

CESAR SCHOOL

FERNANDO GONÇALVES DE OLIVEIRA

**ANÁLISE DO IMPACTO DAS BOAS PRÁTICAS DE GOVERNANÇA
DE DADOS ALIADO A *BIG DATA* NA ADMINISTRAÇÃO PÚBLICA
FEDERAL**

RECIFE

2020

FERNANDO GONÇALVES OLIVEIRA

**ANÁLISE DO IMPACTO DAS BOAS PRÁTICAS DE GOVERNANÇA
DE DADOS ALIADO A *BIG DATA* NA ADMINISTRAÇÃO PÚBLICA
FEDERAL**

Dissertação apresentada ao programa de Mestrado em Engenharia de Software do Centro de Estudos e Sistemas Avançados do Recife – CESAR SCHOOL, como requisito para a obtenção do título de Mestre em Engenharia de Software.

Orientação: Prof. Dr. Felipe Silva Ferraz

RECIFE

2020

Catálogo da Publicação na Fonte (CIP)
Centro de Estudos e Sistemas Avançados do Recife – CESAR School

<xxx> Oliveira, Fernando Gonçalves de

Análise do Impacto das Boas Práticas de Governança de Dados aliado a *Big Data* na Administração Pública Federal

113 p.: il.

Dissertação (Mestrado) – Centro de Estudos e Sistemas Avançados do Recife – CESAR School. Programa de Pós-Graduação em Engenharia de Software.

Orientação: Prof. Dr. Felipe Silva Ferraz

1. Governança de Dados. 2. *Big Data*. 3. Administração Pública Federal. 4. Ecossistema. 5. Survey. I. Ferraz, Felipe (Orientador). II. Título.

<xxx>

Dissertação de Mestrado apresentada por Fernando Gonçalves de Oliveira ao Programa de Mestrado em Engenharia de Software do Centro de Estudos e Sistemas Avançados do Recife – CESAR SCHOOL, intitulada ANÁLISE DO IMPACTO DAS BOAS PRÁTICAS DE GOVERNANÇA DE DADOS ALIADO A *BIG DATA* NA ADMINISTRAÇÃO PÚBLICA FEDERAL, orientada por: Felipe Silva Ferraz. A dissertação foi () pela Banca Avaliadora formada por: () e ().

Felipe Silva Ferraz
Orientador

Avaliador(a) Interno(a)

Avaliador(a) Externo(a)

Visto e permitida a impressão

Recife/PE, de de 2020.

Felipe Santana Furtado Soares
Coordenador do Mestrado Profissional em Engenharia de Software
CESAR – Centro de Estudos e Sistemas Avançados do Recife

DEDICATÓRIA

A minha mãe Oscalina e irmão Júlio César

À minha esposa Marta Joelma.

Ao meu filho e melhor amigo Kaio Felipe.

Aos meus amigos, colegas, professores e familiares por apoiar minhas decisões e por terem me dado condições para alcançar esse objetivo.

AGRADECIMENTOS

Agradeço a Deus.

Agradeço aos professores pelo conhecimento e trocas de ideias, que tornaram esse trabalho melhor. Em especial ao meu orientador Prof. Dr. Felipe Silva Ferraz.

A minha mãe, pelo amor e carinho incondicional.

À minha esposa Marta Joelma, obrigado pelo carinho, pela compreensão e pelo amor que dedica a mim. Eu te amo.

Ao meu filho Kaio Felipe pela inspiração diária de simplicidade e por me lembrar a todo o momento do que vale a pena.

Ao Cesar School que me possibilitou estudar em um ambiente de excelência tecnológico e entidade de referência de ensino superior no Brasil.

A Fábrica de Software Infinity do curso de Mestrado de Engenharia de Software do Cesar School e todos os colegas do curso.

Agradecimentos especiais a José Alberto Sousa Torres, Daniel Alves, Márcio Fernandes, Rodrigo Hagstrom e Washington Henrique por seu constante apoio e amizade e que deram um empurrãozinho necessário para chegar até aqui.

