ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 12, Issue, 04, pp. 55527-55532, April, 2022
https://doi.org/10.37118/ijdr.24305.04.2022

RESEARCH ARTICLE OPEN ACCESS

PROBLEMAS, CAUSAS E SOLUÇÕES DE PRIVACIDADE DE DADOS EM SISTEMAS DE BIG DATA ANALYTICS: UMA REVISÃO SISTEMÁTICA DA LITERATURA

***Danilo Figueiredo de Oliveira and Edmir Parada Vasques Prado**

Área de Sistemas de Informação, Escola de Artes, Ciências e Humanidades (USP), Brasil

ARTICLE INFO

Article History:
Received 20th January, 2022
Received in revised form
09th February, 2022
Accepted 26th March, 2022
Published online 30th April, 2022

Key Words:
Privacidade de dados. Problemas de privacidade. Privacidade em big data analytics. Big data analytics.

****Corresponding author:***
Danilo Figueiredo de Oliveira.

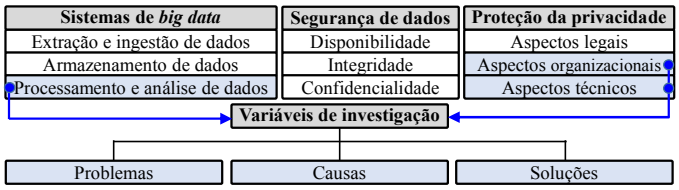
ABSTRACT

A Privacidade é um direito humano fundamental e tem sido um tema cada vez mais importante à medida que aumenta o uso de produtos e serviços digitais. A quantidade de dados gerados a cada momento é grande e exige tecnologias específicas para serem processados e analisados, o que foi denominado como sistemas de *big data analytics* (SIBDA). Como consequência, diversos problemas de privacidade de dados têm ocorrido. Nesse contexto, definiu-se que o objetivo deste estudo é identificar e descrever os problemas de privacidade de dados em SIBDA, bem como as suas causas e soluções, com base na literatura sobre o tema. Esta pesquisa é fundamentada em uma revisão bibliográfica e apresenta característica qualitativa predominante. Foram identificados oito problemas, sete causas e quatro categorias de soluções.

Citation: ***Danilo Figueiredo de Oliveira and Edmir Parada Vasques Prado.*** "Problemas, causas e soluções de privacidade de dados em sistemas de big data analytics: uma revisão sistemática da literatura*", International Journal of Development Research*, 12, (04), 55527-55532.

INTRODUCTION

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). Por essa razão, Wu *et al*. (2014) e Kitchin (2014) alegam que a humanidade vive uma era de transformações tecnológicas rápidas e constantes na análise e tratamento de dados e informações. A consequência desse avanço tecnológico pode ser positiva ou negativa. Porém, isso depende mais de como se aplica a tecnologia do que a sua mera existência. 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 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 e propriedade dos dados, ideias, processos, hardware, código, entre outros. Ocasionalmente, os interesses das organizações podem conflitar com os interesses de privacidade dos usuários, e a partir desse conflito surgem dilemas éticos. Por outro lado, a classificação de uso correto ou ético dos dados é difusa. Porém, pressupõe-se que o uso de dados pessoais com desvio de validade 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). No estudo de Stahl e Wright (2018) sobre ética em sistemas de informação (SI), o tema proteção e privacidade de dados foi o mais proeminente. Além disso, Singh *et al*. (2018) conclui que não há pesquisas suficientes na literatura sobre problemas de confidencialidade e privacidade de dados em Sistemas de Informação de *Big Data Analytics* (SIBDA). Além disso, há a dificuldade de se obter privacidade de dados em SIBDA (YING; GRANDISON, 2017), e muitas áreas do conhecimento carecem de estratégias adequadas (WANG, 2018). Por isso, Joshi e Kadhiwala (2017) concluíram que são necessárias mais análises e pesquisas sobre o gerenciamento da privacidade de dados. Diante deste contexto, o estudo dos problemas de privacidade de dados se mostra relevante. Esse problema se intensifica quando são considerado dois aspectos: privacidade de dados em SIBDA, pois as arquiteturas desses sistemas são bastante variadas e eles são usados para a tomada de decisão nas organizações (SHAYTURA *et al*., 2016); e a realidade brasileira, que possui baixa competitividade digital comparada a países desenvolvidos (IMD WORLD DIGITAL, 2020). Essa realidade é corroborada por Abouelmehdi *et al*. (2017), que alega que as preocupações com privacidade e segurança de dados representam o maior risco em relação a SIBDA. Este estudo tem como objetivo identificar e descrever os problemas de privacidade de dados, bem como as suas causas e soluções, com base na literatura sobre o tema. A Figura 1 detalha o contexto desta pesquisa.



