**UNIVERSIDADE FEDERAL DE MINAS GERAIS**
**ESCOLA DE CIÊNCIA DA INFORMAÇÃO**
**PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO**

Brenner Lopes

# PENTÁGONO DA PRIVACIDADE NO *BIG DATA PRIVACY ANALYTICS:* Proposta de modelo multifacetado de garantia da privacidade e do valor analítico

Belo Horizonte
2024

Brenner Lopes

# PENTÁGONO DA PRIVACIDADE NO *BIG DATA PRIVACY ANALYTICS*: Proposta de modelo multifacetado de garantia da privacidade e do valor analítico

Tese apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Doutor, área de concentração Ciência da Informação.

Linha de Pesquisa: Gestão e Tecnologia da Informação e Comunicação

Orientador: Prof. Dr. Ricardo Rodrigues Barbosa

Belo Horizonte

2024

L864p Lopes, Brenner.
Pentágono da privacidade no big data privacy analytics [recurso eletrônico] : proposta de modelo multifacetado de garantia da privacidade e do valor analítico / Brenner Lopes . - 2024.
1 recurso eletrônico (156 f. : il., color.) : pdf.

Orientador: Ricardo Rodrigues Barbosa.

Tese (doutorado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 137-149.
Apêndice: f. 150-156.
Exigência do sistema: Adobe Acrobat Reader.

1. Ciência da informação – Teses. 2. Big data – Teses. 3. Direito à privacidade – Teses. 4. Proteção de dados – Teses. 5. Lei Geral de Proteção de Dados Pessoais (LGPD) – Teses. I. Barbosa, Ricardo Rodrigues. II. Universidade Federal de Minas Gerais. Escola de Ciência da Informação. III. Título.

CDU: 004.62:343.45

Ficha catalográfica: Elaine Diamantino Oliveira - CRB: 6/2742

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

# FOLHA DE APROVAÇÃO

**PENTÁGONO DA PRIVACIDADE NO BIG DATA PRIVACY ANALYTICS: PROPOSTA DE MODELO MULTIFACETADO DE GARANTIA DA PRIVACIDADE E DO VALOR ANALÍTICO**

## BRENNER LOPES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia da Informação e Comunicação.

Aprovada em 22 de fevereiro de 2024, por videoconferência, pela banca constituída pelos membros:

Prof(a). Ricardo Rodrigues Barbosa (Orientador)
Aposentado/UFMG

Prof(a). George Leal Jamil
Fundação Dom Cabral

Prof(a). Cristiana Fernandes de Muylder
Universidade Federal de Uberlândia

Prof(a). Darly Fernando Andrade
Universidade Federal de Uberlândia

Prof(a). Andre Luiz de Castro Leal
UFRRJ

Belo Horizonte, 22 de fevereiro de 2024.

Documento assinado eletronicamente por **Ricardo Rodrigues Barbosa**, **Membro de comissão**, em 29/02/2024, às 20:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **George Leal Jamil**, **Usuário Externo**, em 01/03/2024, às 14:47, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Darly Fernando Andrade**, **Usuário Externo**, em 01/03/2024, às 17:46, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiana Fernandes de Muylder**, **Usuária Externa**, em 18/03/2024, às 15:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **André Luiz de Castro Leal**, **Usuário Externo**, em 19/03/2024, às 08:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3040452** e o código CRC **8F0BCC2D**.

**Referência:** Processo nº 23072.209310/2024-25 SEI nº 3040452

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

# ATA DA DEFESA DE TESE DO ALUNO

## BRENNER LOPES

Realizou-se, no dia 22 de fevereiro de 2024, às 10:00 horas, por videoconferência, da Universidade Federal de Minas Gerais, a defesa de tese, intitulada *PENTÁGONO DA PRIVACIDADE NO BIG DATA PRIVACY ANALYTICS: PROPOSTA DE MODELO MULTIFACETADO DE GARANTIA DA PRIVACIDADE E DO VALOR ANALÍTICO*, apresentada por BRENNER LOPES, número de registro 2019663796, graduado no curso de CIÊNCIAS ECONÔMICAS, como requisito parcial para a obtenção do grau de Doutor em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Ricardo Rodrigues Barbosa - Aposentado/UFMG (Orientador), Prof(a). George Leal Jamil - Fundação Dom Cabral, Prof(a). Cristiana Fernandes de Muylder - Universidade Federal de Uberlândia, Prof(a). Darly Fernando Andrade - Universidade Federal de Uberlândia, Prof(a). Andre Luiz de Castro Leal - UFRRJ.

A Comissão considerou a tese:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 22 de fevereiro de 2024.

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Ricardo Rodrigues Barbosa**, **Membro de comissão**, em 29/02/2024, às 20:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **George Leal Jamil**, **Usuário Externo**, em 01/03/2024, às 14:47, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Darly Fernando Andrade**, **Usuário Externo**, em 01/03/2024, às 17:45, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristiana Fernandes de Muylder**, **Usuária Externa**, em 18/03/2024, às 15:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **André Luiz de Castro Leal**, **Usuário Externo**, em 19/03/2024, às 08:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3040441** e o código CRC **7BDD1E46**.

**Referência:** Processo nº 23072.209310/2024-25 SEI nº 3040441

## DEDICATÓRIA

Brenner, Bernardo, Sofia, Sheila. Somente a pronúncia dos seus nomes remonta a tudo de bom que a vida me concedeu!

## AGRADECIMENTOS

Aos meus amados filhos e esposa, pelo apoio, confiança e cumplicidade de sempre.

Aos meus pais por terem me dado a vida e pelo companheirismo na caminhada.

Ao meu companheiro de trabalho e de publicações, Luander Falcão, parceiro de todas as horas.

Ao Prof. Ricardo Rodrigues, pela paciência, confiança, crença e lições de humildade.

A todos os colegas e amigos que de uma forma ou de outra contribuíram para essa construção.

“O valor das coisas não está no tempo que elas duram, mas na intensidade com que acontecem. Por isso existem momentos inesquecíveis, coisas inexplicáveis e pessoas incomparáveis.”

Fernando Pessoa

**RESUMO**

O contínuo aumento da quantidade de dados gerados e disponíveis, criou um ambiente caracterizado por uma gigantesca massa de dados, que crescem não só quanto ao seu volume e quantidade, mas também em termos de variedade, sendo criado e transitando em alta velocidade. Se antes, o foco e a disponibilidade de dados, tinha como interface prioritária os chamados dados estruturados, hoje eles estão em quantidade e importância bem menores, e os ajustes e aprimoramentos nas tecnologias e análises foram em parte realizados para se adaptarem a essa nova realidade que se convencionou chamar de *big data*. O valor criado pelas análises de *big data*, têm trazido muitos fatores positivos para diversos campos, como também, por outro lado, têm trazido alguns problemas. Uma das questões de grande relevância nesse contexto que têm suscitado grande preocupação à sociedade são as ameaças à privacidade, trazidas pelas análises avançadas de grandes volumes de dados. A questão posta como resultado de diversas pesquisas, é que os procedimentos, técnicas e tecnologias atualmente disponíveis não conseguem dar garantia plena à privacidade. Assim como as legislações que disciplinam a proteção e privacidade de dados, como no caso brasileiro a Lei Geral de Proteção de Dados (LGPD). Por outro lado, é possível obter um elevadíssimo nível de privacidade, mas ao custo da nulidade das possibilidades de extração de valor do *big data*. Diante desse complexo cenário, o foco dessa pesquisa está em propor um modelo multifacetado (porque precisa ser composto por múltiplas abordagens e construtos) no âmbito do *big data analytics*, que garanta a privacidade, ao mesmo tempo, em que não inviabilize sua extração de valor. A metodologia proposta para esse trabalho, estará baseada na abordagem quantitativa, com o emprego de mineração de texto e aprendizado de máquina não supervisionado com foco na criação de agrupamentos. Após a fase quantitativa foram analisadas a partir de uma abordagem qualitativa as políticas de privacidade de 28 empresas, com dois objetivos principais: encontrar o melhor número de clusters; e obter um entendimento de características distintivas dos mesmos, de forma a aprimorar a análise realizada. Foram selecionadas as políticas de privacidade das mil maiores empresas do Brasil, assim classificadas pelo ranking Valor1000 (2021). Os resultados apontaram para a consistência do modelo do Pentágono da Privacidade no *Big Data Analytics* proposto; assim como para o que foi denominado neste trabalho de “privacidade opaca”.

**Palavras-chave:** Privacidade; *Big Data Analytics*; *Big Data Privacy*; Valor do *big data*; *Clustering*; Aprendizado de máquina não supervisionado; Mineração de texto.

## ABSTRACT

The continuous increase in the amount of data generated and available has created an environment characterized by a gigantic mass of data, which grows not only in terms of volume and quantity, but also in terms of variety, being created and moving at high speed. If before, the focus and availability of data had as a priority interface the so-called structured data, today they are in much smaller quantity and importance, and adjustments and improvements in technologies and analyzes were partly carried out to adapt to this new reality which is conventionally called big data. The value created by big data analysis has brought many positive factors to different fields, but, on the other hand, it has also brought some problems. One of the issues of great relevance in this context that has raised great concern in society are the threats to privacy, brought about by advanced analyzes of large volumes of data. The question raised as a result of several researches is that the procedures, techniques and technologies currently available cannot fully guarantee privacy. As well as legislation that regulates data protection and privacy, such as in the Brazilian case the General Data Protection Law (LGPD). On the other hand, it is possible to obtain a very high level of privacy, but at the cost of nullifying the possibilities of extracting value from big data. Given this complex scenario, the focus of this research is to propose a multifaceted model (because it needs to be composed of multiple approaches and constructs) within the scope of big data analytics, which guarantees privacy, at the same time, in which it does not make it impossible to extract value. The methodology proposed for this work will be based on a quantitative approach, using text mining and unsupervised machine learning with a focus on creating clusters. After the quantitative phase, the privacy policies of 28 companies were analyzed using a qualitative approach, with two main objectives: finding the best number of clusters; and obtain an understanding of their distinctive characteristics, in order to improve the analysis carried out. The privacy policies of the thousand largest companies in Brazil were selected, classified as such by the Valor1000 ranking (2021). The results pointed to the consistency of the Privacy Pentagon model in the proposed Big Data Analytics; as well as what was called “opaque privacy” in this work.



**Keywords:** Privacy; Big Data Analytics; Big Data Privacy; Value of big data; Clustering; *Unsupervised machine learning; Text mining.*

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE QUADROS

## LISTA DE SIGLAS E ABREVIATURAS

| | | |
|---|---|---|
| ABNT | - | Associação Brasileira de Normas Técnicas |
| AI | - | Artificial Intelligence |
| CRISP-DM | - | Cross Industry Standard Process for Data MINING |
| DP | - | Dados Pessoais |
| DPO | - | Data Protection Officer |
| EBITDA | - | Earnings Before Interest, Taxes, Depreciation and Amortization |
| EDA | - | Exploratory Data Analysis |
| GPDR | - | General Data Protection Regulation |
| GPU | - | Graphics Processing Units |
| IASB | - | International Accounting Standards Board |
| IFRS | - | International Financial Reporting Standards |
| IOT | - | Internet das Coisas |
| ISO | - | International Organization for Standartization |
| LGPD | - | Lei Geral de Proteção de Dados |
| NLP | - | Natural Language Processing |
| PDF | - | Portable Document Processing |
| PP | - | Política por Privacidade |
| KDT | - | Knowledge Discovery in Textual Databases |

## SUMÁRIO

# 1 INTRODUÇÃO

O contínuo aumento da quantidade de dados gerados e disponíveis, que com base nas projeções e análises mais recentes só tendem a continuar aumentando, visto o impulso de tecnologias, como, por exemplo, a internet das coisas (*IoT - Internet of Things*); que tem conectado cada vez mais coisas (equipamentos, veículos, máquinas e múltiplos dispositivos) e seres (humanos, animais e mais recentemente humanoides), tem gerado uma quantidade quase infinita e recorrente de dados. Some-se a isso o advento das redes sociais e outras plataformas digitais, onde quem participa é quem gera o conteúdo, está constituído o quadro geral do recurso mais relevante de nosso tempo, a informação.

Uma característica fundamental desse ambiente, caracterizado por uma gigantesca massa de dados, é que ele cresce não só quanto ao seu volume e quantidade, mas também em termos de variedade, sendo criado e transitando em alta velocidade. Se antes o foco e a disponibilidade de dados tinham como interface prioritária os chamados dados estruturados, hoje eles estão em quantidade e importância bem menor, e os ajustes e aprimoramentos nas tecnologias e análises foram em parte realizados para se adaptarem a essa nova realidade, que se convencionou chamar de *big data*.

Segundo apontado por Rani e Dhamodaran (2016, p. 33, grifos nossos, tradução nossa), nos últimos anos, “a quantidade total de dados gerados por humanos teve um aumento explosivo de 300 vezes, de exabytes (1 exabyte = $10^{18}$ = 1.000 petabyte) para octabytes (1 octabyte = $10^{24}$ = 1.000.000 petabytes)”. E complementam a explanação demonstrando que tais dados advém de diversas fontes, como, por exemplo, “pesquisa, governo, finanças e negócios, sociedade, redes, fotografia, vídeos, áudios de telefones celulares”, dentre outros.

Sun, Pambel e Strang (2018, p.6, tradução nossa), relatam alguns exemplos práticos das dimensões quantitativas de dados processados por algumas empresas, segundo os autores “o google processa mais de 20 petabytes de dados diariamente. O Walmart coleta mais de 2,5 petabytes de dados não estruturados de um milhão de

clientes a cada hora". Como colocado por Oostveen (2016, p. 302, tradução nossa), "quanto mais dados são coletados, mais o potencial de interferência com a vida privada pode se dar".

E nesse contexto, é possível acrescentar uma questão que problematiza ainda mais esse cenário, o fato de que as empresas públicas e privadas estão coletando dados, armazenando-os e analisando-os por mais tempo, num patamar sem precedentes até então. (Altman *et al*, 2018; Thomson e Thibadeau, 2016)

Para fazer frente a essa avalanche informacional foi necessário criar alternativas não apenas para extrair valor desses dados e informações, mas antes disso, para processá-los e armazená-los. Nesse âmbito, ocorreram dois eventos fundamentais para sedimentar toda a dinâmica atual que temos vivenciado. De um lado, o advento da chamada "nuvem", que não só possibilitou o processamento e o armazenamento de imensas quantidades de dados em seus múltiplos formatos e advindos de diversas fontes, como também fez cair drasticamente o custo envolvido para a realização dessas atividades.

E não só o custo deve ser observado, mas também e principalmente um novo modelo de negócio que foi gerado, ao estilo "*pay-per-use*" (pague pelo que usar, tradução nossa), ou seja, só se paga pela estrutura e volume de dados que se processa e armazena efetivamente. Esse novo modelo possibilitou o que se pode chamar de uma democratização quanto ao acesso e possibilidades de tratamento, armazenamento e análise de grandes dados, facilitando o acesso a esses serviços por empresas de qualquer porte, assim como pesquisadores, ligados ou não a instituições de pesquisa.

Por outro lado, como a segunda força transformadora, ocorreram avanços também das técnicas analíticas envolvidas na análise desses grandes dados. É importante destacar que o núcleo estruturante dessas técnicas não é algo novo, já estando disponível há diversos anos. Mas muitas das técnicas mais avançadas, só tiveram sua utilização plena com o advento da possibilidade de processamento e

armazenamento de grandes volumes de dados. Novas técnicas e o melhoramento de técnicas mais antigas, foram também fundamentais.

Com toda essa disponibilidade de dados e respectiva possibilidade de coleta, armazenamento, análise e respectiva criação de valor atrelado a múltiplos objetivos e interesses, disponíveis para praticamente qualquer pessoa e empresa, surgiram, por um lado, novos problemas e, por outro, problemas já existentes foram agravados. Assim como, várias novas soluções trouxeram grandes e importantes benefícios às pessoas e empresas.

Como destacado por Oostveen (2016, p. 299, tradução nossa), o *big data* “é o presente e o futuro”. Onde a partir dessa afirmação aponta para as possibilidades de desenvolvimento em diversos campos, como nos negócios, ciência, educação, saúde, governos e outros mais. E conclui destacando que a “enorme coleção e o uso de dados pessoais também levanta uma série de questões”, entre elas e de maneira particular a questão da privacidade.

Um exemplo dessa constatação está nas afirmações de Politou, Alepis e Paysakis (2018, p .2, tradução nossa), quando apontam para quatro tendências recentes, que já têm gerado impactos à esfera da privacidade, “o smartphone, a nuvem, as redes sociais e o *big data*, em conjunto, permitem o surgimento de uma nova sociedade de vigilância poderosa que levanta novas ameaças significativas à privacidade, como rastreamento de localização”.

Sendo assim, o foco desta pesquisa estará assentado no contexto do *big data analytics*, onde os times internos e/ou externos de analítica têm franco acesso aos dados, portanto, não há uma questão tão premente de segurança, mas sim de entender se os times de *analytics*, ou de maneira mais formal e abrangente, as funções de *analytics* nas organizações, conseguem garantir em seus processos, processamentos e análises a privacidade quanto aos dados pessoais acessados e utilizados.

Parte-se do ponto em que se entende que as questões de segurança, apesar de serem questões muito próximas e simbióticas à problemática da privacidade, pois

sem segurança não existe privacidade, foram previamente atendidas e, portanto, o foco se deu estritamente na questão da privacidade. Mas vale o destaque, reforçando o que já foi exposto anteriormente, de que mesmo a privacidade não existindo sem a segurança, a segurança não é capaz de forma isolada de garantir a privacidade. Essa afirmação é validada por Abouelmehdi, Beni-Hessane e Khaloufi (2018, p. 4, tradução nossa), quando trazem uma afirmação direta e contundente, muito da visão que norteou a definição estreita do foco desta pesquisa em privacidade, ao afirmarem que "embora a segurança seja vital para proteger os dados, ela é insuficiente para lidar com a privacidade".

As questões ligadas de forma direta à proteção e garantia da privacidade, por ser considerada um valor universal e inato à qualquer pessoa, fez com que diversas nações, frente aos riscos trazidos pelo chamado *big data privacy*, ou seja, todo o conceito e abordagens ligadas ao *big data analytics* no âmbito das questões atreladas à privacidade, propusessem e aprovassem legislações que tratam diretamente dos riscos e ameaças à privacidade dos cidadãos. Trazidos por essa nova realidade, tem o objetivo de garantir esse direito inalienável de qualquer pessoa.

Esse é um movimento que ocorreu e vem ocorrendo em diversas nações ao redor do globo, em alguns casos a adoção e em outros aprimoramentos a partir das experiências com a adoção das legislações anteriores. Nesse âmbito podem ser citados alguns exemplos, como: *Privacy Ac*t, 1998 (Austrália); *Privacy Principles* (OCDE, 2010); *Fair Information Practices* (FTC, 2000), Proteção de dados das crianças – COPPA (EUA, 1998); *Health Information Act* – HIPAA (EUA, 1996), e a Lei de Proteção de Informações Pessoais e Documentos Eletrônicos – PIPEDA (Canadá, 2000). (Ouazzani e Bakkali, 2020).

No Brasil, no ano de 2018, tivemos a aprovação da chamada LGPD (Lei Geral de Proteção de Dados), que baseada na legislação europeia, o GDPR (Regulamento Geral sobre a Proteção de Dados) / 2016/679, visa regulamentar o acesso e utilização aos dados pessoais, assim como salvaguardar o direito de todo cidadão à privacidade.

Essas iniciativas reforçaram outras questões fundamentais na esfera da garantia da privacidade, como as abordagens tecnológicas e técnicas de preservação da privacidade, que já existiam e eram praticadas por alguns agentes, públicos e privados, mas tomaram novas dimensões e perspectivas.

A contraparte dessas afirmações, não obstante essa nova realidade, também traz um ponto fundamental e que não pode, pelos benefícios que podem gerar para a sociedade de maneira geral, serem desconsiderados, que é o valor possível que pode ser extraído / gerado pelo *big data analytics* em diversas frentes, como melhores produtos, inovações, novos processos, melhorias no atendimento, novas formas de se fazer marketing, e porque não novos empregos e outras formas de geração de renda e postos de trabalho. Também é verdadeira, ou seja, existe um lado muito negativo dessas abordagens, dentre as quais poderíamos citar a discriminação algorítmica, vigilância e divulgação.

Como colocado por Acquisiti, Taylor e Wagman (2016, p.483, tradução nossa), “na verdade, a exploração do valor comercial dos dados pode muitas vezes implicar numa redução de sua utilidade privada, e às vezes, mesmo no bem-estar social em geral”.

Nesse ponto se encontra um grande dilema que integrará a estrutura principal dessa pesquisa, que é a importância de se poder extrair o valor do *big data* ao mesmo tempo, em que se garante a privacidade. Esse é um verdadeiro dilema que merece atenção. É possível garantir privacidade total ao custo da nulidade das possibilidades de extração de valor do *big data analytics*.

A questão central desta pesquisa, está assentada em duas importantes frentes:

1) Como extrair o devido valor do *big data analytics* ao mesmo tempo em que seja possível dar garantias de preservação da privacidade?
2) Como garantir a privacidade frente às análises de grandes quantidades de dados?

A análise dos estudos mais recentes tem demonstrado que essas questões remetem a desafios ainda sem solução, apesar dos avanços em múltiplas frentes, como legislação, política, ética, comportamento, técnica, métodos, metodologias e tecnologia. É possível constatar que o *big data analytics*, e todo o seu arsenal, tem avançado de forma mais acelerada que os avanços nos campos acima citados. (Chanson *et al*, 2019)

Uma primeira confirmação dessa constatação, com foco na segurança e tecnologia, pode ser vista nas afirmações de Pragash e Jayabharathy (2017, p. 95, tradução nossa), quando apontam que as medidas de segurança consideradas tradicionais, como "firewall, antivírus e sistemas de detecção e prevenção de intrusão, não estão mais fornecendo os níveis exigidos de granularidade, proteção e fiscalização", requisitados para o atendimento das regulamentações no campo da segurança e privacidade.

Como afirma Rao, Krishna e Kumar (2018, p. 9, tradução nossa), na discussão dos resultados de sua pesquisa, no que tange à revisão sistemática da literatura que empreenderam, "observou-se que todos os mecanismos existentes de preservação da privacidade dizem respeito a dados estruturados". A questão que se coloca aqui é que mais de 80% dos dados gerados hoje não são estruturados.

Segundo Ohm (2010, p. 1704, tradução nossa), "os dados podem ser úteis ou perfeitamente anônimos, mas nunca ambos". O que já se sabe é que as legislações, técnicas e tecnologias disponíveis ainda não conseguiram desatar esse nó. (Rao, Krishna e Kumar, 2018; Adams, 2017; Wilson, Belliveau e Gray, 2017; Ishii, 2017; Oostveen, 2016; Sun, Pambel e Strang, 2018).

Some-se a isso uma questão relevante, a de que a proteção / garantia da privacidade ao nível das empresas (públicas e/ou privadas), está atrelada, de alguma forma à determinação dessas em seguir não só a legislação é efetivamente utilizar técnicas e tecnologias adequadas, mas também de instituírem códigos internos de ética, comportamento e procedimentos visando esse objetivo, já que a detecção de possíveis descumprimentos, por exemplo, a nível da legislação instituída é algo bem

complicado de ser identificado *a priori*, que só se torna mais visível em casos de grandes proporções ou com base em denúncias (que assim também precisam ser confirmadas).

Diante desse complexo cenário que se apresenta, o foco dessa pesquisa esteve em propor e validar um modelo multifacetado (porque precisa ser composto por múltiplas abordagens e construtos), no âmbito do *big data analytics,* de forma a garantir a privacidade, ao mesmo tempo em que não inviabilize a extração do seu valor.

## 2 JUSTIFICATIVA

A análise de dados, no âmbito do *big data,* é muito útil às organizações em diversas frentes, possibilitando, dentre outros benefícios, decisões mais assertivas; porém ela já apresenta e continuará apresentando graves preocupações quanto à privacidade. Foi o que apontaram Sun, Pambel e Strang (2018, p.4, tradução nossa), quando afirmaram que o "*big data* traz grandes problemas de privacidade e segurança". A fase de análise de dados no *big data*, está sujeita a violações de privacidade e divulgação de dados, devido em parte às características "multiusuário dos ambientes de *big data*". (Al-Zobbi, Shahrestani e Ruan, 2017, p. 1; tradução nossa)

Conforme colocado por Adams (2017, p.1, tradução nossa), privacidade, no âmbito do *big data*, se tornou "uma questão central, que afeta diferentes áreas de tecnologia, pois a conectividade e o compartilhamento de informações ultrapassaram em muito os esforços de proteção de dados". Os mesmos autores ainda colocam que "a privacidade de dados individuais é esperada complexa e multifacetada, estendendo-se por domínios tecnológicos, legais, comerciais e financeiros". (Adams, 2017, p. 15, tradução nossa)

Essa ideia é corroborada por Altman *et al* (2018, p. 39, tradução nossa), quando afirmam que a "expansão da escala de dados e novos usos comerciais estão aumentando os riscos e diminuindo a eficácia dos controles".

Chanson *et al* (2019, p.1, tradução nossa) consolidam essa visão ao declararem que "o potencial dos problemas de segurança e privacidade relacionados aos sistemas de informação (SI) pode afetar os clientes em suas vidas diárias e esferas privadas, tornando esses desafios as principais prioridades dos negócios".

Sinnott *et al* (2016, p. 606, tradução nossa), advogam que "privacidade de dados é uma área de crescente importância de pesquisa, especialmente à medida que mais organizações adotam um modelo terceirizado de fornecimento de TI e compartilhamento de dados por nuvem".

Na visão de Sun, Pambel e Strang (2018, p. 5, tradução nossa), “as soluções de segurança tradicionais não são projetadas para proteger a privacidade individual na era do *big data*”, o que reforça a importância de se pensar e propor modelos que consigam abarcar o maior número de questões determinantes da garantia da privacidade. Esses mesmos autores também destacam, tratar-se de uma temática ainda incipiente, quando apontam que “a maioria da literatura recente afirma que precisamos de mais pesquisas sobre a privacidade e segurança no *big data*” (Sun, Pambel e Strang, 2018, p.2, tradução nossa).

Jain, Gyanchandani e Khare (2016, p. 3, tradução nossa), reforçam essa visão ao afirmar que “embora a segurança seja fundamental para proteger os dados, não é suficiente para abordar a privacidade.” E complementam ao raciocinar, refletir e afirmar que “a análise avançada de dados pode extrair informações valiosas da big data, mas, ao mesmo tempo, representa um grande risco para a privacidade dos usuários”. (Jain, Gyanchandani e Khare, 2016, p. 10, tradução nossa)

E quando se projeta a questão da privacidade como foco principal, como nessa pesquisa, isso é ainda mais relevante, pois como afirmam Hadar *et al* (2017, p. 275, tradução nossa), para o caso de usuários e desenvolvedores de softwares, mas a nosso ver a ideia é válida de maneira geral, existe uma grande confusão de entendimento entre os termos, “os usuários muitas vezes reduzem a noção de privacidade apenas as questões de segurança”.

Partindo da motivação e visão de complementação quanto a afirmação anterior, é cabível a proposição de uma explicação relevante. Quando consideradas em conjunto “*big data analytics* + privacidade”, ou melhor, *big data analytics* no contexto da privacidade (*big data privacy*), tem-se um campo de estudos recente que ainda carece de mais estudos e pesquisas. Já quando analisado o campo de estudos no âmbito da privacidade, é possível detectar que os estudos e pesquisas nesse campo datam de muitos anos e possuem múltiplas nuances de atenção dos pesquisadores.

Apesar dessa constatação, Politou, Alepis e Paysakis (2017, p. 3, tradução nossa), trazem uma clara visão da importância e necessidade de haver ainda um

considerável espaço para se debruçar sobre a temática privacidade, ao afirmarem que ainda “não existe uma definição universalmente aceita de privacidade”;

A constatação anterior pode ser corroborada pelas colocações de Chanson *et al* (2019, p. 5, tradução nossa) quando os autores são categóricos em afirmar que “apesar do corpo de conhecimento existente, faltam soluções viáveis” à problemática da privacidade. Como também pelas constatações de Jain, Gyanchandani e Khare (2016, p. 23, tradução nossa), quando advogam nas conclusões do seu estudo que “como tal, existe uma enorme margem para mais pesquisas sobre métodos de preservação da privacidade em *big data*”.

Na visão trazida por Adams (2017, p. 12, tradução nossa), a questão da privacidade no âmbito do *big data* torna-se uma questão muito relevante, visto que a “conectividade e o compartilhamento de informações ultrapassam em muitos os esforços de proteção de dados”. No que Abouelmehdi, Beni-Hessane e Khaloufi (2018, p. 15, tradução nossa), são taxativos ao declararem que “métodos de privacidade precisam ser aprimorados”.

No que se refere ao *big data*, esse conceito pode ser entendido como um conjunto de dados estruturados e não estruturados, cujo tamanho não pode ser processado/analisado utilizando as ferramentas tradicionais de processamento e análise de dados, necessitando, portanto, de estruturas de processamento distribuídas, como o Haddop, Pig e Spark, por exemplo. Está estruturado pelos chamados 3Vs: volume, velocidade e variedade. É justamente a multiplicidade de fontes que integram a análise de *big data* que traz no seu bojo um dos grandes desafios à privacidade.

Rhoen (2015, p. 51, tradução nossa), expõe uma questão fundamental e de grande relevância, quando de forma sucinta e objetiva coloca em questão um ponto fundamental para a discussão aqui proposta, que é o fato de que o *big data* “coloca a proteção dos dados à prova”. Além disso, complementam, justificando tal posicionamento ao afirmarem que os consumidores concedem sua permissão para

que seus dados pessoais sejam processados “abrindo suas vidas pessoais, graças à identificação de dados”.

Por isso mesmo, na visão de Rao e Mehta (2019, p.1, tradução nossa), “a privacidade é uma das principais preocupações com a análise de *big data*”. E mesmo não havendo dúvida quanto à utilidade e valor da análise de *big data*, ela inevitavelmente demandará muita atenção e preocupação no que diz respeito à privacidade. (Rao, Krishna e Kumar, 2018)

Como reforçam Sun, Pambel e Strang (2018, p. 6, tradução nossa), “as legislações existentes são basicamente baseadas em países, enquanto a era do *big data* é global em essência”. Os mesmos autores complementam sua visão anterior ao afirmarem que “ainda existem lacunas na proteção da privacidade do indivíduo”.

Ouazzani e Bakkali (2020, p. 2, tradução nossa), apontam para uma das questões fundamentais quanto a motivação e justificativa da importância da temática dessa pesquisa, ao afirmarem que “a maioria das técnicas convencionais de privacidade de dados não suporta a escala completa do *big data*”. Portanto, ainda na visão dos autores, “é obrigatório garantir a privacidade, garantindo que todas as tentativas de identificar o indivíduo falhem”.

Segundo destacado por Abouelmehdi, Beni-Hessane e Khaloufi (2018, p. 4, tradução nossa), em janeiro de 2014, a equipe de John Podeste conselheiro do então presidente dos EUA Barak Obama, via Casa Branca, realizou uma profunda revisão sobre as questões de privacidade e *big data*, que trouxe contundentes recomendações objetivando a potencialização dos benefícios e a minimização dos riscos do *big data*. Esses *insights* serviram de grande inspiração e direcionamento para a estruturação desta pesquisa, assim como do seu foco e objetivos:

- A atenção política deve se concentrar mais nos usos reais do *big data* e menos em sua coleta e análise. Essas políticas existentes provavelmente não produzirão estratégias eficazes para melhorar a privacidade ou serão escalonáveis ao longo do tempo;

- A política relativa à proteção da privacidade deve abordar o propósito em vez de prescrever o mecanismo;
- A pesquisa é necessária nas tecnologias que ajudam a proteger a privacidade, nos mecanismos sociais que influenciam o comportamento de preservação da privacidade e nas opções legais que são robustas a mudanças na tecnologia e criam um equilíbrio apropriado entre oportunidades econômicas, prioridades nacionais e proteção de privacidade;
- Aumento das oportunidades de educação e treinamento sobre proteção da privacidade, incluindo planos de carreira para profissionais. Os programas que fornecem educação que conduzem à experiência em privacidade são essenciais e precisam de incentivo.

Pode-se verificar de forma clara que a questão da privacidade está além, apesar de precisar e se estruturar a partir delas, das questões legais e de segurança. Politou, Alepis e Paysakis (2018, p. 15, tradução nossa), reforçam essa constatação ao afirmarem em sua análise sobre a questão da vigência da GPDR na União Europeia que, “poucas organizações ainda são capazes de provar a conformidade real com a GPDR”. E segundo os mesmos autores, num aprofundamento dessa constatação, apontam que um dos principais fatores causadores desse panorama negativo é que a GPDR é um documento legal.

