, J.; HAN, F.; LIU, H. Challenges of Big Data analysis. *National Science Review*, v. 1, n. 2, p. 293–314, 2014. ISSN 2053714X. Citado na página 37.

FARRUGIA, A.; CLAXTON, R.; THOMPSON, S. Towards social network analytics for understanding and managing enterprise data lakes. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, IEEE, p. 1213–1220, 2016. Citado 2 vezes nas páginas 38 e 61.

FIORINIA, P. d. C.; SELES, B. M. R. P.; JABBOUR, C. J. C.; MARIANO, E. B.; JABBOUR, A. B. L. d. S. Management theory and big data literature: From a review to a research agenda. *International Journal of Information Management*, Elsevier, v. 43, n. May, p. 112–129, 2018. ISSN 02684012. Citado 2 vezes nas páginas 34 e 42.

FULLER, M. Big data and the Facebook scandal: Issues and responses. *Theology*, v. 122, n. 1, p. 14–21, jan 2019. ISSN 0040-571X. Disponível em: ⟨http://journals.sagepub.com/doi/10.1177/0040571X18805908⟩. Citado na página 60.

G1. *Megavazamento de dados de 223 milhões de brasileiros: o que se sabe e o que falta saber*. 2021. Disponível em: ⟨https://g1.globo.com/economia/tecnologia/noticia/2021/01/28/vazamento-de-dados-de-223-milhoes-de-brasileiros-o-que-se-sabe-e-o-que-falta-saber.ghtml⟩. Citado na página 20.

GAMBS, S. Privacy and Ethical Challenges in Big Data. In: . [s.n.], 2019. p. 17–26. Disponível em: ⟨http://link.springer.com/10.1007/978-3-030-18419-3_2⟩. Citado 4 vezes nas páginas 53, 56, 68 e 69.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, Elsevier Ltd, v. 35, n. 2, p. 137–144, 2015. ISSN 02684012. Citado 3 vezes nas páginas 33, 34 e 36.

GERHARDT, T. E.; SILVEIRA, T. D. *Métodos de pesquisa*. 1a. ed. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2009. ISBN 978-85-386-0071-8. Citado na página 77.

GHANI, N. A.; HAMID, S.; UDZIR, N. I. Big data and data protection: Issues with purpose limitation principle. *International Journal of Advances in Soft Computing and its Applications*, v. 8, n. 3, p. 116–121, 2016. ISSN 20748523. Citado 4 vezes nas páginas 48, 52, 53 e 55.

GIL, A. C. *Como elaborar projetos de pesquisa*. 4ª. ed. São Paulo: Atlas, 2002. Citado na página 77.

GOOGLE. *Google Search*. 2020. Disponível em: ⟨https://www.google.com/⟩. Citado na página 45.

GOOGLE. *Google Translator*. 2020. Disponível em: ⟨http://translate.google.com/⟩. Citado na página 45.

GOOGLE. *Google Trends*. 2020. Disponível em: ⟨https://trends.google.com/trends/explore?date=today5-y{&}q=BigData,DataAnalyt⟩. Citado na página 33.

GREENLEAF, G. The influence of European data privacy standards outside Europe: Implications for globalization of convention 108. *International Data Privacy Law*, v. 2, n. 2, p. 68–92, 2012. ISSN 20444001. Citado 2 vezes nas páginas 26 e 27.

GRUSCHKA, N.; MAVROEIDIS, V.; VISHI, K.; JENSEN, M. Privacy issues and data protection in big data: A case study analysis under GDPR. *arXiv*, IEEE, p. 5027–5033, 2018. Citado 5 vezes nas páginas 28, 58, 59, 61 e 62.

GUERRIERO, M. Privacy-aware data-intensive applications. In: *ASE 2017 - Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*. [S.l.: s.n.], 2017. p. 1030–1033. ISBN 9781538626849. Citado na página 65.

HARTZOG, W. The Case Against Idealising Control. *European Data Protection Law Review*, v. 4, n. 4, p. 423–432, 2018. ISSN 23642831. Citado 2 vezes nas páginas 25 e 26.

HASHEM, I. A. T.; YAQOOB, I.; ANUAR, N. B.; MOKHTAR, S.; GANI, A.; Ullah Khan, S. The rise of "big data"on cloud computing: Review and open research issues. *Information Systems*, Elsevier, v. 47, p. 98–115, 2015. ISSN 03064379. Citado 6 vezes nas páginas 20, 31, 34, 39, 41 e 56.

HASSON, F.; KEENEY, S.; MCKENNA, H. Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, v. 32, n. 4, p. 1008–1015, 2000. ISSN 03092402. Citado 2 vezes nas páginas 79 e 80.

HEMLATA; GULIA, P. Dci3 model for privacy preserving in big data. *Advances in Intelligent Systems and Computing*, v. 654, p. 351–362, 2018. ISSN 21945357. Citado na página 63.

HSU, C. C.; SANDFORD, B. A. The Delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation*, v. 12, n. 10, p. 1–8, 2007. ISSN 15317714. Citado 2 vezes nas páginas 79 e 80.

HU, Y.; GE, L.; ZHANG, G.; QIN, D. Research on differential privacy for medical health big data processing. *Proceedings - 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2019*, p. 140–145, 2019. Citado 2 vezes nas páginas 49 e 69.

IMD World Digital. IMD World Digital Competitiveness Ranking 2020. *IMD World Competitiveness Center*, p. 180, 2020. Disponível em: ⟨https://www.imd.org/globalassets/wcc/docs/release-2017/world{_}digital{_}competitiveness{_}yearbook⟩. Citado na página 20.

ISSA, D. T.; CHANG, A. V.; ISSA, D. T. Sustainable Business Strategies and PESTEL Framework. *Gstf International Journal on Computing*, v. 1, n. 1, p. 73–80, 2010. ISSN 20102283. Citado na página 73.

JADON, P.; MISHRA, D. K. Security and Privacy Issues in Big Data: A Review. In: . [s.n.], 2019. p. 659–665. Disponível em: ⟨http://link.springer.com/10.1007/978-981-13-2285-3_77⟩. Citado na página 65.

JAGADISH, H.; GEHRKE, J.; LABRINIDIS, A.; PAPAKONSTANTINOU, Y.; PATEL, J.; RAMAKRISHNAN, R.; SHAHABI, C. Big data and its technical challenges. *Communications of the ACM*, v. 57, n. 7, p. 86–94, 2014. Citado 4 vezes nas páginas 36, 38, 41 e 42.

JIANG, R.; SHI, M.; ZHOU, W. A Privacy Security Risk Analysis Method for Medical Big Data in Urban Computing. *IEEE Access*, v. 7, p. 143841–143854, 2019. Citado 5 vezes nas páginas 54, 55, 63, 64 e 65.

JIAO, Y. Necessity and countermeasures of big data security and privacy protection. In: *Advances in Intelligent Systems and Computing*. [S.l.: s.n.], 2021. v. 1244 AISC, p. 1002–1006. ISBN 9783030539795. Citado na página 66.