Fonte: elaborada pelo autor

Figura 1. Contexto da pesquisa

O contexto dessa análise refere-se a SIBDA e não inclui trabalhos que discutam privacidade de dados no contexto de internet das coisas ou *blockchain*, pois estes tópicos têm desafios específicos que diferem de outras aplicações de BDA. Da mesma forma, esta pesquisa só considera problemas de segurança de dados que tenham impacto direto na privacidade de dados. Os aspectos políticos, econômicos, sociais, ambientais e legais foram considerados somente quando houve relação direta com a tecnologia.

REFERENCIAL TEÓRICO

Nesta seção são apresentados conceitos fundamentais utilizados neste estudo. Esses conceitos foram agrupados em três tópicos. O primeiro trata da privacidade de dados e o segundo trata da relação entre privacidades de dados e os SI. O terceiro e último tópico trata do ambiente tecnológico de SIBDA.

Privacidade de Dados: o conceito de privacidade tem definições muito amplas e difusas na literatura (STUTZMAN; HARTZOG, 2012). Ou seja, sua definição não é clara. Além disso, segundo Barker et al., (2009), presume-se com frequência que privacidade é um conceito globalmente uniforme, porém nem sempre isso é verdadeiro. Os conceitos que formam a ideia de privacidade são: direito de ser deixado em paz, sigilo, controle sobre as próprias informações pessoais e intimidade (SOLOVE, 2002).Porém, Hartzog (2018) reconhece que há discordância na definição de privacidade e define este conceito na área de SI como sendo o controle do usuário sobre as configurações dos sistemas, por ser amplamente adotada por acadêmicos, executivos, legisladores, reguladores e juízes. As questões relacionadas à privacidade de dados possuem aspectos legais e regulatórios, que evidenciam sua importância, além de sua associação com SI e segurança da informação nas organizações.

Privacidade de Dados e Sistemas de Informação:Schaub, Konings e Weber (2015) classificaram como onipresentes os SI que são usados em diversas situações do cotidiano dos cidadãos. Esses autores alegam que isso gera inúmeras implicações de privacidade, pois esses sistemas podem reunir e trocar informações extremamente abrangentes dos usuários com pessoas ou empresas em qualquer lugar do mundo. Como consequência, 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 (AHMADIAN*et al*., 2018). A Lei Geral de Proteção de Dados explicita que os dados coletados de terceiros devem ser utilizados apenas para o propósito definido entre as partes (BRASIL, 2018). No entanto, Constantiou e Kallinikos (2015) e Muller *et al*. (2016) alegam que organizações públicas e privadas têm coletado dados para seus sistemas de *big data* sem um proposito pré-estabelecido, esperando que no futuro possam extrair valor desses dados, o que pode ser caracterizado como uma violação de privacidade. Além disso, violações de privacidade podem acontecer mesmo quando a organização está em conformidade com regulações e boas práticas de gestão (WALL; LOWRY; BARLOW, 2016), pois há fatores importantes na proteção de privacidade que vão além do devido cumprimento de regras, leis e boas práticas. Segundo Cavoukian (2012), o projeto de um SI precisa sofrer influência de questões relacionadas à privacidade. A partir dessa ideia surgiu o conceito de *privacy by design*, que é uma abordagem de engenharia de software no qual se exige que a privacidade seja levada em consideração ao longo de todo o processo de engenharia. Outro aspecto importante dos projetos de SI é a segurança da informação. Ela é dividida em três pilares (CHEN; ZHAO, 2012): confidencialidade, quando o acesso às informações estão restritas apenas a quem é necessário; integridade que se refere a ter as informações incorruptíveis; e disponibilidade, ou seja, estar disponível a quem deve acesse-la.