Finalizamos as considerações sobre a importância e relevância do tema proposto para essa pesquisa, trazendo algumas reflexões e considerações de Mills (2018, p. 597, tradução nossa), quando o mesmo argumenta pelo não esquecimento quanto aos problemas afetos à ética e a privacidade, em uma “era em que o *big data* pode ser usado por empresas, estados e nações”. Esse mesmo autor ainda destaca questões sobre um uso secreto do *big data*, apontando, portanto, para um possível lado negro do *big data*, concluindo que as preocupações e garantias no âmbito da privacidade requerem muito mais do que ajustes nos atuais “protocolos”. (Mills, 2018)

Nesse sentido, parece muito relevante que a proteção / garantia da privacidade deverá congregar soluções tecnológicas, legais, sociais, culturais e políticas, tanto ao nível de nações quanto de empresas.

## 3 QUESTÃO DE PESQUISA E OBJETIVOS

Partindo do ponto de entendimento de que os mecanismos atualmente disponíveis para prover a garantia da privacidade no âmbito do *big data analytics*, não são suficientes para prover essa pretensa garantia, a questão de partida que motivou essa proposta de pesquisa está amparada pela seguinte pergunta, para a qual propõe se uma alternativa: “Como garantir a privacidade e o valor potencial do *big data* em análises de *big data*?”, que enseja, portanto, a “proposição e validação de um modelo multifacetado de garantia da privacidade em análises de *big data*”.

Conforme apresentado por Hair *et al* (2009, p. 545), um “modelo é uma representação de uma teoria” (portanto, não pode ser pensado sem uma teoria ou teorias consistentes suportando-a), que poderia ser visualizado como um conjunto lógico de relações em torno de uma determinada questão, um determinado tema, capaz de fornecer uma “explicação consistente e abrangente” sobre a questão inicialmente levantada.

Esta pesquisa esteve assentada no esforço de identificação, análise lógica e validação dos construtos estruturantes de um modelo multifacetado que consiga prover efetiva garantia à preservação da privacidade, no que se refere aos dados pessoais, ao mesmo tempo, em que garante a geração de valor, objetivo fundamental e sustentador da proposta do *big data*.

A literatura tem apontado consistentemente para o fato de que ainda não se possui um modelo suficientemente robusto que consiga garantir a privacidade dos dados em análises de *big data*, ao mesmo tempo que também garanta seu valor. A constatação é que o problema é complexo e envolve múltiplas dimensões como legislação, política e códigos de conduta, sensibilização, comportamento e tecnologia.

O valor, em *big data*, são os possíveis resultados trazidos por suas análises, como aumento de produtividade, de faturamento, melhoria de produtos, maior assertividade das campanhas de marketing, dentre outros. Sem dados consistentes e

em volumes elevados, muitas das mais representativas análises em *big data* serviriam para pouca coisa, pois não geram o valor esperado.

Tem-se, na atualidade, um ponto de inflexão e altamente desafiador no que tange à privacidade e ao *big data*. Mesmo não sendo possível garantir a total privacidade de um indivíduo, a aplicação de múltiplos mecanismos com foco na obtenção dessa pretensa garantia, se em graus elevados, impossibilita a geração de valor nas análises de *big data*.

Como obter o importante valor gerado pelo *big data* ao mesmo tempo que se garanta a privacidade? Talvez essa seja uma das questões mais desafiadoras dos próximos anos. Por enquanto, o que se tem até o momento são processos, técnicas, tecnologias e legislações, que na maioria das vezes pendem mais fortemente para um lado ou para o outro, sem encontrar um espaço onde essa questão possa ser pacificada. Com certeza esse espaço deverá ter características múltiplas e unir diversas visões e dimensões.

Privacidade é um direito universal inquestionável. O valor trazido pelo *big data* tem provido a humanidade de muitos ganhos em inúmeras áreas. A análise de *big data* como qualquer outra abordagem de impacto tão profundo, sempre traz consigo alguns perigos quando da sua utilização inadequada e sem compromisso com a ética e a legalidade. Mas, como pontuado, seu valor é inquestionável. A busca por soluções que equacionem ou tragam algum avanço quanto a esse desafio é uma questão de primeira ordem.

## 3.1 Objetivo Geral

Propor um modelo multifacetado, e sua validação, para garantia da privacidade e geração de valor no *big data* no âmbito das organizações brasileiras.

## 3.2 Objetivos Específicos

São os objetivos específicos que sustentaram o objetivo geral:

- Identificar qual é o status e o estado da arte da pesquisa de privacidade e geração de valor no paradigma do *big data analytics*;
- Identificar, analisar e sistematizar as teorias ou fragmentos dessas a partir das pesquisas e publicações sobre a temática;
- Identificar, analisar e selecionar o conjunto de variáveis suficientes e abrangentes para proposição de um modelo multifacetado de garantia da privacidade e geração de valor no *big data analytics*;
- Coletar, preparar e analisar textos, através da abordagem de aprendizado de máquina não supervisionado, para realização de mineração de texto semiautomizada, com análise por agrupamentos realizada com base na metodologia de *clustering*, tendo como foco as políticas de privacidade das mil maiores organizações do Brasil, segundo o ranking Valor 1000 (2021);
- Validar o modelo teórico constituído.

# 4 ESTRATÉGIA DE BUSCA PARA REVISÃO DA LITERATURA

Visando a obtenção de entendimento e conhecimento sobre a temática da privacidade no âmbito do *big data analytics* foi realizada uma busca avançada no Portal Capes (Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

Após pesquisas e análises preliminares para verificação e confirmação da relevância e atualidade do tema desta pesquisa, foi possível ter clareza sobre um importante aspecto que foi um direcionador importante para o delineamento da estratégia de revisão da literatura, que foi o caráter multifacetado do tema.

Ou seja, foi possível verificar a existência de muitas pesquisas sobre *big data* e sobre privacidade. Inclusive no que tange à privacidade, a existência de muitas pesquisas e estudos de longa data. Por outro lado, pesquisas específicas sobre a questão da privacidade no seio da análise de *big data* no âmbito das organizações, se mostrou um campo mais restrito no que tange ao quantitativo de estudos e pesquisas, e estes foram empreendidos em anos mais recentes.

Partindo desse ponto, na revisão de literatura executada nesta pesquisa, optou-se por utilizar o Portal Capes não como uma primeira camada de busca para posterior refinamento, mas como a principal ferramenta de busca e posterior seleção e refinamento dos resultados dessas buscas. Como exposto acima, isso se deu principalmente com base na característica mais visível e impactante da temática, que é seu contorno multifacetado, sendo discutido de um lado (*big data analytics*) ou de outro (*big data privacy*) em múltiplos repositórios, não possuindo uma predominância, a princípio, nem mesmo naqueles com forte posicionamento multidisciplinar.

Essa estratégia possibilitou abarcar de maneira mais ampla as possibilidades de recuperação de materiais relevantes, independentemente do repositório em que se encontravam, desde que, claro, estivessem abarcados pelo Portal Capes.

Segundo consta nas explicações do seu site, o Portal de Periódicos, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), é uma

biblioteca virtual que reúne e disponibiliza a instituições de ensino e pesquisa no Brasil o melhor da produção científica internacional. Ele conta com um acervo de mais de 45 mil títulos com texto completo, 130 bases referenciais, 12 bases dedicadas exclusivamente a patentes, além de livros, enciclopédias e obras de referência, normas técnicas, estatísticas e conteúdo audiovisual.

Por todas as questões expostas acima, o processo de revisão de literatura, esteve eivado de uma complexidade no que tange à seleção dos termos que balizaram as buscas pelos periódicos. Isso se deu, devido ao fato de que os termos centrais da pesquisa, "*big data*" e "privacidade", não são temáticas tão novas assim; já no que tange ao foco desta pesquisa, a temática privacidade no âmbito do *big data analytics*, passa a ser um tema cujas pesquisas começaram nos últimos anos a serem publicadas, tratando-se, portanto, de uma temática ainda no começo de seu processo exploratório.

Com o objetivo de se chegar aos termos mais adequados e que pudessem recuperar os conteúdos mais aderentes à temática deste trabalho, foram realizadas diversas incursões na busca avançada do Portal Capes, testando um conjunto de termos simples e compostos, com a subsequente análise dos periódicos extraídos, realizada através de seleção de amostras desses periódicos e de verificação quanto a presença dos termos no resumo e palavras-chave, assim como a leitura dos respectivos resumos.

Nesse sentido foram realizadas um total de seis buscas e respectivas extrações e análises. Os termos testados e as configurações utilizadas nas buscas e seus respectivos resultados são os que seguem abaixo:

1. Portal Capes – sintaxe busca avançada: "*big data*" AND "*privacy*"; Termos: "é exato para os dois termos"; Qualquer campo: "título, autor, assunto"; Data Publicação: "últimos 5 anos"; Tipo de material: "artigos"; Idioma: "inglês", Filtro: "ordenação por relevância". Retornou um total de 27.747 documentos.

Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 24.245 documentos;

2. Portal Capes – sintaxe busca avançada: “*big data*” AND “*privacy issues*”; Termos: “é exato para os dois termos”; Qualquer campo: “título, autor, assunto”; Data Publicação: “últimos 5 anos”; Tipo de material: “artigos”; Idioma: “inglês”, Filtro: “ordenação por relevância”. Retornou um total de 19.803 documentos. Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 18.029 documentos;

3. Portal Capes – sintaxe busca avançada: “*big data analytics*” AND “*privacy issues*”; Termos: “é exato para os dois termos”; Qualquer campo: “título, autor, assunto”; Data Publicação: “últimos 5 anos”; Tipo de material: “artigos”; Idioma: “inglês”, Filtro: “ordenação por relevância”. Retornou um total de 4.888 documentos. Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 4.396 documentos;

4. Portal Capes – sintaxe busca avançada: “*big data analytics*” AND “*privacy problems*”; Termos: “é exato para os dois termos”; Qualquer campo: “título, autor, assunto”; Data Publicação: “últimos 5 anos”; Tipo de material: “artigos”; Idioma: “inglês”, Filtro: “ordenação por relevância”. Retornou um total de 3.940 documentos. Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 2.623 documentos;

5. Portal Capes – sintaxe busca avançada: “*big data analytics*” AND “p*rivacy risks*”; Termos: “é exato para os dois termos”; Qualquer campo: “título, autor, assunto”; Data Publicação: “últimos 5 anos”; Tipo de material: “artigos”; Idioma: “inglês”, Filtro: “ordenação por relevância”. Retornou um total de 2.245 documentos. Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 2.031 documentos;

6. Portal Capes – sintaxe busca avançada: “*big data analytics*” AND “*big data privacy*”; Termos: “é exato para os dois termos”; Qualquer campo: “título, autor, assunto”; Data Publicação: “últimos 2 anos”; Tipo de material: “artigos”;

Idioma: “inglês”, Filtro: “ordenação por relevância”. Retornou um total de 37 documentos. Com a aplicação do filtro “mostrar somente periódicos revisados por pares”, restaram 28 documentos.

Nas sintaxes criadas e pesquisadas, segundo os parâmetros estipulados até a quinta tentativa, obteve-se um retorno não só de um conjunto muito grande de materiais, mas o mais relevante e definidor, a análise desses materiais mostrou pouquíssima aderência à temática desse trabalho “privacidade no âmbito do *big data analytics*”.

Na sexta tentativa, foi possível encontrar a sintaxe ideal, que pôde ser confirmada através da análise dos materiais retornados pela busca. Por outro lado, verificou-se um conjunto restrito de materiais, em termos quantitativos, apesar da garantia de alinhamento com uma questão fundamental que estruturou todas as buscas, que foi a questão da “atualidade” dos materiais.

A pesquisa realizada por Sun, Pambel e Strang (2018, p. 11, tradução nossa), trouxe grande ajuda para a estrutura lógica e o refinamento das buscas realizadas sobre publicações cobrindo o campo do *big data*. Segundo os autores, “há mais artigos de conceito do tipo genérico totalmente teórico nos anais de conferências (3109 ou 42%) em comparação com periódicos (748 ou 5%)”. O que se verificou, como os próprios autores apontam é que “mais de um terço dos artigos de conferências são de natureza conceitual, e apenas 5% dos manuscritos de periódicos são puramente teóricos”. Uma outra constatação fundamental dos pesquisadores foi que “a maior parte dos resultados do corpo de conhecimentos sobre *big data* está em periódicos (2011-2016, N=13029)”.

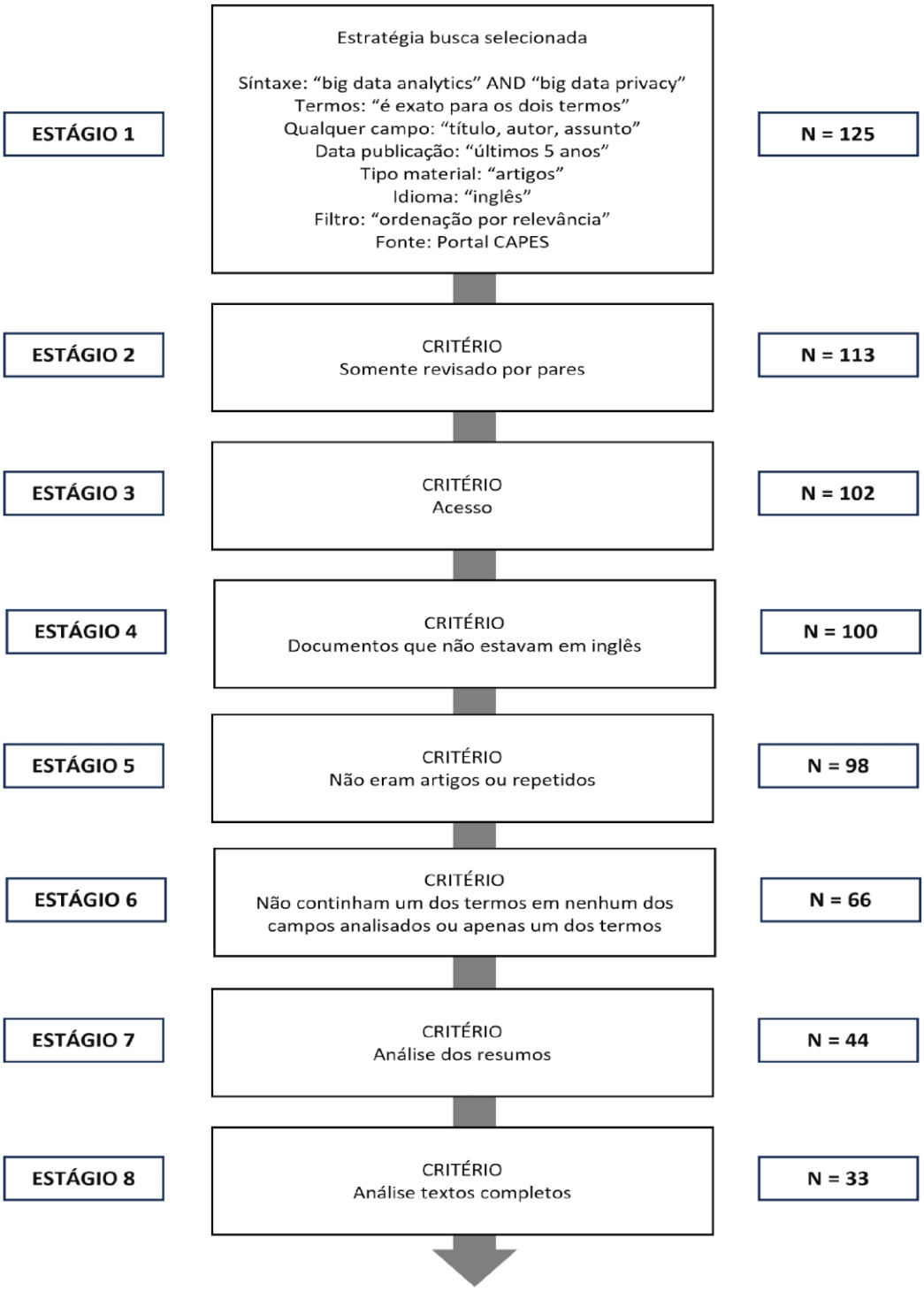
Sendo assim e partindo desse histórico e achados, no dia nove de julho de 2020, foi realizada uma nova busca, mantendo todos os elementos da sexta tentativa, alterando apenas o tempo de cobertura, ou seja, no lugar do parâmetro “data publicação: últimos 2 anos”, foi realizada a alteração para “Data Publicação: últimos 5 anos”.

Essa última busca retornou 125 documentos classificados como artigos. Mais da metade (55%) dos materiais recuperados foram originários de duas bases principais, Scopus (31%) e *Aerospace Database Technology Researc*h (25%). Numa proporção próxima, foram recuperados também no âmbito da base de documentos *Database*, 16%; sendo que a *Web Of Science* e a *Onelife* contribuíram na mesma magnitude, com 14% cada uma delas.

A partir desse ponto, os artigos foram analisados a partir dos seguintes critérios: a) condições de acesso aos artigos; b) confirmação de que todos os artigos estivessem em inglês; c) confirmação de que todos os documentos fossem efetivamente artigos; d) confirmação de não repetição dos documentos, visto que os mesmos poderiam estar publicados em mais de um repositório; e) análise da presença dos termos-chave de busca, principalmente do termo "*privacy*"; e) análise dos resumos; f) leitura completa dos artigos.

Na figura 1 abaixo, apresenta-se de forma esquemática a lógica de busca, análise e seleção dos materiais selecionados:

**FIGURA 1 - Lógica de busca, análise e seleção da literatura de interesse**



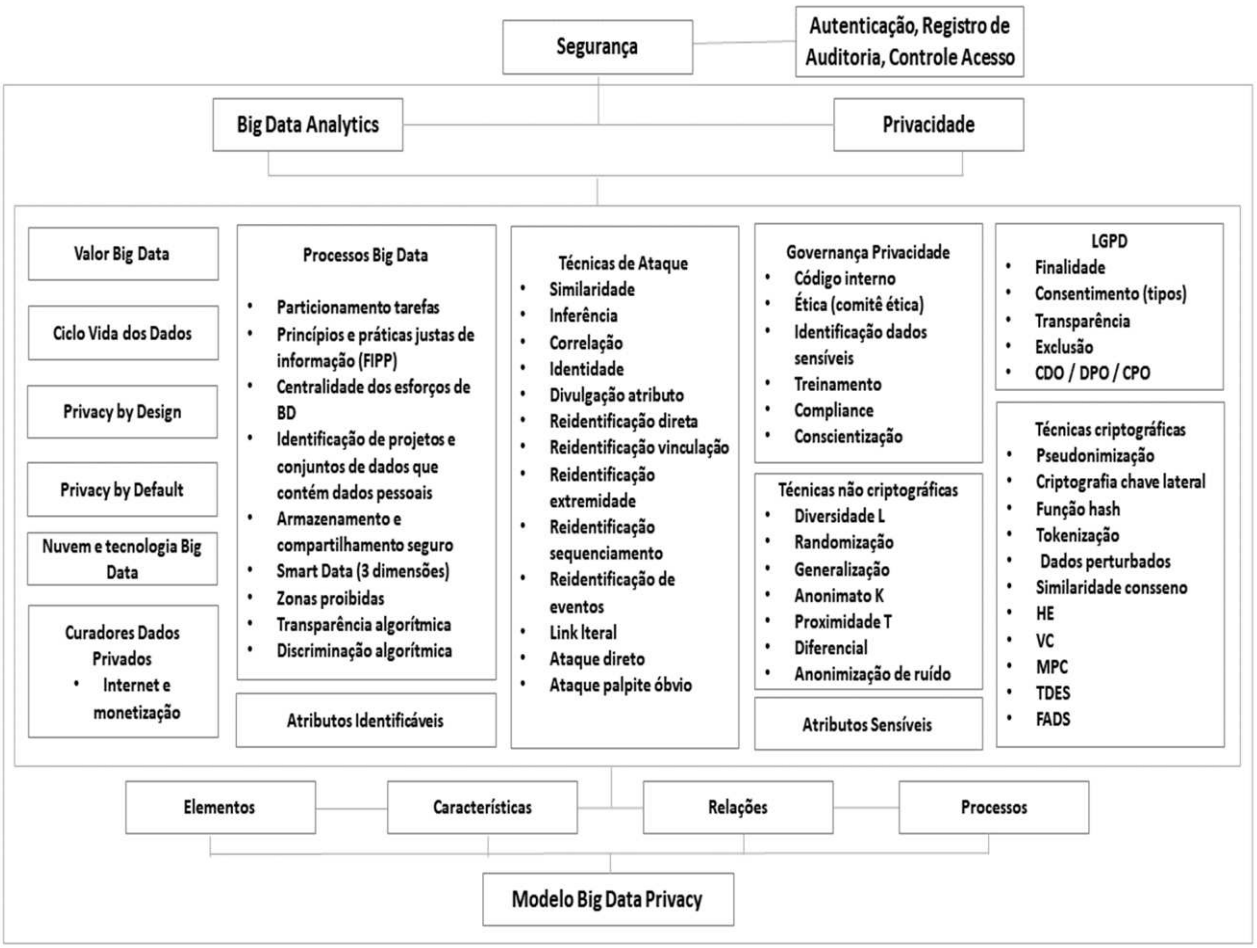

Fonte: Elaborado pelo autor, 2020

Ao final de todo o processo criterioso de análise dos documentos chegou-se a um conjunto de trinta e três artigos, revisados pelos pares, através dos quais se firmou a convicção, por toda argumentação exposta anteriormente, de que foi acessado um material aderente e alinhado à desafiadora proposta dessa pesquisa, assim como pelo fato de trazer uma consistente garantia da extensão e exaustividade que a análise desses materiais trouxe em termos da revisão da literatura que foi empreendida.

## 5 MAPA DA LITERATURA

A análise dos materiais selecionados, permitiu construir o seguinte mapa da literatura, que deu um claro direcionamento para a realização de uma consistente revisão da literatura.

**FIGURA 2 - Mapa da literatura**




Fonte: elaboração própria, 2021

## 6 REVISÃO DA LITERATURA

O trabalho de revisão da literatura cumpre diversos objetivos de grande relevância para todas as pesquisas, e de maneira específica para o caso da pesquisa em questão foram fundamentais para dar o direcionamento e consistência necessários, para dentre outros, estabelecer um corpo consistente de forma a mostrar a importância dessa pesquisa, além de possibilitar a estipulação de uma referência que possibilitou a comparação com outros estudos; proporcionando a possibilidade de diálogo com a literatura sobre o tema, preenchendo lacunas e ampliando e/ou aprofundando estudos anteriores; assim como possibilitou o compartilhamento dos achados de outros estudos já realizados sobre a temática diretamente e/ou sobre temáticas adjacentes que compõe uma visão ampliada do campo de pesquisa proposto. (Creswell, 2010)

Como definido por Noronha e Ferreira (2000, p. 182) a revisão da literatura pode ser entendida como “estudos que analisam a produção bibliográfica em determinada área temática, dentro de um recorte de tempo, fornecendo uma visão geral ou um relatório do estado-da-arte sobre um tópico específico”. E tais análises cumprem um conjunto de objetivos, como a evidenciação de novas ideias, técnicas, métodos, metodologias, apontamento de opiniões e visões em comum a diversos pesquisadores, dentre outros. As revisões de literatura, podem ser classificadas segundo “seu propósito, abrangência, função e tipo de análise desenvolvida”. (Noronha e Ferreira, 2000, p. 184)

Com base nessa classificação proposta por Noronha e Ferreira (2000), a revisão de literatura que ora se apresenta, pode ser classificada quanto ao seu propósito como de base, que se advoga como aquela que visa suportar a comprovação ou não de hipóteses e questões de pesquisa; quanto à sua abrangência ela mescla a abordagem temática com a temporal, por estar focada em um tema específico e dentro de um recorte temporal estabelecido (últimos cinco anos); quanto à sua função, ela se classifica como de avaliação, pelo seu foco no levantamento e

análise de estudos recentes, identificando informações, visões e resultados de outros trabalhos para o desenvolvimento de toda a lógica do modelo desta pesquisa; por fim, quanto ao tratamento e abordagem dada aos trabalhos analisados, ela se enquadra como bibliográfica, ou seja, por ter servido para comparação das abordagens realizadas no âmbito dos múltiplos trabalhos analisados, possibilitando a identificação daqueles de maior relevância e aderência ao tema proposto para essa pesquisa.

A criação de um modelo de privacidade no *big data analytics*, necessitou da identificação e análise das hipóteses e construtos já evidenciados em trabalhos anteriores, de forma a possibilitar uma base inicial e com respaldo em múltiplos trabalhos e autores de referência, formatando assim o corpo inicial da estrutura lógica do modelo.

Por outro lado, também possibilitou, através da análise crítica e dos apontamentos relatados em estudos anteriores, a identificação e proposição “lógica” de novas hipóteses e construtos, formatando assim, o desenho final do modelo atestado.

Bryman (1989, p. 17, tradução nossa), traz algumas reflexões e colocações sobre esse processo, apontando que “muitas vezes as hipóteses e seus conceitos associados são o produto de deliberações em conexão com a literatura relacionada a um campo substantivo”. Complementando o raciocínio e fazendo o vínculo direto à abordagem aqui realizada, afirma que “questões teóricas anteriores podem surgir como justificativas para a inclusão de variáveis específicas ou para os padrões descritos nas hipóteses”.

Como é uma ação frequente em pesquisas de cunho quantitativo, a revisão da literatura sobre o tema da pesquisa em epígrafe, foi realizada nessa seção, separada das demais seções.

Por ser uma temática ainda incipiente, a revisão de literatura, que envolve o tema *big data privacy* (*big data analytics* + privacidade), necessitou de uma estrutura que lhe fornecesse a robustez, abrangência e assertividade necessárias para suportar, não só um melhor entendimento do tema e suas características adjacentes,

como também das etapas posteriores dessa pesquisa, sendo assim, com base no mapa da literatura constituído, estruturou-se a revisão da literatura baseada nos seguintes macros temas: segurança, privacidade, *big data analytics*, e Lei Geral de Proteção de Dados (LGPD).

## 6.1 Segurança

A análise de grandes volumes de dados é um problema até então sem solução definitiva não só para a privacidade, mas também para a segurança. É praticamente impossível, hoje em dia, afirmar que um determinado sistema é totalmente seguro, principalmente se esse tiver interação com o ambiente da internet (a web superficial, que é a que a maioria das pessoas conhece e utiliza; a web profunda, a web opaca, a web privada, web proprietária, web invisível e a *dark web*, que compõe o que é conhecido como *deep web)*.

E à medida que novas tecnologias e aplicativos vão surgindo, a régua dos desafios da segurança vai também se alargando em proporção muito maior. “Para salvaguardar as informações pessoais, é crucial que os processos de armazenamento e transporte sejam integrados com medidas de segurança”. (Adams, 2017, p. 17, tradução nossa)

Segurança e privacidade são questões siamesas, ou seja, estão, em geral sempre juntas, apesar de que, ao mesmo tempo em que não existe privacidade sem uma ou mais camadas de segurança, a existência dessa última não consegue garantir em sua plenitude a privacidade. Tem-se ainda uma grande confusão entre os termos, que são utilizados em muitos casos como se fossem sinônimos.

As plataformas que possibilitaram o *big data analytics*, como o Hadoop, por exemplo, de maneira geral são desprovidas de mecanismos eficazes de segurança. Esse é um dos grandes problemas do *big data*, que tendem a refletir obviamente nas questões de privacidade. (Kapil *et al*, 2020)

Como apontado por Thomson e Thibadeau (2016, p. 2, tradução nossa), “a segurança é tão forte quanto seu elo mais fraco, e isso é essencialmente verdadeiro para o *big data*”. Os autores fazem essa afirmação baseados na arquitetura das soluções de *big data*, que tem a característica de serem distribuídas, o que apresenta muitos pontos vulneráveis nas fases de processamento e armazenamento.

Os mesmos autores dão um pouco mais de profundidade à afirmação anterior e apontam para o fato de que no processo de desenvolvimento das arquiteturas que suportam o *big data analytics*, em sua quase totalidade são soluções de código aberto (atualmente é possível verificar um aumento das soluções proprietárias, que também passam a ter integração com as soluções de código aberto), cujo foco da segurança foi estabelecido no chamado perímetro do sistema, “mas o que está por trás do perímetro não é seguro. Ataques internos são uma ameaça particular”. (Thomson e Thibadeau, 2016, p.2, tradução nossa)

Tendo como pano de fundo esse desafio de segurança no âmbito do *big data*, Thomson e Thibadeau (2016, p. 4, tradução nossa) propõe três princípios a serem seguidos:

1) A segurança deve ser incorporada à arquitetura e ao design (design por privacidade e design por padrão) dos sistemas de informação da organização e aos ativos de tecnologia da informação;
2) Uma organização deve empregar uma estratégia de defesa em profundidade para lidar com todas as vulnerabilidades das soluções no campo do *big data*;
3) Não se deve implementar soluções com vulnerabilidades já previamente conhecidas.

Jain, Gyanchandani e Khare (2019), colocam que os tradicionais mecanismos de segurança, podem ser agrupados em quatro instâncias: a) a nível dos arquivos; b) ao nível dos bancos de dados; c) ao nível de mídia e; d) ao nível dos aplicativos.

Segurança refere-se às políticas, procedimentos e medidas técnicas usadas para evitar acesso não autorizado, alternância, roubo de dados ou danos físicos a dispositivos e sistemas. (Sun, Pambel e Strang, 2018)

Segurança é a prática de “defesa de informações e ativos de informação através do uso de tecnologia, processos e treinamento contra: acesso não autorizado, divulgação, interrupção, modificação, inspeção, gravação, destruição”. (Jain, Gyanchandani e Khare, 2016, p. 3, tradução nossa)

Segurança é confidencialidade, integridade e disponibilidade de dados. É a proteção contra o acesso não autorizado, permanecendo, portanto, confiável, íntegra e precisa, estando acessível sempre que solicitada, mas protegida contra ataques e roubos com finalidade escusa. (Jain, Gyanchandani e Khare, 2019; Ouazzani Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018)

### 6.2 Privacidade

Essa subseção se iniciará com uma ponderação sobre uma importante e ampla reflexão trazida por Westin (1970, p. 34, tradução nossa), quando afirma que “o desenvolvimento da individualidade é particularmente importante nas sociedades democráticas, uma vez que qualidades de pensamento independente, diversidade de pontos de vista e não conformidade são consideradas características desejáveis [...]”.

Talvez a característica mais relevante do conceito de privacidade e que lhe confere contornos múltiplos, sob muitos olhares, seja possuir um conceito dinâmico, que pode sofrer variações / alterações, consoante com as tecnologias e normas sociais vigentes. (Hadar *et al*, 2017, tradução nossa)

Conforme colocado por Politou, Alepis e Paysakis (2018, p.2, tradução nossa) a privacidade foi inserida como um direito em “1890 por Waren e Brandeis”, sendo que somente nas “últimas três décadas que ela foi amplamente discutida em suas várias formas e contextos, principalmente devido à computação e ciência da informação”.

No final da década de 1960, o conceito de privacidade surgiu em um contexto mais filosófico, e como apontado por Politou, Alepis e Paysakis (2018, p. 2, tradução nossa), desde então, “é discutida em grande controvérsia entre os círculos filosóficos, jurídicos, sociais e científicos”.

E mesmo hoje, como apontam esses mesmos autores, “não existe uma definição universalmente aceita de privacidade”. Ela pode ser vista sob um caleidoscópio, ou seja, sob múltiplas visões e pontos de vista, como o controle sobre os dados de uma pessoa por ela própria, o direito de não ser monitorado e nem mesmo identificado, de ser deixado em “paz” e até mesmo ser “esquecido”. (Politou, Alepis e Paysakis, 2018, p. 2, tradução nossa)

Esses fatos de maneira distinta, mas não de forma exaustiva, conferem em diversos momentos uma certa confusão entre privacidade e outros conceitos, sendo uma das piores confusões aquela existente em relação ao conceito de segurança. Segurança e privacidade são termos confundidos com frequência, apesar de terem significados bem distintos.

Eles possuem uma relação simbiótica, onde a existência do primeiro (privacidade), depende da existência do segundo (segurança). Estão, portanto, em geral, quase sempre “juntos”, quando o contexto de privacidade é discutido. Ouazzani e Bakkali *et al* (2020, p. 143, tradução), também comungam dessa afirmação, quando apontam que “privacidade é frequentemente confundida com segurança”. A privacidade “consiste em dois pontos principais: confidencialidade e uso justo”. (Sun, Pambel e Strang, 2018, p. 4, tradução nossa).

A garantia de privacidade, ou a tentativa dessa garantia, se refere ainda à possibilidade de ocultar a verdadeira identidade de uma pessoa. Por sua vez a “segurança, lida com confidencialidade, integridade e disponibilidade”. Esse trabalho adotou essa como a definição de referência para privacidade, balizando, portanto, todas as referências à mesma dispostas nesta pesquisa. (Ouazzani e Bakkali, 2020, p. 143, tradução nossa)

Frequentemente a privacidade é definida como a capacidade de proteger informações confidenciais sobre as informações pessoalmente identificáveis, ou seja, dados que direta ou indiretamente possam identificar um indivíduo. (Abouelmehdi, Beni-Hessane e Khaloufi, 2018)

Esses mesmos autores construíram um quadro, muito sucinto e objetivo, onde apontam as principais diferenças entre privacidade e segurança, que está reproduzido no Quadro 1 abaixo.