JOSHI, N.; KADHIWALA, B. Big data security and privacy issues — A survey. *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, p. 1–5, 2017. Citado 3 vezes nas páginas 22, 31 e 54.

JURKIEWICZ, C. L. Big Data, Big Concerns: Ethics in the Digital Age. *Public Integrity*, Taylor Francis, v. 0, n. 0, p. 1–14, 2018. ISSN 15580989. Disponível em: ⟨https://doi.org/10.1080/10999922.2018.1448218⟩. Citado na página 67.

KEIM, D.; QU, H.; MA, K. L. Big-data visualization. *IEEE Computer Graphics and Applications*, IEEE, v. 33, n. 4, p. 20–21, 2013. ISSN 02721716. Citado na página 39.

KESHAV, S. How to read a paper. *ACM SIGCOMM Computer Communication Review*, v. 37, n. 3, p. 83–84, jul 2007. ISSN 0146-4833. Citado na página 47.

KHAN, S. I.; HOQUE, A. S. M. L. Privacy and security problems of national health data warehouse: A convenient solution for developing countries. *Proceedings of 2016 International Conference on Networking Systems and Security, NSysS 2016*, IEEE, 2016. Citado 3 vezes nas páginas 26, 53 e 66.

KHANAN, A.; ABDULLAH, S.; MOHAMED, A. H. H. M.; MEHMOOD, A.; ARIFFIN, K. A. Z. Big Data Security and Privacy Concerns: A Review. In: . [s.n.], 2019. p. 55–61. Disponível em: ⟨http://link.springer.com/10.1007/978-3-030-01659-3_8⟩. Citado na página 50.

KIERKEGAARD, S.; WATERS, N.; GREENLEAF, G.; BYGRAVE, L.; LLOYD, I.; SAXBY, S. 30 years on - the review of the council of europe data protection convention 108. *Computer Law and Security Review*, v. 27, p. 223–231, 2011. Citado na página 27.

KITCHENHAM, B. Procedures for Performing Systematic Reviews. *Keele University Technical Report*, v. 33, n. 2004, p. 1–26, 2004. ISSN 09754466. Citado na página 44.

KITCHIN, R. Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, SAGE Publications Ltd, v. 1, n. 1, 2014. ISSN 20539517. Citado na página 18.

KOKOTT, J.; SOBOTTA, C. The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR. *International Data Privacy Law*, v. 3, n. 4, p. 222–228, 2013. ISSN 20444001. Citado na página 31.

La Torre, M.; DUMAY, J.; REA, M. A. Breaching intellectual capital: critical reflections on Big Data security. *Meditari Accountancy Research*, v. 26, n. 3, p. 463–482, 2018. ISSN 20493738. Citado na página 68.

LINSTONE, H.; TUROFF, M. The Delphi Method: Techniques and Applications. *Journal of Marketing Research*, v. 13, n. 3, p. 317–318, 1976. Citado na página 81.

LIU, H. Research on privacy protection framework design and key technologies in large data environment. *Proceedings - 2019 International Conference on Robots and Intelligent System, ICRIS 2019*, IEEE, p. 327–330, 2019. Citado na página 63.

MACCORMACK, A.; BALDWIN, C.; RUSNAK, J. Exploring the duality between product and organizational architectures: A test of the "mirroring" hypothesis. *Research Policy*, v. 41, n. 8, p. 1309–1324, 2012. ISSN 0048-7333. Citado na página 30.

MADDEN, S. From databases to big data. *IEEE Internet Computing*, IEEE, v. 16, n. 3, p. 4–6, 2012. ISSN 10897801. Citado 2 vezes nas páginas 34 e 38.

MAOHONG, Z.; AIHUA, Y.; HUI, L. Research on security and privacy of big data under cloud computing environment. *ACM International Conference Proceeding Series*, p. 52–55, 2018. Citado na página 56.

MARTIN, K. D.; KIM, J. J.; PALMATIER, R. W.; STEINHOFF, L.; STEWART, D. W.; WALKER, B. A.; WANG, Y.; WEAVEN, S. K. Data Privacy in Retail. *Journal of Retailing*, New York University, v. 96, n. 4, p. 474–489, 2020. ISSN 00224359. Disponível em: ⟨https://doi.org/10.1016/j.jretai.2020.08.003⟩. Citado na página 67.

MCAFEE, A.; BRYNJOLFSSON, E. Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017. *Harvard Business Review*, n. October, p. 1–9, 2012. Citado 3 vezes nas páginas 32, 34 e 42.

MEHROTRA, S.; SHARMA, S.; ULLMAN, J. D.; GHOSH, D.; GUPTA, P.; MISHRA, A. PANDA: Partitioned Data Security on Outsourced Sensitive and Non-sensitive Data. *ACM Transactions on Management Information Systems*, v. 11, n. 4, 2020. ISSN 21586578. Citado na página 57.

MEHTA, B. B.; RAO, U. P. Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges. *Physics Procedia*, v. 78, p. 120–124, 2016. ISSN 18753892. Citado na página 65.

MICHAELIS. *Dicionário prático língua portuguesa*. 3a edição. ed. [S.l.]: Melhoramentos, 2016. ISBN 8506078598. Citado 2 vezes nas páginas 25 e 43.

MITHAS, S.; LEE, M. R.; EARLEY, S.; MURUGESAN, S.; DJAVANSHIR, R. Leveraging big data and business analytics. *IT Professional*, IEEE, v. 15, n. 6, p. 18–20, 2013. ISSN 15209202. Citado na página 33.

MÜLLER, O.; JUNGLAS, I.; BROCKE, J. V.; DEBORTOLI, S. Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, v. 25, n. 4, p. 289–302, 2016. ISSN 14769344. Citado 6 vezes nas páginas 18, 21, 29, 34, 37 e 43.

NEMITZ, P. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 376, n. 2133, 2018. ISSN 1364503X. Citado na página 19.

NORRIS, C.; SOLOWAY, E. A disruption is coming. A primer for educators on the mobile technology revolution. *Mobile Technology for Children*, p. 83–98, 2009. Citado na página 18.

Olhar Digital. *Dados vendidos na web incluem WhatsApp, profissão e salário de 112 milhões de brasileiros*. 2021. Disponível em: ⟨https://olhardigital.com.br/2021/03/19/seguranca/dados-vendidos-na-web-expoem-112-milhoes-de-brasileiros/⟩. Citado na página 20.

Oxford University. *Oxford English Dictionary*. 2020. Disponível em: ⟨https://en.oxforddictionaries.com/definition/analytics⟩. Citado 2 vezes nas páginas 33 e 43.

PATEL, K.; PATEL, B.; MISHRA, M.; PATEL, N. Privacy issues in big data. *2017 2nd International Conference for Convergence in Technology, I2CT 2017*, v. 2017-Janua, p. 259–264, 2017. Citado na página 50.

POWELL, C. The Delphi technique: Myths and realities. *Journal of Advanced Nursing*, v. 41, n. 4, p. 376–382, 2003. ISSN 03092402. Citado 3 vezes nas páginas 79, 80 e 81.