A todas as pessoas que, direta e indiretamente, me ajudaram na conquista de todos os meus objetivos pessoais e na concretização desta dissertação, o meu sincero agradecimento.

Fernando Gonçalves de Oliveira

EPÍGRAFE

“O que sabemos é uma gota. O que
ignoramos é um oceano”.

Isaac Newton

RESUMO

Com o avanço da tecnologia, as organizações buscam cada vez mais atentar-se à análise de dados com o intuito de explorar fraquezas e riquezas que contribuam com o crescimento diário das mesmas. Assim sendo, existem correlações entre as demandas técnicas da análise de dados e suas disseminações: quase todos os órgãos administrativos do setor público utilizam ferramentas de TI simples para análise, e apenas uma pequena parcela está envolvida em análises avançadas de dados usando novas tecnologias. No entanto, as organizações estão discutindo seu uso. Agências governamentais (principalmente agências alemãs) são pioneiras nessa área. Quanto mais complexa a análise dos dados, maior a satisfação demonstrada. Tais administrações acreditam que já está se convertendo suas descobertas de análise de dados em benefícios demonstráveis e específicos. Outras administrações observaram apenas efeitos diretos, como critérios aprimorados para a tomada de decisões, riscos reduzidos ou serviços individuais em alguns casos. Desta forma, esta dissertação realiza a Análise do Impacto das Boas Práticas de Governança de Dados aliado a *Big Data* na Administração Pública Federal - APF, baseando-se no princípio da governança de dados. Além disso, exploram-se subsídios para responder quais os objetivos e maiores desafios quanto à implantação de *Big Data* e/ou Ecossistemas de *Big Data* nas administrações públicas federais, apontando diretrizes (*guidelines*). A metodologia é concentrada em três fases: mapeamento sistemático das aplicações de *Big Data*, pesquisa *survey* aplicada em instituições públicas federais e, por fim, uma análise baseada em grupo focal especializado em *Big Data* para definição juntamente com a aplicação de um *guideline* como proposta para resolução de barreiras apontadas pelo grupo focal. A metodologia se relaciona com o conhecimento da literatura correlata que pode auxiliar no entendimento da questão em epígrafe. Compreende-se que a implantação de *Big Data* no setor público precisa estar alinhada com a missão institucional do órgão que se propõe a implantar *Big Data*, mas que os gestores públicos nem sempre percebem suas instituições como aptas a implantarem projetos como este, o que pode, por sua vez, demandar que seja apresentado todo um conjunto de diferentes projetos de *Big Data* em diferentes esferas e missões institucionais. Finalmente, se percebe que há pouca literatura relacionada com o caso brasileiro, o que pode destacar a importância da pesquisa para esta dissertação.

Palavras-Chave: Governança de Dados, *Big Data*, Administração Pública Federal, Ecossistema. *Survey*.

ABSTRACT

With the advancement of technology, organizations increasingly seek to pay attention to data analysis in order to explore weaknesses and riches that contribute to their daily growth. As such, there are correlations between the technical demands of data analysis and their dissemination: almost all public sector administrative bodies use simple IT tools for analysis, and only a small portion is involved in advanced data analysis using new technologies. However, organizations are discussing its use. Government agencies (mainly German agencies) are pioneers in this area. The more complex the analysis of the data, the greater the satisfaction shown. Such administrations believe that their data analysis findings are already converting into demonstrable and specific benefits. Other administrations observed only direct effects, such as improved criteria for decision making, reduced risks or individual services in some cases. Thus, this dissertation carries out the Analysis of the Impact of Good Data Governance Practices combined with *Big Data* in the Federal Public Administration - APF, based on the principle of data governance. In addition, subsidies are explored to answer the objectives and major challenges regarding the implementation of *Big Data* and / or *Big Data* Ecosystems in federal public administrations, pointing out guidelines. The methodology is concentrated in three phases: systematic mapping of *Big Data* applications, survey research applied in federal public institutions and, finally, an analysis based on a focus group specialized in *Big Data* for definition together with the application of a guideline as a proposal for resolution of barriers identified by the focus group. The methodology relates to the knowledge of the related literature that can help in understanding the question in question. It is understood that the implementation of *Big Data* in the public sector needs to be aligned with the institutional mission of the body that proposes to implement *Big Data*, but that public managers do not always perceive their institutions as able to implement projects like this, which can, in turn, demand that a whole set of different *Big Data* projects be presented in different spheres and institutional missions. Finally, it is noticed that there is little literature related to the Brazilian case, which can highlight the importance of research for this dissertation.