**QUADRO 1 - Diferenças entre segurança e privacidade**

| Segurança | Privacidade |
|---|---|
| Segurança é a “confidencialidade, integridade e disponibilidade” de dados | Privacidade é o uso apropriado das informações do usuário |
| Várias técnicas como criptografia, firewall, etc. são usados para evitar o comprometimento dos dados de tecnologia ou vulnerabilidades na rede de uma organização | A organização não pode vender as informações de seus pacientes / usuários / clientes a um terceiro sem consentimento prévio do usuário |
| Pode fornecer confidencialidade ou proteger uma empresa ou agência | Preocupa-se com o direito do indivíduo de proteger suas informações de quaisquer outras partes |
| A segurança oferece a capacidade de ter certeza de que as decisões são respeitadas | Privacidade é a capacidade de decidir quais informações de um indivíduo vão e para onde vão |

Fonte: adaptado de Abouelmehdi, Beni-Hessane e Khaloufi, 2018, p. 5

Para a proteção da confidencialidade, tem-se diversas alternativas, como tecnologias e sistemas que aumentam a privacidade e “podem ser usados para permitir aos usuários criptografar e-mails, ocultar seu endereço IP para evitar rastreamento por servidor web, ocultar sua localização geográfica ao usar telefones celulares”, e até mesmo recorrer à utilização de “credenciais anônimas”, tornando “a consulta a banco de dados não rastreável” e permitindo a publicação de “documentos anonimamente”. (Sun, Pambel e Strang, 2018, p.4, tradução nossa)

Aprofundando o entendimento do que seja privacidade, pode se dizer que privacidade é o querer do indivíduo de ser deixado em paz, livres de qualquer interferência, vigilância de outros indivíduos, organizações ou sistemas. Pode também ser considerada como o uso adequado das informações sobre um indivíduo, de forma que pessoas ou organizações que venham a deter esses dados saibam utilizá-los, com um fim específico e somente esse fim, determinado no momento anterior até mesmo à coleta ou aquisição desses dados por meio dos corretores de dados ou outros fornecedores de dados. Por fim, também pode ser compreendida como a

determinação de um indivíduo de deliberar sobre quais dados seus poderiam ser compartilhados, ao mesmo tempo, em que possui algum controle sobre esses. Seria ainda a não divulgação de informações pessoais de forma pública. (Sun, Pambel e Strang, 2018; Rao, Krishna e Kumar, 2018; Jain, Gyanchandani e Khare, 2016; Westin, 1970)

Solove (2002, p. 1129, tradução nossa), apresenta uma alternativa para essa situação, ao afirmar que a "maioria dos teóricos tenta conceituar a privacidade isolando um ou mais aspectos essenciais e comuns". Nesse sentido propõe que a discussão e definição desse termo deveria ser discutida sob a ótica de seis temas conjuntamente: "1) o direito de ser deixado em paz; 2) acesso limitado a si mesmo; 3) sigilo; 4) controle de informações pessoais; 5) personalidade; 6) intimidade [...]" (Solove, 2002, p. 1094, tradução nossa).

A privacidade, também, tem foco no uso e na governança de dados pessoais dos indivíduos, no sentido da necessidade do estabelecimento de políticas e requisitos de autorização, garantindo, portanto, que as informações pessoais sejam coletadas, compartilhadas, processadas e analisadas de maneira adequada. (Abouelmehdi, Beni-Hessane e Khaloufi, 2018)

Não seria razoável finalizar essa seção, sem apontar outro termo também muito relacionado e mais ainda confundido e utilizado de maneira indistinta e intercambiável quando da referência à privacidade que é a questão do conceito de "proteção" e "proteção de dados", ambos com forte vínculo à questão da privacidade; mas que, na verdade, constituem duas noções distintas. (Politou, Alepis e Paysakis, 2018; Rhoen, 2015)

Com foco num maior entendimento dessa questão, Politou, Alepis e Paysakis (2018, p. 1, tradução nossa), definem a questão da privacidade como geralmente se referindo "à proteção do espaço pessoal de um indivíduo, enquanto a proteção de dados se refere a limitações ou condições no processamento de dados relativos a um indivíduo identificável".

O que é reforçado por Kapil *et al* (2020, p. 4, tradução nossa), quando destacam e detalham parte dessa questão, ao indicar que a “proteção de dados é um processo para proteger os dados em repouso e armazenamento e durante a transmissão com a ajuda da criptografia e mascaramento”.

Por fim, e com o intuito precípuo de apontar o emaranhado de conceitos superpostos que muitas vezes tornam o aprofundamento das discussões sobre temas mais específicos nesse rol um grande desafio, destacamos as colocações feitas por Hadar *et al* (2017, p. 260, tradução nossa), quando destacam que “o significado de privacidade informacional”, no contexto do entendimento europeu, é entendido como proteção de dados.

Na tentativa de introduzir e vincular essa seção à próxima, assim como colocar em evidência as motivações e pontos focais estruturantes dessa pesquisa, seria importante destacar as colocações feitas por Altman *et al* (2018, p. 42, tradução nossa), quando relatam que as abordagens tradicionais de privacidade “como controle individual, consentimento e transparência, falham em abordar adequadamente os problemas de discriminação” oriundos das análises baseadas em algoritmos no âmbito do *big data analytics*. A mesma realidade, segundo os autores, vale para “técnicas baseadas em modelos de privacidade formais, como privacidade diferencial”. Ou seja, é preciso galgar caminhos e construir alternativas que reforcem, ao mesmo tempo que vão além e consigam efetivamente dar garantias à privacidade, por meio de um modelo mais amplo e completo, que pense e proponha alternativas para todas as facetas da privacidade e do *big data analytics* e seu processo de criação de valor.

## 6.3 *Big Data* e *Big Data Analytics*

Como conceito, *big data* não é um conceito tão novo. Segundo Adams (2017, p. 13, tradução nossa), ele existe há pelo menos duas décadas, “desde que foi usado

por Cox e Ellsworth, em 1997”, quando esses autores se referiram ao termo no contexto de um conjunto imenso de dados científicos.

E justamente o que transformou todo o cenário da análise avançada de dados em nossos dias foi a motivação para esses pesquisadores se verem perplexos e sem muitas alternativas para armazenar, processar e analisar esse grande conjunto de dados. Ou seja, ainda não estavam disponíveis soluções como o armazenamento distribuído baseado em nuvem e as soluções de processamento também distribuído e paralelo, materializados em soluções como Hadoop, considerado a maior referência e o grande divisor de águas para o que se concebe hoje como *big data*. (Adams, 2017)

Como princípio, o *big data* descreve uma “ampla disponibilidade de dados em formato digital, com uma presença concomitante de mineração de dados e capacidade de geração de conhecimento em várias redes”. (Adams, 2017, p.14, tradução nossa)

Refere-se, portanto, de maneira mais direta à infinidade de dados e informações digitais, contínua e amplamente coletadas pelas mais diversas organizações e instituições, públicas, privadas ou do terceiro setor. Jain, Gyanchandani e Khare (2019, p. 2, tradução nossa), dão uma ideia da dimensão desse fluxo de dados, ao apontarem que “todos os dias quintilhões de bytes de dados são criados, ou seja, 90% dos dados do mundo hoje foram criados apenas nos últimos dois anos”.

Sun, Pambel e Strang (2018, p. 4, tradução nossa), trazem exemplos concretos de algumas das fontes do *big data* quando declaram que sua gigantesca quantidade de dados é gerada “a partir de vários instrumentos, bilhões de telefones, sistemas de pagamento, câmeras, sensores, transações na internet, e-mails, vídeos, fluxos de cliques, serviços de redes sociais” e diversas outras fontes, se tornando até mesmo difícil exaurir sua citação num rol imenso de fontes.

Sarkar (2017), faz a comparação entre as diferenças entre os dados tradicionais e os dados no âmbito do *big data*. Tal representação possibilita um aprofundamento e maior entendimento do fosso existente e das grandes mudanças

necessárias embarcadas no conceito do *big data*. O Quadro 2 abaixo, reproduz as comparações realizadas pelo autor.

**QUADRO 2 - Comparação entre dados tradicionais e *big data***

| Dados tradicionais | *Big Data* |
|---|---|
| São usualmente mensurados em GB (gigabytes) | Requer TB ou PB para grandes medições de dados |
| O crescimento dos dados tradicionais é medido por hora ou dia | Esse período não é fixado para *big data* |
| No formato tradicional os dados são considerados estruturados | *Big data* pode ser estruturado, não estruturado ou semiestruturado |
| A integração de dados no contexto de dados tradicionais é simples | É bastante difícil e demorado para *big data* |
| Em geral, RDBMS é usado para gerenciar dados tradicionais | Arquiteturas como o sistema de arquivos baseado em Hadoop com MapReduce, NoSQL (não apenas SQL), sistema de computação de alto desempenho (HPCS) são usadas para armazenar e analisar grandes conjuntos de dados de forma confiável |
| O acesso aos dados tradicionais é interativo | O sistema em lote ou quase em tempo real é necessário para gerenciar *big data* |

Fonte: Sarkar, 2017, p. 136, tradução nossa

Uma definição básica e mais comum de *big data* é a que o define como um conjunto de dados que são tão grandes e complexos que as aplicações tradicionais de processamento de dados não são suficientes, para armazená-los, processá-los e analisá-los. Na atualidade, o gerenciamento dos grandes fluxos de dados, por meio de ambientes físicos e virtuais, o processamento de *big data* se tornou um fator de criticidade para praticamente todas as organizações. E o mesmo tem apontado para uma nova realidade destacada não só pelo tamanho dos dados, mas também pela sua complexidade de organização e análise. (Sun, Pambel e Strang, 2018; Jain, Gyanchandani e Khare, 2016; Rani e Dhamadaran, 2016; Adams, 2017; Ouazzani e Bakkali, 2020)

Mas não se pode obter uma compreensão completa sobre a dimensão e complexidade do *big data*, sem se entender minimamente algumas de suas características mais relevantes, que lhe dão os contornos e a singularidade que possui. Nesse âmbito, o *big data* pode ser descrito pelas características ou dimensões

dadas por seus V's. Originalmente estavam caracterizadas por seu volume, variedade e velocidade. Mais tarde, novos estudos e reflexões apontaram para a insuficiência dessas três características como definidoras do *big data* e, portanto, passaram a considerar nesse conjunto novas características, como valor, veracidade e variabilidade. (Ristevski e Chen, 2018; Jain, Gyanchandani e Khare, 2016; Sun, Pambel e Strang, 2018; Sarkar, 2017; Wilson, Belliveaus e Gray, 2017; Chauhan, Agarwal e Kar, 2016; Ouazzani e Bakkali, 2020; Mehta e Rao, 2019)

Nesse contexto, volume se refere à abundância de dados, já citada anteriormente, ao passo que velocidade está vinculada ao fluxo e frequência de criação, armazenamento, processamento e análise desses dados. O valor, uma dimensão tão importante e impactante para a estrutura e balizamento da concepção e escopo dessa pesquisa, é dado pelo que a análise de *big data* gera enquanto resultado, que pode ser materializado como aumento de produtividade, redução de riscos, aumento de faturamento, testagem de novas drogas, redução de fraudes, dentre muitos outros.

Variedade, como uma das principais dimensões componentes e estruturantes do conceito de *big data*, dá conta da multiplicidade, heterogeneidade e complexidade dos dados em *big data*, no que tange a serem estruturados, semiestruturados ou não estruturados. Essas características e as soluções para fazerem frente a elas são estruturantes e definidoras do que se chama *big data*.

Variabilidade, diz respeito à robustez e consistência dos dados considerados ao longo do tempo; e por fim, uma característica também de grande relevância em *big data* que é a veracidade, vinculada às questões como relevância e qualidade dos dados. Como em todo processo analítico, sua consistência e capacidade preditiva está diretamente ligada à qualidade dos dados e a sua relevância (principalmente se são dados que guardam algum vínculo ou lógica de interferência no âmbito que se quer analisar), quando imputados no início do processo. Ou seja, de maneira bem simples, se não está garantida a veracidade dos dados, os resultados obtidos não se prestarão ao seu objetivo. (Ristevski e Chen, 2018; Rani e Dhamodaran, 2016)

Muitos pesquisadores apontam para a existência e relevância de outros V's, para uma caracterização completa das dimensões fundamentais do *big data*. Para essa pesquisa os V's explicitados e explicados anteriormente serão entendidos como completos e suficientes para esse fim.

A estrutura para essa nova realidade foi dada pela lógica que vem se consolidando, há pelo menos uma década, da própria sociedade, que tornou-se cada vez mais dependente da tecnologia e comunicação. (ADAMS, 2017, p.12)

Jain, Gyanchandani e Khare (2016, p. 1, tradução nossa), complementam esse conceito geral e aceito de *big data* ao pontuarem que o *big data* poder ser definido, "como uma geração de tecnologias e arquiteturas, projetadas para volumes muito grandes" de um amplo e variado espectro de dados (quantitativos, qualitativos, estruturados, semiestruturados e desestruturados), possibilitando elevada velocidade de coleta, processamento, análise e geração de outputs.

Oostveen (2016, p. 302, tradução nossa), coloca que num nível mais básico, o *big data*, "entra em conflito com a privacidade e proteção de dados". Isso porque, segundo os autores, "a coleta de dados na fase de aquisição pode revelar detalhes íntimos sobre a vida de uma pessoa", ou seja, infringe sua privacidade.

Mills (2018, p.598, tradução nossa), complementa e aprofunda as colocações feitas por Oostveen (2016), aos destacar que o *big data* "é usado para objetivos específicos e usa algoritmos e análises específicas que podem causar desconforto", assim como "servir a interesses particulares que podem ou não estar alinhados" com os interesses dos detentores dos dados analisados. É justamente nesse ponto em que muitas questões no âmbito da privacidade podem não ser levadas em conta.

Na esfera da privacidade, o *big data* "foi definido como dados sobre um indivíduo ou um grupo de indivíduos, que podem ser analisados para fazer inferências sobre esses indivíduos". (Wilson, Belliveau e Gray, 2017, p.3, tradução nossa)

No Quadro 3 abaixo, proposto por Sun, Pambel e Strang (2018), os autores apontam para os impactos do *big data* nos campos da privacidade e da segurança.

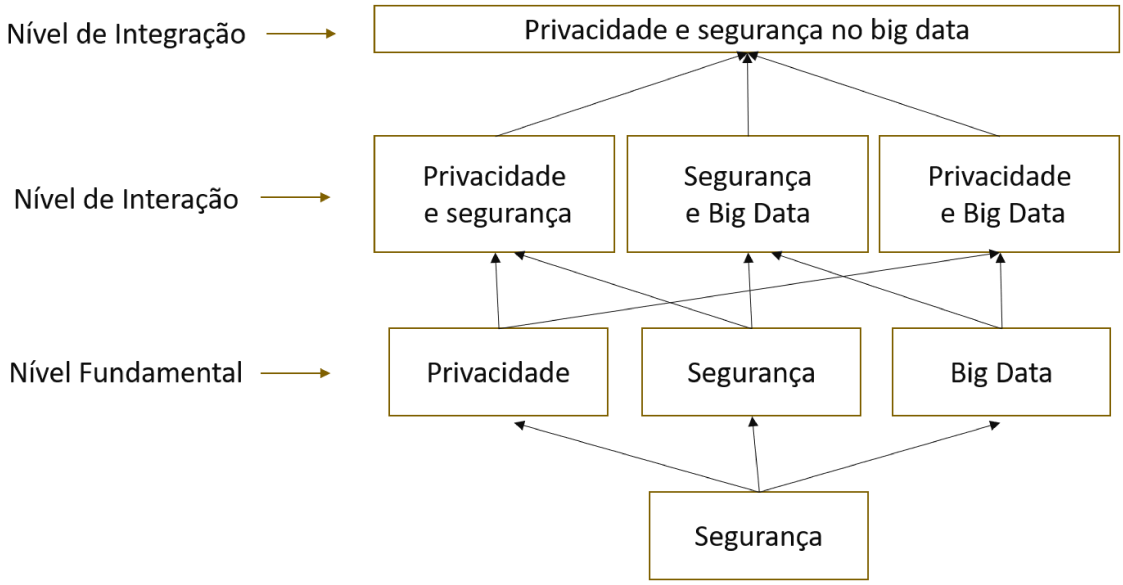
**QUADRO 3 - Características da privacidade e segurança no *big data***

| Características do *big data* | Privacidade | Segurança |
|---|---|---|
| Volume | Criar grande valor é poder! Os dados são dinheiro | Contribui para um grande número de cibercriminosos |
| Velocidade | Dados de localização em tempo real | Riscos de segurança física<br>Gerar perfil de clique e posição do indivíduo |
| Variedade | Não pode gerenciar de forma eficaz dados contendo informações sensíveis | Muitas empresas não deram segurança e protegeram adequadamente os dados não estruturados |
| Veracidade | Dados de variação de tempo de indivíduos são preocupações relacionadas à privacidade | Violação de segurança relacionada a um grande número de cartões de crédito |

Fonte: Sun, Pambel e Strang, 2018 p.6, tradução nossa

Sun, Pambel e Strang (2018), constituíram ainda um quadro analítico baseado na lógica do modelo booleano, onde de uma maneira muito clara, densa, ao mesmo tempo que sutil, constroem a visão geral das interligações existentes e consistentes entre *big data*, privacidade e segurança, reforçando o arcabouço no qual se baseia esse trabalho. A Figura 3 na sequência, retrata a lógica trazida pelos autores.

**FIGURA 3 - Modelo booleano para *big data*, privacidade e segurança**




Fonte: Sun, Sun e Strang, 2016, p.4, tradução nossa

Nesse âmbito, é irrefutável a conclusão colocada por Oostveen (2016, p. 302, tradução nossa), quando analisa o fenômeno do *big data*, ao afirmar que "as possibilidades do *big data* parecem infinitas". Um dos motivos que suportaram essa indicação do autor, deriva da capacidade do *big data* de construir respostas às questões que ainda não foram constituídas, ou seja, mesmo que ninguém tenha feito uma determinada pergunta, o *big data analytics* é capaz de construir respostas e antecipá-las.

Gupta e Rani (2019, p.322, tradução nossa), destacam que os dados cada vez maiores "oferecem uma onda gigantesca de oportunidades e desafios em termos de captura, armazenamento, manipulação, gerenciamento, análise, extração de conhecimento, segurança, privacidade e visualização de dados".

*Big data analytics*, portanto, também pode ser compreendido como uma versão mais poderosa de uma antiga e renomada abordagem no campo da mineração de dados originalmente conhecida como "descoberta de conhecimento em bases de

dados” (KDD - *knowledge data descovery*), definida, como colocado por Pragash e Jayabharathy (2017, p. 95, tradução nossa), “como a extração não trivial de informações implícitas, previamente desconhecidas e potencialmente úteis dos dados”.

Sob essa ótica, sua abordagem possibilita a descoberta de conhecimento a partir de um conjunto considerável de dados, que cobrem desde a “integração de dados heterogêneos, até o controle de dados, análise, modelagem, interpretação e validação” dos mesmos. (Ristevski e Chen, p. 1, tradução nossa)

Shanmugapriya e Kavitha (2019, p.1, tradução nossa), veem o *big data* como uma “filosofia de exame de dados potencializado por outras tecnologias e arquitetura que suportam captura, armazenamento e investigação de dados de alta velocidade”.

Assim como em diversas outras inovações, tecnologias e processos que vêm causando impacto na nossa sociedade, o *big data analytics* possui claramente um ponto de alerta para não ultrapassar determinados limites e passar a causar problemas graves para a sociedade; mas também como apontado por Salas e Domingo-Ferrer (2018, p. 270, tradução nossa), o *big data* “representa uma grande oportunidade para aprimorar nosso conhecimento como sociedade e como indivíduos”.

Nesse sentido, ele possibilita melhores decisões, ampliando o campo e a assertividade da análise das questões envolvidas num determinado processo decisório, ao mesmo tempo, portanto, que reduz os níveis de risco sempre embarcados em qualquer processo decisório.

Sendo assim, com base em melhores decisões, é possível gerar impactos em diversas esferas, como melhor utilização dos recursos, redução de custos, redução e/ou eliminação de fraudes, maior eficiência dos processos, maior impacto positivo das decisões, mais velocidade e maior amplitude, dentre outros. (Oostveen, 2016)

Sun, Pambel e Strang (2018, p.4, tradução nossa), reforçam a materialidade dessa ideia ao apontar que o *big data* tornou-se um “ativo estratégico para explorar percepções de negócios e economia de serviços”.

Para o fechamento desta seção é fundamental construir um vínculo mais consistente entre *big data* e *big data analytics*. Apesar de serem termos difíceis de se "separar", pois se não há a análise dos grandes dados, na verdade, não se têm nada, apenas grandes dados. Não faz nenhum sentido (em termos de estratégia, financeiro, processos, etc.) coletar, tratar e armazenar grandes quantidades de dados (*big data*), se esses dados não são analisados. Por isso mesmo, não é possível, sob uma condição lógica, ter se *"big data"* se não se tem junto o *"analytics".*

De qualquer forma, no contexto desse trabalho de pesquisa considerou-se de grande pertinência tais definições para dirimir possíveis dúvidas e confusões criadas pelo uso alternado e em contextos diferentes dos termos em muitos momentos.

Sun, Sun e Strang (2016, p. 2, tradução nossa), apresentam uma consistente definição para a análise de *big data* (*big data analytics*), onde sustentam que essa poderia ser definida como "o processo de coleta, organização e análise de *big data* para descobrir, visualizar e exibir padrões, conhecimento e inteligência".

Esses mesmos autores também destacam o fato de a análise de *big data* ser uma abordagem recente e multidisciplinar, envolvendo "tecnologia da informação e comunicação (TIC), matemática, pesquisa operacional (OR) e aprendizado de máquina (ML)". Seu contexto é constituído por diferentes abordagens analíticas, quais sejam: descritiva, preditiva e prescritiva. (Sun, Sun e Strang, 2016, p. 2, tradução nossa)

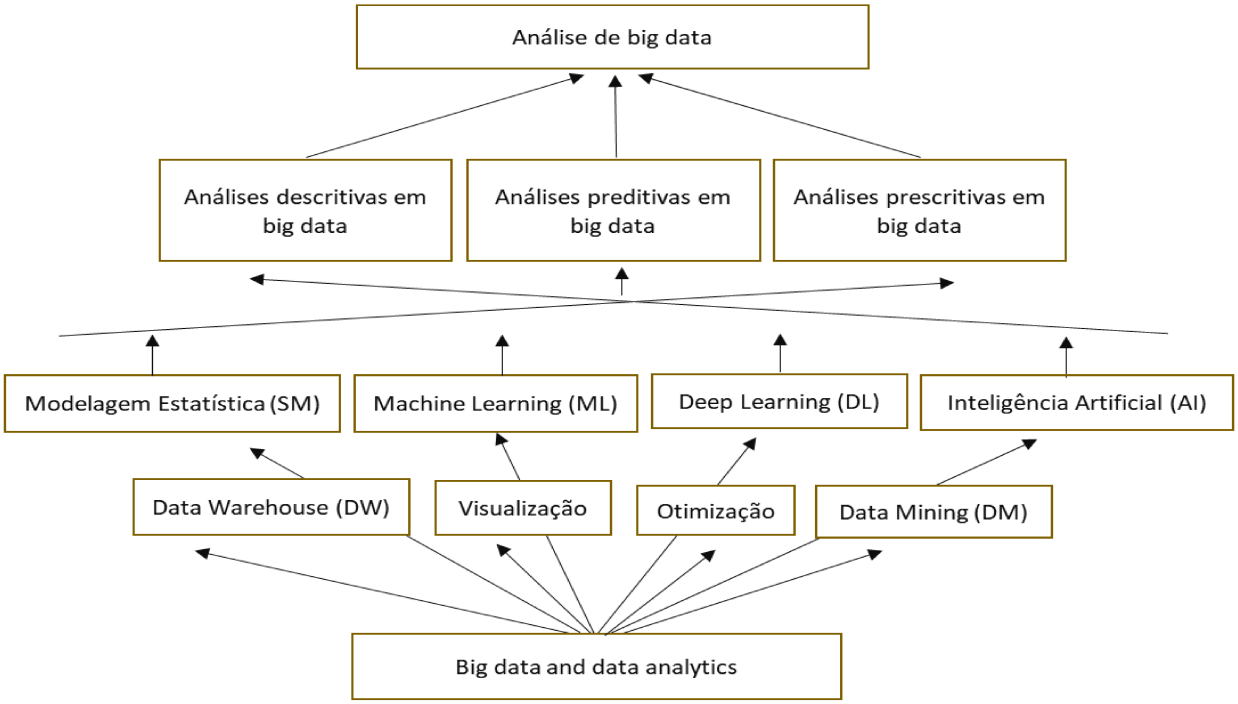
Pramanika *et al* (2020, p. 8, tradução nossa), definem *big data analytics*, numa análise dos seus impactos sobre a saúde, mas que poderiam ser ampliados para qualquer outro segmento ou campo de estudo, como se referindo às tecnologias avançadas "envolvidas na análise de conjuntos de dados heterogêneos em grande escala, mineração de *big data* e análise estatística".

E complementam sua explanação dando contornos mais vívidos ao conceito declarado, adicionando que "em ambientes de *big data*, os conjuntos de dados são divididos e executados em vários nós em paralelo". São, como já discutidas aqui, as novas formas de processamento que tornaram possível realizar o armazenamento, o

processamento e a análise em *big data*, sendo o grande símbolo dessas novas soluções a plataforma Hadoop. (Pramanika *et al*, 2020, p. 8, tradução nossa)

De forma a explicitar os conceitos e as relações no campo do *big data analytics*, Sun, Sun e Strang (2016, p. 4, tradução nossa), propõem uma ontologia, que é justamente a "definição de conceitos e suas inter-relações que existem para um determinado domínio particular do discurso", no caso para o *big data analytics*, conforme retratado pela Figura 4 abaixo.

**FIGURA 4 - Uma ontologia para *big data analytics***




Fonte: Adaptado de Sun, Sun e Strang, 2016, p.4, tradução nossa

Tal proposta é complementada e praticamente exaurida, pelo menos num contexto lógico e geral, por Pramanika *et al* (2020, p. 10, tradução nossa), quando explicitam e relacionam as características principais, pesquisas emergentes, tecnologias, ferramentas e sistemas no âmbito do *big data analytics*:

- Características principais / desafios de conjuntos de dados: volume, velocidade, variedade, veracidade e valor;
- Pesquisas emergentes: computação em nuvem; análises de redes sociais; modelos estatísticos preditivos; análise de privacidade e segurança; sistemas de saúde em tempo real; aprendizagem de máquina, redes neurais e mineração de textos; Hadoop, MapReduce baseados em análise de dados em saúde; saúde inteligente; análise de sentimento;
- Tecnologias: computação distribuída; clusterização, classificação, segmentação e integração de dados; algoritmos genéticos; detecção de anomalias; mineração de regras de associação; otimização;
- Ferramentas e sistemas: hadoop; MapReduce; Cassandra; Zookeeper; Mahout.

## 6.4 Legislação de Proteção de Dados (LGPD)

Em diversos países ao redor do mundo o debate sobre a questão da privacidade e da proteção de dados está muito ativo, isto porque apesar das legislações que passaram a vigorar nos últimos anos e mesmo algumas leis que já vigoram há bastante tempo, o que os países e suas lideranças têm visto é que a legislação não tem sido suficiente para fazer frente aos avanços do *big data analytics*.

Esses instrumentos são imensamente relevantes, mas com certeza deverão ser continuamente revistos e melhorados. E mesmo essa revisão contínua não será suficiente para equacionar e/ou antecipar e prever todas as questões dada a velocidade do *big data analytics*.

Assim como no caso da LGPD, Ishii (2017, p. 572, tradução nossa), aponta que embora a legislação da União Europeia, denominada pela sigla GPDR (*General Data Protection Regulation*) tenha o “potencial de melhorar a proteção de dados”, ainda carece de “mais trabalho para formular diretrizes normativas e mecanismos práticos para colocar em prática os novos direitos e responsabilidades”. Talvez aqui esteja a

semente inicial que motivou a busca pela compreensão desse fenômeno e por conseguinte os estudos que motivaram a proposição desta pesquisa.

No Brasil a LGPD (Lei Geral de Proteção de Dados), Lei 13.709/18, foi sancionada em 14 de agosto de 2018, após vários anos de discussão. A lei de proteção de dados brasileira se inspirou de forma bem ampla no Regulamento Geral de Proteção de Dados (GPDR em inglês) da União Europeia.

A LGPD se aplica a qualquer pessoa física ou jurídica (pública ou privada), que desenvolva projetos, iniciativas, ações que envolvam o tratamento de dados pessoais. Basicamente, ela estabelece regras no que diz respeito à coleta, armazenamento e compartilhamento de dados pessoais.

A ênfase nos dados pessoais é relevante, porque todos os demais dados, por exemplo, referentes às pessoas jurídicas, não estão abarcados por essa legislação. Politou, Alepis e Paysakis (2018, p.3, tradução nossa), de maneira simples e direta, definem dados pessoais como “aqueles referentes a um indivíduo”. Numa visão mais ampla, “dados pessoais, significa qualquer informação relativa a uma pessoa física identificada ou identificável (titular dos dados); uma pessoa singular identificável é aquela que pode ser identificada, direta ou indiretamente”. (GPDR, 2016, artigo 4, p. 3, tradução nossa).

Tem relevância no âmbito dos dados pessoais, entendendo-se os conceitos e diferenças existentes entre dados sensíveis, dados diretamente identificáveis e dados indiretamente identificáveis. Segundo a LGPD (2018, artigo 4º), denomina-se dado pessoal sensível como sendo aqueles vinculados a uma pessoa natural e aos aspectos raciais, étnicos, “convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico”.

Na visão de Oostveen (2016, p. 305, tradução nossa), dados diretamente identificáveis são aqueles onde a “informação imediatamente aponta um indivíduo específico”, já aqueles indiretamente identificáveis se dão “quando uma pessoa é destacada por meio de uma combinação única de dados”.

A LGPD, como não poderia deixar de ser, vai na mesma direção (como foi dito, até mesmo por ser inspirada na legislação europeia), no seu artigo 5°, e dá o seguinte entendimento para dados pessoais, “informação relacionada a pessoa natural identificada ou identificável”.

Uma das principais motivações que balizaram a discussão e construção dos atuais arcabouços legais sobre privacidade e proteção de dados, estiveram baseadas no fato de que a maioria das pessoas ainda não fazem a mínima ideia dos procedimentos e processamentos aos quais seus dados são submetidos. (Politou, Alepis e Paysakis, 2018)

Como já apontado, a questão da privacidade não é um tema de debate recente, mas sua concepção e garantia, em tempos de *big data analytics*, é com certeza tema recente de pesquisas. Por conseguinte, tornou-se uma questão de preocupação mundial. E nesse interim é possível perceber esses dois momentos, ou seja, temos diversas legislações que incidem sobre a questão da privacidade e são mais antigas e diversas, outras que surgiram ou foram modificadas a partir da realidade do *big data.*

Conforme destacado por Sun, Pambel e Strang (2018, p. 6, tradução nossa), “quase todos os países têm suas próprias leis que regem a privacidade e a segurança para proteger a privacidade dos indivíduos e proteger seus dados privados”. Abaixo são relacionados alguns exemplos dessas legislações, relacionadas por Ouazzani e Bakkali (2020) e Sun, Pambel e Strang (2018):

- Austrália: *Privacy Act*, 1988;
- União Europeia, 1995;
- Canadá: PIPEDA, 2000;
- EUA: Health Information Act: HIPAA, 1996;
- EUA: Proteção de dados das crianças (COPPA, 1998);
- EUA: Finanças (Gramm-Leach-Bliley, 1999);
- EUA: Fair Information Practices Principles (FIPP, 1998);
- CDE: *Privacy Principles*, 2010;

- UE: Registro Geral de Proteção de Dados (GPDR, 2016);

No âmbito da lei brasileira de proteção de dados, em seu artigo 6º, estão listadas as atividades de tratamento de dados pessoais que deverão ser observadas e congregam os pontos estratégicos sobre o qual sua estrutura está assentada, são eles:

> - Finalidade: realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;
> - Adequação: compatibilidade do tratamento com as finalidades informadas ao titular, conforme o contexto do tratamento;
> - Necessidade: limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;
> - Livre acesso: garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;
> - Qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;
> - Transparência: garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;
> - Segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;
> - Prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;
> - Não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;
> - Responsabilização e prestação de contas: demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

Esses são os pontos fundamentais sobre os quais uma boa parte da modelagem de garantia da privacidade deve estar alicerçada, mas como já foi demonstrado, são questões mais relevantes, mas não suficientes.