Big Data Analytics: Analytics 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 para Gandomi e Haider (2015) o big data por si só tem pouca ou nenhuma utilidade. Isso é, seu potencial só é aproveitado se for utilizado na tomada de decisão.No entanto, não existe consenso na literatura sobre a definição de big data. Uma definição comum e bastante aceita de big data são os 3 Vs, citado por Chen e Storey (2018): volume, variedade e velocidade. Ou seja, como explicado por McAfee e Brynjolfsson (2012), big data tem como contexto um grande volume de dados, vindos de diversas fontes diferentes e em intervalos de tempo extremamente curtos. *Analytics*, por sua vez, frequentemente está associada a *insights*. Cooper (2012) 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 Jagadish *et al*. (2014) entende BDA como um subprocesso da extração de *insights* do *big data*. Segundo Ranjan e Foropon (2021), apesar do número crescente de organizações que lançam iniciativas de BDA, elas têm limitações ao tentar converter o potencial dessas iniciativas em valor para os negócios. Estes autores concluíram que organizações de diversos tamanhos, estruturas e setores têm grandes dificuldades com BDA e destacam que essa realidade é um desafio intrínseco a esse contexto. A análise do ambiente de BDA não se restringe à quantidade de dados e seus processos de extração de *insights*. Fan, Han e Liu (2014) alegam que parte das causas ou das soluções dos problemas de privacidade de dados podem estar atrelados ao desenho da arquitetura ou às tecnologias adotadas. Isto porque, segundo esses autores, as características da *big data* geram desafios complexos, tais como alto custo computacional, instabilidade algorítmica e dificuldade na agregação de dados de fontes múltiplas que usam tecnologias distintas entre si. Por esses motivos, o desenho arquitetural é uma atividade que exige técnica, principalmente à medida que a velocidade e o volume dos dados aumentam e diferentes arquiteturas podem facilitar ou dificultar a proteção de privacidade (JAGADISH *et al*., 2014).

MÉTODO DE PESQUISA

Esta pesquisa é fundamentada em uma revisão bibliográfica e apresenta característica qualitativa predominante. Além disso, trata-se de um estudo bibliográfico e exploratório. Segundo Creswell e Creswell (2021), a pesquisa exploratória visa oferecer informações sobre o objeto de estudo e orientar a formulação de proposições para futuras pesquisas. Dessa forma, esta pesquisa explora as bases de dados científicas para identificar os problemas, causas e soluções de privacidade de dados em SIBDA, e que possam servir de base para futuras pesquisas descritivas e explicativas, que permitam aprimorar os SI, adequando-os às necessidades de privacidade dos cidadãos e às normas legais. A seguir são apresentadas as fases da pesquisa, bem como os procedimentos metodológicos para coleta e tratamento dos dados.

Fases da Pesquisa: esta pesquisa foi desenvolvida em quatro fases como ilustra a figura 2. A primeira fase constituiu o levantamento na literatura dos conceitos fundamentais da pesquisa. Esse levantamento foi realizado na literatura da área de SI. Na segunda fase foi realizada uma revisão sistemática da literatura (RSL) nas bases de dados científicas. Em seguida, na terceira fase, foram analisados os artigos selecionados pela RSL, usando a técnica de análise de conteúdo semântica (BARDIN, 2011), e foram identificados os problemas, causas e soluções de privacidade de dados em SIBDA. Por último, na quarta fase, foram feitas proposições para futuras pesquisas.

Procedimentos de Coleta e Análise dos Dados: os dados foram coletados das bases científicas no primeiro semestre de 2021.

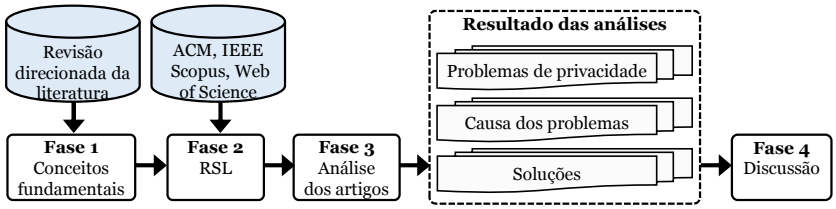
Inicialmente foram feitas análises com técnicas qualitativas e em seguida com técnicas quantitativas. A seguir estão apresentados os procedimentos para coleta de dados por meio da RSL, e em seguida são apresentadas as técnicas qualitativa e quantitativa de tratamento dos dados.