PURANDHAR, N.; Saravana Kumar, N. M. Review of data extraction, segregation privacy with big data analytics in the online health care systems. *Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019*, IEEE, n. Iciss, p. 193–197, 2019. Citado na página 69.

Rama Devi, G.; Rajesh Babu, Y. A research on security, privacy issues and privacy preserving techniques - Big data. *International Journal of Innovative Technology and Exploring Engineering*, v. 8, n. 6 Special Issue 4, p. 1571–1576, 2019. ISSN 22783075. Citado 2 vezes nas páginas 52 e 54.

Ramya Devi, R.; Vijaya Chamundeeswari, V. Triple DES: Privacy Preserving in Big Data Healthcare. *International Journal of Parallel Programming*, Springer US, v. 48, n. 3, p. 515–533, 2020. ISSN 15737640. Citado na página 57.

RANJAN, J.; FOROPON, C. Big Data Analytics in Building the Competitive Intelligence of Organizations. *International Journal of Information Management*, Elsevier Ltd, v. 56, n. August 2020, p. 102231, 2021. ISSN 0268-4012. Citado 8 vezes nas páginas 31, 32, 33, 34, 37, 41, 42 e 43.

ROSS, P.; MCGOWAN, C. G.; STYGER, L. E. J. A Comparison of Theory and Practice in Market Intelligence Gathering for Australian Micro-Businesses and SMEs. *SSRN Electronic Journal*, p. 1–17, 2013. Citado na página 32.

ROWE, G.; WRIGHT, G.; BOLGER, F. Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change*, v. 39, n. 3, p. 235–251, 1991. ISSN 00401625. Citado na página 80.

SALLEH, K. A.; JANCZEWSKI, L. Technological, Organizational and Environmental Security and Privacy Issues of Big Data: A Literature Review. *Procedia Computer Science*, The Author(s), v. 100, p. 19–28, 2016. ISSN 18770509. Citado na página 68.

SANTOS, A. S. dos. *INSIGHT: qual é o significado e a tradução desse anglicismo?* 2018. Disponível em: ⟨https://www.teclasap.com.br/insight/⟩. Citado na página 36.

SAXENA, S. Privacy concerns in integrating big data in "e-Oman". *Journal of Information, Communication and Ethics in Society*, v. 15, n. 4, p. 385–396, 2017. ISSN 17588871. Citado na página 58.

SCHADT, E. E.; LINDERMAN, M. D.; SORENSON, J.; LEE, L.; NOLAN, G. P. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, Nature Publishing Group, v. 11, n. 9, p. 647–657, 2010. ISSN 14710056. Citado 2 vezes nas páginas 21 e 35.

SCHAUB, F.; KONINGS, B.; WEBER, M. Context-Adaptive Privacy: Leveraging Context Awareness to Support Privacy Decision Making. *IEEE Pervasive Computing*, v. 14, n. 1, p. 34–43, jan 2015. ISSN 1536-1268. Citado na página 29.

SCHMEELK, S. E. Where is the risk? Analysis of government reported patient medical data breaches. *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, WI 2019 Companion*, p. 269–272, 2019. Citado na página 67.

SCHMIDT, R. C. Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, v. 28, n. 3, p. 763–774, 1997. ISSN 00117315. Citado 2 vezes nas páginas 81 e 82.

SCHWARTZ, P. M.; SOLOVE, D. J. The PII problem: Privacy and a new concept of personally identifiable information. *New York University Law Review*, v. 86, n. 6, p. 1814–1894, 2011. ISSN 00287881. Citado na página 28.

SCOTTI, V. Big data or big (privacy) problem? *IEEE Instrumentation Measurement Magazine*, v. 20, n. 5, p. 23–26, oct 2017. ISSN 1094-6969. Disponível em: ⟨http://ieeexplore.ieee.org/document/8036692/⟩. Citado 2 vezes nas páginas 67 e 68.

SERRADO, J.; PEREIRA, R. F.; Mira da Silva, M.; Scalabrin Bianchi, I. Information security frameworks for assisting GDPR compliance in banking industry. *Digital Policy, Regulation and Governance*, v. 22, n. 3, p. 227–244, 2020. ISSN 23985038. Citado na página 20.

SHAMSI, J.; KHOJAYE, M. Understanding privacy violations in big data systems. *IT Professional*, v. 20, n. 3, p. 73–81, 2018. Citado na página 50.

SHAYTURA, S. V.; STEPANOVA, M. G.; SHAYTURA, A. S.; ORDOV, K. V.; GALKIN, N. A. APPLICATION OF INFORMATION-ANALYTICAL SYSTEMS. *Journal of Theoretical and Applied Information Technology*, v. 90, n. 2, 2016. Citado na página 19.

SHOZI, N. A.; MTSWENI, J. Big data privacy in social media sites. *2017 IST-Africa Week Conference, IST-Africa 2017*, 2017. Citado 2 vezes nas páginas 61 e 62.

SINGH, M.; HALGAMUGE, M.; EKICI, G.; JAYASEKARA, C. A review on security and privacy challenges of big data. In: *Lecture Notes on Data Engineering and Communications Technologies*. [S.l.: s.n.], 2018. v. 14, p. 175–200. Citado 8 vezes nas páginas 21, 48, 52, 54, 55, 58, 59 e 61.

SKINNER, R.; NELSON, R. R.; CHIN, W. W.; LAND, L. The Delphi method research strategy in studies of information systems. *Communications of the Association for Information Systems*, v. 37, p. 31–63, 2015. ISSN 15293181. Citado na página 79.

SOLOVE, D. J. Conceptualizing privacy. *California Law Review*, v. 90, n. 4, p. 1087–1155, 2002. ISSN 00081221. Citado na página 25.

STAHL, B. C.; WRIGHT, D. Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security and Privacy*, IEEE, v. 16, n. 3, p. 26–33, 2018. ISSN 15584046. Citado na página 21.

STRANG, K. D.; SUN, Z. Hidden big data analytics issues in the healthcare industry. *Health Informatics Journal*, v. 26, n. 2, p. 981–998, 2020. ISSN 17412811. Citado na página 66.

STUTZMAN, F.; HARTZOG, W. Boundary regulation in social media. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, p. 769–778, 2012. Citado na página 25.

TERZI, R.; SINANC, D.; SAGIROGLU, S. A survey on security and privacy issues in big data. In: *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*. [S.l.: s.n.], 2015. Citado 5 vezes nas páginas 19, 34, 35, 52 e 57.

TIWARI, A.; SHARMA, N.; KAUSHIK, I.; TIWARI, R. Privacy Issues Security Techniques in Big Data. In: *Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019*. [S.l.: s.n.], 2019. v. 2019-Janua, p. 51–56. ISBN 9781728148267. Citado 2 vezes nas páginas 54 e 68.

VARSHNEY, S.; MUNJAL, D.; BHATTACHARYA, O.; SABOO, S.; AGGARWAL, N. Big data privacy breach prevention strategies. *Proceedings - 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, iSSSC 2020*, 2020. Citado na página 57.