# 7 CIÊNCIA DA INFORMAÇÃO, *BIG DATA ANALYTICS* E PRIVACIDADE

Nesse ponto, parte-se de uma das poucas questões de consenso quando nos referimos à Ciência da Informação, que é seu caráter multidisciplinar. (Saracevic, 1999; Lenzi e Brambila, 2006; Chang e Huang, 2012; Xiao, Zhang e Li, 2015; Chang, 2018; Hjørland, 2018; Seadle e Havelka, 2023)

Saracevic (1999), traz um forte e contundente argumento que escancara essa característica nativa de interdisciplinaridade da Ciência da Informação, ao afirmar que os problemas discutidos neste campo não podem ser equacionados com abordagens e construções de qualquer campo isoladamente, por esse motivo, a interdisciplinaridade seria uma característica inerente ao campo da Ciência da Informação.

De maneira rasa e objetiva, a Ciência da Informação estuda e analisa de forma sistemática “um conceito chamado informação”, que se encontra difundido na base e no contexto de todas as dimensões da sociedade, o que muitas vezes faz parecer que ela desparece “nas sombras de outras disciplinas”. A Ciência da Informação é “um campo complexo de raízes”, onde muitas vezes a própria definição do seu campo de atuação parece não existir, ou ter contornos indeléveis difíceis de definir e delimitar. (Seadle e Havelka, 2023, p.1, tradução nossa)

As características da sociedade em que vivemos, ou seja, uma sociedade onde a informação e o conhecimento são o principal recurso de valor, tornam natural que múltiplos campos de estudos, projetos, as esferas do comércio, da política e as demais esferas e aspectos da vida que compõe nossa sociedade, de alguma forma tenham de lidar com a informação e o conhecimento. (Saracevic, 1999)

Seadle e Havelka (2023, p. 2, tradução nossa), apontam que a informação “é fundamentalmente parte da forma como os humanos compreendem o seu mundo”. Os autores complementam e reforçam essa ideia ao destacar que “nenhum campo pode existir inteiramente fora da ideia de informação”. Esses mesmos autores concluem que “a informação está em toda parte e é parte integrante de todas as

disciplinas acadêmicas e formas de pesquisa". (Seadle e Havelka, 2023, p. 4, tradução nossa)

Saracevic (1999), destaca três características inerentes à evolução e existência da Ciência da Informação, compartilhadas com diversos campos de estudo:

> 1) Ela é interdisciplinar por natureza, e essa característica está em evolução;
> 2) Ela está intrinsecamente ligada à tecnologia da informação, e essa ligação pode impor limites ou potencializar sua evolução;
> 3) Ela é juntamente com outros campos de estudo um elo relevante na evolução da Sociedade da Informação.

De maneira específica quanto ao item 2 postulado por Saracevic (1999), Seadle e Havelka (2023, p.4, tradução nossa), apontam para uma condicionante de extrema relevância para as Escolas de Ciência da informação, ao defenderem que "habilidades genuínas de computação são hoje uma das muitas ferramentas do currículo moderno da ciência da informação".

Sob a ótica do conjunto de características trazidas por Saracevic (1999), Zains (2007, p.335, tradução nossa) parece trazer uma espécie de resultado ou consequência dessas características, quando advoga que o campo da Ciência da Informação, "está em constante mudança. Portanto, os cientistas da informação são obrigados a revisar regularmente, e se necessário, redefinir seus princípios fundamentais".

Para Maa e Marchioninld (2023, p. 1, tradução nossa), novas técnicas e soluções foram desenvolvidas para análise de resultados no campo da matemática e da estatística, da mesma forma no campo da computação, com a chegada de novos algoritmos, hardwares e softwares. Assim também, no campo da Ciência da Informação, novas técnicas e práticas foram desenvolvidas para "monitorar, gerenciar e preservar resultados de processos de máquina que ficam ao lado dos resultados tradicionais de processos humanos".

Expandindo a discussão na direção de quais são as principais atividades que compõem a abordagem em Ciência da Informação, Zhang, Wolfram e Ma (2023),

propõem quatro principais atividades: seleção, organização, recuperação, disseminação e uso da informação.

Dando sequência, complementando e aprofundando o raciocínio exposto por Zhang, Wolfram e Ma (2023); Zains (2007, p. 335, tradução nossa), destaca três conceitos inerentes ao campo da Ciência da Informação, quais, sejam, “dados, informação e conhecimento que estão incorporados no conceito de ciência da informação e estão inter-relacionados”. Portanto, fica claro que os fenômenos envolvendo dados, informações e conhecimento por si só integram o campo de estudo abarcado pela Ciência da Informação.

Como destacado por Seadle e Havelka (2023, p.1, tradução nossa), os dados “são os eunucos do mundo da informação, porque os dados e os conjuntos de dados são fundamentalmente estéreis e sem sentido, sem referência no ambiente intelectual, espacial, temporal e social que represente o seu contexto”.

Esse trabalho de pesquisa trouxe vários conceitos e frentes diretas, perpendiculares e cruzadas, que integram não só a essência, mas também trazem uma perspectiva da evolução da Ciência da Informação. Esse conjunto teórico e analítico prático está baseado em pelo menos quatro vertentes sobre as quais essa pesquisa esteve estruturada:

1) Reforçando, talvez, a única afirmação de conceito consolidado no campo da Ciência da Informação, essa pesquisa está no âmago da interdisciplinaridade. Como colocado por Zins (2007, p. 338, tradução nossa), “a ciência da informação é o mutável e disciplina transitória na confluência da biblioteconomia, documentação, mídia, comunicações, computação e filosofia aplicada”;
2) Utiliza e se apoia em ferramental no campo da tecnologia da informação e da estatística, como soluções fundamentais para o enfrentamento à complexidade e volume informacional trazidos pela Sociedade da Informação, em específico a amplitude tomada pelo digital. Como apontado por Zhang, Wolfram e Ma (2023, p.1, tradução nossa), “um

grande conjunto de dados não é autoexplicativo e necessita de metodologias adequadas de análise e processamento";

3) Teve como foco dois importantes fenômenos hodiernos, a questão da privacidade e a abordagem trazida pelo *big data analytics*. Com o crescimento exponencial da quantidade de dados e informações na Sociedade da Informação, não o *big data*, pois grandes dados por si só, como colocado por Seadle e Havelka (2023), são estéreis e sem sentido; mas o *big data analytics* chega como a solução para apoio à compreensão e análise de uma gigantesca quantidade de dados e, portanto, passa a ser um instrumento fundamental e estratégico para a evolução da Ciência da Informação. Zhang, Wolfram e Ma (2023, p. 6, tradução nossa), reforçam essa ideia ao defenderem que "métodos de pesquisa e técnicas de análise de dados estabelecidos e emergentes na ciência da informação têm muito potencial para a análise de *big data*, em múltiplas áreas do campo". Esses mesmos autores advogam, como um fechamento dessas provocações, que "a tendência emergente de *big data* impactou estudos de pesquisa empírica baseados em dados na ciência da informação". (Zhang, Wolfram e Ma, 2023, p. 1, tradução nossa)
4) Na essência o foco esteve em dados e informações relativas a pessoas e essa relação e possíveis utilizações desses dados e informações estão vinculados ao campo da privacidade. Como colocado por Zhang, Wolfram e Ma (2023, p. 6, tradução nossa), "na ciência da informação, muitos dos dados que estudamos têm um elemento direto ou indireto que pode levantar questões de privacidade". Campo esse ainda incipiente no contexto da Ciência da Informação, mas que necessariamente deverá ser alargado e aprofundado dadas as demandas da nossa sociedade.

Reforçando as colocações do item 4, acima, foram discutidos os resultados dos trabalhos desenvolvidos por Grisoto, Sant'ana e Segundo (2015) e Lott e Cianconi (2018). Esses autores realizaram pesquisas sobre as questões de privacidade e *big data* no âmbito da produção científica em Ciência da Informação.

Grisoto, Sant'ana e Segundo (2015), realizaram levantamento com foco específico na identificação da ocorrência do tema privacidade no âmbito das dissertações e teses defendidas no Programa de Pós-Graduação em Ciência da Informação (PPGCI) da UNESP de Marília. Foram recuperadas 133 dissertações, para o período de 2001 a 2014. No caso das teses, foram recuperadas 65, no período de 2007 a 2014.

Para o caso das dissertações, após as tratativas realizadas pelos autores, apenas quatro apresentaram o termo privacidade de forma considerada relevante para os objetivos da pesquisa. Esses trabalhos trataram das seguintes temáticas:

1) Discussão da privacidade com o uso da internet e questões autorais de 2002;
2) Privacidade na ética profissional, de 2006
3) Privacidade sobre os direitos de acesso à informação em 2006;
4) Privacidade da divulgação de informações de prontuários de pacientes de 2014;

No âmbito da análise das teses, apenas dois trabalhos foram considerados alinhados com os objetivos da pesquisa, que tratavam das seguintes temáticas:

1) Privacidade no acesso à informação de e-mail e no uso da internet, de 2008;
2) Privacidade de usuários em ambientes informacionais digitais, de 2010.

Por sua vez, Lott e Cianconi (2018), expandiram o universo de análise; nesse caso os autores objetivaram identificar como a Ciência da Informação, no Brasil, vem

tratando de temas como vigilância, privacidade, *big data* e dados pessoais em suas pesquisas. A base de pesquisa utilizada foram os Anais do Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB. De maneira mais específica foram feitas consultas a três fontes de informação: a) repositório BENANCIB, com pesquisas publicadas de 1994 a 2014; b) Anais do XVI ENANCIB; c) Anais do XVII ENANCIB.

De forma específica quanto ao termo “privacidade” os resultados das buscas empreendidas foram:

- Base BENANCIB: cinco trabalhos;
- Anais do ENANCIB XVI e XVII: um trabalho.

Esses resultados corroboram as colocações realizadas nessa pesquisa quanto à privacidade ainda ser uma temática de estudo incipiente no campo da Ciência da Informação, cenário esse que deve se alterar tendo em vista a recente legislação colocada em vigor pela Lei Geral de Proteção de Dados (LGPD), assim como os avanços no campo da analítica avançada trazidas pelas várias frentes do *big data analytics* e o aumento das preocupações dos indivíduos quanto a manipulação e utilização de suas informações para múltiplos fins, muitas vezes, inclusive, de forma prejudicial a seus interesses.

## 8 METODOLOGIA

A palavra método, do qual deriva a palavra metodologia, tem origem no grego "*methodos*", composta pelo termo "*meta*" (por meio de, através de) e pelo termo "*hodos*" (caminho); e deve ser interpretada como um caminho para se chegar a determinado objetivo ou resultado. Ou seja, o método dá a ordenação e cadência necessária para se alcançar com maior assertividade um determinado objetivo. (Dicionário Etimológico)

Segundo Vergara (2005, p.9), pode-se entender a metodologia como "método de intervenção do pesquisador, sua atividade mental consciente para realizar o papel cognitivo da teoria". A mesma autora assevera que "o método aproxima o investigador do fenômeno estudado". Adicionalmente, Vergara (2005, p. 10) aponta que o método se ancora em "regras e procedimentos que operacionalizam a posição epistemológica do pesquisador".

Conforme destacado por Creswell (2010, p. 35), o pesquisador "não apenas seleciona um estudo" quantitativo, qualitativo ou misto, como também "decide sobre um tipo de estudo dentro destas três escolhas". Nesse ínterim quanto a estratégia de investigação, esse projeto de pesquisa se utilizará, como já pontuado anteriormente, da estratégia mista, quantitativa e qualitativa.

Ao propor uma estrutura para projetos de pesquisa, Creswell (2010, p. 28) une três perspectivas como condição para tal: concepções filosóficas; estratégias de investigação selecionadas e métodos de pesquisa, conforme demonstrado pela Figura 5 abaixo.

**FIGURA 5 - Uma estrutura para o projeto**


Fonte: Elaborado pelo autor, a partir de Creswell, 2010, p.28

Com base na lógica proposta por Creswell (2010, p. 34), essa pesquisa teve por concepção filosófica (conjunto de crenças lógicas que guiam a ação), a concepção pragmática, que surge mais vinculada às ações, situações e consequências, do que derivada das condições antecedentes. Nesse caso, no lugar de se concentrar nos métodos, o pesquisador dá ênfase ao problema de pesquisa constituído e utiliza de todos os métodos, técnicas e abordagens disponíveis com vistas ao entendimento desse problema.

A metodologia proposta para a pesquisa desta tese foi por concepção conclusiva. Segundo Malhotra (2006), essa classificação conceitual possui algumas características essenciais como, por exemplo: o processo de pesquisa é formal e estruturado e, está estruturado com base na análise quantitativa a partir de uma amostra representativa e com a amplitude necessária.

Por conseguinte, pode ser classificada ainda como causal, onde esteve presente a manipulação de uma ou mais variáveis independentes, objetivando a determinação de relações de causa e efeito (Malhotra, 2006).

Segundo Aaker *et al* (2001, p. 96), a pesquisa causal é apropriada quando “é necessário mostrar que uma variável causa ou determina o valor de outras variáveis”. De maneira geral os requisitos para comprovação da relação de causalidade são mais exigentes, o que por sua vez, demanda a seleção de variáveis mais específicas (Malhotra, 2006; Aaker *et al*, 2001).

Para se alcançar o objetivo proposto para essa pesquisa, que foi a proposição de um modelo multifacetado de garantia da privacidade e do valor do *big data analytics*, dentro do contexto das maiores empresas do Brasil, o método utilizado foi o método misto.

O principal fator para a escolha da estratégia do método misto, esteve baseada na necessidade de entendimento dos elementos necessários e suficientes, e suas características e relações, para que fosse possível obter uma construção robusta de uma proposta de modelo, a partir de uma abordagem quantitativa inicial, para na sequência realizar uma análise qualitativa de elementos selecionados para que os resultados alcançados na primeira etapa pudessem ser compreendidos de forma mais abrangente e aprofundada, obtendo-se, assim, o melhor resultado possível. Havendo a possibilidade, inclusive, de ajustes ao modelo obtido como resultados da primeira fase.

Para atendimento ao raciocínio exposto acima, este trabalho de pesquisa se utilizará da denominada “estratégia transformativa sequencial”. Segundo Creswell (2010, p. 248), tal estratégia está baseada em duas fases, com um olhar teórico se sobrepondo “aos procedimentos sequenciais”. É composta por uma primeira fase, quantitativa no caso dessa pesquisa, seguida de uma segunda fase, qualitativa, “a qual se desenvolve sobre a fase anterior”.

Vieira e Zouain (2005, p. 124. Org.), em uma síntese dos pressupostos da abordagem quantitativa de pesquisa, sob a ótica dos pressupostos metodológicos,

apontam que essa abordagem está centrada na causa e efeito (processo dedutivo, da teoria para os dados), em que as “generalizações levam à predição, explanação e ao entendimento” (descontextualização), “mediante validade e confiabilidade” (acurácia e consistência).

No caso da abordagem qualitativa, o foco está nas “inter-relações de fatores” (processo indutivo dos dados para a teoria), com base em “padrões e teorias desenvolvidas para o entendimento” (contextualização), “mediante verificação e força da argumentação teórica” (acurácia e consistência). (Vieira e Zouain, 2005, p. 124. Org.)

De início foi observada uma questão crucial e importante, mas de não tão simples transposição. O cerne dessa questão esteve baseado em como levantar informações sobre os procedimentos e processos internos das organizações no que tange ao tratamento e análise de dados, principalmente aqueles de cunho pessoal, no âmbito de seus clientes, parceiros e fornecedores. Como obter informações sobre questões tão críticas, ainda mais, no contexto do início da validade das questões legais previstas na LGPD?

Após uma análise extensa e diversas tentativas em múltiplas frentes, chegou-se à conclusão objetiva e pragmática que o melhor acesso a tais informações, mesmo entendendo que a LGPD ainda dá seus primeiros passos, e, portanto, seus instrumentos e orientações ainda estariam num estágio embrionário e com espaços consideráveis para melhorias, seria o acesso e análise das políticas de privacidade divulgadas pelas organizações no Brasil, a partir da disponibilização de tão importantes documentos em seus respectivos sites corporativos.

Segundo a norma ISO 29100, a política de privacidade pode ser definida como uma intenção e orientação geral, regras e compromissos, formalmente expressos pelo controlador de dados pessoais (DP) relativos ao tratamento de DP em uma configuração específica. (ABNT/NBR/ISO 29100: 2020, p.3).

A mesma norma aponta ainda ser conveniente que as organizações documentem suas respectivas políticas de privacidade por escrito e que esta

seja complementada por regras e obrigações “mais detalhadas das diferentes partes interessadas envolvidas no tratamento” de dados pessoais; fornecendo às pessoas de fora da organização um aviso das práticas de privacidade de uma determinada organização; incluindo-se aqui como os dados pessoais são tratados e analisados.

Segundo o “Guia de elaboração de termo de uso e política de privacidade”, do Ministério da Gestão e da Inovação em Serviços Públicos (2023, p. 36), uma política de privacidade deve:

> • ser editada em linguagem acessível, clara e simples;
> • apresentar informações precisas sobre a realização do tratamento dos dados pessoais do cidadão;
> • ser exposta em local de fácil acesso e visualização;
> • deixar de forma clara como o usuário pode apresentar eventual manifestação sobre as finalidades de coleta, uso, armazenamento, tratamento e proteção dos dados pessoais dos usuários; e
> • ser constantemente atualizada.

A partir desse ponto, a segunda questão estratégica para viabilização de uma pesquisa consistente, foi a estipulação de uma lógica robusta para seleção de um conjunto de organizações que teriam suas políticas de privacidade analisadas.

Como resposta a tal questão, após extensa busca e pesquisa, encontrou-se e utilizou-se uma alternativa ótima, mesmo que com limitações, principalmente, como dito anteriormente devido à imaturidade dos preceitos trazidos pela LGPD. Utilizou-se, portanto, o ranking de um importante veículo de comunicação brasileiro, com foco em negócios, que publica anualmente, segundo metodologia própria, a lista das mil maiores empresas no Brasil.

Foram eleitos os resultados desse ranking para o ano de 2021, por uma questão principal. Como, pela metodologia do veículo, a cada ano de avaliação para composição do ranking, são consideradas as demonstrações contábeis encerradas nos últimos dois anos, anteriores ao ano de análise das informações, foi minimizado, ou pelo menos balanceado, os efeitos nos mais diversos setores econômicos trazidos pela pandemia causada pelo vírus SARS-CoV-2, vulgarmente conhecido como Covid-

19, já que nesse caso seriam analisados para composição do ranking as demonstrações dos anos de 2019, antes da pandemia, e 2020 durante a pandemia. Foi garantida assim certa normalidade para considerarmos o conjunto das 1.000 maiores empresas do Brasil, conforme metodologia própria do veículo de comunicação.

Por se tratar de documentos públicos disponibilizados nos respectivos *sites* das organizações integrantes do *ranking* das maiores empresas do Brasil para o ano de 2021, em geral, estamos falando de textos contendo tais políticas. Em sua grande maioria tais documentos são encontrados em arquivos PDF (*Portable Document Format*) e no idioma português.

Nesse ponto foi necessária a realização de algumas adequações, ou seja, todos os documentos e conteúdos que não estavam no formato PDF, foram transformados para esse formato. Assim como, todos os documentos encontrados em outro idioma que não fosse o português foram traduzidos para esse idioma. Dessa forma foi possível garantir uma uniformidade inicial e estrutural dos elementos a serem analisados.

A análise de textos de forma automática é uma das frentes de análise que mais têm crescido nos últimos anos. Como colocado por Dele e Crossland (2008, p.1, tradução nossa), os pesquisadores e outros profissionais, têm encontrado um grande desafio, pois ao ampliarem “o corpo de conhecimento relevante, sempre foi importante trabalhar arduamente para reunir, organizar, analisar e assimilar peças existentes da literatura”, ou como no caso deste trabalho, reunir, organizar, analisar e assimilar o conteúdo textual das políticas de privacidade das mil maiores empresas do Brasil.

Nesse caso adentrou-se no que foi nomeado por processamento de linguagem natural (NLP – *Natural Language Processing*). Conforme Melo Júnior (2018, p. 15), NLP é uma “subárea da ciência da computação, inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais”.

Segundo Kaczmarek *et al* (2022, p. 4, tradução nossa), “é um campo que combina linguística tradicional com algoritmos de mineração de dados para análise automática de textos”. É a mesma perspectiva apontada por Zhao *et al* (2021, p.4, tradução nossa) quando defendem que o processamento de linguagem natural “é um campo que emprega técnicas computacionais com o propósito de aprender, compreender e produzir conteúdo em linguagem humana”.

É importante destacar que, por se tratar de dados não estruturados, a análise desses é uma questão muito mais complexa do que quando se trata de dados estruturados. Segundo Kaczmarek *et al* (2022, p. 4, tradução nossa), um dos “problemas é a heterogeneidade dos dados textuais expressos na ambiguidade das línguas naturais”.

Quanto ao método de coleta dos dados, ou seja, a coleta das Políticas de Privacidade das 1.000 maiores empresas do Brasil, conforme ranking 2021 do veículo de comunicação Valor Econômico em parceria com a Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas (FGV) e Serasa Experian; se deu por visita, busca e *download* dos documentos e/ou conteúdos relativos às políticas de privacidade disponíveis no site de cada uma das organizações ranqueadas entre as mil maiores, no Brasil, no ano de 2021. O esforço de coleta de todo esse material ocorreu entre o período de 03/03/2023 a 19/05/2023.

Quanto ao método de análise selecionado, por se tratar de documentos textuais, portanto, de informação não estruturada; e tendo em vista a quantidade dos mesmos, uma análise de cunho mais qualitativo a partir da leitura, análise e consolidação de todas as políticas das empresas ranqueadas acarretaria uma grande possibilidade de imprecisão e perda de informação e entendimento sobre esse conjunto de informações. Ou seja, quando nos deparamos com um grande volume de informações, nesse caso textuais, fica infrutífera e imprecisa uma análise que utilize consultas estruturadas ou amostrais. Nesses casos a utilização de algoritmos com foco no agrupamento de suas questões principais e estruturantes são um caminho

mais assertivo, confiável e produtivo, capaz de garantir uma análise ampla ao mesmo tempo que altamente efetiva e eficaz.

Nesse sentido, optou-se por uma abordagem no âmbito da analítica avançada, a partir da utilização de um algoritmo de *machine learning* não supervisionado, focado numa análise de *text mining*, com aplicação da técnica de *clustering* para agrupamento das principais dimensões, identificadas a partir da revisão da literatura realizada. Ou de forma resumida, utilizou-se uma abordagem de aprendizado de máquina de agrupamento *k-means*.

Foi utilizada a plataforma "*colaboratory*", ou "colab" do google, onde foi possível acessar os bancos de dados disponibilizados pela empresa, assim como rodar o algoritmo selecionado. Segundo as informações disponíveis na página inicial da plataforma, ela possibilita escrever e executar códigos Python no navegador, possuindo as seguintes e relevantes características: não requer nenhuma configuração; possibilita o acesso a GPUs (Unidades de Processamento Gráfico) sem custo financeiro, assim como seu compartilhamento.

O Colab é uma solução genericamente conhecida como "notebook", que são ambientes onde é possível realizar todas as atividades no âmbito da análise avançada de dados, como análise exploratória de dados (EDA), limpeza e transformação de dados, visualização de dados, modelagem estatística, *machine learning*, *deep learning*, dentre outras possibilidades. No caso específico do Colab, seus "notebooks" executam código de servidores em nuvem do Google.

O objetivo dessa aplicação foi justamente entender se as principais questões apontadas pela literatura de privacidade, a partir das quais se constitui uma proposta inicial de modelo multifacetado, no âmbito do *big data analytics* que garanta a privacidade, ao mesmo tempo que a extração de valor derivada desse processo; estão refletidas nas políticas de privacidade das empresas selecionadas. Onde se entendeu refletiriam diretamente as práticas das organizações nesse âmbito.

De maneira geral e resumida, o método de pesquisa utilizado neste trabalho, pode ser dividido em cinco etapas:

1. Definições das principais estruturas/conceitos referentes a um possível modelo de garantia da privacidade no âmbito do *big data analytics*, no campo teórico, baseadas na literatura analisada;
2. Seleção e preparação dos dados a serem analisados;
3. Aplicação de uma abordagem de aprendizado de máquina não supervisionado e mineração de textos;
4. Aplicação de técnica de agrupamentos, para construção de conjuntos definidos pela similaridade;
5. Análise dos resultados encontrados.

Com base nos resultados dessa primeira fase, terá início uma segunda fase, de cunho qualitativo. Após a execução da fase quantitativa, foram selecionadas 28 organizações e suas respectivas políticas de privacidade, perfazendo 3% do total de empresas integrantes do ranking Valor 1000 com políticas de privacidade disponíveis e acessíveis em seus respectivos sites, sobre as quais será realizada uma leitura analítica aprofundada das mesmas, de forma a melhorar a visão do pesquisador sobre fatores relevantes e distintivos no âmbito dos clusters que foram constituídos.

O principal motivo dessa análise foi obter insights para uma possível decisão envolvendo o número de clusters ideal para os objetivos deste trabalho; assim como um melhor entendimento sobre determinadas características dos clusters a serem definidos. Como colocado por Gonçalves *et al* (2018, p. 9), é necessária uma "análise qualitativa do detalhamento de cada agrupamento para que sejam obtidas tais conclusões".

Roter, Ninkovic e Dordevic (2022, p. 2, tradução nossa), realizaram a mesma abordagem em sua pesquisa, e apontaram que "a melhor forma de estimar o número ideal de aglomerados, é combinar cálculos de aprendizagem de máquina com conhecimento humano" sobre o tema objeto do estudo.

## 8.1 Aprendizado de Máquina e Análise Automatizada de Textos

Mariani, Navrotska e Mancini (2023, p. 3, tradução nossa), afirmam que “o aprendizado de máquina envolve o desenvolvimento de abordagens computacionais para analisar padrões automaticamente, aprender com os dados e tomar decisões com assistência humana mínima ou não explicada”.

Para Machado (2018, p.94), aprendizado de máquina é um “método de análise de dados que busca a automatização do desenvolvimento de modelos analíticos”. Com base em algoritmos que aprendem de forma interativa a partir de dados, que possibilitam a geração de insights “ocultos sem serem explicitamente programados para procurar uma informação oculta específica”.

Segundo Ozaydin *et al* (2017, p.1, tradução nossa), “os desenvolvimentos mais recentes em aprendizado de máquina e mineração de texto oferecem algumas soluções potenciais para enfrentar” os desafios analíticos trazidos pelas novas tecnologias e que, de maneira geral, se traduzem em “grandes volumes de textos através de processos semiautomáticos”.

Os algoritmos de aprendizado de máquina são classificados de duas formas, aprendizado de máquina supervisionado e aprendizado de máquina não supervisionado. Neste trabalho utilizamos a abordagem não supervisionada. De forma resumida, quando não sabemos antecipadamente o que estamos procurando e/ou quando não damos um direcionamento de busca à máquina, estamos falando de aprendizagem de máquina não supervisionada.

Já no caso do aprendizado de máquina supervisionado, o algoritmo aprende a partir de dados rotulados anteriormente, com base nesse histórico o algoritmo aprende e consegue predizer uma determinada questão/circunstância. Foreman (2016, p.30), traz um claro exemplo para facilitar a diferenciação das abordagens ao descrever a seguinte situação, se uma empresa tem como objetivo dividir seus clientes em dois grupos, “digamos possíveis compradores e possíveis não compradores e eu forneço exemplos de históricos de tais clientes ao computador e digo a ele para atribuir todas

as novas direções a um desses dois grupos"; temos uma abordagem supervisionada.

"A partir da análise de um conjunto de dados de treinamento conhecido, os algoritmos de aprendizado de máquina supervisionado produzem uma função inferida para fazer previsões sobre os valores de saída". (Mariani, Navrotska e Mancini, 2023, p. 3, tradução nossa)

Neste trabalho foi utilizada a abordagem de aprendizado de máquina não supervisionado, indicada para abordagens de cunho exploratório, quando não se quer impor um direcionamento anterior ao algoritmo. Ou de maneira mais específica para o caso em questão, quando se tenciona entender a partir da análise das políticas de privacidade das organizações selecionadas se serão encontradas na base constituída a partir dessas políticas, as dimensões de garantia de privacidade no âmbito do *big data analytics*, como forma de analisar se são as mesmas que estão presentes no modelo teórico constituído nesse trabalho.

Ou seja, o interesse da análise no contexto dessa pesquisa foi descobrir e descrever quais dimensões no âmbito da garantia de privacidade no *big data analytics*, construídas a partir da teoria, estavam ou não presentes nas políticas de privacidade das organizações selecionadas; de grande representatividade por se tratar das mil maiores empresas do Brasil, e a partir daí poder confirmar ou refutar o modelo proposto.

Como já pontuado, foram analisados textos, configurados nas políticas de privacidade das organizações selecionadas. A análise de textos ou mineração de textos, como é mais conhecida, pode ser definida, segundo Kaczmarek *et al* (2022, p. 4, tradução nossa) "como um processo para obter informações úteis e de alta qualidade a partir de dados de texto". Ainda segundo os mesmos autores, a abordagem de mineração de texto, "contém um conjunto de técnicas de aprendizado de máquina, linguística e estatística usadas para extrair informações do texto".

Para Delen e Crossland (2008, p. 1708, tradução nossa), “é o processo de descobrir informações novas, até então desconhecidas e potencialmente úteis, a partir de uma variedade de fontes de dados não estruturados”.

Na aplicação da lógica de mineração de dados à mineração de textos, tem-se o que se denomina como “descoberta de conhecimento em textos (KDT)”, que pode ser entendido como um processo não “trivial de identificação de padrões implícitos, a partir de dados textuais”, composto, de maneira geral por várias etapas, sendo as principais: coleta, preparação dos dados, análise e construção dos insights, avaliação e refinamento. (Gonçalves *et al*, 2018, p.3)

De forma mais específica, esse trabalho se referenciou no processo de mineração de texto dado pelo método do Processo Padrão Inter Indústria para Mineração de Dados - CRISP-DM (*Cross Industry Standard Process for Data Mining*), composto pelas seguintes etapas: a) determinação do objetivo do estudo/análise; b) entendimento da natureza e disponibilidade dos dados; c) preparação dos dados; d) modelagem dos dados; e) avaliação do modelo constituído; f) desenvolvimento/aplicação.

Suganya e Porkodi (2017, p. 122, tradução nossa), destacam que a mineração de texto “é uma área de pesquisa que inclui processamento de linguagem natural, aprendizado de máquina, mineração de dados e recuperação da informação”. Apesar da mineração de textos estar muitíssimo próxima da mineração de dados, da qual se originou, essas abordagens se diferem de forma distintiva pelo tipo de informação que analisam; onde a primeira trabalha com textos em linguagem natural, portanto, informação não estruturada; a segunda está focada e preparada para a análise de dados estruturados, em geral, aqueles que contém uma grande quantidade de registro de números, valores.

Como reforça Raghupathi, Ren e Raghupathi (2020, p. 5, tradução nossa), “ao usar o processamento de linguagem natural, a análise de texto pode transformar dados não estruturados em um formato estruturado adequado para análise e aplicação de algoritmos de aprendizado de máquina”. A utilização da mineração de

textos permite realizar a avaliação de múltiplas “dimensões dos conceitos centrais que desejam buscar nos dados não estruturados” presentes nos textos.

A mineração de textos, a partir da intensificação de sua utilização nos últimos anos, tem aumentado imensamente seu campo de aplicação. São algumas dessas possibilidades: tratamento de reclamações de clientes; filtragem de spans e outros lixos eletrônicos em e-mails; *feedback* a clientes; pesquisa de mercado; personalização do cliente em aplicações de comércio eletrônico, perfis de usuários, sumarização de textos, análise de mídia social, triagem de dados em setores como jurídico, saúde, detecção de fraudes, dentre muitos outros. (Delen e Crossland, 2008; Aoun, 2023).

Na visão de Aoun (2023, p. 53, tradução nossa), são alguns dos principais objetivos da mineração e análise de textos:

> 1. Identificação de temas e conceitos-chave presentes numa coleção de dados de texto;
> 2. Extração de informações úteis, como reconhecimento de entidades e análise de sentimentos;
> 3. Melhorar a tomada de decisões, a partir do fornecimento de insights sobre clientes e tendências de mercado;
> 4. Melhorar a eficiência operacional, pela rapidez e eficiência com as quais consegue analisar e gerar insights sobre grandes quantidades de informação não estruturada.