- Aplicação da Revisão Sistemática da Literatura: segundo Kitchenham *et al*. (2009), a RSL possui três etapas principais: (1) planejamento, na qual são identificadas as questões de pesquisa e elaborado o protocolo de revisão; (2) condução, na qual são selecionados os estudos seguindo o protocolo especificado; e (3) relatório, na qual são sumarizados os dados e analisados os resultados.
- Planejamento da Revisão Sistemática da Literatura. O protocolo da RSL iniciou com a seleção das bases de dados científicas. As bases utilizadas foram: ACM Digital Library (https://dl.acm.org), IEEXplore (http://ieeexplore.ieee.org), Scopus (http://www.scopus.com) e *Web of Science* (https://apps.webofknowledge.com). A RSL tem como objetivo identificar os problemas, causas e soluções de privacidade de dados em SIBDA. Assim, foram definidas as seguintes questões para pesquisa nas bases de dados: (1) Quais são os problemas encontrados no tratamento da privacidade de dados em SIBDA? (2) Quais são as causas identificadas ou sugeridas pelos problemas no devido tratamento da privacidade de dados em SIBDA? (3) Quais são as soluções identificadas para resolver ou mitigar as causas dos problemas de privacidade de dados em SIBDA?

objetivo de estudar causas ou problemas de privacidade de dados. A estratégia de extração e síntese de dados foi baseada na abordagem sugerida por Keshav (2007), ou seja, a leitura passou por três etapas: na primeira etapa foi lido o título, palavras-chaves e resumo, em seguida foi lida a introdução, os títulos das seções e subseções e a conclusão; na segunda etapa foi feita a leitura de diagramas, ilustrações e eventualmente quadros e tabelas; e na terceira etapa foi feita a leitura cuidadosa de todo o artigo. Os dados extraídos do texto foram registrados em planilhas e em documentos de texto.

Condução da Revisão Bibliográfica. Os procedimentos para seleção dos artigos da RSL ocorreram em maio de 2021, e foram encontrados 745 resultados aplicando-se os termos de busca e os critérios de inclusão. Os critérios de exclusão foram aplicados por etapas, restando 61 artigos selecionados. Os resultados estão sumarizados na Tabela 1.

Tratamento dos Dados: Foram utilizadas técnicas qualitativa e quantitativa para análise e tratamento dos dados. As causas e problemas de privacidade de dados em SIBDA, coletados na RSL, foram tratados por meio de análise de conteúdo (BARDIN, 2011). Foi utilizada especificamente a análise semântica, que consiste em analisar o sentido das palavras encontradas nos textos, para sintetizar problemas, causas e soluções encontradas. Em seguida foram utilizadas técnicas estatísticas descritivas para comparar frequência de causas e problemas.



Fonte: elaborada pelo autor

Figura 2. Fases da pesquisa

Tabela 1. Artigos selecionados na revisão sistemática da literatura

Etapa	Eliminados							Restante
	2016	2017	2018	2019	2020	2021	Total	
CE1	1	22	17	22	12	2	76	669
CE2	4	8	6	3	13	8	42	627
CE3	15	18	8	6	17	11	75	552
CE4	12	16	18	21	17	5	89	463
CE5	6	15	18	18	14	9	80	383
CE6	3	7	5	4	4	2	25	358
CE7	32	81	80	63	36	5	297	61

Fonte: dados da pesquisa

Para uma pesquisa ser selecionada pela RSL, foi obrigatório atender a todos os nove critérios de inclusão (CI): 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 ou conferência; CI6, a publicação deve estar em inglês (padrão de língua científica); CI7, a publicação deve estar no estágio final de publicação; CI8, a área de estudo da publicação deve incluir SI; e CI9, ter sido publicado a partir de 2016, para garantir pesquisas recentes. Se uma pesquisa atender a qualquer critério de exclusão (CE), ela é desconsiderada. Foram definidos sete CE: CE1, documento repetido; CE2, não permitir acesso e não ser encontrado em outras fontes; CE3, não abordar problemas ou causas de problemas de privacidade de dados; CE4, abordar o tema privacidade, mas não ter como objeto de estudo a privacidade em SIBDA; CE5, 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; e CE7, abordar privacidade em SIBDA, mas não ter o

APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Nesta seção é apresentado o resultado da RSL. Primeiramente foram descritos os problemas de privacidadee sua frequência de citação na literatura. Em seguida, se procedeu de forma semelhante para as causas desses problemas, e por último foram descritas as soluções, agrupadas em quatro categorias.