VATSALAN, D.; KARAPIPERIS, D.; GKOULALAS-DIVANIS, A. An Overview of Big Data Issues in Privacy-Preserving Record Linkage. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.l.]: Springer International Publishing, 2019. v. 11409 LNCS, p. 118–136. ISBN 9783030197582. Citado na página 57.

VENKATRAMAN, S.; VENKATRAMAN, R. Big data security challenges and strategies. *AIMS Mathematics*, v. 4, n. 3, p. 860–879, 2019. ISSN 24736988. Citado na página 61.

VONITSANOS, G.; DRITSAS, E.; KANAVOS, A.; MYLONAS, P.; SIOUTAS, S. Security and Privacy Solutions associated with NoSQL Data Stores. *SMAP 2020 - 15th International Workshop on Semantic and Social Media Adaptation and Personalization*, 2020. Citado na página 65.

WALL, J. D.; LOWRY, P. B.; BARLOW, J. B. Organizational violations of externally governed privacyand security rules: Explaining and predicting selective violations under conditions of strain and excess. *Journal of the Association for Information Systems*, v. 17, n. 1, p. 39–76, 2016. ISSN 15583457. Citado 2 vezes nas páginas 26 e 29.

WANG, K. A survey on risks of big data privacy. In: *Advances in Intelligent Systems and Computing*. [S.l.: s.n.], 2018. v. 580, p. 161–167. ISBN 9783319670706. Citado 5 vezes nas páginas 19, 22, 51, 52 e 56.

WIERINGA, J.; KANNAN, P. K.; MA, X.; REUTTERER, T.; RISSELADA, H.; SKIERA, B. Data analytics in a privacy-concerned world. *Journal of Business Research*, Elsevier, v. 122, n. May 2019, p. 915–925, 2021. ISSN 01482963. Disponível em: ⟨https://doi.org/10.1016/j.jbusres.2019.05.005⟩. Citado na página 51.

WU, E.; BATTLE, L.; MADDEN, S. R. The case for data visualization management systems [vision paper]. *Proceedings of the VLDB Endowment*, v. 7, n. 10, p. 903–906, 2014. ISSN 21508097. Citado na página 40.

WU, X.; ZHU, X.; WU, G. Q.; DING, W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 26, n. 1, p. 97–107, 2014. ISSN 10414347. Citado 2 vezes nas páginas 18 e 20.

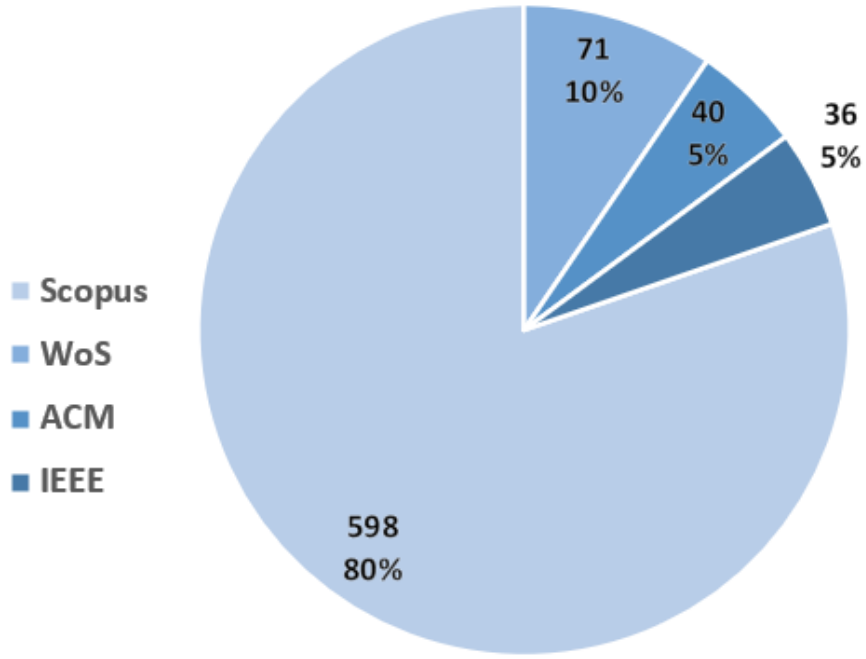
YING, S.; GRANDISON, T. Big data privacy risk: Connecting many large data sets. *Proceedings - 2016 IEEE 2nd International Conference on Collaboration and Internet Computing, IEEE CIC 2016*, IEEE, p. 86–91, 2017. Citado 4 vezes nas páginas 22, 48, 49 e 55.

YU, Y. C.; TSAI, D. R. A privacy weaving pipeline for open big data. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, IEEE, p. 997–998, 2016. Citado na página 60.

Apêndice A – Detalhes da busca da RSL

Foram utilizados quatro motores de busca de publicações científicas. No total, 745 resultados foram obtidos a partir das buscas nos quatro motores, incluindo os resultados repetidos. A figura 8 apresenta a quantidade de artigos por motor de busca obtidos a partir dos parâmetros configurados, antes da aplicação de qualquer CE. Nota-se que a maior parte das pesquisas foram encontradas pelo Scopus, seguido por Web of Science, ACM e IEEE, nessa ordem.