De maneira específica, no que tange aos objetivos da mineração de texto, esse trabalho se valeu tanto da identificação de temas e conceitos-chave como da extração de informações úteis. Quanto às aplicações, utilizou-se o processo de *clustering,* que trata “do agrupamento de documentos semelhantes entre si sem ter um conjunto predefinido de categorias”. (Delen e Crossland, 2008, p. 1707)

## 8.2 Agrupamento *k-means*

Segundo definido por Fahim (2021, p. 1, tradução nossa), o agrupamento de dados, ou no inglês “*clustering*” de dados, no caso desse trabalho elementos textuais, é um importante método de aprendizado de máquina e mineração de dados. “Tem como objetivo descobrir conhecimento latente a partir dos dados”, não estruturados.

Conforme Hair *et al* (2009, p. 430), a análise de agrupamentos é um “grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem”. Os autores ainda esclarecem que a análise de agrupamentos “tem sido chamada de análise Q, construção de tipologia, análise de classificação e taxonomia numérica”.

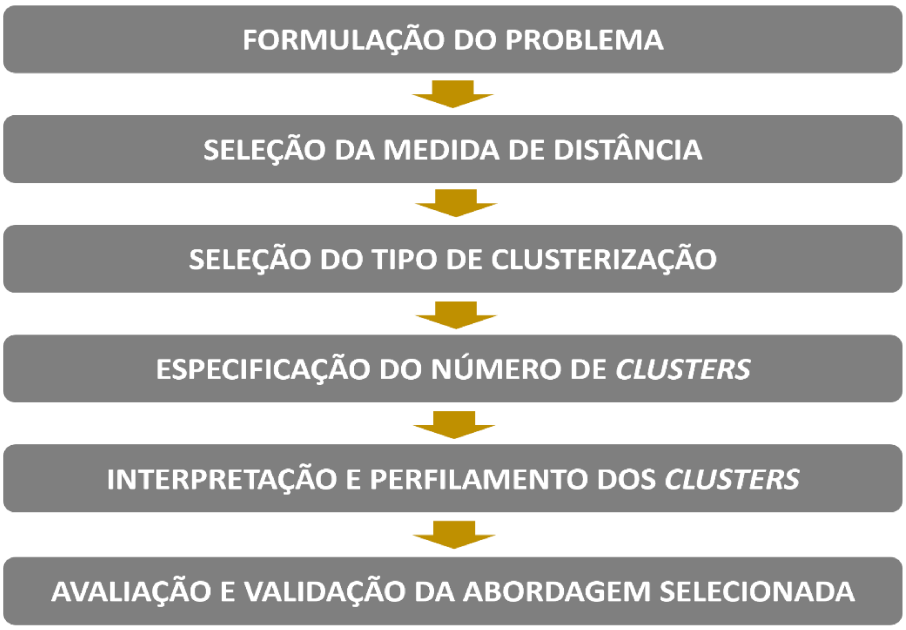
O principal objetivo da análise de agrupamento “é agregar os dados com alta similaridade e separar os dados com baixa similaridade, o que é chamado de divisão da estrutura de dados”. Ou seja, cada cluster tende a ter objetos/casos semelhantes entre si, mas que diferem dos objetos/casos de outros clusters. (Hu *et al*, 2021, p. 1, tradução nossa)

*Clustering* é uma técnica exploratória de dados, sendo ainda uma das soluções mais usuais no âmbito do aprendizado de máquina não supervisionado. (Roter, Ninkovic e Dordevic, 2022)

Segundo Malhotra (2006, p. 572), a análise de agrupamentos (*clusters*) é uma “técnica utilizada para classificar objetos ou casos em grupos relativamente homogêneos”, portanto, por definição é uma abordagem analítica de classificação, ou “taxonomia numérica”. A maioria “dos métodos de aglomeração é heurística, baseada em algoritmos”. (Malhotra, 2006, p. 574)

Nesse trabalho a análise para aplicação da abordagem de cluster, seguiu as seguintes etapas explicitadas na Figura 6 abaixo.

**FIGURA 6 - Análise para aplicação da abordagem de *cluster***




Fonte: Malhotra, 2006, p. 575

Existem inúmeras abordagens no âmbito da análise de *cluster*, dentre as quais poderíamos citar: *k-means*, *k-methods*, hierárquicos, misturas gaussianas, DBSCAN, Jarvis-Patrick, LDA (*Latent Dirichlet Allocation*), CLIQUE, MNCut (Modified Normalized), *Clustering* Teórico de Grafos (*Spectral Clustering*). Nesse trabalho foi utilizado o método *k-means*. (Roter, Ninkovic e Dordevic, 2022; Jain, 2010)

*K-means* é uma das abordagens mais tradicionais e menos complexas de agrupamento, tendo suas primeiras aplicações ocorrido na década de 1950 (Jain, 2010; Foreman, 2016).

De maneira geral é um algoritmo escolhido por sua "capacidade de alto desempenho com grandes conjuntos de dados e à sua capacidade de criar *clusters* que agrupam porções substanciais de dados", além de isolarem os *outliers*. (Zelasky *et al*, 2023, p. 4, tradução nossa)

Embora o *algoritmo k-means* tenha sido, como visto acima, desenvolvido a décadas atrás, ele ainda é amplamente utilizado devido a sua simplicidade, conveniência, facilidade de implementação, sucesso empírico e alta eficiência. (Hu *et al*, 2023; Jain, 2010)

O método *k-means* objetiva "dividir um conjunto de "N" objetos em "K" *clusters*, onde cada *cluster* é representado pelo vetor médio de seus objetos" Uma das principais e mais complexas questões na análise de clusters é a definição do número de clusters a serem considerados para a análise. (Fahim, 2021, p. 1, tradução nossa)

Numa visão pragmática, mas, ao mesmo tempo não reconhecida por grande parte dos autores do assunto, Roter, Ninkovic e Dordevic (2022, p. 2, tradução nossa), apregoam que "o agrupamento é uma técnica exploratória e o número exato de agrupamentos não existe". Na visão desses autores, a solução ideal para estimação do número de *clusters* "é combinar cálculos de aprendizagem automática com conhecimento humano". A partir dessa combinação de abordagens, os autores afirmam que é possível produzir o "agrupamento mais significativo".

Conforme postulado por Jain (2010, p. 655, tradução nossa), o "agrupamento continua a ser um problema difícil. Isso pode ser atribuído à imprecisão inerente à definição de um *cluster* e a dificuldade de definir uma medida de similaridade apropriada e função objetiva".

Aprofundando no entendimento dessa complexa questão, Jain (2010, p. 656, tradução nossa), trazendo uma visão de solução ótima, apregoa que "o melhor valor de k é então escolhido com base em um critério predefinido". O autor defende que "considerando o mesmo conjunto de *corpus* de documentos, diferentes grupos de usuários podem estar interessados em gerar partições de documentos com base em suas respectivas necessidades".

Jain (2010, p. 656, tradução nossa) vai além e apregoa que um "método de agrupamento que satisfaça os requisitos de um grupo de usuários pode não satisfazer os requisitos de outro". Nesse sentido, conclui afirmando um ponto fundamental e cujo princípio foi orientador das definições dessa pesquisa ao afirmar que "o *clustering* está

nos olhos de quem vê". Dessa forma, essa abordagem deve envolver as "necessidades do usuário ou da aplicação".

Neste trabalho o número de *clusters* partiu da estrutura inicial teórica visualizada a partir da revisão de literatura e, portanto, aderente às prerrogativas dadas pela necessidade do usuário (pesquisador) e pela aplicação (abordagem constituída). É fundamental explicitar que mesmo partindo dessa robusta premissa, alinhada aos conceitos de mais alto nível na abordagem proposta, visando a confirmação dessa visão, foram realizadas diversas tentativas e suas respectivas análises de consistência.

Nesse sentido, foram testados diversos números de "K"; números de clusters, com dez, sete, cinco e três *clusters*, sendo o penúltimo (cinco *clusters*) o parâmetro dado pela literatura do tema. Os resultados, como se verá à frente em capítulo específico que contempla essa discussão, mas que é antecipado aqui de forma a demonstrar a assertividade dos parâmetros utilizados, corroboraram a teoria, onde o número de cinco clusters, que foi o parâmetro utilizado a partir da revisão da literatura para a constituição de um modelo teórico de garantia da privacidade no âmbito do *big data analytics* e da perspectiva do pesquisador, foram aqueles que mais aderência e lógica apresentaram, no contexto dos objetivos esperados.

Segundo Malhotra (2006, p. 589), as variáveis para construção do *cluster* "devem ser selecionadas com base em pesquisa passada, na teoria, nas hipóteses que estão sendo testadas ou no julgamento do pesquisador". Neste trabalho, assim como indicado por Malhotra (2006), realizou-se a construção dos *clusters* a partir de uma profunda e analítica revisão da literatura, complementada e refinada pelo julgamento do pesquisador.

Malhotra (2006, p. 589), ainda preceitua que "deve-se escolher uma medida adequada de distância ou de semelhança". Visando atender a esse preceito. Neste trabalho foi selecionado o método centróide, que segundo esse mesmo autor é um "método de variância de aglomeração hierárquica em que a distância entre dois

aglomerados é a distância entre seus centroides (médias para todas as variáveis). (Malhotra, 2006, p. 578)

Os algoritmos de *clustering* podem ser divididos em dois grupos: hierárquicos e não hierárquicos. No âmbito dos algoritmos não hierárquicos, K-means é o algoritmo de particionamento mais simples e mais popular. Conforme postulado por Jain (2010, p. 653, tradução nossa), algoritmos de *cluster* particionado, “encontram todos os clusters, simultaneamente como uma partição de dados e não impõem uma estrutura hierárquica”. São em geral aqueles elegidos quando o pesquisador tenciona o reconhecimento de padrões, tendo em vista a natureza dos dados utilizados. Neste trabalho foi utilizada a abordagem de *clustering* particionado.

São algumas das aplicações possíveis da abordagem de agrupamento: detecção de *outliers*; processamento de gráficos; reconhecimento de padrões; análise de densidade espectral; integração de dados de parques eólicos; previsão climática; agrupamento de amostras de populações ou áreas geográficas com características demográficas, ou sociais semelhantes; segmentação de marketing; dentre muitas outras. (Hu *et al*, 2021; Mariani, Navrotska e Mancini, 2023)

Hair *et al* (2009, p. 431), apontam que a abordagem de agrupamentos sofre um conjunto de críticas muito comuns, mas muito importantes e pertinentes de serem consideradas, e que essas deveriam ser resolvidas, como foi o procedimento adotado neste trabalho, por suporte conceitual e não empírico:

- A análise de agrupamentos é descritiva, não teórica e não inferencial. Não possui uma base estatística sobre a qual possa realizar inferências (de uma amostra para a população) sendo, portanto, denominada como uma técnica exploratória;
- A análise de agrupamentos sempre criará agrupamentos, independentemente da existência real de alguma estrutura nos dados. Quando o pesquisador usa a análise de agrupamentos, ele está fazendo uma suposição sobre alguma estrutura entre os objetos. Somente com

forte suporte conceitual e validação, os agrupamentos são potencialmente significativos e relevantes;

- A análise de agrupamentos não é generalizável, pois é totalmente dependente das variáveis usadas como base para a medida de similaridade. Além disso, como apontado no primeiro item ela não possui base estatística que possibilite a realização de inferências.

## 8.3 Preparação dos dados

Para a efetiva implementação da clusterização de textos existe uma etapa de enorme relevância, que ditará muito da qualidade final dos resultados obtidos. Não será realizada aqui uma descrição pormenorizada e altamente técnica, por não ser o objetivo deste trabalho, mas serão demonstradas todas as etapas e esforços realizados para a preparação dos dados para a realização da análise executada.

Conforme apontado por Gonçalves *et al* (2018, p.6), o processo de clusterização com base em dados de texto, requer uma série de intervenções prévias para serem processados e analisados. “Como a maioria dos processos de *data mining*, os dados precisam ser tratados, limpos, trabalhados, transformados, indexados e trazidos a uma base comum para que finalmente seja aplicado o algoritmo de *k-means* sobre esse resultado”. Aoun (2023, p.54, tradução nossa), reforça essa indicação ao afirmar que “as técnicas de pré-processamento de texto são o primeiro passo na mineração e análise de texto”.

No caso dessa pesquisa, para a preparação e pré-processamento dos textos analisados, foram utilizados os processos básicos tradicionais que são os processos de tokenização, remoção de *stop word* e *stemming*.

Inicialmente foi realizado o processo de retirada das “*stop words*”. Como explicitado por Delen e Crossland (2008, p. 1711), alguns elementos do texto, “como artigos, verbos auxiliares e termos usados em quase todos os documentos do *corpus*, não têm poder diferenciador”; por isso mesmo a necessidade de excluí-los de forma

a que não contaminem a análise. Esses termos de parada (*stop words*), são termos específicos do “domínio de estudo e devem ser identificados pelos especialistas do domínio”.

A transformação de todas as palavras em letras minúsculas e a retirada de todos os caracteres especiais (como por exemplo, pontuações, colchetes, etc.), foi uma importante padronização e mais um passo na preparação dos dados.

Além disso, foi acoplado a esse processo de preparação a retirada dos nomes das organizações integrantes do ranking Valor1000 no âmbito de suas respectivas “Políticas de Privacidade” para não influenciar na contagem das palavras, não mascarando palavras mais relevantes para os objetivos da pesquisa. Esse foi um passo fundamental na preparação dos dados, pois nas primeiras vezes que o modelo foi executado, o nome das organizações que se repetia diversas vezes nos documentos trazia uma forte contaminação.

Na sequência foi realizado o tratamento conhecido como “*stemming*”, que trata da retirada dos chamados afixos, que se referem aos sufixos e prefixos de todas as palavras. Um dos principais objetivos dessa etapa é a redução do tamanho da estrutura de indexação, pois “os termos de índices distintos são reduzidos”. (Sunganya e Porkodi, 2017, p. 113)

De forma complementar, Delen e Crossland (2008, p. 1712, tradução nossa) destacam que o processo de “*stemming*” é recomendado para que se consiga criar índices com precisão, reduzindo as palavras às suas raízes. “Para que, por exemplo, diferentes formas gramaticais ou declinação de verbos sejam identificadas e indexadas como uma mesma palavra”.

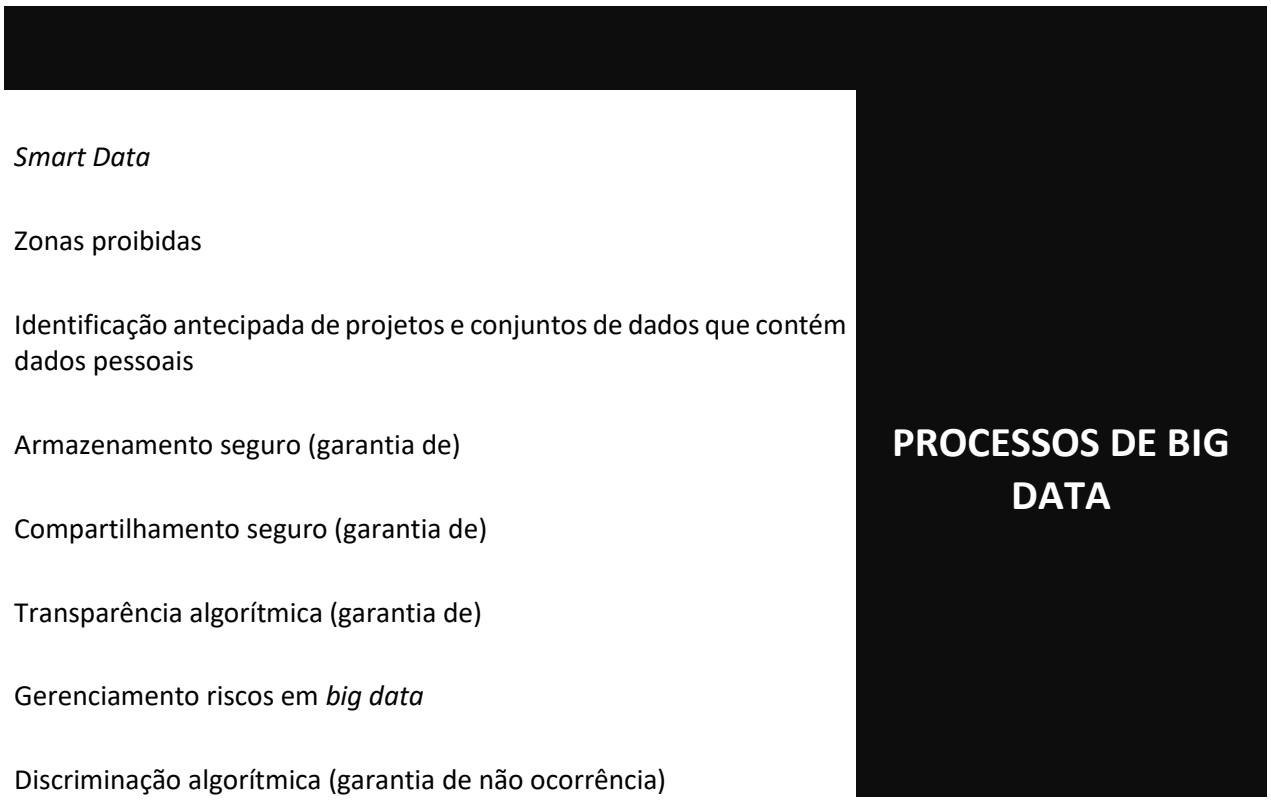
Em particular pelas características do domínio de que trata essa pesquisa foi necessário a realização de algumas adequações ao processo clássico de *stemming*, com foco no maior aproveitamento dos termos-chave, que dão na sua forma composta na maioria das vezes o melhor entendimento e suporte para obtenção de maior consistência pelo algoritmo utilizado.

No que tange ao processo de tokenização, que seria o processo seguinte, o texto é dividido em palavras distintas chamadas tokens. Ou seja, é o processo de dividir o texto em unidades menores, nesse caso as unidades são as palavras constantes no texto, mas em muitos casos podem ser símbolos, frases, dentre outros elementos. (Duan *et al*, 2021; Suganya e Porkodi, 2017)

Com base nos termos-chave advindos da análise da revisão bibliográfica, foi constituído um rol desses termos, base para a criação do chamado “saco de palavras”. Como afirma Suganya e Porkodi (2017, p. 113, tradução nossa), “as técnicas de extração de informações são amplamente utilizadas no modelo de saco de palavras para tarefas que incluem correspondência, classificação e agrupamento”.

Na Figura 7 abaixo, são apresentados os termos-chave utilizados como base para a constituição do “saco de palavras” utilizado neste trabalho:

**FIGURA 7 - Lista de termos-chave para constituição do “saco de palavras”**



Princípio da finalidade (é observado)

Princípio da adequação (é observado)

Princípio da necessidade (é observado)

Dados portadores de política

Parâmetros para permissões de transmissão (são observados)

Parâmetros para uso e descarte (são observados)

Procedimentos ciclo do *big data* (que garantem a proteção da privacidade)

Políticas ciclo do *big data* (que garantem a proteção da privacidade)

Prática ciclo *big data* (que garantem a proteção da privacidade)

Políticas internas (possui políticas internas que ordenam a garantia da privacidade)

Controle do Proprietário (é garantido)

Acesso ao titular de dados (é garantido)

Treinamento (constante sobre privacidade no que tange aos profissionais da organização)

Encarregado de dados (possui um encarregado de dados, próprio ou terceiro)

*Privacy by design* -Privacidade por princípio/design (abordagem onde a privacidade é a base de qualquer iniciativa, ex.: desenvolvimento solução, software, etc.)

**GOVERNANÇA EM PRIVACIDADE**

| | |
|---|---|
| Comitê de Ética e Revisão (no âmbito do *big data*, a organização possui) | |
| Inventário de dados (a organização realiza) | |
| Corretores de dados (a empresa possui ações para seu uso adequado) | |
| Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade) | |
| | |
| Desidentificação | **TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS** |
| Pseudonimização | |
| Anonimização | |
| Generalização | |
| Randomização | |
| Triple DES | |
| Modelo de execução híbrida - HybrEx | |
| Modelos de classificação de dados sensíveis | |
| Técnicas de limitação de divulgação estatística | |
| | |
| *Structured Map Reduced Layer* | **TÉCNICAS E PROCESSOS DE SEGURANÇA** |
| Segurança dos dados (dados protegidos por salvaguardas de segurança) | |

| | |
|---|---|
| Controle de acesso (autenticação, autorização auditoria, MAC) | **ADOTADAS NA ORGANIZAÇÃO** |
| Segurança da informação | |
| Técnicas de ataque em *big data* (link lateral, palpite óbvio, reidentificação, etc.) | |
| | |
| Confidencialidade | |
| Especificação de propósito | |
| Processos de comunicação vazamento ou roubo de dados | |
| Consentimento (garante que o indivíduo dê seu consentimento) | |
| Aviso (informa ao titular sobre a coleta de dados) | **PRIVACIDADE POR POLÍTICA - FIPP (*Fair Information Practices Principals*)** |
| Minimização de dados (impõe limitações sobre o tipo de informações a serem coletadas) | |
| Cuidado análise dados de populações vulneráveis (crianças, grávidas, prisioneiros, LGTB, etc.) | |
| Políticas que disciplinam quem pode fazer tratamento dos dados | |
| Retificação (permite que o titular realize a retificação dos seus dados) | |

Fonte: elaboração própria, 2023

Com base nessa lista de termos e “no conjunto de palavras do *Natural Language Toolkit* (NLTK)” para filtragem das palavras mais relevantes, o “saco de palavras” foi constituído a partir da contagem da incidência desses termos no conjunto de textos selecionados. E com base no “saco de palavras” são constituídos os *clusters*. (Duan *et al*, 2021, p. 6, tradução nossa)

## 8.4 Anuário Valor 1000 – Edição 2021

No que tange ao universo de realização da pesquisa, foram consideradas as declarações constantes nas “Políticas de Privacidade” das 1.000 maiores empresas do Brasil, conforme o “Anuário Valor 1000 – Edição 2021”.

Segundo as informações disponíveis no regulamento direcionador das regras de participação e metodologia utilizada para o ranqueamento das empresas do Brasil, esse anuário é um esforço conjunto do jornal Valor Econômico, da Escola de Administração de Empresas de São Paulo, da Fundação Getúlio Vargas e da Serasa Experian.

A metodologia utilizada está basicamente apoiada na análise de questionários respondidos pelas empresas participantes, assim como na análise de suas demonstrações contábeis encerradas nos dois últimos anos anteriores ao período de análise, que devem estar alinhadas ao padrão internacional IFRS (*International Financial Reporting Standards*), emitidas pelo IASB (*International Accounting Standards Board*). Sendo o ranking ordenado por classe decrescente de receita líquida, independentemente do segmento ao qual pertence uma determinada organização.

Na edição de 2021, foram analisados 29 segmentos econômicos: Açúcar e Álcool; Agropecuária; Água e Saneamento; Alimentos e Bebidas; Comércio Atacadista e Exterior; Comércio Varejista; Comunicação e Gráfica; Construção e Engenharia; Educação e Ensino; Eletroeletrônica; Empreendimentos Imobiliários; Energia Elétrica, Farmacêutica e Cosméticos; Fumo; Materiais de Construção e de Decoração; Mecânica; Metalurgia e Mineração (inclui Siderurgia); Papel e Celulose; Petróleo e Gás; Plásticos e Borracha; Química e Petroquímica; Serviços Ambientais (Engenharia Ambiental); Serviços Especializados; Serviços Financeiros (exceto bancos, seguradoras, resseguradoras, companhias de previdência e de capitalização e planos de saúde); Serviços Médicos; Tecnologia da Informação e Telecomunicações; Têxtil, Couro e Vestuário; Transportes e Logística; Veículos e Peças.

A análise para construção do ranking leva em conta os seguintes critérios e seus respectivos pesos: Receita Líquida (2,5); Margem Ebitda (2); Rentabilidade (1,5); Giro do Ativo (1); Liquidez Corrente (1); Margem da Atividade (1); Cobertura de Juros (0,5); Crescimento Sustentável (0,5).

# 9 ANÁLISE DOS RESULTADOS

## 9.1 As primeiras conclusões trazidas pelo processo de coleta dos dados

A etapa de coleta das informações, realizada no período de 29/03/2023 a 19/05/2023, demonstrou um fato até certo ponto esperado, devido à precocidade de uma nova realidade trazida pela efetiva implementação da LGPD, que ainda necessitará de algum tempo para obter seus melhores resultados. Mesmo tendo em vista os avanços ocorridos, principalmente nos países europeus, que já demonstravam uma tendência inequívoca, as maiores empresas do Brasil ao que parece esperaram toda a situação legal, jurídica e processual se estabelecer para efetivamente pensarem a sério a questão da privacidade.

Das mil maiores empresas que integraram o ranking Valor 1000 (2021), 88,2% possuíam uma política de privacidade, independente nesse momento de qualquer análise quanto a qualidade, validade, alinhamento à legislação em vigor e completude. Um total de 108 (10,8%) dessas empresas não possuíam uma política de privacidade disponível no site da organização; e 1% (10 empresas) possuíam um link ou alguma outra referência a uma política de privacidade que não permitia o acesso à mesma (links quebrados, links sem vínculo com um documento, etc.).

A Tabela 1, apresentada abaixo, faz uma consolidação desses dados para disponibilizar uma visão clara da situação encontrada na fase de coleta dos dados:

**TABELA 1 - Política de Privacidade (PP) x Disponibilização da PP no site das organizações analisadas**

| Situação | Total | Percentual (%) |
|---|---|---|
| § Possuíam uma Política de Privacidade disponível no site [illegible] organização | 882 | 88,2 |
| § Não possuíam uma Política de Privacidade disponível no site [illegible] organização | 108 | 10,8 |
| § Existe a indicação de um link ou outro espaço onde estaria disponível a Política de Privacidade, mas não é possível acessar [illegible] PP | 10 | 1 |
| TOTAL | 1.000 | 100 |

Fonte: dados da pesquisa, elaboração própria, 2023

Salta aos olhos essa primeira constatação de que das 1000 maiores empresas do Brasil (conforme as regras de medidas estabelecidas pela publicação Valor 1000), 118 (11,8%) dessas empresas não possibilitavam o acesso às suas Políticas de Privacidade, quando essas eram buscadas nos sites dessas organizações.

Tendo em vista o porte dessas organizações, independente dos seus respectivos segmentos econômicos, não deixa de ser um sinal de alerta inequívoco da imaturidade, despreparo ou falta de acreditação de que a questão da privacidade é uma coisa a ser levada a sério também no Brasil. Uma forte tendência oriunda principalmente da Europa aponta para o fato de que os principais fundos de investimento internacionais começam a analisar os negócios em que investem com muito mais vigor no que tange à questão da privacidade. É um movimento parecido ao que as questões de sustentabilidade e equidade tem provocado em mercados de *commodities*, como o café, onde os grandes players desse mercado passam a exigir tais questões e imprimem uma forte pressão “cadeia abaixo”.

Isso é demonstrado não só pelo quantitativo analisado anteriormente, mas também através de uma análise de cunho mais qualitativo que foi realizada em uma

amostra dessas políticas. Ou seja, ao se analisar um grupo de 20% dessas políticas no âmbito das 1000 maiores empresas do Brasil, o que se verificou foi um conjunto de documentos que parecem cópias uns dos outros na maioria dos casos, e isso é incrível quando observado em um conjunto de empresas de setores e atuações bem diferentes.

O que se identificou nessa análise qualitativa foi que grande parte das organizações se preocupam apenas em cumprir um limite mínimo de alinhamento com algumas questões principais da LGPD, sem qualquer preocupação ou explicitação do que fazem internamente, em termos analíticos, com os dados de seus clientes, fornecedores e parceiros. O que inviabiliza totalmente uma mínima visão sobre possíveis problemas nessas abordagens como os tão proclamados, mas nem tanto divulgados ou conhecidos (claro, pois não se conhece minimamente as ferramentas e análises realizadas com os dados disponíveis internamente nessas organizações) vieses analíticos, que resultam na chamada discriminação algorítmica.

Como foi colocado por O'Neil (2020, p.19), ao afirmar que um algoritmo e um grande conjunto de dados e informações sobre um determinado indivíduo gera "uma probabilidade" sobre este indivíduo, que pode ser o indicativo de uma contratação de empréstimo duvidoso relativo a essa pessoa, ou seja, um devedor de alto risco, ou um indicativo de potencial terrorista ou até de um "péssimo professor". A autora argumenta que essa "probabilidade é destilada numa pontuação, que pode pôr a vida de alguém de ponta-cabeça". E ainda conforme apregoado pela autora, "mesmo quando a pessoa reage, evidências sugestivas do contrário simplesmente não bastam".

## 9.2 Caracterização da amostra

Do conjunto das mil maiores empresas do Brasil dado pelo ranking da revista Valor 1000, do ano de 2021, com "Políticas de Privacidade" disponibilizadas e acessíveis junto aos seus respectivos sites, foi possível identificar, por exemplo, os Estados brasileiros onde as mesmas estão presentes. Identificou-se que as

organizações analisadas neste estudo estão presentes em 22 dos 27 estados brasileiros, em todas as cinco regiões. Desse total, 66% tem sua sede na região sudeste, nos estados de São Paulo, Minas Gerais e Rio de Janeiro; ao passo que 20,1% tem sua sede na região sul, nos estados do Paraná, Rio Grande do Sul e Santa Catarina. O restante das empresas (13,9%), se encontram nos demais estados e regiões do país. Tal distribuição guarda uma lógica próxima da real distribuição de empresas no território brasileiro, o que denota um equilíbrio e balanceamento adequado da amostra utilizada. A Tabela 2 abaixo, traz um consolidado dos resultados encontrados.

**TABELA 2 - Distribuição das empresas da amostra por estados e regiões**

| Estado | Região | Total | Percentual (%) |
|---|---|---|---|
| São Paulo | Sudeste | 403 | 45,7 |
| Minas Gerais | Sudeste | 93 | 10,5 |
| Rio de Janeiro | Sudeste | 86 | 9,8 |
| Paraná | Sul | 71 | 8 |
| Rio Grande do Sul | Sul | 65 | 7,4 |
| Santa Catarina | Sul | 41 | 4,6 |
| Demais Estados | Demais Regiões | 123 | 13,9 |
| Total | | 882 | 100 |

Fonte: dados da pesquisa, elaboração própria, 2023

O conjunto de empresas analisadas, dado pelo ranking Valor1000 (2021), com “Políticas de Privacidade” disponibilizadas e acessíveis junto aos seus respectivos sites, estão presentes em vinte e oito segmentos econômicos. Dessas 882 empresas, aproximadamente 50% (49,3%), estão concentradas nos seguintes segmentos, nessa ordem: Comércio Varejista, Agronegócio, Transporte e Logística, Alimentos e Bebidas, Energia Elétrica, Química e Petroquímica, Metalurgia e Siderurgia e Serviços Especializados. Na Tabela 3 abaixo, é possível visualizar essa distribuição:

**TABELA 3 - Distribuição das empresas por segmento econômico**

| Setor | Total de Empresas | Percentual |
|---|---|---|
| Comércio Varejista | 78 | 8,8 |
| Agronegócio | 64 | 7,3 |
| Transporte e Logística | 63 | 7,1 |
| Alimentos e Bebidas | 61 | 6,9 |
| Energia Elétrica | 47 | 5,3 |
| Química e Petroquímica | 44 | 5 |
| Metalurgia e Siderurgia | 40 | 4,5 |
| Serviços Especializados | 38 | 4,3 |
| Bioenergia | 36 | 4,1 |
| Petróleo e Gás | 36 | 4,1 |
| TI e Telecom | 35 | 4 |
| Demais Segmentos Econômicos | 340 | 38,5 |
| TOTAL | 882 | 100 |

Fonte: dados da pesquisa, elaboração própria, 2023

No conjunto, as empresas integrantes do ranking Valor1000, obtiveram uma receita líquida no ano de 2021 de R$ 6, 33 trilhões. Tais valores dão a importância e magnitude em termos do país no que tange a esse conjunto de empresas analisadas e consideradas neste estudo.

## 9.3 Construção e seleção dos *clusters*

É fundamental dar se início a essa questão com o apontamento de que a identificação do número ideal de *clusters* é um desafio ainda sem solução definitiva. Como já foi pontuado anteriormente, o agrupamento é uma técnica exploratória, nesse ínterim, “o número correto de agrupamentos não existe”. É como advoga Fahim (2021, p. 2, tradução nossa), “até agora não há nenhuma pesquisa proposta para resolver” essa questão.

Na verdade, não seria totalmente correto dizer que o número ideal de agrupamentos não existe, mas sim, que a identificação antecipada desse “número

ideal” é praticamente impossível, mesmo contando com algumas abordagens quantitativas que podem contribuir para a atenuação dessa realidade complexa, como, por exemplo, “critério de Calinski-Harabasz, índice de Davies-Bouldin, gráficos de silhueta e denograma”, dentre outros. (Roter, Ninkovic e Dordevic, 2022, p. 2, tradução nossa)

Jain (2010, p. 655, tradução nossa), reforça a lógica dessa problemática, ao afirmar que “o agrupamento continua a ser um problema difícil. Isto pode ser atribuído à imprecisão inerente à definição de um *cluster*, e a dificuldade em definir uma medida de similaridade apropriada e função objetiva.” Esse mesmo autor conclui suas colocações sobre esse amplo paradoxo ao afirmar que “o *clustering* está nos olhos de quem vê”, sendo assim o *clustering* deve “envolver as necessidades do usuário ou da aplicação”. Ou seja, os agrupamentos (*clustering*) devem servir aos interesses específicos da pesquisa a que se destinam.