Problemas de privacidade de dados em SIBDA: usando a análise de conteúdo semântica (BARDIN, 2011), foi possível realizar uma síntese dos problemas de privacidade em SIBDA com base nas classes que emergiram dessa análise. Neste trabalho, problema é entendido como uma situação considerada indesejável ou prejudicial aos indivíduos que tenha sido causada por violação de privacidade em SIBDA. A síntese dos problemas está apresentada na Tabela 2.

P1 -Vida e liberdade. Este problema refere-se às ameaças à vida ou à liberdade que indivíduos podem sofrer devido a incidentes de privacidade de dados. Essas ameaças podem vir, inclusive, do

próprio governo, em casos de países totalitários ou com democracia frágil.

Tabela 2. Número de citações dos problemas na literatura

Ano	Problemas							
	P1 5	P2 5	P4 4	P6 4	P7 4	P5 2	P3 2	P8 2
2019	●		●				●	
2018	●			●				●
2018		●		●				●
2017	●	●				●		
2021				●	●			
2021		●	●					
2020		●					●	
2019				●		●		
2018			●		●			
2018	●	●						
2017	●				●			
2017			●		●			

Fonte: dados da pesquisa

P2 **-Discriminação.** Assédio moral ou discriminação contra indivíduos é outro problema que pode ser desencadeado por incidentes de privacidade de dados. É possível que um indivíduo sofra violência psicológica, tenha tratamento de segregação, seja tratado de maneira diferente e parcial, por motivos de diferenças sexuais, raciais, religiosa por conta de vazamento de SPI (*Sensitive Personal Information*).

P3 **-Negociações.** Desvantagens em negociações podem ocorrer em casos de incidentes de privacidade de dados, por exemplo, como na compra ou venda de ativos, ou negociação salarial, entre outros.

P4 **-Fraudes.** Fraudes e outros crimes contra as vítimas podem ser facilitadas ou totalmente viabilizadas por incidentes de privacidade de dados. Indivíduos podem ser enganados ou levados ao erro em benefício de outro indivíduo ou organizações terceiras.

P5 **-Anonimidade.** Um indivíduo pode perder a viabilidade de manter-se anônimo por conta de incidentes de privacidade de dados, ou seja, é possível que indivíduos percam totalmente a opção de manter sua anonimidade em um determinado momento ou situação.

P6 **-Re-identificação.** A re-identificação de dados anonimizados pode ocorrer de diversas maneiras, mas algumas das principais formas envolve o cruzamento de diferentes conjuntos de dados para enriquecimento. Incidentes de privacidade de dados pode facilitar a re-identificação de dados anonimizados pela disponibilização de novas informações sigilosas de indivíduos.

P7 **-Acesso.** Roubo ou acesso não autorizado a dados é um dos problemas de privacidade de dados identificados. Neste caso, trata-se deincidentes relacionados ao roubo de dados ou a acessos indevidos por qualquer motivo.

P8 **-Vigilância ilegal.** Indivíduos podem sofrer de vigilância ilegal por outros indivíduos ou organizações, como quadrilhas, empresas e governos, em decorrência de incidentes de privacidade de dados.

Causa dos problemas de privacidade de dados em SIBDA: neste trabalho, causa é aquilo que é origem de um problema de privacidade de dados, desde que esteja no contexto de SIBDA. As diversas causas encontradas na literatura foram categorizadas em conjuntos de acordo com a análise de conteúdo semântica, conforme tabela 3 e estão descritas a seguir.

C1 **- Ataques e vulnerabilidade de segurança.**Representam causas associadas à vulnerabilidade e falta de segurança de organizações e áreas que tratam dados.Nesta categoria, incluem-se fatores como ataques externos por hackers ou malwares e chaves de criptografia fracas, além de outros fatores que estimulam ou facilitamataques como alto valor de mercado de dados médicos, baixa preocupação com segurança em setores específicos, e uso de ferramentas de terceiros.

C2 **- Deficiência da gestão de BDA.** Causas associadas à gestão ineficiente de dados e não-conformidade de boas práticas e regulamentos nas organizações e áreas que tratam dados.Isso inclui falta de propósito de uso dos dados, falta de transparência do uso dos dados, mudança no propósito do uso dos dados, e retenção indevida de dados.