Figura 8 – Quantidade de resultados por motor de busca

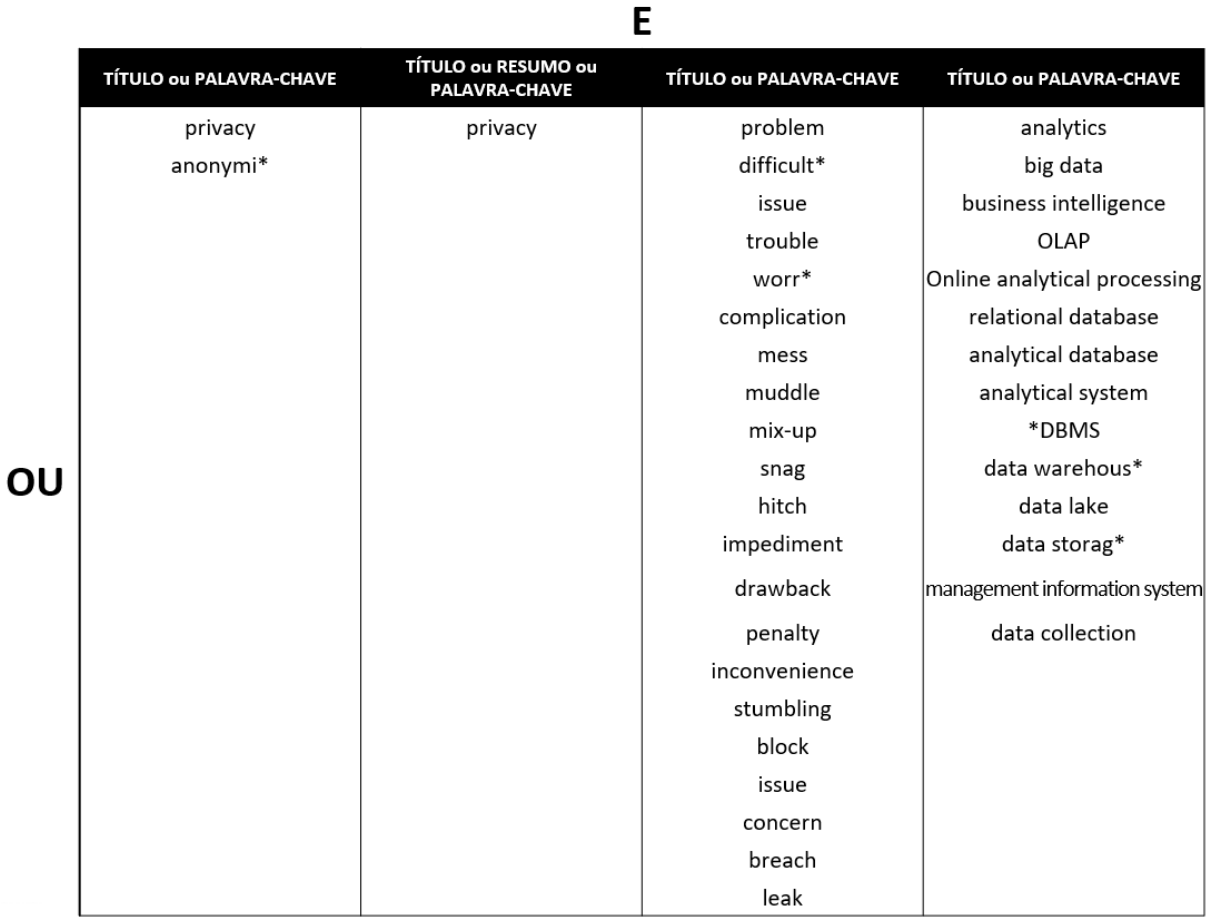


Fonte: Danilo Figueiredo de Oliveira, 2023

Para facilitar a compreensão das *strings* de busca utilizadas para pesquisar por publicações que atendessem aos critérios de inclusão, a figura 9 apresenta a estrutura lógica que serviu de base às *strings*.

As colunas representam o operador lógico **E** e as linhas representam o operador lógico **OU**. Sendo assim, era necessário que no **título ou palavra-chave** existisse a palavra "*privacy*" **OU** o radical "*anonymi*" **E** no **título ou resumo ou palavra-chave** tivesse a palavra "*privacy*" **E** no **título ou palavra-chave** existisse uma das palavras ou radicais análogos a "*problem*" **E** no **título ou palavra-chave** tivesse uma das palavras ou radicais relacionados a "*big data analytics*".

Figura 9 – Estrutura lógica da string de busca



Fonte: Danilo Figueiredo de Oliveira, 2023

As *strings* de busca estão apresentadas, de forma canônica, a seguir.

A.1 ACM Digital Library

- Base: The ACM Guide to Computing Literature

- *Link*: https://dl.acm.org/

- Caminho: opção "*Advanced Search*"

- *String* de busca: (Title:(privacy OR anonymi*) OR Keyword:(privacy OR anonymi*)) AND (Title:(privacy) OR Abstract:(privacy) OR Keyword:(privacy)) AND Title:((problem OR difficult* OR issue OR trouble OR worr* OR complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience OR stumbling OR block OR issue OR concern OR breach OR leak) OR Keyword:(problem OR difficult* OR issue OR trouble OR worr* OR

complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience OR stumbling OR block OR issue OR concern OR breach OR leak)) AND (Title:(analytics OR "big data"OR "business intelligence"OR "OLAP"OR "Online analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "DBMS"OR "RDBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection") OR Keyword:("analytics"OR "big data"OR "business intelligence"OR "OLAP"OR "Online analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "DBMS"OR "RDBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection"))

- Filtros: [Publication Date: (01/01/2017 TO 03/31/2021)]. Tipo de publicação e idioma foram filtrados manualmente.

A.2 IEEE Xplore

- Base: todas as bases

- *Link*: https://ieeexplore.ieee.org/

- Caminho: opção "*Advanced Search*" e, em seguida, "*Command Search*"

- *String* de busca: (("Document Title":privacy OR "Document Title":anonymity) OR ("Author Keywords": privacy OR "Author Keywords": anonymity)) AND ("Document Title":privacy OR "Abstract":privacy OR "Author Keywords": privacy) AND (("Document Title":problem OR "Document Title":"difficult*"OR "Document Title":issue OR "Document Title":trouble OR "Document Title":"worr*"OR "Document Title":complication OR "Document Title":mess OR "Document Title":muddle OR "Document Title":"mix-up"OR "Document Title":snag OR "Document Title":hitch OR "Document Title":impediment OR "Document Title":drawback OR "Document Title":penalty OR "Document Title":inconvenience OR "Document Title":stumbling OR "Document Title":block OR "Document Title":issue OR "Document Title":concern OR "Document Title":breach OR "Document Title":leak) OR ("Author Keywords": problem OR "Author Keywords": "difficult*"OR "Author Keywords": issue OR "Author Keywords": trouble OR "Author Keywords": "worr*"OR "Author Keywords": complication OR "Author Keywords": mess OR "Author Keywords": muddle OR "Author Keywords": "mix-up"OR "Author Keywords": snag OR "Author Keywords": hitch OR "Author Keywords": impediment OR

"Author Keywords": drawback OR "Author Keywords": penalty OR "Author Keywords": inconvenience OR "Author Keywords": stumbling OR "Author Keywords": block OR "Author Keywords": issue OR "Author Keywords": concern OR "Author Keywords": breach OR "Author Keywords": leak)) AND (("Document Title":analytics OR "Document Title":"big data"OR "Document Title":"business intelligence"OR "Document Title":"OLAP"OR "Document Title":"Online analytical processing"OR "Document Title":"relational database"OR "Document Title":"analytical database"OR "Document Title":"analytical system"OR "Document Title":"DBMS"OR "Document Title":"RDBMS"OR "Document Title":"data warehous*"OR "Document Title":"data lake"OR "Document Title":"data storage"OR "Document Title":"management information system"OR "Document Title":"data collection") OR ("Author Keywords": analytics OR "Author Keywords": "big data"OR "Author Keywords": "business intelligence"OR "Author Keywords": "OLAP"OR "Author Keywords": "Online analytical processing"OR "Author Keywords": "relational database"OR "Author Keywords": "analytical database"OR "Author Keywords": "analytical system"OR "Author Keywords": "DBMS"OR "Author Keywords": "RDBMS"OR "Author Keywords": "data warehous*"OR "Author Keywords": "data lake"OR "Author Keywords": "data storage"OR "Author Keywords": "management information system"OR "Author Keywords": "data collection"))

- Filtros: ano de publicação de 2017 a 2021. Tipo de publicação e idioma foram filtrados manualmente.