Jain (2010, p. 656, tradução nossa), sustenta que em se tratando da análise de dados a partir de textos, essa questão é ainda mais relevante, problemática e difícil, e nesse caso “infelizmente não existe uma representação universalmente boa” e aponta que a escolha do número de *clusters* “deve ser orientada pelo conhecimento do domínio”.

A partir dessa realidade, para a escolha do número de *clusters* adequado a ser trabalhado nessa pesquisa, de forma a obter se os melhores resultados possíveis, utilizou-se de duas abordagens, que juntas possibilitaram a identificação, senão do número ideal, do número de clusters mais próximo possível dessa visão. Para isso lançou-se mão, de um lado, das questões suportadas pelo conhecimento do domínio adquiridas ao longo de todo o processo de levantamento e análise de dados dessa pesquisa; e de outro lado foram realizadas várias execuções experimentais conduzidas com o objetivo de juntamente com a lógica do conhecimento de domínio encontrar o melhor número de *clusters* possível. (Jain, 2010)

Nesse sentido, foram realizadas quatro experimentações, com a seguinte quantidade de clusters: dez, sete, cinco e três. Os resultados encontrados estão demonstrados nos Quadros 4 e 5.

**QUADRO 4 - Resultados dos experimentos 1 e 2**

| **EXPERIMENTO 1 - 10 *CLUSTERS*** | | |
|---|---|---|
| Cluster | Qte Políticas | % |
| 0 | 221 | 25,1 |
| 1 | 88 | 10 |
| 2 | 3 | 0,3 |
| 3 | 150 | 17 |
| 4 | 35 | 4 |
| 5 | 185 | 21 |
| 6 | 5 | 0,6 |
| 7 | 174 | 19,7 |
| 8 | 19 | 2,2 |
| 9 | 2 | 0,2 |
| Total | 882 | 100 |
| ***SILHOUETTE SCORE*: 0,34** | | |

| **EXPERIMENTO 2 - 7 *CLUSTERS*** | | |
|---|---|---|
| Cluster | Qte Políticas | % |
| 0 | 212 | 24 |
| 1 | 132 | 15 |
| 2 | 44 | 5 |
| 3 | 238 | 27 |
| 4 | 6 | 0,7 |
| 5 | 247 | 28 |
| 6 | 3 | 0,3 |
| Total | 882 | 100 |
| ***SILHOUETTE SCORE*: 0,39** | | |

Fonte: elaboração própria

**QUADRO 5 - Resultados dos experimentos 3 e 4**

| **EXPERIMENTO 3 - 5 *CLUSTERS*** | | |
|---|---|---|
| Cluster | Qte Políticas | % |
| 0 | 323 | 36,6 |
| 1 | 53 | 6 |
| 2 | 305 | 34,6 |
| 3 | 194 | 22 |
| 4 | 7 | 0,8 |
| Total | 882 | 100 |
| ***SILHOUETTE SCORE*: 0,43** | | |

| **EXPERIMENTO COM 3 *CLUSTERS*** | | |
|---|---|---|
| Cluster | Qte Políticas | % |
| 0 | 349 | 39,6 |
| 1 | 477 | 54,1 |
| 2 | 56 | 6,3 |
| Total | 882 | 100 |
| ***SILHOUETTE SCORE*: 0,50** | | |

Fonte: elaboração própria

Na análise dos resultados dos experimentos expostos de forma sintética acima, sob os dois aspectos utilizados para seleção do número de clusters (conhecimento do domínio e realização e análise dos experimentos), foi selecionado como o mais adequado para os objetivos dessa pesquisa o experimento com cinco *clusters* (experimento 3).

Foi conduzida uma rápida análise qualitativa ao nível dos resultados encontrados pelos experimentos, que ajudou na confirmação de que a abordagem realizada a partir de cinco *clusters*, guardava um nível mais adequado de similaridade entre seus elementos, que ao mesmo tempo, não estavam dispersos demais e nem concentrados demais, o que proporcionaria uma menor perda da capacidade de análise; além do fato de estarem mais alinhados ao nosso conhecimento do domínio. Conforme colocado por Gonçalves *et al* (2018, p.9), para chegar-se a essas e outras conclusões é necessária uma análise qualitativa "do detalhamento de cada agrupamento".

Outro ponto fundamental analisado foi o indicador "*silhouette score*". O "*silhouette score*" varia de -1 a 1. De maneira resumida, este indicador é analisado da seguinte forma: 1) o valor da pontuação igual a 1, o *cluster* é denso e bem separado dos outros *clusters*; 2) valor da pontuação próximo a 0, remete a *clusters* sobrepostos, ou seja, com amostras muito próximas ao limite de decisão dos *clusters* vizinhos; 3) o

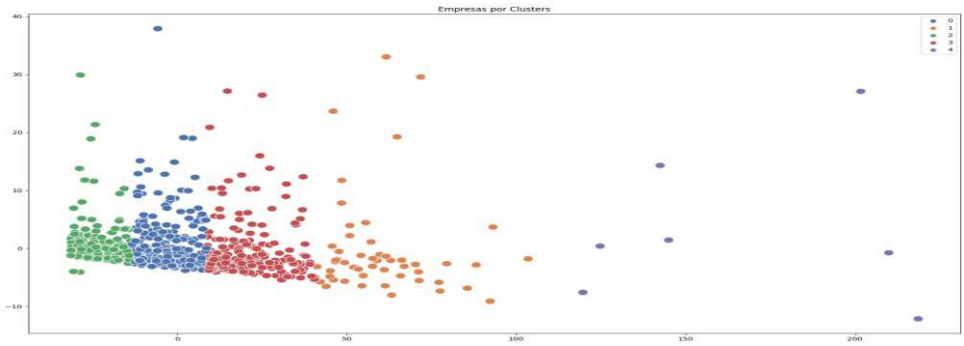
valor da pontuação é -1, nesse caso temos um indicativo de que as amostras podem ter sido atribuídas a *clusters* errados.

Conforme defendido por Mariani, Navrotska e Mancini *et al* (2023, p. 4, tradução nossa), “o coeficiente de silhueta é uma medida interna para validação de *cluster* que considera as distâncias intra-cluster e inter-cluster”. Em outras palavras, poderíamos entender como uma medida do quanto um “elemento analisado” se encaixa em um determinado *cluster*, ou ainda como uma avaliação da distância de cada *cluster* (a princípio quanto maior a distância, melhores os resultados).

Pela análise dos indicadores “*silhouette score*” de todos os experimentos, é possível verificar uma característica inerente a todos eles, que é a sobreposição, ou seja, praticamente todos estão próximos a zero. Nesse sentido, como indicado acima, na explicação desse indicador, as amostras estão próximas ao limite de decisão dos clusters vizinhos. Essa constatação possibilitou diversas conclusões, assim como ajudou a explicar em muito uma realidade já percebida pela análise qualitativa.

Na Figura 8 abaixo, é possível visualizar essa constatação, ou seja, é possível verificar de forma visual a proximidade e relativa sobreposição dos clusters presentes no experimento cinco:

**FIGURA 8 - Visualização da proximidade e relativa sobreposição dos clusters do experimento cinco**




Fonte: dados da pesquisa, elaboração própria, 2023

Cabe ressaltar que tal proximidade não significa que os *clusters* não têm relevância ou não podem ser interpretados, mas sim que nesses casos é necessário exatamente o que foi realizado aqui, ou seja, uma análise humana (conhecimento do domínio, análise qualitativa e interpretação do indicador "*silhouette score*") para se chegar à decisão do número ideal de *clusters*, para os objetivos dessa pesquisa.

Um achado importante dessa análise e seleção dos *clusters*, foi a visualização do que será denominado nesse trabalho de "privacidade opaca", ou seja, as maiores empresas do Brasil, independente do seu setor, tamanho ou localidade, de maneira geral tem ações, atividades e processos declarados em suas "Políticas de Privacidade", sem muita diferenciação, remetendo a uma sensação de "*copy* cola" com base na referência às questões mais genéricas e/ou mais básicas das diretrizes constantes na LGPD. A questão aqui colocada, como forma de materialização do raciocínio que se quer construir, remete ao seguinte pensamento/questão, "o que precisamos fazer constar na Política de Privacidade da nossa organização de forma a ficarmos livres de problemas, processo e/ou multas das principais questões constantes na LGPD?".

Salvo honrosas exceções, não é possível ter a mínima noção do que as organizações fazem com os dados de seus clientes, fornecedores, parceiros, etc.; no âmbito das análises trazidas pela abordagem do *big data analytics*. E mesmo no caso de algumas poucas organizações que conseguem trazer alguns indicativos disso, eles não são suficientes para possibilitar esse entendimento. Por isso, pode-se concluir, tomando por base o conjunto das maiores empresas do Brasil, que se tem uma situação de "privacidade opaca" no Brasil, ou seja, temos alguns indicativos aderentes à legislação em vigor, mas não é possível saber efetivamente se tais organizações conseguem ou não dar garantias reais à privacidade no contexto das análises avançadas de dados internamente realizadas.

Em graus diferentes para as diversas realidades, isso parece não ser um problema exclusivamente das grandes empresas do Brasil, como mostram os resultados de um estudo anual realizado pela empresa de tecnologia Cisco, "*Data Privacy Benchmark Study: Privacy Becomes Mission Critical (2022)*", onde, por um lado, 87% das empresas entrevistadas acreditam que já possuem processos em vigor para garantir que as decisões tomadas de forma automatizada estão alinhadas às expectativas de seus clientes. Por outro lado, 46% dos consumidores, ou seja, quase a metade dos consumidores entrevistados (com base em outra pesquisa da Cisco, sobre privacidade do consumidor, realizada no ano de 2021) não se sentem protegidos adequadamente quanto a seus dados. A principal razão disso é que não entendem de forma clara o que as organizações estão coletando e fazendo com seus dados.

Sob essa ótica, a "opacidade" seria uma forma de estar oficialmente alinhado à legislação em vigor, nas questões eminentemente necessárias, ao mesmo tempo que impossibilita de forma proposital o entendimento de como estão sendo utilizados os dados de seus consumidores, principalmente.

Pensando no grupo de empresas do Brasil analisados neste trabalho, se existe alguma forma de garantia da privacidade no âmbito do *big data analytics*, refletida

principalmente nas decisões automatizadas, por que tais processos, ações e práticas não estão refletidas em suas Políticas de Privacidade?

A conclusão, baseada em fatos trazidos por essa pesquisa, é que elas não dispõem de tais práticas, e, portanto, não há qualquer garantia de privacidade quanto a utilização dos dados pessoais de consumidores, principalmente, mas de muitas outras classes. E se o foco do entendimento for a garantia da privacidade no âmbito específico do *big data analytics*, aí então esse trabalho corrobora a convicção de que essa garantia não existe, ou não é efetivamente demonstrada.

### 9.3.1 Análise do conjunto de termos constituídos para a construção dos clusters

Após os múltiplos processos de preparação e principalmente de construção dos termos-chave para suporte ao processo de agrupamento, chegou-se ao seguinte conjunto de termos-chave, que serão apresentados abaixo, conforme a dimensão do modelo do “Pentágono da Privacidade” a que pertencem, e na sequência serão analisados e apontados seus principais resultados.

No primeiro agrupamento todos os termos integram a dimensão “Processos de *Big Data Analytics*”, conforme demonstrado na sequência pela Figura 9.

**FIGURA 9 - Termos-chave no âmbito da dimensão "Processos de *Big Data Analytics*"**

| Termos de Pesquisa | Palavras Chaves | Processos |
|---|---|---|
| Smart Data | Smart Data | PROCESSOS DE BIG DATA |
| Zonas proibidas | Zonas proibidas | PROCESSOS DE BIG DATA |
| Zona proibida | Zonas proibidas | PROCESSOS DE BIG DATA |
| Identificação antecipada | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| antecipada projetos | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| antecipada projeto | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| antecipada conjuntos | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| antecipada conjunto | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| dados pessoais | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| dado pessoal | Identificação antecipada de projetos e conjuntos de dados que contém dados pessoais | PROCESSOS DE BIG DATA |
| Armazenamento seguro | Armazenamento seguro (garantia de) | PROCESSOS DE BIG DATA |
| garantia Armazenamento | Armazenamento seguro (garantia de) | PROCESSOS DE BIG DATA |
| Compartilhamento seguro | Compartilhamento seguro (garantia de) | PROCESSOS DE BIG DATA |
| garantia Compartilhamento | Compartilhamento seguro (garantia de) | PROCESSOS DE BIG DATA |
| Transparência algorítmica | Transparência algortimica (garantia de) | PROCESSOS DE BIG DATA |
| garantia Transparência | Transparência algortimica (garantia de) | PROCESSOS DE BIG DATA |
| Gerenciamento riscos | Gerenciamento riscos em big data | PROCESSOS DE BIG DATA |
| Riscos big data | Gerenciamento riscos em big data | PROCESSOS DE BIG DATA |
| Discriminação algorítmica | Discriminação algoritmica (garantia de não ocorrência) | PROCESSOS DE BIG DATA |
| garantia ocorrência | Discriminação algoritmica (garantia de não ocorrência) | PROCESSOS DE BIG DATA |
| Princípio finalidade | Princípio da finalidade (é observado) | PROCESSOS DE BIG DATA |
| Princípio adequação | Princípio da adequação (é observado) | PROCESSOS DE BIG DATA |
| Princípio necessidade | Princípio da necessidade (é observado) | PROCESSOS DE BIG DATA |
| Dados portadores | Dados portadores de política | PROCESSOS DE BIG DATA |
| Dado portador | Dados portadores de política | PROCESSOS DE BIG DATA |
| portadores política | Dados portadores de política | PROCESSOS DE BIG DATA |
| portador política | Dados portadores de política | PROCESSOS DE BIG DATA |
| Parâmetros permissões | Parâmetros para permissões de transmissão (são observados) | PROCESSOS DE BIG DATA |
| Parâmetro permissão | Parâmetros para permissões de transmissão (são observados) | PROCESSOS DE BIG DATA |
| permissões transmissão | Parâmetros para permissões de transmissão (são observados) | PROCESSOS DE BIG DATA |
| permissão transmissão | Parâmetros para permissões de transmissão (são observados) | PROCESSOS DE BIG DATA |
| Parâmetros uso | Parâmetros para uso e descarte (são observados) | PROCESSOS DE BIG DATA |
| Parâmetro uso | Parâmetros para uso e descarte (são observados) | PROCESSOS DE BIG DATA |
| Parâmetros descarte | Parâmetros para uso e descarte (são observados) | PROCESSOS DE BIG DATA |
| Parâmetro descarte | Parâmetros para uso e descarte (são observados) | PROCESSOS DE BIG DATA |
| Procedimentos ciclo | Procedimentos ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Procedimento ciclo | Procedimentos ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| ciclo big data | Procedimentos ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| proteção privacidade | Procedimentos ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| garant proteção | Procedimentos ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Políticas ciclo | Políticas ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Política ciclo | Políticas ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| ciclo big data | Políticas ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Políticas ciclo big data | Políticas ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Política ciclo big data | Políticas ciclo do big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Prática ciclo | Prática ciclo big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |
| Prática ciclo big data | Prática ciclo big data (que garantem a proteção da privacidade) | PROCESSOS DE BIG DATA |

Fonte: dados da pesquisa, elaboração própria, 2023

No segundo agrupamento todos os termos integram a dimensão “Governança em Privacidade”, conforme demonstrado na sequência pela Figura 10.

**FIGURA 10 - Termos-chave no âmbito da dimensão “Governança em Privacidade”**

| Termos de Pesquisa | Palavras Chaves | Processos |
|---|---|---|
| Políticas internas | Políticas internas (possui políticas internas que ordenam a garantia da privacidade | GOVERNANÇA EM PRIVACIDADE |
| Política interna | Políticas internas (possui políticas internas que ordenam a garantia da privacidade | GOVERNANÇA EM PRIVACIDADE |
| garantia privacidade | Políticas internas (possui políticas internas que ordenam a garantia da privacidade | GOVERNANÇA EM PRIVACIDADE |
| Controle Proprietário | Controle do Proprietário (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| Acesso titular | Acesso ao titular de dados (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| titular dados | Acesso ao titular de dados (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| titular dado | Acesso ao titular de dados (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| Acesso titular dados | Acesso ao titular de dados (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| Acesso titular dado | Acesso ao titular de dados (é garantido) | GOVERNANÇA EM PRIVACIDADE |
| Treinamento | Treinamento (constante sobre privacidade no que tange aos profissionais da organização) | GOVERNANÇA EM PRIVACIDADE |
| Treinamento constante | Treinamento (constante sobre privacidade no que tange aos profissionais da organização) | GOVERNANÇA EM PRIVACIDADE |
| Treinamento privacidade | Treinamento (constante sobre privacidade no que tange aos profissionais da organização) | GOVERNANÇA EM PRIVACIDADE |
| Encarregado dados | Encarregado de dados (possui um encarregado de dados, próprio ou terceiro) | GOVERNANÇA EM PRIVACIDADE |
| Encarregado dado | Encarregado de dados (possui um encarregado de dados, próprio ou terceiro) | GOVERNANÇA EM PRIVACIDADE |
| DPO | Encarregado de dados (possui um encarregado de dados, próprio ou terceiro) | GOVERNANÇA EM PRIVACIDADE |
| Privacy by design | Privacy by design -Privacidade por princípio/design (abordagem onde a privacidade é a base de qualquer iniciativa, ex.: desenvolvimento solução, software, etc.) | GOVERNANÇA EM PRIVACIDADE |
| Privacidade princípio | Privacy by design -Privacidade por princípio/design (abordagem onde a privacidade é a base de qualquer iniciativa, ex.: desenvolvimento solução, software, etc.) | GOVERNANÇA EM PRIVACIDADE |
| Privacidade design | Privacy by design -Privacidade por princípio/design (abordagem onde a privacidade é a base de qualquer iniciativa, ex.: desenvolvimento solução, software, etc.) | GOVERNANÇA EM PRIVACIDADE |
| Comitê Ética Revisão | Comitê de Ética e Revisão (no âmbito do big data, a organização possui) | GOVERNANÇA EM PRIVACIDADE |
| Comitê Ética | Comitê de Ética e Revisão (no âmbito do big data, a organização possui) | GOVERNANÇA EM PRIVACIDADE |
| Comitê Revisão | Comitê de Ética e Revisão (no âmbito do big data, a organização possui) | GOVERNANÇA EM PRIVACIDADE |
| Inventário dados | Investário de dados (a organização realiza) | GOVERNANÇA EM PRIVACIDADE |
| Inventário dado | Investário de dados (a organização realiza) | GOVERNANÇA EM PRIVACIDADE |
| Corretores dados | Corretores de dados (a empresa possui ações para seu uso adequado) | GOVERNANÇA EM PRIVACIDADE |
| Corretor dado | Corretores de dados (a empresa possui ações para seu uso adequado) | GOVERNANÇA EM PRIVACIDADE |
| Corretores dado | Corretores de dados (a empresa possui ações para seu uso adequado) | GOVERNANÇA EM PRIVACIDADE |
| Protocolos segurança fornec | Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade | GOVERNANÇA EM PRIVACIDADE |
| Protocolo segurança fornece | Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade | GOVERNANÇA EM PRIVACIDADE |
| Protocolos segurança | Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade | GOVERNANÇA EM PRIVACIDADE |
| segurança fornecedores | Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade | GOVERNANÇA EM PRIVACIDADE |
| segurança fornecedor | Protocolos de segurança quanto aos fornecedores (de forma a garantir a privacidade | GOVERNANÇA EM PRIVACIDADE |

Fonte: dados da pesquisa, elaboração própria, 2023

No terceiro agrupamento todos os termos integram a dimensão “Técnicas criptográficas e não criptográficas”, conforme demonstrado na sequência pela Figura 11.

**FIGURA 11 - Termos-chave no âmbito da dimensão “Técnicas criptográficas e não criptográficas”**

| Termos de Pesquisa | Palavras Chaves | Processos |
|---|---|---|
| Desidentificação | Desidentificação | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Pseudonimização | Pseudonimização | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Anonimização | Anonimização | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Generalização | Generalização | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Randomização | Randomização | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Triple DES | Triple DES | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Triple | Triple DES | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| DES | Triple DES | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelo execução híbrida | Modelo de execução híbrida - HybrEx | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelo execução | Modelo de execução híbrida - HybrEx | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| execução híbrida | Modelo de execução híbrida - HybrEx | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| HybrEx | Modelo de execução híbrida - HybrEx | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| hybrid | Modelo de execução híbrida - HybrEx | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelos classificação dados | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelo classificação dado | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| classificação dado sensível | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| classificação dados sensíveis | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelo classificação | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| classificação dado | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| dado sensível | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Modelos classificação | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| classificação dados | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| dados sensíveis | Modelos de classificação de dados sensíveis | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Técnicas limitação divulgaçã | Técnicas de limitação de divulgação estatística | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| limitação divulgação estatísti | Técnicas de limitação de divulgação estatística | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| Técnica limitação | Técnicas de limitação de divulgação estatística | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| limitação divulgação | Técnicas de limitação de divulgação estatística | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |
| divulgação estatística | Técnicas de limitação de divulgação estatística | TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS |

Fonte: dados da pesquisa, elaboração própria, 2023

No quarto agrupamento todos os termos integram a dimensão “Técnicas e processos de segurança, adotadas na organização”, conforme demonstrado na sequência pela Figura 12.

**FIGURA 12 - Termos-chave no âmbito da dimensão "Técnicas e processos de segurança, adotadas na organização"**

| Termos de Pesquisa | Palavras Chaves | Processos |
|---|---|---|
| Structured Map Reduced Layer | Structured Map Reduced Layer | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Segurança dados | Segurança dos dados (dados protegidos por salvaguardas de segurança) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Segurança dado | Segurança dos dados (dados protegidos por salvaguardas de segurança) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Controle acesso | Controle de acesso (autenticaçao, autorizaçao auditoria, MAC) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| autenticação | Controle de acesso (autenticaçao, autorizaçao auditoria, MAC) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| autorização auditoria | Controle de acesso (autenticaçao, autorizaçao auditoria, MAC) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| MAC | Controle de acesso (autenticaçao, autorizaçao auditoria, MAC) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Segurança informação | Segurança da informação | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Técnicas ataque big data | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| Técnicas ataque | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| ataque big data | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| link lateral | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| palpite óbvio | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |
| reidentificação | Técnicas de ataque em big data (link lateral, palpite óbvio, reidentificação, etc.) | TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADAS NA ORGANIZAÇÃO |

Fonte: dados da pesquisa, elaboração própria, 2023

No quinto agrupamento todos os termos integram a dimensão "Privacidade por Política", conforme demonstrado na sequência pela Figura 13.

**FIGURA 13 - Termos-chave no âmbito da dimensão "Privacidade por Política"**

| Termos de Pesquisa | Palavras Chaves | Processos |
|---|---|---|
| Confidencialidade | Confidencialidade | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Especificação propósito | Especificação de propósito | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Processos comunicação | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Processo comunicação | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| comunicação roubo dados | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| comunicação roubo dado | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| comunicação vazamento | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| roubo dados | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| roubo dado | Processos de comunicação vazamento ou roubo de dados | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Consentimento | Consentimento (garante que o indivíduo dê seu | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Aviso | Aviso (informa ao titular sobre a coleta de dados) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Minimização dados | Minimização de dados (impõe limitações sobre o tipo de informações a serem coletadas) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Minimização dado | Minimização de dados (impõe limitações sobre o tipo de informações a serem coletadas) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| populações vulneráveis | Minimização de dados (impõe limitações sobre o tipo de informações a serem coletadas) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| população vulnerável | Cuidado análise dados de populações vulneráveis (crianças, grávidas, prisioneiros, LGTB, etc.) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Políticas disciplinam | Políticas que disciplinam quem pode fazer tratamento dos | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| fazer tratamento dados | Políticas que disciplinam quem pode fazer tratamento dos | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Retificação | Retificação (permite que o titular realize a retificação dos seus dados) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |
| Retifica | Retificação (permite que o titular realize a retificação dos seus dados) | PRIVACIDADE POR POLÍTICA - FIPP (Fair Information Practices Principals) |

Fonte: dados da pesquisa, elaboração própria, 2023

A consolidação e análise das citações dos termos-chave presentes nas "Políticas de Privacidade" das organizações integrantes do ranking Valor1000 (2021) resultou na construção da Tabela 4, abaixo:

**TABELA 4 - Termos x Nº Citações x Nº Empresas x Dimensão do Pentágono da Privacidade no *Big Data Analytics***

| TERMO | CITAÇÕES | Nº EMPRESAS | DIMENSÃO |
|---|---|---|---|
| dado pessoal | 27964 | 856 | *PROCESSOS DE BIG DATA* |
| consentimento | 4349 | 771 | PRIVACIDADE POR POLÍTICA |
| aviso | 1862 | 382 | PRIVACIDADE POR POLÍTICA |
| dpo | 836 | 333 | GOVERNANÇA EM PRIVACIDADE |
| seguranca informacao | 814 | 349 | TÉCNICAS E PROCESSOS DE SEGURANÇA |
| anonimizacao | 800 | 509 | TÉCNICAS CRIPTOGRÁFICAS |
| confidencialidade | 747 | 420 | PRIVACIDADE POR POLÍTICA |
| seguranca dado | 338 | 218 | TÉCNICAS E PROCESSOS DE SEGURANÇA |
| retifica | 272 | 194 | PRIVACIDADE POR POLÍTICA |
| autenticacao | 255 | 164 | TÉCNICAS E PROCESSOS DE SEGURANÇA |
| dado sensivel | 190 | 106 | TÉCNICAS CRIPTOGRÁFICAS |
| protecao privacidade | 156 | 121 | PROCESSOS DE *BIG DATA* |
| controle acesso | 138 | 100 | TÉCNICAS E PROCESSOS DE SEGURANÇA |
| disciplina | 138 | 101 | PRIVACIDADE POR POLÍTICA |
| politica interna | 77 | 64 | GOVERNANÇA EM PRIVACIDADE |
| acesso titular dado | 21 | 7 | GOVERNANÇA EM PRIVACIDADE |
| minimizacao dado | 21 | 20 | PRIVACIDADE POR POLÍTICA |
| protocolo seguranca | 20 | 18 | GOVERNANÇA EM PRIVACIDADE |
| pseudonimizacao | 17 | 12 | TÉCNICAS CRIPTOGRÁFICAS |
| principio necessidade | 14 | 13 | PROCESSOS DE *BIG DATA* |
| comite etica | 12 | 11 | GOVERNANÇA EM PRIVACIDADE |
| gerenciamento risco | 11 | 6 | PROCESSOS DE *BIG DATA* |
| armazenamento seguro | 10 | 9 | PROCESSOS DE *BIG DATA* |
| garantia privacidade | 7 | 7 | GOVERNANÇA EM PRIVACIDADE |
| privacidade design | 7 | 7 | GOVERNANÇA EM PRIVACIDADE |
| privacidade principio | 7 | 7 | GOVERNANÇA EM PRIVACIDADE |
| principio finalidade | 6 | 6 | PROCESSOS DE *BIG DATA* |
| seguranca fornecedor | 4 | 4 | GOVERNANÇA EM PRIVACIDADE |
| inventario dado | 4 | 2 | GOVERNANÇA EM PRIVACIDADE |
| compartilhamento seguro | 3 | 1 | PROCESSOS DE *BIG DATA* |
| desidentificacao | 3 | 3 | TÉCNICAS CRIPTOGRÁFICAS |
| classificacao dado | 2 | 2 | TÉCNICAS CRIPTOGRÁFICAS |
| corretor dado | 2 | 2 | GOVERNANÇA EM PRIVACIDADE |
| processo comunicacao | 2 | 2 | PRIVACIDADE POR POLÍTICA |
| treinamento privacidade | 2 | 2 | GOVERNANÇA EM PRIVACIDADE |
| generalizacao | 1 | 1 | TÉCNICAS CRIPTOGRÁFICAS |

| | | | |
|---|---|---|---|
| protocolo | 1 | 1 | GOVERNANÇA EM PRIVACIDADE |
| roubo dado | 1 | 1 | PRIVACIDADE POR POLÍTICA |
| politica ciclo | 1 | 1 | PROCESSOS DE *BIG DATA* |
| principio adequacao | 1 | 1 | PROCESSOS DE *BIG DATA* |

Fonte: dados da pesquisa, elaboração própria, 2023

Uma análise atenta da Tabela 4, nos permite apontar achados muito relevantes e que dentre outras questões caras a essa pesquisa, remetem diretamente ao modelo teórico proposto pela lógica do "Pentágono da Privacidade no *Big Data Analytics*"; assim como remete à questão da opacidade no âmbito da privacidade. No que tange à primeira questão será feita análise em tópico específico mais à frente.

No que tange à questão da aqui denominada "privacidade opaca", foi possível verificar que os termos que remetem a um maior e mais destacado número de citações (dado pessoal, consentimento, aviso e "dpo") são aqueles usualmente introdutórios ou mais "básicos", se pudéssemos assim dizer, no contexto da LGPD e também da garantia da privacidade, principalmente, como é o foco deste trabalho, se a análise for realizada tendo como pano de fundo o *big data analytics*. São as questões iniciais que conseguem fazer um vínculo direto à legislação e que naturalmente tendem a ser muito citadas.

Por exemplo, "dado pessoal", que é o cerne de todo o foco da privacidade, ou seja, tanto a abordagem "lato sensu", quanto a legislação em vigor só se aplicam aos dados pessoais, referentes a uma pessoa física. Pessoas jurídicas, não são contempladas por questões de privacidade e, portanto, muito menos pela legislação que tenta protegê-la.

No caso do "consentimento", não só, mas principalmente, no âmbito da coleta de dados pessoais pelos sites e outras plataformas das organizações, foi uma das primeiras adequações realizadas pelas organizações. Ou seja, grande parte delas solicita o consentimento dos indivíduos para captura de diversos dados sobre ele, por meio dos mecanismos tecnológicos disponíveis em seus sites e demais plataformas. Mesmo que em muitos casos não cumpram adequadamente as condicionantes de

obtenção deste “consentimento”. Como foi possível verificar, 97% das Políticas de Privacidade das empresas pesquisadas neste trabalho fazem uma ou mais citações ao termo.

Analisando a lista dos “termos-chave” utilizados para a construção dos agrupamentos (figuras 7 a 11) *vis a vis* a lista daqueles que estão presentes nas Políticas de Privacidade das organizações, pelo menos uma vez, como demonstrado pela Tabela 4, é possível verificar mais duas questões muito relevantes, a primeira é que nem todos os “termos-chave” inicialmente listados estão presentes nas “Políticas de Privacidade” das organizações analisadas; ou seja, eles não apareceram nem uma vez nessas políticas.

Fica nítido, portanto, que do conjunto de proposições que compõem as dimensões do “Pentágono da Privacidade no *Big Data Analytics*”, apenas uma parte delas estão presentes, mesmo que de maneira incipiente na maioria dos casos, na realidade das grandes empresas do Brasil, isso tomando por base, claro, a análise de suas respectivas “Políticas de Privacidade”.

Além disso, a maioria dos termos-chave estão presentes nas Políticas de Privacidade em uma frequência baixa, ou seja, estão presentes, mas de maneira pouco representativa, o que denota uma certa fragilidade e efetiva aplicação dos conceitos e práticas a que remetem esses termos e que constituem as dimensões do Polígono da Privacidade no *Big Data Analytics*.

Ao que parece as organizações ainda não se atentaram para a nova frente que avança e que colocará a privacidade novamente sob severa ameaça, caso alternativas não sejam construídas desde já, que é a frente do *big data analytics*: *Inteligência Artificial (AI), machine learning e deep learning*, principalmente, que tem possibilitado múltiplas frentes e soluções baseadas em tomadas de decisões automatizadas.

### 9.3.2 Termos mais relevantes por *cluster*

Os principais termos por *cluster*, ou seja, aqueles mais representativos, que mais apareceram nas políticas de privacidade das organizações presentes em cada *cluster*, foram constituídos a partir do centroide de seus respectivos clusters.