Tabela 3. Número de citações das causas na literatura

Ano	Causas						
	C7 17	C5 14	C3 13	C2 9	C4 9	C6 9	C1 5
2019	●	●	●		●		●
2019		●	●	●		●	●
2019	●	●			●	●	
2019	●	●	●				●
2018	●	●	●				●
2018		●	●	●		●	
2021		●	●			●	
2021	●	●				●	
2019	●		●	●			
2019		●		●		●	
2019	●				●	●	
2017		●	●	●			
2016		●	●			●	
2016		●				●	●
2016	●			●	●		
2016			●	●	●		
2019	●		●				
2019	●			●			
2019		●	●				
2019	●				●		
2018	●				●		
2018	●				●		
2017	●	●					
2017	●				●		
2017	●			●			
2017	●		●				

Fonte: dados da pesquisa

C3 **- Desafios técnicos de BDA.** Causas associadas às dificuldades técnicas de proteção de privacidade, principalmente devido ao volume, velocidade, e variedade dos dados. Tratar muitos dados vindos de diversas fontes diferentes é um grande desafio técnico relacionado ao BDA. Além disso, anonimizar os dados oriundos de SIBDA é altamente complexo.Outros desafios técnicos são uso de chave natural PII (*Personally Identifiable Information*) como chave de negócio, falta de mensuração de privacidade, metadados contendo PII, ou mesmo publicação indevida ou inadequada de dados.

C4 **- Empoderamento e comunicação com o usuário.** Causas associadas à inaptidão de usuários do sistema, como desinformação, desconhecimento ou despreocupação dos usuários em relação à privacidade, o que pode levar ao controle inadequadodos seus dados. Além disso, falta de controle do usuário nos sistemas oufalta de consentimento de uso de dados pelo usuário também fazem parte desta categoria.

C5 **- Gestão de acesso inadequada.** Causas associadas à má gestão de acesso a dados, como acesso excessivamente granular, ilegal, não autorizado, ou excessivo. Além de acesso inadequado por terceiros, ou mesmo total falta de controle de acesso sobre os dados da organização.

C6 **- Problemas de gestão organizacional.**Esta categoria inclui responsabilização inadequada ou falta de regulamentos internos, comportamento malicioso de funcionários ou terceiros, cultura fraca em privacidade, falta de apoio da alta gestão, ou falta de capacitação técnica das equipes.

C7 **- Revelação ou inferência de dados não autorizados.** Causas associadas à re-identificação de dados supostamente anonimados. Isso inclui inferência, a partir dos dados originais, com ou sem cruzamento com outras bases, de dados mais

granulares, dados PIIs, dados novos, ou re-identificaçãoindevida de dados.

Soluções para privacidade de dados em SIBDA: nesta pesquisa as soluções compreendem as principais ações e práticas que podem ser adotadas para evitar, minimizar ou resolver os problemas identificados na RSL. Ying e Grandison (2017) destacam que soluções definitivas de privacidade de dados pode ser extremamente difíceis. Além disso, diferentes áreas podem ter diferentes problemas, causas e soluções: no setor financeiro, Singh et al. (2018) citam a autenticação e a autorização como dois dos primeiros desafios que exigem atenção imediata das organizações; e Ghani, Hamid e Udzir (2016) citam que na área de saúde a coleta de dados e um ponto crítico. As soluções identificadas na literatura foram agrupadas em quatrocategorias. As primeiras três se relacionam mais a questões técnicas e a última mais a questões de gestão e governança. Cabe destacar que, segundo Bertino (2016), as soluções da categoria de criptografia são amplamente investigadas pela literatura.

S1 - Anonimização. Este grupo é composto por métodos e técnicas que visam garantir matematicamente que nenhum indivíduo pode ser reconhecido, geralmente chega-se a esse estado ao aplicar técnicas de de-identificação, principalmente ruído. É irreversível, não pode ser desfeita. A anonimização visaproteger o cidadãopelo desfazimento de qualquer tipo de vínculo capaz de associar um dado ao seu respectivo titular. Singh *et al*. (2018)sugere técnicas de anonimização para que dados de cidadãos não sejam divulgados, recuperados facilmente, ou vazados por*hackers*.K-anonimato é 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, por sua vez, 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.