A.3 Scopus

- Base: todas as bases

- *Link*: https://www.scopus.com/

- Caminho: opção "*Document Search*" e, em seguida, "*Advanced*"

- *String* de busca: ((TITLE (privacy OR anonymi*) OR KEY (privacy OR anonymi*)) AND TITLE-ABS-KEY (privacy)) AND (TITLE (problem OR difficult* OR issue OR trouble OR worr* OR complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience OR stumbling OR block OR issue OR concern OR breach OR leak) OR KEY (problem OR difficult* OR issue OR trouble OR worr* OR complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience

OR stumbling OR block OR issue OR concern OR breach OR leak)) AND (TITLE (analytics OR "big data"OR "business intelligence"OR "OLAP"OR "Online analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "*DBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection") OR KEY (analytics OR "big data"OR "business intelligence"OR "OLAP"OR "Online analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "*DBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection"))

- Filtros: PUBYEAR > 2015 AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "re") OR LIMIT-TO (DOCTYPE, "ch")) AND (LIMIT-TO (SUBJAREA, "COMP")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SRCTYPE, "p") OR LIMIT-TO (SRCTYPE, "j") OR LIMIT-TO (SRCTYPE, "k"))

A.4 Web of Science

- Base: todas as bases

- *Link*: http://apps.webofknowledge.com/

- Caminho: opção "*Pesquisa avançada*"

- *String* de busca: (((TI=(privacy OR anonymi*) OR AK=(privacy OR anonymi*)) AND (TI=(privacy) OR AB=(privacy) OR AK=(privacy)) AND (TI=(problem OR difficult* OR issue OR trouble OR worr* OR complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience OR stumbling OR block OR issue OR concern OR breach OR leak) OR AK=(problem OR difficult* OR issue OR trouble OR worr* OR complication OR mess OR muddle OR "mix-up"OR snag OR hitch OR impediment OR drawback OR penalty OR inconvenience OR stumbling OR block OR issue OR concern OR breach OR leak)) AND (TI=(analytics OR "big data"OR "business intelligence"OR "OLAP"OR "Online analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "DBMS"OR "RDBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection") OR AK=("analytics"OR "big data"OR "business intelligence"OR "OLAP"OR "Online

analytical processing"OR "relational database"OR "analytical database"OR "analytical system"OR "DBMS"OR "RDBMS"OR "data warehous*"OR "data lake"OR "data storag*"OR "management information system"OR "data collection"))))

- Filtros: Últimos 5 anos. AND IDIOMA: (English) AND TIPOS DE DOCUMENTO: (Article OR Database Review OR Proceedings Paper OR Review)

Apêndice B – Relação de referências primárias da RSL

O quadro 31 apresenta a relação das 61 pesquisas que foram selecionadas para a RSL após a aplicação dos CE e CI e seus respectivos autores e ano de publicação.

Quadro 31 – Referências primárias

Título original	**Autor(es)**	**Ano**
A Privacy Security Risk Analysis Method for Medical Big Data in Urban Computing	JIANG; SHI; ZHOU	2019
A privacy weaving pipeline for open big data	YU; TSAI	2016
A research on security, privacy issues and privacy preserving techniques - Big data	RAMA; RAJESH	2019
A review on security and privacy challenges of big data	SINGH et al.	2018
A survey on risks of big data privacy	WANG	2018
A survey on security and privacy issues in big data	TERZI; TERZI; SAGIROGLU	2016
An experimental technique on potential issues and prospective solution for preserving privacy in big data	CHOUDHARY; GARG	2019
An Overview of Big Data Issues in Privacy-Preserving Record Linkage	VATSALAN; KARAPIPERIS; GKOULALASDIVANIS	2019
Big data analytics for security and privacy challenges	DEV; BEER	2017
Big data and data protection: Issues with purpose limitation principle	GHANI; HAMID; UDZIR	2016
Big data and the Facebook scandal: Issues and responses	FULLER	2019
Big data emerging issues: Hadoop security and privacy	ABOUELMEHDI et al.	2017
Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes	ALOYSIUS et al.	2018
Big data or big (privacy) problem?	SCOTTI	2017
Big Data Privacy Breach Prevention Strategies	VARSHNEY et al.	2020
Big data privacy in social media sites	SHOZI; MTSWENI	2017
Big Data Privacy Issues Solutions	AGARWAL; GUPTA; SHARMA	2019
Big data privacy risk: Connecting many large data sets	YING; GRANDISON	2017
Big Data Security and Privacy Concerns: A Review	KHANAN et al.	2019
Big data security and privacy in healthcare: A Review	ABOUELMEHDI et al.	2017
Big data security and privacy issues-A survey	JOSHI; KADHIWALA	2017
Big data security challenges and strategies	VENKATRAMAN	2019
Big Data, Big Concerns: Ethics in the Digital Age	JURKIEWICZ	2018
Breaching intellectual capital: critical reflections on Big Data security	LA TORRE; DUMAY; REA	2018
Data analytics in a privacy-concerned world	WIERINGA et al.	2021
Data entities e its privacy with big data techniques in e-health systems	ALI; SUDHAKAR; MANOJ	2019
Data Privacy in Retail	MARTIN et al.	2020
Data Security and Privacy: Concepts, Approaches, and Research Directions	BERTINO	2016

continua

Quadro 31 – Referências primárias. (*Continuação*)

Título original	**Autor(es)**	**Ano**
Dci3 model for privacy preserving in big data	HEMLATA; GULIA	2018
Development of national health data warehouse Bangladesh: Privacy issues and a practical solution	KHAN; HOQUE	2016
Enhancing the data privacy for public data lakes	CHEN; CHEN; HUANG	2018
Exploratory strategic roadmapping framework for big data privacy issues	DABAB; CRAVEN; BARHAM; GIBSON	2018
Familiarity with big data, privacy concerns, and self-disclosure accuracy in social networking websites: An APCO model	ALASHOOR; HAN; JOSEPH	2017
Hidden big data analytics issues in the healthcare industry	STRANG; SUN	2020
Living in a big data world: Predicting mobile commerce activity through privacy concerns	EASTIN et al.	2016
Necessity and countermeasures of big data security and privacy protection	JIAO	2021
PANDA: Partitioned Data Security on Outsourced Sensitive and Non-sensitive Data	MEHROTRA et al.	2020
Privacy and ethical issues of big data in the airline industry	CHANG; JI; ARAMI	2019
Privacy and security problems of national health data warehouse: A convenient solution for developing countries	KHAN; HOQUE	2016
Privacy concerns in integrating big data in “e-Oman”	SAXENA	2017
Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR	GRUSCHKA et al.	2019
Privacy issues in big data	PATEL et al.	2017
Privacy Issues in Big Data from Collection to Use	ALWABEL	2020
Privacy and Ethical Challenges in Big Data	GAMBS	2019
Privacy Issues Security Techniques in Big Data	TIWARI et al.	2019
Privacy Preserving Big Data Publishing	CANBAY; VURAL; SAGIROGLU	2019
Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges	MEHTA; RAO	2016
Privacy-aware data-intensive applications	GUERRIERO	2017
Research on differential privacy for medical health big data processing	HU et al.	2019
Research on privacy protection framework design and key technologies in large data environment	LIU	2019
Research on security and privacy of big data under cloud computing environment	MAOHONG; AIHUA; HUI	2018
Review of data extraction, segregation privacy with big data analytics in the online health care systems	PURANDHAR; SARAVANA	2019
Rise of Big Data - Issues and Challenges	ALABDULLAH; BELOFF; WHITE	2018
Security and Privacy Issues in Big Data: A Review	JADON; MISHRA	2019
Security and Privacy Solutions associated with NoSQL Data Stores	VONITSANOS et al.	2020
Technological, Organizational and Environmental Security and Privacy Issues of Big Data: A Literature Review	SALLEH; JANCZEWSKI	2016

continua

Quadro 31 – Referências primárias. (*Continuação*)