**FIGURA 14 - Top termos por *cluster***

| Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| dado pessoal | dado pessoal | dado pessoal | dado pessoal | dado pessoal |
| consentimento | consentimento | consentimento | consentimento | consentimento |
| aviso | aviso | aviso | aviso | aviso |
| anonimização | anonimização | anonimização | anonimização | anonimização |
| dpo | dpo | dpo | dpo | dpo |
| confidencialidade | confidencialidade | confidencialidade | confidencialidade | confidencialidade |
| segurança informação | segurança informação | segurança informação | segurança informação | segurança informação |
| segurança dado | segurança dado | segurança dado | segurança dado | segurança dado |
| autenticação | autenticação | | autenticação | |
| retifica | | retifica | retifica | retifica |
| | dado sensível | | | dado sensível |
| | | disciplina | | |

Fonte: dados da pesquisa, elaboração própria, 2023

A Figura 14 acima, mostra os dez termos mais presentes, de maior relevância no conjunto de empresas de cada um dos *clusters* analisados. A análise dos principais termos presentes nos *clusters* analisados, reforça de maneira clara e objetiva as constatações feitas anteriormente, mostrando a clara proximidade dos clusters, que se diferem pela presença ou ausência de poucos termos diferenciados, no conjunto de seus "*top ten*" termos.

Podemos constatar uma certa pasteurização dos conteúdos presentes nas "Políticas de Privacidade", como já pontuado, com honrosas exceções, mas essas também ainda muito prematuras e insuficientes no que tange à garantia da privacidade no *big data analytics*.

Numa rápida análise dos "*top ten*" termos, podemos constatar que 60% dos termos mais utilizados nas "Políticas de Privacidade" no conjunto de empresas

pesquisadas são os mesmos, apresentando apenas pequenas variações na intensidade da utilização e presença das mesmas em suas respectivas “Políticas de Privacidade”. Eles são a síntese geral da lógica proposta da opacidade. Ou seja, de forma geral as empresas analisadas “falam” a mesma coisa, a diferença está no grau de importância dado a uma determinada prática ou outra, vistas como exceções.

### 9.3.3 Um olhar mais aprofundado sobre os *clusters*

Devido à proximidade e relativa sobreposição dos *clusters* analisados, conforme já explicado anteriormente, no experimento 5 e nos demais experimentos, a tarefa de nominá-los e caracterizá-los de uma maneira única é bem complexa. De qualquer forma, é possível através de suas pequenas diferenças, diferenciá-los.

O foco aqui foram os clusters do experimento 5, que foi aquele escolhido como a melhor representação para os objetivos dessa pesquisa e sobre os quais toda a análise e discussão dos resultados esteve baseada.

A análise de algumas dessas pequenas diferenças permitiu criar uma denominação específica para cada um dos *clusters* do experimento 5, como colocado abaixo:

- *Cluster* 0: Pentágono Opaco 50
- *Cluster* 1: Pentágono Opaco 30
- *Cluster* 2: Pentágono Opaco 40
- *Cluster* 3: Pentágono Opaco 51
- *Cluster* 4: Pentágono Opaco 20

A nominação data a cada um dos *clusters* pretende dar a exata medida da maior ou menor contribuição para a confirmação do modelo dado pelo Pentágono da Privacidade no *Big Data Analytics*, assim como do maior ou menor grau de opacidade de cada cluster.

Um olhar mais atento sobre a Figura 8, nos permite concluir que dos cinco clusters analisados, no experimento 5, as empresas que compõem o cluster 4 (Pentágono Opaco 20), foram agrupadas nesse cluster por se tratarem dos outliers,

ou seja, o algoritmo agrupou nesse cluster todas as empresas com centróides mais dispersos que não conseguiram ser agrupados nos demais clusters. Individualmente e em conjunto são o grupo mais diferente dos demais.

A análise dos segmentos econômicos das empresas que compunham cada um dos *clusters*, permitiu obter uma visão também muito rica das características de cada grupo, assim como um reforço da assertividade da nomenclatura atribuída a cada um deles.

Foram os segmentos mais representativos em cada cluster, em ordem decrescente quanto ao seu grau de opacidade:

- Pentágono Opaco 50 (cluster 0): Comércio Varejista, Transporte e Logística, Agronegócio
- Pentágono Opaco 51 (cluster 3): Comércio Varejista, Transporte e Logísticas, Agronegócio
- Pentágono Opaco 40 (cluster 2): Alimentos e Bebidas, Agronegócio, Comércio Varejista
- Pentágono Opaco 30 (cluster 1): Química e Petroquímica, Agronegócio, Serviços Especializados

No caso do Pentágono Opaco 4, ele é composto por 7 empresas, cada uma de um segmento e todas com a mesma representatividade: Bioenergia, Comércio Varejista, Metalurgia e Siderurgia, Mineração, Petróleo e Gás, Química e Petroquímica, Veículos e Peças.

Sendo assim, foi possível destacar alguns achados:

- O segmento do Agronegócio está presente em todos os *clusters*, à exceção do *Cluster* 4 (Pentágono Opaco 20);
- Os *clusters* 0 e 3 (Pentágono Opaco 50 e Pentágono Opaco 51), possuem os mesmos segmentos de destaque;

- À exceção do Pentágono Opaco 30 (*cluster* 1), onde não está presente, e do Pentágono Opaco 20 (*cluster* 4), onde ele tem a mesma intensidade dos demais; o comércio varejista é um dos setores de destaque nos *clusters*, 0 (Pentágono Opaco 50), 3 (Pentágono Opaco 51) e 2 (Pentágono Opaco 40).

É possível verificar que não existe um padrão com base nos respectivos segmentos econômicos das empresas analisadas. E que, as diferenças existentes, independente do seu setor, se dão ao nível das empresas vistas de forma individual, com base em suas características de gestão, visão e estratégia; e principalmente com base na forma como vêem e lidam com a questão da privacidade, conforme pode ser verificado pela análise qualitativa realizada de diversas políticas de privacidade.

A lógica de escalonamento dada aos clusters se confirma quando analisamos os termos de referência, derivados das práticas previstas em cada dimensão do modelo de privacidade proposto pelo Pentágono da Privacidade no *Big Data Analytics*, que aparecem de forma diferenciada em cada cluster. No Quadro 6, na sequência foi realizada a identificação de cada termo relacionando-o à respectiva dimensão do modelo proposto pelo Pentágono da Privacidade no *Big Data Analytics* e depois realizaremos uma análise dos mesmos *vis a vis* o foco nas diferenças entre os *clusters*:

**QUADRO 6 - Relação termo busca x prática x dimensão modelo**

| TERMO | PRÁTICA | DIMENSÃO |
|---|---|---|
| **AUTENTICAÇÃO** | Controle de acesso (autenticação, autorização, auditoria, MAC | Técnicas e processos de segurança adotadas na organização |
| **RETIFICA** | Retificação (permite que o titular realize a retificação dos seus dados) | Privacidade por Política – FIPP (*Fair Information Practices Principals*) |
| **DADO SENSÍVEL** | Modelos de classificação de dados sensíveis | Técnicas criptográficas e não criptográficas |
| **DISCIPLINA** | Políticas que disciplinam quem pode fazer tratamento dos dados | Privacidade por Política – FIPP (*Fair Information Practices Principals*) |

Fonte: dados da pesquisa, elaboração própria, 2023

É possível classificar os termos constantes no Quadro 6, como mais sofisticados e menos sofisticados no que tange à visão da garantia da privacidade no âmbito *do big data analytics*. Nesse sentido, os termos “dado sensível” e “disciplina” estão mais alinhados a questões mais sofisticadas no sentido das garantias à privacidade no âmbito do *big data analytics*, e, portanto, os *clusters* que os contém tendem a ser menos “opacos” e mais “avançados” no que tange às garantias à privacidade no *big data analytics;* além de garantirem maior alinhamento ao modelo do Pentágono da Privacidade no *Big Data Analytics*.

Essas colocações ajudam a entender porque os clusters 1 (Pentágono Opaco 30) e 4 (Pentágono Opaco 20) possuem um menor número de empresas que poderia-se induzir serem mais avançadas na perspectiva das garantias da privacidade no âmbito do *big data analytics* e também diferem dos demais em termos de relevância dos setores que os compõem, mas nesse caso (dos segmentos econômicos) seria muito temerário arriscar algum vínculo mais nítido com a questão da garantia da privacidade no âmbito do *big data analytics*.

## 9.4 Análise do modelo proposto pelo Pentágono da Privacidade no *Big Data Analytics*

A análise dos estudos mais recentes, como já demonstrado nesse estudo, tem apontado que as questões aqui discutidas remetem a desafios ainda sem solução, pelo menos com as soluções e alternativas constituídas até o momento, apesar, claro, dos avanços em múltiplas frentes como legislação, política, ética, comportamento, técnica, métodos, metodologias e tecnologias, no âmbito da garantia de privacidade. Apesar disso, os avanços no campo *do big data analytics* têm suplantado em velocidade e amplitude os avanços nos campos citados acima.

A literatura tem apontado consistentemente para o fato de que ainda não se possui um modelo suficientemente robusto que consiga garantir a privacidade dos dados em análises de *big data*, ao mesmo tempo que também garanta a extração do seu valor. A constatação é que o problema é complexo e envolve múltiplas dimensões como legislação, política e códigos de conduta, sensibilização, comportamento e tecnologia.

A revisão da literatura empreendida nesse trabalho, esteve assentada no esforço de identificação, análise lógica e validação dos possíveis construtos estruturantes de um modelo multifacetado que consiga prover efetiva garantia à preservação da privacidade, no que se refere aos dados pessoais, ao mesmo tempo, em que garante a geração de valor, objetivo fundamental e sustentador da proposta do *big data analytics*.

Nesse sentido, parece muito claro que a proteção / garantia da privacidade deverá congregar soluções tecnológicas, legais, sociais, culturais e políticas, tanto ao nível de nações quanto de empresas. As possibilidades teóricas encontradas indicam que a lógica geral do modelo se aplica quando todos os elementos deste estão presentes, atuantes e integrando a cultura das organizações.

Cabe destacar que os construtos teóricos balizadores dessa proposta inicial de modelo estiveram baseadas em diversos estudos empíricos realizados por seus

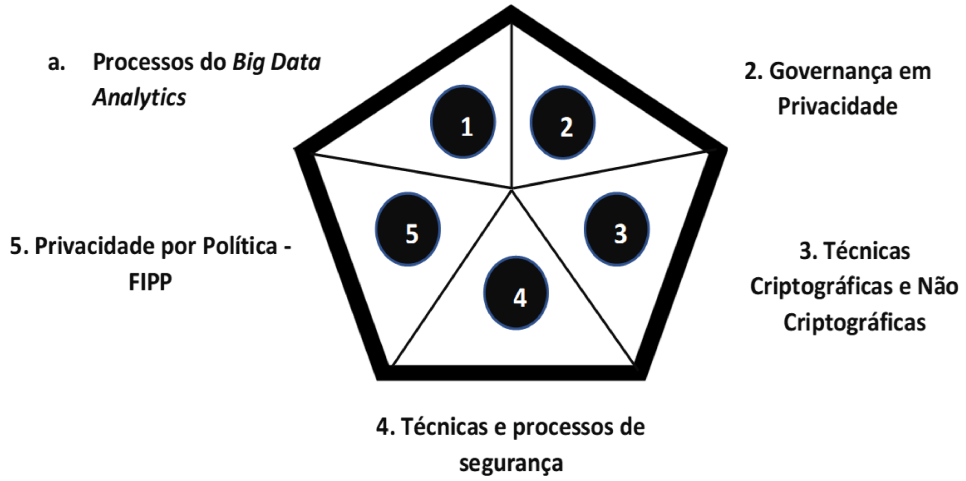
respectivos pesquisadores responsáveis. Mesmo assim, é preciso testá-lo frente a realidade de um conjunto significativo das maiores empresas brasileiras a partir das declarações divulgadas em um importante documento no âmbito da privacidade, a "Política de Privacidade" declarada dessas organizações.

Na Figura 15 abaixo, é apresentado o modelo estruturado com base nos levantamentos e achados teóricos dessa pesquisa, que se denominou de "Pentágono da Privacidade no *Big Data Analytics*", modelo esse que contempla um caleidoscópio de soluções capazes de garantir a privacidade ao mesmo tempo que também dá garantias para a extração de valor do *big data analytics*.

O modelo do "Pentágono da Privacidade no *Big Data Analytics*", está estruturado sobre as cinco dimensões fundamentais e suas respectivas práticas visando garantir a privacidade no *Big Data Analytics*, ao mesmo tempo, em que também consegue garantir a extração de valor a partir dessa abordagem. É, portanto, um modelo complexo, amplo e em constante evolução, pois sempre deverá ser revisitado e atualizado, principalmente no que tange a suas práticas, *vis a vis,* principalmente, os avanços das tecnologias e dos modelos analíticos no *Big Data Analytics*.

Na Figura 15 abaixo, se apresenta uma visão geral do modelo estruturado com base nos levantamentos e achados dessa pesquisa; modelo esse que contempla um caleidoscópio de soluções capazes, como pontuado acima, de garantir a privacidade ao mesmo tempo que também dá garantias para a extração de valor no *big data analytics.*

**FIGURA 15 - Pentágono da Privacidade no *Big Data Analytics***




Fonte: dados da pesquisa, elaboração própria, 2023

Em linhas gerais as dimensões apresentadas no modelo podem ser entendidas da seguinte forma:

1) Processos do *Big Data Analytics*: essa dimensão contempla processos e práticas visando a garantia da privacidade em todo o ciclo de vida do *Big Data Analytics*: coleta, armazenamento, processamento e análise;

2) Governança em Privacidade: essa dimensão contempla questões como a existência de comitês de ética e revisão no âmbito do ciclo de vida do *Big Data Analytics*, privacidade por design, treinamento, inventário de ativos digitais, dentre outros;

3) Técnicas Criptográficas e Não Criptográficas: essa dimensão contempla o nível de utilização e maturidade da organização quanto a utilização de técnicas criptográficas e não criptográficas;

4) Técnicas e processos de segurança: essa dimensão contempla o nível de utilização e maturidade da organização quanto a utilização de técnicas e processos em segurança e proteção da informação;

5) Privacidade por Política-FIPP: essa dimensão contempla tanto o enfoque da privacidade por política, quanto à privacidade por arquitetura e a garantia dos princípios de prática de informação e privacidade justas. Em linhas gerais contempla a proteção da privacidade por meio de políticas, semelhante à lógica das legislações; assim como visam garantir a incorporação de funcionalidades de preservação da privacidade nos estágios iniciais do desenvolvimento de sistemas e processos.

No Quadro 7, abaixo são apresentados os principais elementos constituintes (práticas) de cada uma das dimensões do Pentágono da Privacidade no *Big Data Analytics:*

**Quadro 7 - Elementos e práticas constituintes das dimensões do Pentágono da Privacidade no *Big Data Analytics***

| **PROCESSOS DO *BIG DATA ANALYTICS*** | **GOVERNANÇA EM PRIVACIDADE** | **TÉCNICAS CRIPTOGRÁFICAS E NÃO CRIPTOGRÁFICAS ADOTADAS PELA ORGANIZAÇÃO** | **TÉCNICAS E PROCESSOS DE SEGURANÇA ADOTADOS PELA ORGANIZAÇÃO** | **PRIVACIDADE POR POLÍTICA - FIPP (*FAIR INFORMATION PRACTICES PRINCIPLES*)** |
|---|---|---|---|---|
| A organização trabalha sob a ótica do *smart* | A organização protege a privacidade de | Desidentificação / Pseu-donimização (ex.: criptografia | *Structured Map Reduced Layer* | Confidencialidade – as informações de identificação |

| | | | | |
|---|---|---|---|---|
| *data* (proteção de dados pessoais; incorporação de regras de acesso aos dados; respostas às solicitações de informações dependentes de suas regras de acesso) | dados individuais por meio de políticas internas | com chave secreta, função hash, função hash com chave armazenada, criptografia determinística, tokenização, dentre outros) | | pessoal são protegidas com salvaguardas administrativas, técnicas e físicas |
| São identificados antecipadamente projetos e conjuntos de dados que contém dados pessoais, criando "zonas proibidas" para os mesmos | O controle do proprietário dos dados é um valor central nas soluções analíticas da organização (acesso ao titular dos dados) | Anonimização / Generalização (ex.: diversidade L, proximidade T, k-anonimato baseado em generalização, dentre outros) | Segurança de dados – os dados pessoais são protegidos por salvaguardas de segurança | Especificação de propósito – as informações coletadas são usadas para um propósito específico para o qual foram coletadas e não são utilizadas para outros fins sem a autorização devida |
| Utilizam-se combinações de controles (computação x inferência x uso) | A organização possui um profissional responsável pelo tratamento dos dados pessoais, cujos contatos | Anonimização/Randomização (ex.: K-Anominato) baseado em supressão, permutação, substituição, | Controle de acesso (ex.: Autenticação, autorização, auditoria, MAC – Integração e controle de | A organização possui uma política clara e de conhecimento de todos quanto aos procedimentos de comunicação ao |

| | | | | |
|---|---|---|---|---|
| para gerenciar riscos em *big data* | estão publicamente divulgados em seu site também em outros canais de comunicação | privacidade diferencial, *nulling out*, mascaramento de dados, embaralhamento, dentre outros | acesso obrigatório) | mercado e stakeholders, sobre possíveis incidentes de vazamento ou roubo de dados |
| Há a garantia quanto ao armazenamento e compartilhamento seguro de dados | A organização incorpora funcionalidades de preservação da privacidade nos estágios iniciais do desenvolvimento de sistemas e soluções (*Privacy by Design*) | Outras técnicas criptográficas: triple DES, anonimização baseada em identidade, homomórfica | Segurança da informação | Consentimento – busca o consentimento do indivíduo para coleta, processamento, uso e transferência de seus dados e manutenção dos consentimentos individuais |
| É garantida a transparência algorítmica nos processos de análise | A organização possui um comitê de ética e revisão no âmbito da coleta, processamento e análise de dados em *Big Data* | Modelo de execução híbrida – HybrEx (confidencialidade e privacidade em computação em nuvem) | Processo de capacitação e certificação de todo o time da organização, que deve conhecer e trabalhar para conter as principais técnicas de ataque (link lateral, palpite óbvio, reidentificação por | Aviso – Informa ao titular sobre a coleta de seus dados |

| | | | | |
|---|---|---|---|---|
| | | | meio de extremidades, dentre outros) | |
| A discriminação algorítmica não é um resultado analítico aceito pela organização | A organização tem estabelecido um processo e constante revisão e atualização de sua política de privacidade, refletindo o acompanhamento e as melhorias nesse âmbito | Modelos de classificação de dados sensíveis (ex.: SVM Multi-Kernel) | | Minimização de dados – impõe limitações sobre os tipos de informações e organização para coletar dados sobre um indivíduo |
| As atividades de tratamento de dados na organização observam a boa-fé quanto aos princípios da finalidade, adequação e necessidade | A organização realiza anualmente um inventário de ativos digitais | Técnicas de limitação de divulgação estatística (agregação, supressão, perturbação) | | Existem políticas e procedimentos diferenciados quanto aos dados e análises de populações vulneráveis (ex.: crianças, mulheres grávidas, prisioneiros, etc.) |
| São definidos parâmetros para as permissões de transmissão, | A organização possui políticas e práticas que consideram os | | | Existência de políticas que disciplinam não só quem pode |

| | | | | |
|---|---|---|---|---|
| armazenamento, uso e descarte dos dados (“dados portadores de políticas”) | cuidados e limites necessários quando da utilização de dados fornecidos por “corretores de dados” | | | analisar, mas que tipos de dados podem ser analisados |
| Efetiva existência e consistência quanto aos procedimentos, políticas e práticas visando a proteção e preservação da privacidade em todo o ciclo de vida do *Big Data* (coleta, armazenamento, processamento e análise) | A organização requer e garante que seus fornecedores possuam protocolos de segurança capazes de garantir a privacidade e a confidencialidade dos dados | | | Retificação – permite que o titular dos dados exija retificação dos mesmos caso estejam imprecisos |

Fonte: dados da pesquisa, elaboração própria, 2023

Uma análise pormenorizada dos resultados encontrados pela abordagem analítica realizada nesse estudo referente às Políticas de Privacidade das empresas integrantes do ranking Valor1000 (2021), mostra que a grande maioria das práticas citadas pelas organizações em suas respectivas Políticas de Privacidade integram o modelo do Pentágono da Privacidade. Tal fato pode ser visualizado no Quadro 8, abaixo;

**QUADRO 8 - Proposta do Pentágono da Privacidade no *Big Data Analytics* x Realidade das Práticas nas Organizações Pesquisadas**

| DIMENSÃO | NÚMERO PRÁTICAS QUE COMPÕE O MODELO | NÚMERO PRÁTICAS PRESENTES/CITADAS NAS POLÍTICAS DE PRIVACIDADE DAS ORGANIZAÇÕES PESQUISADAS |
|---|---|---|
| Processos de *big data* | 17 | 9 (53%) |
| Governança em privacidade | 10 | 8 (80%) |
| Técnicas criptográficas e não criptográficas | 9 | 5 (56%) |
| Técnias e processos de segurança adotadas na organização | 5 | 3 (60%) |
| Privacidade por política – FIPP (*Fair Information Practices Principals*) | 9 | 7 (78%) |

Fonte: dados da pesquisa, elaboração própria, 2023

Pelos dados constantes no Quadro 8, é possível constatar o alinhamento das práticas previstas no modelo proposto com a realidade das organizações. Por esse ângulo, o modelo proposto pelo Pentágono da Privacidade no *Big Data Analytics* reflete as práticas das maiores organizações do Brasil e, portanto, teria a validade necessária como um modelo consistente.

Mas, a partir do aprofundamento e detalhamento dessa análise, a realidade não se reflete de maneira tão contundente. E a principal questão posta aqui é o que se denominou de “Privacidade Opaca”, que é resultado de diversas questões, mas principalmente da ainda incipiente cultura de privacidade nas empresas e da preocupação patente das mesmas no sentido de estarem mais focadas em dar respostas às questões mais básicas e menos complexas do que ordena a legislação brasileira, ao mesmo tempo que parecem evitar aprofundamentos nas declarações em suas Políticas de Privacidade, dos tratamentos e análises realizadas com os dados pessoais, principalmente de seus clientes que estão em suas bases de dados.

Ou seja, com base nos dados coletados e analisados nesta pesquisa, o modelo geral do Pentágono da Privacidade no *Big Data Analytics* em termos de sua estrutura

geral foi validado, mas de maneira mais particularizada quando foram analisadas suas práticas individualmente e em conjunto, segundo suas respectivas dimensões, foi possível verificar que por todos os motivos já colocados anteriormente verificou-se um baixo nível de concentração das citações das práticas previstas no modelo proposto no âmbito das Políticas de Privacidade das organizações estudadas.

No Quadro 9 abaixo, são apresentados os dados consolidados por dimensão do modelo proposto, de forma a poder-se constatar esse resultado.

**QUADRO 9 - Consolidação das citações das práticas por dimensão do modelo proposto**

| DIMENSÃO | ACUMULADO DAS PRÁTICAS CITADAS POR DIMENSÃO |
|---|---|
| Pocessos de *big data* | 28.964 |
| Governança em privacidade | 35.536 |
| Técnicas criptográficas e não criptográficas | 28.166 |
| Técnicas e processos de segurança adotadas na organização | 28.977 |
| Privacidade por política – FIPP (Fair Information Practices Principals) | 29.509 |

Fonte: dados da pesquisa, elaboração própria, 2023

É possível verificar que existe um número razoável de citações relativas às práticas previstas no Pentágono da Privacidade no *Big Data Analytics.* Mas, como está demonstrado no Tabela 4, essas citações estão enormemente concentradas em um pequeno número de práticas, apesar da expressiva quantidade de empresas. Por outro lado, observaram-se termos muito expressivos no que tange à análise da privacidade *no big data analytics* citados pouquíssimas vezes e por pouquíssimas empresas.

Ou seja, a maioria das práticas em todas as dimensões do modelo proposto são citadas nas políticas de privacidade das empresas analisadas, a questão é que são citadas num número pequeno delas em diversos extratos, em práticas que

guardam grande consistência e importância para a garantia da privacidade no âmbito do *big data analytics*.

Portanto, o modelo proposto aponta para a assertividade do seu conjunto. O avanço na importância da privacidade é que dará a devida consistência ao mesmo, no tempo.

A lógica geral do modelo se aplica quando todos esses elementos estão presentes, atuantes e integrando as práticas e a cultura das organizações. Em última instância, a depender do estágio de maturidade em que se encontra a organização, pode ser visto como um roteiro de melhores práticas, um caminho para se alcançar a maturidade em garantia da privacidade no âmbito do *Big Data Analytics*.

# 10 CONCLUSÕES

Apesar da preocupação com a privacidade não ser uma questão tão recente, a partir dos avanços no âmbito do *big data analytics*, é possível afirmar que passa a ser uma das grandes preocupações não só de governos, mas também das organizações públicas, privadas e do terceiro setor.

Como pontuado por Gupta e Rani (2019, p. 337, tradução nossa) “a privacidade de dados é um problema técnico e também sociológico”. E de forma contundente isso pode ser visualizado e demonstrado pelas facetas das práticas dispostas nas dimensões do Pentágono da Privacidade no *Big Data Analytics*.

A questão da privacidade nas maiores empresas do Brasil, no seu conjunto, é permeada por uma opacidade que não permite o entendimento e também não dá qualquer garantia de que a privacidade possa ser minimamente garantida.

De maneira individual, foi possível, a partir da análise qualitativa de um conjunto de empresas, visualizar algumas práticas mais avançadas e condizentes com o modelo aqui proposto, mas ainda sim muito incipientes, o que não permite uma diferenciação mais consistente e, portanto, permanece a situação geral detectada do que se chamou neste trabalho de “privacidade opaca”.

A análise dos dados demonstrou a validade do modelo proposto pelo Pentágono da Privacidade no *Big Data Analytics*, que pode ser aprimorado a partir desses primeiros achados. O modelo traz uma primeira proposta sobre uma importante questão verificada neste trabalho e ratificada por Gupta e Rani (2019, p. 337, tradução nossa), onde afirmam que “não existem diretrizes e padrões claros para proteger a privacidade individual”.

O modelo proposto neste trabalho pode e deve ser melhorado, tendo como base pelo menos duas frentes, a expansão da base de análise, ou seja, coletar um número maior de Políticas de Privacidade declaradas pelas empresas e submetê-las ao tratamento analítico dado neste trabalho.

Por outro lado, o avanço do trabalho de pesquisa, permitiu identificar que novas práticas podem e devem integrar e reforçar o modelo proposto pelo Pentágono da Privacidade no *Big Data Analytics*. Como, por exemplo, uma prática referente ao conceito de dados sintéticos, que são dados gerados artificialmente para o treinamento de modelos de Inteligência Artificial, evitando inúmeras questões problemáticas como o "*profiling*" e seus respectivos derivados no âmbito da discriminação algorítmica. Outra prática relevante identificada e que poderia ser testada de forma a tornar o modelo mais consistente seria o estabelecimento de métricas de privacidade, especialmente quanto alinhadas, aprovadas e reportadas à alta direção.

Um modelo que garanta a proteção da privacidade no âmbito do *big data analytics*, no limite, frente aos avanços cada vez mais rápidos, trazidos pelos novos modelos de análise de dados e seus respectivos algoritmos, quebra a possibilidade de uma realidade "pan-óptica", onde um único vigilante poderia observar a todos, ao mesmo tempo, em que os observados não teriam a menor ideia de que estão sendo vigiados, observados, controlados e manipulados.

É como colocou Davenport e Harris (2010, p. 40), as organizações altamente analíticas, aquelas em estágio mais avançado, "seguem o juramento de Hipócrates sobre a privacidade da informação: acima de tudo, nunca causar mal ou danos a alguém". Ou seja, o *big data analytics* deve ter como parâmetro básico e norteador de toda sua ampla abordagem, a garantia da privacidade.

Pelos achados desta pesquisa e resultados discutidos, ainda temos no Brasil um longo caminho pela frente no campo da privacidade no âmbito do *Big Data Analytics*.

## REFERÊNCIAS

AAKER,David A,; KUMAR, V.; DAY, George S. **Pesquisa de Marketing**. São Paulo: Atlas, 2001.

ABOUELMEHDI, Karim; BENI-HESSANE, Abderrahim; KHALOUFI, Hayat. *Big healthcare data*: preserving security and privacy. **Journal of Big Data**, v. 5, n. 1, p. 1-18, Jan. 2018. DOI: https://doi.org/10.1186/s40537-017-0110-7. Disponível em: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0110-7. Acesso em: 11 set. 2020.

ABNT - Associação Brasileira de Normas Técnicas. **ABNT NBR ISO/IEC 29100**: Tecnologia da informação – Técnicas de segurança – Estrutura de Privacidade. Rio de Janeiro: ABNT, 2020. 26 p.

ACQUISTI, Alessandro; TAYLOR, Curtis; WAGMAN, Liad. The economics of privacy. **Journal of Economic Literature**, v. 54, n. 2, p. 442-492, 2016. DOI: http://dx.doi.org/10.1257/jel.54.2.442. Disponível em: https://pubs.aeaweb.org/doi/pdfplus/10.1257/jel.54.2.442. Acesso em: 09 set. 2020.

ADAMS, Mackenzie. Big Data and Individual Privacy in the Age of the Internet of Things. **Technology Innovation Management Review**, v. 7, n. 4, p. 1-24, Apr. 2017. Disponível em: https://timreview.ca/article/1067. Acesso em: 31 ago. 2020.

ALTMAN, Micah; WOOD, Alexandra; O'BRIEN, David R.; GASSER, Urs. Practical approaches to big data privacy over time. **International Data Privacy Law**, 2018, Vol. 8, No. 1. DOI: https://doi.org/10.1093/idpl/ipx027. Disponível em: https://academic.oup.com/idpl/article/8/1/29/4930711. Acesso em: 06 nov. 2020.

AL-ZOBBI, Mohammed; SHAHRESTANI, Seyed; RUAN, Shahrestani; Chun. Improving Map Reduce privacy by implementing multi-dimensional sensitivity-based anonymization. Journal of Big Data, 2017. DOI: https://doi.org/10.1186/s40537-017-0104-5. Disponível em: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0104-5. Acesso em: 20 jan. 2021.

AOUN, Muhammad. Comparative Analysis of Text Mining Techniques for News Article Summarization. **LC International Journal of Stem**, Volume 04, Issue 01, March 2023. **DOI:** https://doi.org/10.5281/zenodo.7893329. Disponível em: http://www.lcjstem.com/index.php/jstem/article/view/177. Acesso em: 09 maio. 2023.

ARTICLE 29 DATA PROTECTION WORKING PARTY. **Opinion 5/2014 on Anonymisation techniques**. Bruxelas: [s. n.], 2014. Disponível em: https://www.statewatch.org/media/documents/news/2014/apr/eu-art-29-dp-wp-216.pdf. Acesso em: 20 fev. 2021.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário oficial da União**, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm. Acesso em: 03 set. 2020.

BRASIL. Ministério da Educação. **Portal de Periódicos, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes)**. 2023. Disponível em: www.periodicos.capes.gov.br. Acesso em: 31 ago. 2020.

BRASIL. Ministério da Educação. **Guia e Elaboração de Termo de Uso e Política de Privacidade. Programa de Privacidade e Segurança da Informação (PPSI)**. Brasília, versão 2.0, março de 2023. Disponível em: https://www.gov.br/governodigital/pt-br/seguranca-e-protecao-de-dados/ppsi/guia_termo_uso_politica_privacidade.pdf. Acesso em: 02 ago. 2023.

BRYMAN, Alan. **Research Methods and Organziation Studies**. London: Routledge, 1989. (Contemporary Social Research, 20).

CHANG, Yu-Wei; HUANG, Mu-Hsuan. A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. **Journal of the American Society for Information Science and Technology**, *[S. l.]*, v. 63, n. 1, p. 22–33, 2012. DOI: https://doi.org/10.1002/asi.21649. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21649. Acesso em: 23 nov. 2022.

CHANG, Yu-Wei. Examining interdisciplinarity of library and information science (LIS) based on LIS articles contributed by non-LIS authors. **Scientometrics**, *[S. l.]*, v. 116, n. 3, p. 1589–1613, 2018. DOI: https://doi.org/10.1007/s11192-018-2822-7. Disponível

em: https://link.springer.com/article/10.1007/s11192-018-2822-7. Acesso em: 25 nov. 2022.

CHANSON, Mathieu; BOGNER, Andreas; BILGERI, Dominik; FLEISCH, Elgar. Blockchain for the IoT: Privacy-Preserving Protection of Sensor Data**. Journal of the Association for Information Systems**, v. 20, n. 9, p. 1274-1309, Mar. 2019. DOI: http://dx.doi.org/10.17705/1jais.00567. Disponível em: https://aisel.aisnet.org/jais/vol20/iss9/10/. Acesso em: 08 set. 2020.

CHAUHAN, Sumedha; AGARWAL, Neetima; KAR, Arpan Kumar. Addressing big data challenges in smart cities: a systematic literature review. **Emerald Group Publishing Limited**. VOL. 18, N°. 4, 2016, pp. 73-90. DOI: https://doi.org/10.1108/info-03-2016-0012. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/info-03-2016-0012/full/html. Acesso em: 17 jun. 2021.