S2 - De-identificação. Os métodos de de-identificação são abordagens que dificultam a restauração do vínculo entre um indivíduo e seus dados, removendo ou transformando pontos de dados específicos (KUSHIDA et al., 2012). Ela pode ser classificada em três tipos: por ruído e perturbação; por generalização; e por supressão ou mascaramento de dados. Alguns métodos dessa categoria são citados por Shozi e Mtsweni (2017): (1) supressão, na qual identificadores diretos são removidos por completo dos conjuntos de dados; (2) generalização, tal como remover uma data de nascimento e substituir apenas pelo ano; (3) agregação, usada para aumentar a cardinalidade de conjuntos. (4) troca de dados, tal como substituir informações de idade e etnia para proteger registros sensíveis e de risco; (5) ruído aleatório, que consiste em adicionar ruído aos dados para reduzir as possibilidades de re-identificação, geralmente aplicado a dados numéricos; e (6) dados sintéticos, que consiste na substituição dos valores originais por valores simulados de distribuições de probabilidade.

S3 - Criptografia. A criptografia é conversão de dados de um formato legível em um formato codificado com o objetivo fundamental de permitir que as entidades se comuniquem através de um canal inseguro de tal forma que um oponente não possa entender o que está sendo comunicado(STINSON; PATERSON, 2018). Entre as técnicas citadas destacam-se criptografia funcional, criptografia por chave simétrica (*Advanced Encryption Standard*), criptografia homomórfica, calculo verificável, e computação multipartidária segura.

S4 - 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). Os métodos e técnicas desta categoria podem ser classificados em quatro tipos:

CONCLUSÃO

O objetivo deste estudo foi objetivo identificar e descrever os problemas de privacidade de dados em SIBDA, bem como as suas causas, com base na literatura sobre o tema. A pesquisa tem limitações das quais podem ser destacadas: (1) os problemas e causas foram classificados em função do número de citações na literatura, o que não representa o grau de importância tanto na academia como na indústria de SI; e (2) a análise e síntese dos problemas e causas, realizada por meio de uma análise de conteúdo semântica apresenta subjetividade.

As conclusões podem ser tecidas a partir dos três tópicos analisados:

- **Problemas**. Foram identificados nove problemas de privacidade de dados, dos quais dois se destacaram: ameaça à vida e à liberdade (P1) e discriminação (P2). Pode-se inferir que o problema P1 tem destaque na literatura em função de governos de países totalitários ou com democracia frágil. No caso do problema P2, é notório a quantidade de casos de discriminação desencadeado por incidentes de privacidade de dados, que independem do regime político ou condição economia de uma nação.
- **Causas**. Foram identificadas sete causas de problemas de privacidade de dados, dos quais três se destacaram: desafios técnicos (C3), gestão de acesso (C5) e re-identificação (C7). Nota-se que há causas técnicas e gerenciais. Foram destacados não só os desafios técnicos, no contexto de SIBDA, devido ao volume, velocidade, e variedade dos dados, mas também causas relacionadas à má gestão de acesso de dados. Este último refere-se à acesso ilegal, não autorizado, ou excessivo, além de acesso inadequado por terceiros. Inclui-se nas causas a re-identificação de dados anonimados. Isso sugere que a solução de problemas de privacidade vai além de questões técnicas e incluem ações gerenciais. Vai ao encontro dessa realidade a crescente importância que as organizações têm atribuído à governança de dados.
- **Soluções**. Foram identificados quatro grandes grupos de soluções. A anonimização foi diferenciada da de-identificação sendo que o objetivo do primeiro é garantir matematicamente a proteção da privacidade, inclusive aplicando técnicas de de-identificação no processo, e o segundo é composto por técnicas e métodosde transformação do dado a fim de impedir a associação direta de um registro a uma pessoa específica. O grupo de de-identificação é composto por dezenas de técnicas e métodos possíveis, como adição de ruído, pseudoanonimização, supressão, mascaramento e generalização. Frequentemente a de-identificação é reversível. Outro grande grupo inclui técnicas e métodos relacionados à governança de dados, incluindo políticas, cultura organizacional, controle de acessos e sanitização. A devida aplicação dessas técnicas e métodos envolvem pouca complexidade técnica, mas exigem grande competência organizacional. Por fim, foram classificadas em um grupo as técnicas e métodos de criptografia que se diferencia de anonimização e de-identificação por se tratar de práticasaplicadas na comunicação dos dados. Esta pesquisa contribui para a prática gerencial na medida que identifica e classifica as causas e os problemas de privacidade de dados em SIBDA. A partir da literatura foi possível observar que privacidade de dados em SIBDA é um desafio, seja pela complexidade técnica, ou pelas dificuldades organizacionais. Além disso, as causas dos problemas de privacidade são diversos. Deste modo, a superação das dificuldades organizacionais depende também do patrocínio da alta gestão da organização, e a superação dos desafios técnicos passa pela capacitação da equipe de SI, gestão adequada de acesso, boa comunicação com o usuário, e boa gestão de dados. Esta pesquisa faz parte de um projeto maior que visa identificar

soluções para essas causas e problemas. Isto porque, apesar desta pesquisa ter identificado as causas e problemas de privacidades de dados em SIBDA, as técnicas e ferramentas para mitigar esses problemas carecem de pesquisas.