The use of de-identification methods for secure and privacy-enhancing big data analytics in cloud environments	BONDEL et al.	2020
Towards social network analytics for understanding and managing enterprise data lakes	FARRUGIA; CLAXTON; THOMPSON	2016
Triple DES: Privacy Preserving in Big Data Healthcare	RAMYA; VIJAYA	2020
Understanding privacy violations in big data systems	SHAMSI; KHOJAYE	2018
Where is the Risk? Analysis of Government Reported Patient Medical Data Breaches	SCHMEELK	2019

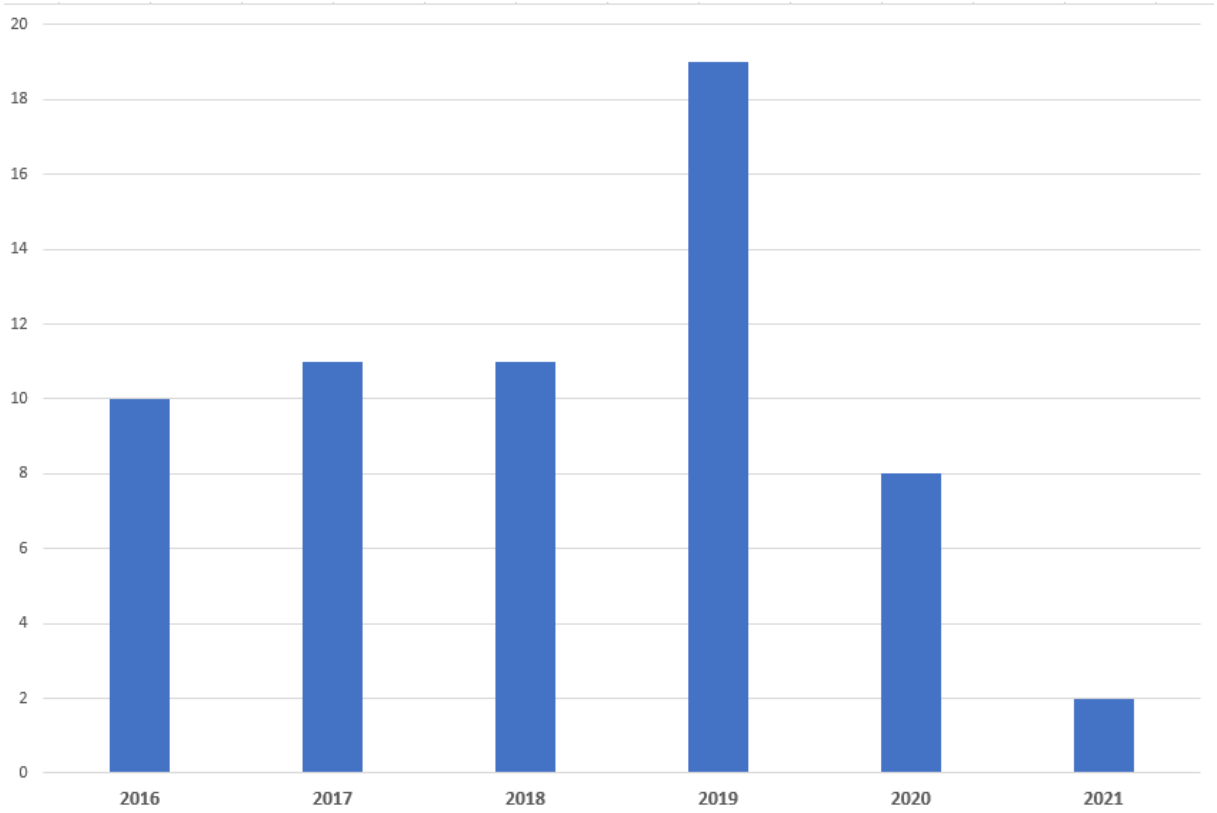
Fonte: Danilo Figueiredo de Oliveira, 2023

Apêndice C – Resultados complementares da RSL

Após a aplicação dos critérios de inclusão e exclusão, 61 pesquisas foram selecionadas para serem esmiuçadas na RSL.

A figura 10 apresenta o ano de publicação dos estudos selecionados. Nota-se que a maior parte das pesquisas selecionadas foram publicadas em 2019. Antes de 2019 é possível notar uma estabilidade da quantidade de publicações selecionadas. Porém, houve menos estudos selecionados publicados em 2020 em comparação aos anos anteriores. A última busca por publicações foi realizada em fevereiro de 2021, por consequência poucos artigos de 2021 foram selecionados.