CHEUNG, Anne, S.Y. Moving Beyond Consent For Citizen Science in Big Data Health and Medical Research, 16 Nw. **J. Tech. & Intell**. Pro p. 15 (2018). Disponível em:https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1300&context=njtip. Acesso em: 03 dez. 2020.

CISCO SECURE. **Privacy Becomes Mission Critical. Data Privacy Benchmark Study**, **2022**. Disponível em: https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-privacy-benchmark-study-2022.pdf?CCID=cc000742&DTID=esootr000515. Acesso em: 25 out. 2023.

COX, Michael; ELLSWORTH, David. Application Controled Demand Paging for Out-or-Core Visualization. NASA, **Ames Research Center**, 1997. Disponível em: https://www.nas.nasa.gov/assets/pdf/.../1997/nas-97-010.pdf. Acesso em: 10 jul. 2019.

CRESWELL, John W. **Projeto de Pesquisa**: métodos qualitativo, quantitativo e misto. Porto Alegre: Artmed, 2010.

DAVENPORT, Thomas; HARRIS, Jeanne. **Inteligência analítica nos negócios**: como usar a análise de informações para obter resultados superiores. Rio de Janeiro: Elsevier, 2010.

DELEN, Dursun; CROSSLAND, Martin D. Seeding the survey and analysis of research literature with text mining. **Science Direct, Expert Systems with Applications**. Elsevier, 2008. DOI: https://doi.org/10.1016/j.eswa.2007.01.035. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417407000486?via%3Dihub. Acesso em: 08 jan. 2023.

Dicionário Etimológico. **Método**. Disponível em: https://www.dicionarioetimologico.com.br/metodo/. Acesso em 25 out. 2023.

DUAN, Huijue Kelly; VASARHELYI, Miklos A.; CODESSO, Maurício; ALZAMIL, Zamil. Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. **Revista Internacional de Contabilidade Sistemas de Informação**. Elsevier, 2021. DOI: 10.1016/j.accinf.2022.100600. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1467089522000525?via%3Dihub. Acesso em: 14 dez. 2021.

EL OUAZZANI, Zakariae; EL BAKKALI, Hanan. A classification of non-cryptographic anonymization techniques ensuring privacy in big data. **International Journal of Communication Networks and Information Security (IJCNIS),** v. 12, n. 1, p. 142-152, April 2020. Disponível em: https://www.researchgate.net/publication/342009466_A_Classification_of_non-Cryptographic_Anonymization_Techniques_Ensuring_Privacy_in_Big_Data. Acesso em: 02 set. 2020.

FAHIM, Ahmed. K and starting means for k-means algorithm. **Journal of Computational Science**. Elsevier, 2021. DOI: https://doi.org/10.1016/j.jocs.2021.101445. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1877750321001277?via%3Dihub. Acesso em: 10 jul.2023.

FOREMAN, John W. **Data Smart**: usando data Science para transformar informação em insights. Rio de Janeiro: Alta Books, 2016.

GONÇALVES, Alexandre Leopoldo; FARACO, Fernando Melo; SOUZA, João Artur de; TODESCO, José Leomar; NUNES, Ronnie Carlos Tavares. Análise de agrupamentos sobre textos: um estudo dos resumos do banco de teses e dissertações

da CAPES. Hábitats de Innovación y Economía del Conocimiento: una apuesta para el futuro. **VIII Congresso Internacional de Conocimiento y Innovación**. Guadalajara, 24 y 25 de septiembre, 2018. Disponível em: https://kmeducationhub.de/congreso-internacional-de-conocimiento-e-innovacin-ciki/. Acesso em: 20 set. 2023.

GRISOTO, Ana Paula; SANT'ANA, Ricardo Cesar Gonçalves; SEGUNDO, José Eduardo Santarem. A questão da privacidade no contexto da Ciência da Informação: uma análise das Teses e Dissertações do Programa de Pós-Graduação em Ciência da Informação da UNESP Campus de Marília. **Revista Ibero-americana de Ciência da Informação – RCI**. Brasília, v. 8, n.2, p.165-181, jul. / dez. 2015. DOI: https://doi.org/10.26512/rici.v8.n2.2015.2066. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/2066. Acesso em: 30 jul. 2023.

GUPTA, Deepak; RANI, Rinkle. A study of big data Evolution and research challenges. **Journal of Information Science (JIS)**, 2019. DOI: https://doi.org/10.1177/0165551518789880. Disponível em: https://journals.sagepub.com/doi/10.1177/0165551518789880. Acesso em: 22 abr. 2022.

HADAR, Irit; HASSON, Tomer; AYALON, Oshrat; TOCH, Eran; BIRNHACK, Michael; SHERMAN, Sofia; BALISSA, Arod. Privacy by designers: software developer's privacy mindset. **Empirical Software Engineering,** v. 23, n. 1, p. 259-289, Feb. 2018. DOI: https://doi.org/10.1007/s10664-017-9517-1. Disponível em: https://link.springer.com/article/10.1007/s10664-017-9517-1. Acesso em: 31 ago. 2020.

HAIR, Joseph F. Jr.; BLACK, William C.; BABIN, Barry J.; ANDERSON, Rolph E.; TATHAM, Ronald L. **Análise multivariada de dados**. Tradução: Adonai Schlup Sant'Anna. 6 ed. Porto Alegre: Bookman, 2009.

HJØRLAND, Birger. Library and Information Science (LIS), Part 1. **Knowledge Organization**, *[S. l.]*, v. 45, n. 3, p. 232–254, 2018. DOI: https://10.5771/0943-7444-2018-3-232. Disponível em: https://www.researchgate.net/publication/325773742_Library_and_Information_Science_LIS_Part_1. Acesso em: 04 dez. 2022.

HU, Haize; LIU, Jianxun; ZHANG, Xianpging; FANG, Mengge. An Effective and Adaptable K-means Algorithm fog Big Data Cluster. **Patter Recognition**. Elsevier, 2023. DOI: https://10.1016/j.patcog.2023.109404. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S003132032300105X?via%3Dihub. Acesso em: 05 set. 2023.

ISHII, Kaori. Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects. **AI & Society**, v. 34, p. 509-533, Aug. 2019. DOI: https://doi.org/10.1007/s00146-017-0758-8. Disponível em: https://link.springer.com/article/10.1007/s00146-017-0758-8. Acesso em: 02 set. 2020.

JAIN, Anil K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters** 31. Elsevier, 2010. DOI: https://doi.org/10.1016/j.patrec.2009.09.011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323. Acesso em: 23 jul. 2023.

JAIN, Priyank; GYANCHANDANI, Manasi; KHARE, Nilav. Big data privacy: a technological perspective and review. **Journal of Big Data**, v. 3, n. 1, p. 1-25, 2016. DOI: https://doi.org/10.1186/s40537-016-0059-y. Disponível em: https://journalofbigdata.springeropen.com/counter/pdf/10.1186/s40537-016-0059-y.pdf . Acesso em: 10 set. 2020.

JAIN, Priyank; GYANCHANDANI, Manasi; KHARE, Nilay. Enhanced Secured Map Reduce layer for Big Data privacy and security. **Journal of Big Data**, v. 6, n. 60, p. 1-17, Mar. 2019. DOI: https://doi.org/10.1186/s40537-019-0193-4. Disponível em: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0193-4. Acesso em: 01 set. 2020.

KACZMAREK, Iwona; IWANIAK, Adam; SWIETLICKA, Aleksandra; PIWOWARCZYK, Mateusz; NADOLNY, Adam. A machine learning approach for integration of spatial development plans based on natural language processing. **Sustainable Cities and Society**, Elsevier, 2022. DOI: https://doi.org/10.1016/j.scs.2021.103479. Disponível em: https://www.sciencedirect.com/science/article/pii/S2210670721007460?via%3Dihub. Acesso em: 31 ago. 2023.

KAPIL, Gayatri; AGRAWAL, Alka; ATTAALLAH, Abbulaziz; ALGARNI, Abdullah; KUMAR, Rajeev; KHAN, Raees Ahmad. Attribute based honey encryption algorithm for securing big data: Hadoop distributed file system perspective. **Peer J. Computer Science**, 2020. DOI: https://doi.org/10.7717/peerj-cs.259. Disponível em: https://peerj.com/articles/cs-259/. Acesso em: 05 dez. 2020.

LENZI, Lívia Aparecida Ferreira; BRAMBILA, Ednéa Zandonadi. Ciência da Informação, Ciência e Revolução Científica: breve histórico e reflexões. **Informação & Informação**, v. 11, n. 1, jan. / jun., 2006. DOI: https://doi.org/10.5433/1981-8920.2006v11n1p26. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/1679. Acesso em: 30 maio. 2022.

LOTT, Yuri Monnerat; CIANCONI, Regina de Barros. Vigilância e privacidade, no contexto do big data e dados pessoais: análise da produção da Ciência da Informação no Brasil. **Perspectivas em Ciência da Informação**, v.23, n.4, p.117-132, out. / dez. 2018. DOI: https://doi.org/10.1590/1981-5344/3313. Disponível em: https://www.scielo.br/j/pci/a/BXMsD73NL5dpYQWqGm8YrBN/?lang=pt. Acesso em: 19 ago. 2020.

MAA, Feicheng; MARCHIONINI, Gary. Introduction to the special issue on data Science and information science. **Data and Information Management**, 2023. DOI: https://doi.org/10.1016/j.dim.2023.100034. Disponível em: https://www.sciencedirect.com/journal/data-and-information-management/vol/7/issue/1. Acesso em: 18 jun. 2023

MACHADO, Felipe Nery Rodrigues. **Big Data**: o futuro dos dados e aplicações. São Paulo: Ética/Saraiva, 2018.

MALHOTRA, Naresh K. **Pesquisa de Marketing**: uma orientação aplicada. 3ª edição. Porto Alegre: Bookman, 2006. 720 p.

MARIANI, Costanza; NAVROTSKA, Yuliya; MANCINI, Mauro. Unsupervised machine learning for Project stakeholder classification: Benefits and limitations. **Project Leadership and Society**. Elsevier, 2023. DOI: https://doi.org/10.1016/j.plas.2023.100093. Disponível em:

https://www.sciencedirect.com/science/article/pii/S2666721523000145. Acesso em: 13 dez. 2023.

MELO JÚNIOR, Cleuton Sampaio de. **Data Science para Profissionais – Utilizando R**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2018.

MEHTA, Brijesh B.; RAO, Udai Pratap. Improved l-diversity: Scalable anonymization approach for Privacy Preserving Big Data Publishing. **Journal of King Saud University** – Computer and Information Sciences, 2019. DOI: https://doi.org/10.1016/j.jksuci.2019.08.006. Disponível em: https://www.sciencedirect.com/science/article/pii/S1319157819304173. Acesso em: 19 ago. 2020.

MILLS, Kathy A. What are the threats and R potentials of big data for qualitative research? **Qualitative Research**, v. 18, n. 6, p. 591–603, 2018. DOI: https://doi.org/10.1177/1468794117743465. Disponível em: https://journals.sagepub.com/doi/full/10.1177/1468794117743465. Acesso em: 04 set. 2020.

NORONHA, Daisy Pires; FERREIRA, Sueli Mara S. P. Revisões de literatura. *In*: CAMPELLO, Bernadete Santos; CONDÓN, Beatriz Valadares; KREMER, Jeannette Marguerite (org.) **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: UFMG, 2000.

OHM, Paul. Broken promisses of privacy: responding to the surprising failure of anonymization. **UCLA Law Review**, v. 57, p. 1702-1777, Aug. 2010. Disponível em: https://www.uclalawreview.org/pdf/57-6-3.pdf. Acesso em: 05 set. 2020.

O'NEIL, Cathy. **Algoritmos de destruição em massa**: como o big data aumenta a desigualdade e a ameaça à democracia. Santo André, SP: Editora Rua do Sabão, 2020.

OOSTVEEN, Manon. Identifiability and the applicability of data protection to big data. **International Data Privacy Law**, v. 6, n. 4, p. 299-309, Nov. 2016. DOI: https://doi.org/10.1093/idpl/ipw012. Disponível em: https://academic.oup.com/idpl/article-abstract/6/4/299/2525426?login=false. Acesso em: 01 set. 2020.

OZAYDIN, Bunyamin; ZENGUL, Ferhat; ONER, Nurettin; DELEN, Dursun. Text-mining analysis of mHealth research. **mHealth**, 2017. DOI: http://dx.doi.org/10.21037/mhealth.20. Disponível em: https://mhealth.amegroups.org/article/view/17842/pdf. Acesso em: 29 jul. 2023.

OUAZZANI, Zacariae El; BAKKALI, Hanan El. A classification of non-cryptographic anonymization techniques ensuring privacy in big data. International Journal of Communication Networks and Information Security (IJCNIS) Vol. 12, No. 1, April 2020. DOI: DOI: https://doi.org/10.17762/ijcnis.v12i1.4401. Disponível em: https://www.ijcnis.org/index.php/ijcnis/article/view/4401. Acesso em: 13 nov. 2020.

POLITOU, Eugenia; ALEPIS, Efthimios; PAYSAKIS, Constantinos. Forgetting personal data and revoking consente under the GDPR: Challenges and proposed solutions. **Journal of Cybersecurity**, v. 4, n. 1, p. 1–20, 2018. DOI: https://doi.org/10.1093/cybsec/tyy001. Disponível em: https://academic.oup.com/cybersecurity/article/4/1/tyy001/4954056. Acesso em: 10 set. 2020.

PRAGASH, K.; JAYABHARATHY, J. A Survey on Big Data Privacy and Security Issues in Healthcare Information System. **Advanced in Natural and Applied Sciences**, v. 11, n. 12, p. 95-99, Oct. 2017. Disponível em: https://www.aensiweb.net/AENSIWEB/anas/anas/2017/October/95-99.pdf. Acesso em: 03 set. 2020.

PRAMANIKA, Ileas; LAUB, Raymond Y. K.; AZADA, Abul Kalam; HOSSAINC, Sakir; CHOWDHURY, Kamal Hossain; KARMAKER, B. K. **Healthcare informatics and analytics in big data**. Elsevier, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.113388. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417420302128?via%3Dihub. Acesso em: 17 abr. 2021.

RAGHUPATHI, Viju; REN, Jie; RAGHUPATHI, Wullianallur. Identifying Corporate Sustainability Issues by Analyzing Shareholder Resolutions: **A Machine-Learning Text Analytics Approach, Sustainability**, 2020. DOI: https://doi.org/10.3390/su12114753. Disponível em: https://www.mdpi.com/2071-1050/12/11/4753. Acesso em: 23 nov. 2023

RANI, Pacha Shobha; DHAMODARAN, Vigneswari. Security and Privacy in Big Data Analytics. **International Journal on Intelligent Electronic System**, Vol.10 No.2, july 2016.

RAO, P. Ram Mohan; KRISHNA, S. Murali; KUMAR, A. P. Silva. Privacy preservation techniques in big data analytics: a survey. **Jornal of big data**, v. 5, n. 22, p. 1-12, Mar. 2018. DOI: https://doi.org/10.1186/s40537-018-0141-8. Disponível em: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0141-8. Acesso em: 08 set. 2020.

RHOEN, Michiel. Big Data and Consumer Participation in Privacy Contracts: Deciding who Decides on Privacy. **Utrecht Journal Of International and European Law**. Researche article, 2015. DOI: https://doi.org/10.5334/ujiel.cu. Disponível em: https://utrechtjournal.org/articles/10.5334/ujiel.cu. Acesso em: 29 set. 2020.

RISTEVSKI, Blagoj; CHEN, Ming. Big Data Analytics in Medicine and Healthcare. **Journal of Integrative Bioinformatics**, 2018. DOI: https://doi:org/10.1515/jib-2017-0030. Disponível em: https://www.degruyter.com/document/doi/10.1515/jib-2017-0030/html. Acesso em: 18 jun. 2021.

ROTER, B.; NINKOVIC, N.; DORDEVIC, S. V. Clustering superconductors using unsupervised machine learning. **Physica C: Superconductivity and its applications**. Elsevier, 2022. DOI: https://doi.org/10.1016/j.physc.2022.1354078. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0921453422000661?via%3Dihub. Acesso em: 24 ago. 2023.

SALAS, Julián; DOMINGO-FERRER, Josep. Some Basics on Privacy Techniques, Anonymization and their **Big Data Challenges. Mathematics in Computer Science**, 2018. DOI: https://doi.org/10.1007/s11786-018-0344-6. Disponível em: https://link.springer.com/article/10.1007/s11786-018-0344-6. Acesso em: 15 fev. 2021.

SARACEVIC, Tefko. **Information Science. Journal of the American Society for Information Science**, 1999.

SARKAR, Bikash Kanti. Big data for secure healthcare system: a conceptual design. **Complex Intell. Syst.**, p. 133-151, 2017. DOI: https://doi.org/10.1007/s40747-017-0040-1. Disponível em: https://link.springer.com/article/10.1007/s40747-017-0040-1#citeas. Acesso em: 02 set. 2020.

SEADLE, Michael; HAVELKA, Stefanie. Information Science: Why it is not data Science. **Data and Information Management**, 2023. DOI: https://doi.org/10.1016/j.dim.2023.100027. Disponível em: https://www.sciencedirect.com/science/article/pii/S2543925123000013?via%3Dihub. Acesso em: 19 ago. 2023.

SHANMUGAPRIYA, E.; KAVITHA, R. Efficient and Secure Privacy Analysis for Medical Big Data Using TDES and MKSVM with Access Control in Cloud. **Journal of Medical Systems** (2019) 43: 265. DOI: http://doi.org/10.1007/s10916-019-1374-6. Disponível em: https://link.springer.com/article/10.1007/s10916-019-1374-6. Acesso em: 15 set. 2020.

SINNOTT, Richard O.; BAYLISS, Christophter; BROMAGE, Andrew; GALANG, Gerson; GONG, Yikai; GREENWOOD, Philip; JAYAPUTERA, Glenn; MAQUES, Davis; MORANDINI, Luca; NOGOORANI, Ghazal; PURSULTANI, Hossein; SARWAR, Muhammad; VOORSLUYS, William; WIDJAJA, Ivo. Privacy Preserving Geo-Linkage in the Big Urban Data Era. **J. Grid Computing**, 2016. DOI: https://doi.org/10.1007/s10723-016-9372-0. Disponível em: https://link.springer.com/article/10.1007/s10723-016-9372-0. Acesso em: 14 set. 2020.

SOLOVE, Daniel J. Conceptualizing Privacy. **California Law Review**, v. 90, n. 4, p. 1087-1155, July 2002. Disponível em: https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2086&context=faculty_publications. Acesso em: 17 dez. 2023.

SUGANYA, G.; PORKODI, R. A Survey on Text Mining Tools and Techniques. **Ipasj International Journal of Computer Science (IIJCS)**; Volume 5, Issue 7, july 2017. Disponível em: https://www.researchgate.net/publication/320621435_A_SURVEY_ON_TEXT_MINING_TOOLS_AND_TECHNIQUES. Acesso em: 25 jun. 2023.

SUN, Zhaohao; SUN, Lee Lizhe; STRANG, Kenneth. Big data analytics services for enhancing usiness Intelligence. **Journal of Computer Information Systems**, v. 58, n. 2, p. 162-168, 2016. DOI: http://dx.doi.org/10.1080/08874417.2016.1220239. Disponível em: https://www.researchgate.net/publication/309389413_Big_Data_Analytics_Services_for_Enhancing_Business_Intelligence. Acesso em: 01 set. 2020.

SUN, Zhaohao; PAMBEL, Francisca; STRANG, Kenneth David. Privacy and security in the big data paradigm. **Journal of Computer Information Systems**, v. 60, n. 3, p. 1-10, Feb. 2018. DOI: https://doi.org/10.1080/08874417.2017.1418631. Disponível em:https://www.researchgate.net/publication/323056850_Privacy_and_security_in_the_big_data_paradigm. Acesso em: 02 set. 2020.

THOMSON, Lucy L.; THIBADEAU, Robert. American Bar Association. Security challenges of the big data ecosystem require a laser like focus on risk. **The SciTech Lawyer**, v. 12, n. 2, Jan. 2016. Disponível em: https://www.kiip.re.kr/webzine/1610/files/library_paper3.pdf. Acesso em: 17 nov. 2021.

VERGARA, Sylvia Constant. **Métodos de Pesquisa em Administração**. São Paulo: Atlas, 2005. 287 p.

VIEIRA, Marcelo Milano Falcão; ZOUAIN, Deborah Moraes (Org.). **Pesquisa qualitativa em administração**: teoria e prática. Rio de Janeiro: Ed. FGV, 2005. 237 p.

XIAO, Ximing; ZHANG, Fangyuan; LI, Jinrui. Library and Information Science Research in China – A Survey Based Analysis of 10 LIS Educational Institutes. **The Journal of Academic Librarianship**, p. 330 – 340, 2015. DOI: https://doi.org/10.1016/j.acalib.2015.02.012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0099133315000166?via%3Dihub. Acesso em: 13 nov. 2021.

WARREN, Samuel D.; BRANDEIS, Louis D. The Right to Privacy. **Harvard Law Review**, v. 4, n. 5, p. 193-220, 1890. Disponível em: https://www.cs.cornell.edu/~shmat/courses/cs5436/warren-brandeis.pdf. Acesso em: 29 set. 2021.

WESTIN, Alan. **Privacy and Freedom**. Nova York: Atheneum, 1970.

WILSON, Rebecca J.; BELLIVEAU, Kiley M.; GRAY, Leigh Ellen. Busting the Black Box: Big Data, Employment and Privacy. **Defense Counsel Journal**, v. 84, n. 3, p. 1-34, July 2017. Disponível em: https://www.iadclaw.org/defensecounseljournal/busting-the-black-box-big-data-employment-and-privacy/. Acesso em: 11 set. 2020.

ZHANG, Jin; WOLFRAM, Dietmar; MA, Feicheng. The impact of big data on research methods in information science. **Data and Information Management**, 2023. DOI: https://doi.org/10.1016/j.dim.2023.100038. Disponível em: https://www.sciencedirect.com/science/article/pii/S2543925123000128. Acesso em: 30 out. 2023.

ZELASKY, Sarah; MARTIN, Chantel L.; WEAVER, Christopher; BAXTER, Lisa K.; RAPPAZZO, Kristen M. Identifying groups of children's social Mobility opportunity for public helath applications using k-means clustering. **Heliyon Journal Cell Press**, 2023. DOI: https://doi.org/10.1016/j.heliyon.2023.e20250. Disponível em: https://www.cell.com/heliyon/fulltext/S2405-8440(23)07458-3?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405844023074583%3Fshowall%3Dtrue. Acesso em: 04 nov. 2023.

ZHAO, Liping; ALHOSHAN, Waad; FERRARI, Alessio; LETSHOLO, Keletso, J.; AJAGBE, Muideen A.; CHIOASCA, Erol-Valeriu; BATISTA-NAVARRO, Riza T. **Natural Language for Requirements Engineering**: A Systematic Mapping Study. ACM Computing Surveys, vol. 54, n° 3, article 55, april 2021. DOI: https://doi.org/10.48550/arXiv.2004.01099. Disponível em: https://arxiv.org/abs/2004.01099. Acesso em: 18 jul. 2023.

ZINS, Chaim. Conceptions of Information Science. **Journal of the American Society for Information Science and Technology**, p. 335-350, 2007. DOI: https://doi.org/10.1002/asi.20507. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.20507. Acesso em: 28 abr. 2022.

## APÊNDICE A – CORRELAÇÃO ENTRE OS TEMAS PESQUISADOS E RESPECTIVOS AUTORES

No quadro abaixo estão as referências teóricas originais e seus respectivos autores, que possibilitaram a construção do modelo teórico de garantia da privacidade no âmbito do *big data analytics*.

| Dimensões – Construtos Teóricos | Autores |
|---|---|
| Dados portadores de política | Mackenzie Adams, 2017 |
| Proprietário dos dados – acesso ao titular dos dados | Mackenzie Adams, 2017; Hadar *et al*, 2017; LGPD, 2018; Wilson, Belliveau e Gray, 2017 |
| Privacidade por estrutura (= privacidade por design) | Mackenzie Adams, 2017; Jain *et al*, 2019; Ishii, 2017; Politou, Alepis e Paysakis, 2017 |
| Privacidade por política (aviso / escolha) = FIPP (Fair Information Practices Principles) | Mackenzie Adams, 2017; Hadar *et al*, 2017; Altman *et al*, 2018; Pragash e Jayabharathy, 2017 |
| Criptografia com chave secreta | Ouazzani e Bakkali, 2020; Article 29 Data Protection Working Party, 2014 |
| Função Hash | Ouazzani e Bakkali, 2020; Article 29 Data Protection Working Party, 2014 |
| Criptografia Determinística | Quazzani e Bakkali, 2020; Article 29 Data Protection Working Party, 2014 |
| Criptografia Homomórfica | Shanmugapriya e Kavitha, 2019 |
| Modelo classificação de dados sensíveis: SVM Multi-Kernel | Shanmugapriya e Kavitha, 2019 |

| | |
|---|---|
| Tokenização | Ouazzani e Bakkali, 2020; Article 29 Data Protection Working Party, 2014 |
| Ataque de correspondência não classificado (técnica ataque) | Abouelmehdi, Beni-Hessane e Khaloufi, 2018 |
| Ataque de Inferência | Rao, Krishna e Kumar, 2018 |
| Vinculação de registro (técnica de ataque) | Mehta e Rao, 2019 |
| Link Lateral (técnica de ataque) | Al-Zobbi, Shahrestani e Ruan, 2017 |
| Palpite óbvio (técnica de ataque) | Al-Zobbi, Shahrestani e Ruan, 2017 |
| Reidentificação direta (técnica de ataque) | Devi e Chamundeeswari, 2018; Politou, Alepis e Paysakis, 2017 |
| Reidentificação por meio de identificação (técnica ataque) | Devi e Chamundeeswari, 2018 |
| Reidentificação por meio de extremidades (técnica ataque) | Devi e Chamundeeswari, 2018 |
| Reidentificação por meio de sequenciamento de eventos (técnica ataque) | Devi e Chamundeeswari, 2018 |
| K-anonimato baseado em generalização | Ouazzani e Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Al-Zobbi, Shahrestani e Ruan, 2017; Snnott *et al*, 2016; Devi e Chamundeeswari, 2018; Kapil *et al*, 2020; Rani e Dhamodaran, 2016; Mehta e Rao, 2019; Rao, Krishna e Kumar, 2018; Article 29 Data Protection Working Party, 2014 |
| K-anonimato | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Jain *et al*, 2016; Rao *et al*, 2018; Metha e Rao, 2019; Rani e Dhamodaran, 2016 |

| | |
|---|---|
| Diversidade L | Ouazzani e Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Al-Zobbi, Shahrestani e Ruan, 2017; Devi e Chamundeeswari, 2018; Kapil *et al*, 2020; Mehta e Rao, 2019; Rao, Krishna e Kumar, 2018; Article 29 Data Protection Working Party, 2014; Jain *et al*, 2016 |
| Proximidade T | Ouazzani e Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Al-Zobbi, Shahrestani e Ruan, 2017; Devi e Chamundeeswari, 2018; Kapil *et al*, 2020; Rao, Krishna e Kumar, 2018; Article 29 Data Protection Working Party, 2014; Jain *et al*, 2016 |
| Substituição | Ouazzani e Bakkali, 2020; Al-Zobbi, Shahrestani e Ruan, 2017; Article 29 Data Protection Working Party, 2014 |
| Embaralhamento | Ouazzani e Bakkali, 2020; Al-Zobbi, Shahrestani e Ruan, 2017; Article 29 Data Protection Working Party, 2014 |
| HybrEx | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Jain *et al,* 2016 |
| K-anonimato baseado em supressão | Ouazzani e Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Al-Zobbi, Shahrestani e Ruan, 2017; Snnott *et al*, 2016; Devi e Chamundeeswari, 2018; Article 29 Data Protection Working Party, 2014 |
| Permutação | Ouazzani e Bakkali, 2020; Al-Zobbi, Shahrestani e Ruan, 2017; Article 29 Data Protection Working Party, 2014 |
| Randomização | Rao, Krishna e Kumar, 2018; Quazzani e Bakkali, 2020 |
| Privacidade diferencial | Ouazzani e Bakkali, 2020; Altman *et al,* 2018; Al-Zobbi, Shahrestani e Ruan, |

| | |
|---|---|
| | 2017; Jain *et al*, 2019; Snnott *et al*, 2016; Kapil *et al*, 2020; Salas e Domingo-Ferrer, 2018; Article 29 Data Protection Working Party, 2014 |
| Nulling out | Ouazzani e Bakkali, 2020; Al-Zobbi, Shahrestani e Ruan, 2017; Article 29 Data Protection Working Party, 2014 |
| Mascaramento de dados | Ouazzani e Bakkali, 2020; Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Pragash e Jayabharathy, 2017; Al-Zobbi, Shahrestani e Ruan, 2017; Thomson e Thibadeau, 2016; Article 29 Data Protection Working Party, 2014; Jain *et al*, 2016 |
| Geração de dados sintéticos | Altman *et al*, 2018 |
| Anonimização | Ouazzani e Bakkali, 2020 |
| Anonimização baseada em identidade | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Salas e Domingo-Ferrer, 2018 |
| Anonimização de ruído | Devi e Chamundeeswari, 2018 |
| Anonimização baseada em identidade | Abouelmehdi, Beni-Hessane e Khaloufi, 2018 |
| Segurança de dados | Abouelmehdi, Beni-Hessane e Khaloufi, 2018 |
| Controle de acesso (MAC – Integração e controle de acesso obrigatório) | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Pragash e Jayabharathy, 2017, Jain *et al*, 2019 |
| Segurança da informação | Abouelmehdi, Beni-Hessane e Khaloufi, 2018 |
| Autenticação | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Thomson e Thibadeau, 2016; Kapil *et al*, 2020; Salas e Domingo-Ferrer, 2018 |

| | |
|---|---|
| Criptografia | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Altman *et al,* 2018; Pragash e Jayabharathy, 2017; Jain *et al*, 2019; Shanmugapriya e Kavitha, 2019; Devi e Chamundeeswari, 2018; Thomson e Thibadeau, 2016; Kapil *et al*, 2020; Rao, Krishna e Kumar, 2018 |
| Aviso | Hadar *et al*, 2017; LGPD, 2018 |
| Confidencialidade | Hadar *et al*, 2017; Thomson e Thibadeau, 2016 |
| Especificação de propósito | Hadar *et al*, 2017; LGPD, 2018 |
| Consentimento (aberto, amplo, dinâmico, legal portátil, meta consentimento) | Hadar *et al*, 2017; LGPD, 2018; Altman, 2018; Ishii, 2017; Cheung, 2018; White *et al*, 2019 |
| Minimização de dados | Hadar *et al*, 2017; LGPD, 2018; Altman *et al*, 2018 |
| Structured Map Reduced Layer (SMR) | Jain *et al*, 2019 |
| Retificação | Hadar *et al*, 2017; LGPD, 2018 |
| Transparência algorítmica | Oostveen, 2016 |
| Discriminação algorítmica | Altman *et al*, 2018 |
| Comitê Ética *Big Data* | Altman *et al*, 2018 |
| Políticas de privacidade | Altman *et al*, 2018; LGPD; Pragash e Jayabharathy, 2017 |
| Técnicas de limitação de divulgação estatística (agregação, supressão, perturbação) | Altman *et al*, 2018 |

| | |
|---|---|
| Comitê de ética e revisão | Altman *et al*, 2018; Pragash e Jayabharathy, 2017 |
| Códigos de conduta | |
| Populações vulneráveis (ex.: crianças, mulheres grávidas, prisioneiros, etc.) | Altman *et al*, 2018 |
| Desidentificação | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Altman *et al*, 2018; Jain *et a*l, 2016 |
| Combinações de controle usadas para gerenciar riscos em *big data* | Altman *et* al, 2018 |
| Atendimento aos critérios da finalidade, adequação e necessidade | LGPD, 2018 |
| Política de comunicação quanto ao vazamento/roubo de dados | LGPD, 2018; Altman *et al*, 2018; Thomson e Thibadeau, 2016; Adams, 2017 |
| Instituição de um profissional responsável pelo tratamento de dados na organização | LGPD, 2018 |
| Estipulação de quem pode analisar quais dados | Pragash e Jayabharthy, 2017 |
| Smart Data | Ishii, 2017 |
| Transparência | Ishii, 2017; LGPD, 2018 |
| Zonas Proibidas – proibição de processamento de dados pessoais | Ishiii, 2017 |
| Inventário de ativos digitais | Thomson e Thibadeau, 2016 |
| Corretores de dados | Wilson, Belliveau e Gray, 2017 |
| Protocolos segurança e privacidade de prestadores de serviço | Wilson, Belliveau e Gray, 2017 |

| | |
|---|---|
| Ciclo de vida dos dados | Abouelmehdi, Beni-Hessane e Khaloufi, 2018; Sinnott *et al,* 2016 |
| Ciclo de vida do b*ig data* | Sarkar, 2017; Abouelmehdi, Beni-Hessane e Khaloufi, 2018 |