REFERÊNCIAS

Ahmadian, A. S.; Struber, D.; Riediger, V.; Jurjens, J. Supporting privacy impact assessment by model-based privacy analysis. Proceedings of the ACM Symposium on Applied Computing, p. 1467-1474, 2018.

Bardin, L.Analise de conteúdo. 2011.

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.

Brasil. Lei Geral de Proteção de Dados Pessoais. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

Cavoukian, A. Privacy by design [leading edge]. IEEE Technology and Society Magazine, IEEE, v. 31, n. 4, p. 18{19, 2012. ISSN 02780097.

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.

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.

Conger, S.; Loch, K. D.; Helft, B. L. Ethics and information technology use: a factor analysis of attitudes tocomputer use. Information Systems Journal, v. 5, n. 3, p. 161-183, 1995. ISSN 13652575.

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.

Cooper, A. What is “Analytics"? Definition and Essential Characteristics. CETIS Analytics Series, v. 1, n. 5, p. 1-10, 2012. ISSN 2051-9214.

Creswell, J. W.; Creswell, J. D.. Projeto de pesquisa: Métodos qualitativo, quantitativo e misto. Penso, 2021.

Dama International. Data Management Body of Knowledge (DMBOK). 2a edição. ed. Basking Ridge: Technics Publications, 2017. ISBN 9781634622349.

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=EN.

Fan, J.; Han, F.; Liu, H. Challenges of Big Data analysis. National Science Review, v. 1, n. 2, p. 293{314, 2014. ISSN 2053714X.

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.

Google. Google Translator. 2020. Disponível em: http://translate.google.com/.

Google. Google Trends. 2020. Disponível em: https://trends.google.com/trends/explore?date=today5-y&q=Big Data,DataAnalytics.

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.

Hartzog, W. The Case Against Idealising Control. European Data Protection LawReview, v. 4, n. 4, p. 423-432, 2018. ISSN 23642831.

IMD World Digital. IMD World Digital Competitiveness Ranking 2020. IMD World CompetitivenessCenter, p. 180, 2020. Disponível em: https://www.imd.org/globalassets/wcc/docs/release-2017/world_digital_competitiveness_yearbook_2017.pdf.

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.

Keshav, S. How to read a paper. ACM SIGCOMM Computer Communication Review, v. 37, n. 3, p. 83-84, 2007. ISSN 0146-4833.

Kitchenham, B;Brereton, O. P.; Budgen, D.; Turner, M.; BAILEY, J.; Linkman, S.Systematic literature reviews in software engineering – A systematic literature review. Information and Software Technology, 51(1), 7–15, 2009.

Kitchin, R. Big Data, new epistemologies and paradigm shifts. Big Data and Society, SAGE Publications Ltd, v. 1, n. 1, 2014. ISSN 20539517.

Kushida, C. A.; nichols, D. A.;Jadrnicek, R.; Miller, R.; Walsh, J. K.; Grifin, K. Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. Medical Care, v. 50, p. 82-101, 2012.

Mcafee, A.; Brynjolfsson, E. Spotlight on Big Data Big Data: The Management Revolution, 2012. Harvard Business Review, n. October, p. 1-9, 2012.

Muller, 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.

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.

Oxford University. Oxford English Dictionary. 2020. Disponível em: https://en.oxforddictionaries.com/definition/analytics.

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.

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, 2015. ISSN 1536-1268.

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.

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.

Solove, D. J. Conceptualizing privacy. California Law Review, v. 90, n. 4, p. 1087-1155, 2002. ISSN 00081221.

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.

Stinson, D. R.; Paterson, M. B.Cryptography: Theory and Practice. 2018. Disponívelem: https://www.taylorfrancis.com/books/9781315282480.

Stutzman, F.; Hartzog, W. Boundary regulation in social media. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, p. 769-778, 2012.

Wall, J. D.; Lowry, P. B.; Barlow, J. B. Organizational violations of externally governed privacy and 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.

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.

****