Figura 10 – Número de pesquisa por ano de publicação

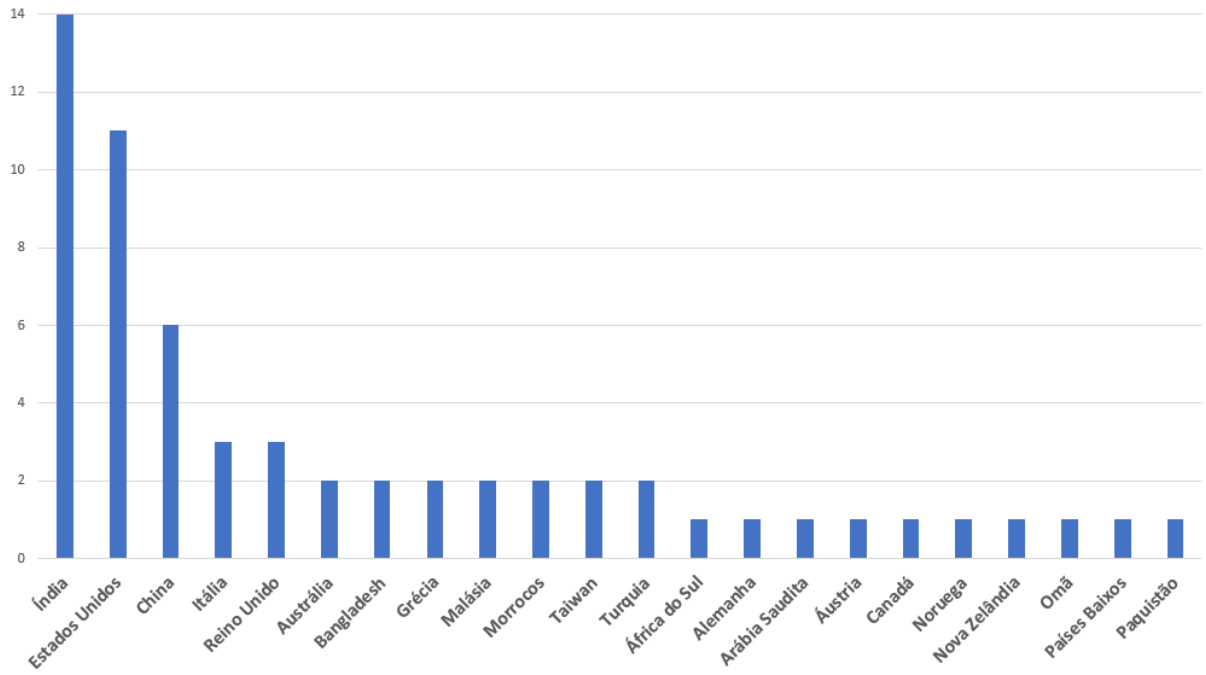


Fonte: Danilo Figueiredo de Oliveira, 2023

A figura 11 apresenta a quantidade de publicações por país. Em publicações com pesquisadores de mais de uma nacionalidade, foi considerado o país do primeiro autor. Observa-se que Índia, Estados Unidos e China são os países com mais publicações selecionadas. De fato, esses são alguns dos países que estão na vanguarda dos estudos em *big data*. É importante destacar a variedade de países, são 22 países representados

com pelo menos uma pesquisa, sem contar os demais autores de outros países também envolvidos em algumas dessas pesquisas.

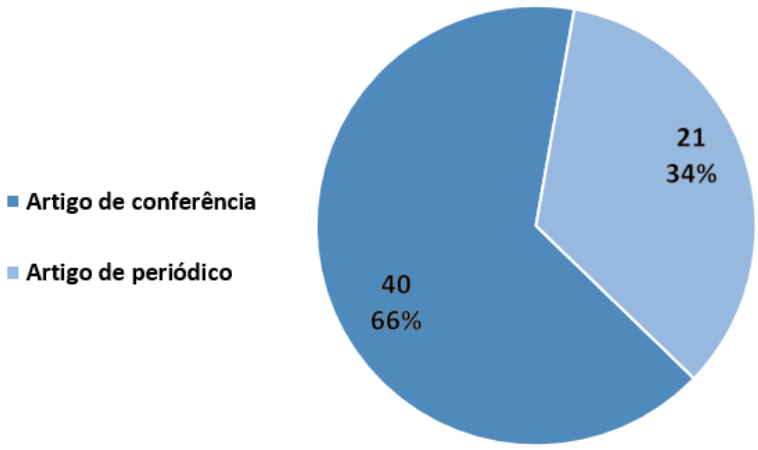
Figura 11 – Número de pesquisa por país



Fonte: Danilo Figueiredo de Oliveira, 2023

Percebe-se por meio da figura 12 que a maior parte dos artigos foram apresentados em conferências, sendo 40 no total. Os outros 21 documentos são artigos de periódicos.

Figura 12 – Número de pesquisa por tipo do documento



Fonte: Danilo Figueiredo de Oliveira, 2023

Apêndice D – Questionário

1. Por favor, forneça as informações de contato (somente o pesquisador responsável terá acesso a essas informações).

   - Nome:
   - E-mail:
   - Telefone (opcional):

2. Por favor, assinale com X a opção mais adequada a respeito da sua atuação profissional atual e anteriores.

   a) Em quais subáreas você tem familiaridade (pode selecionar mais de uma): [ ] Ciência de dados
   [ ] Engenharia de dados
   [ ] Análise de dados
   [ ] Governança de dados
   [ ] Outras áreas relacionadas a dados. Quais?

   b) Tempo de experiência em dados:
   [ ] $<$ 5 anos
   [ ] Entre 5 e 10 anos
   [ ] $>$ 10 anos

   Para as questões 3 e 4, por favor utilize a seguinte notação:
   0 - Não conheço 1 – Muito baixa 2 - Baixa 3 - Média 4 - Alta 5 – Muito alta

3. Na sua percepção, o quanto cada conjunto de causas contribui para ocasionar cada um dos problemas de privacidade abaixo? (figura 13)
4. Na sua percepção, o quanto cada conjunto de técnicas (soluções) contribui para mitigar os riscos de cada conjunto de causas? (figura 14)

Figura 13 – Matriz da questão 3 do questionário

		Problemas									
		P1 - Ameaça à vida ou à liberdade	P2 - Assédio moral ou discriminação	P3 - Constrangimento ou dano reputacional	P4 - Desvantagens em negociações	P5 - Fraudes e outros crimes contra as vítimas	P6 - Inviabilidade de manter-se anônimo	P7 - Re-identificação de dados anonimizados	P8 - Roubo ou acesso não autorizado a dados	P9 - Vigilância ilegal	P10 - Outros
Conjunto de causas	C1 - Ataques e vulnerabilidade de segurança	[	[	[	[	[	[	[	[	[	[
	C2 - Deficiência da gestão de BDA	-	-	-	-	-	-	-	-	-	-
	C3 - Desafios técnicos de BDA	-	-	-	-	-	-	-	-	-	-
	C4 - Empoderamento e comunicação com o usuário	-	-	-	-	-	-	-	-	-	-
	C5 - Gestão de acesso inadequada	-	-	-	-	-	-	-	-	-	-
	C6 - Problemas de gestão organizacional	-	-	-	-	-	-	-	-	-	-
	C7 - Revelação ou inferência de dados não autorizados	-	-	-	-	-	-	-	-	-	-

Fonte: Danilo Figueiredo de Oliveira, 2023

Figura 14 – Matriz da questão 4 do questionário

		Conjunto de técnicas (soluções)										
		S1 - Anonimização	S2 - De-identificação por ruído e perturbação	S3 - De-identificação por generalização	S4 - De-identificação por pseudoanonimização, supressão e mascaramento	S5 - Governança de dados	S6 - Políticas internas de proteção de privacidade	S7 - Controle pelo usuário de seus dados	S8 - Criptografia	S9 - Controle de acesso	S10 - Sanitização de dados	S11 - Outros
Conjunto de causas	C1 - Ataques e vulnerabilidade de segurança	-	-	-	-	-	-	-	-	-	-	-
	C2 - Deficiência da gestão de BDA	-	-	-	-	-	-	-	-	-	-	-
	C3 - Desafios técnicos de BDA	-	-	-	-	-	-	-	-	-	-	-
	C4 - Empoderamento e comunicação com o usuário	-	-	-	-	-	-	-	-	-	-	-
	C5 - Gestão de acesso inadequada	-	-	-	-	-	-	-	-	-	-	-
	C6 - Problemas de gestão organizacional	-	-	-	-	-	-	-	-	-	-	-
	C7 - Revelação ou inferência de dados não autorizados	-	-	-	-	-	-	-	-	-	-	-

Fonte: Danilo Figueiredo de Oliveira, 2023