Keywords: Data Governance, *Big Data*, Federal Public Administration, Ecosystem. Survey.

SUMÁRIO

1. INTRODUÇÃO.....	17
1.1. Motivação.....	18
1.1.1 Motivação de mercado e barreiras	18
1.1.2 Motivação técnica	19
1.2. Problemática e Justificativa.....	20
1.3. Objetivos	22
1.3.1 Objetivo Geral.....	22
1.3.2 Objetivos específicos	22
1.4. Contribuições	22
1.5. Ilustração metodológica	23
1.6. Estrutura da dissertação	24
2. REFERENCIAL TEÓRICO	24
2.1. Governança de Dados	24
2.1.1 Concepção histórica e sua relevância para o cenário	25
2.1.2 Riscos aplicados à Governança de Dados	28
2.1.3 Entendendo sobre o DAMA-DMBOK: Conceitos e Diretrizes.....	30
2.1.4 Aplicação da governança de dados.....	32
2.2. <i>Big Data</i> e suas aplicações	34
2.2.1 Conceito histórico e atribuições.....	35
2.2.2 Aplicações potenciais de <i>Big Data</i> para a Administração Pública Federal e suas restrições.....	39
2.2.3 Recursos de dados disponíveis para a Administração Pública.....	41
2.2.4 Requisitos da Administração Pública: Técnicos e não técnicos.	43
2.2.5 A importância do <i>Big Data</i> na Administração Pública Federal.....	45
2.3. Governança de Dados aliado ao <i>Big Data</i>	48
2.3.1 Por que a Governança de Dados é crucial para ambientes de <i>Big Data</i> ? 49	
2.3.2 O papel integral da Governança de Dados em ambientes de <i>Big Data</i> .51	
2.3.3 Como as diferenças de <i>Big Data</i> podem afetar a Governança de Dados 54	
2.4. Aplicações de Governança de Dados e <i>Big Data</i> em Instituições Públicas Federais: Uma visão globalizada	55
3. METODOLOGIA E APLICAÇÃO DA PESQUISA	59
3.1. Compreendendo as etapas metodológicas	59

3.2.	Apresentação do Mapeamento Sistemático.....	60
3.3.	Apresentação da pesquisa <i>Survey</i>	62
3.3.1	Etapas da aplicação do questionário	65
3.4.	Apresentação da pesquisa referente ao Grupo Focal	67
4.	RESULTADOS E DISCUSSÕES.....	68
4.1.	Análise da aplicação da pesquisa <i>Survey</i>	68
4.2.	Análise da aplicação da pesquisa de grupo focal	90
4.2.1	Perfil dos participantes.....	90
4.2.2	Avaliação de Prontidão para <i>Big Data</i> – Grupo Focal.....	90
4.2.3	Entendendo o cenário atual	93
4.2.4	<i>Guideline</i> para implantação de Governança de Dados e <i>Big Data</i> em Instituições Públicas Federais	95
5.	CONSIDERAÇÕES FINAIS.....	103
5.1.	Contribuições e Trabalhos Futuros	104
	REFERENCIAL BIBLIOGRÁFICO	108
	APÊNDICE A – MAPEAMENTO SISTEMÁTICO	114
	APÊNDICE B – QUESTIONÁRIO APLICADO	126
	APÊNDICE C – QUESTIONÁRIO APLICADO – GRUPO FOCAL	130

LISTA DE IMAGENS

Figura 1 – Apresentação ilustrativa da metodologia.....	23
Figura 2 - Demonstrativo da utilização de governança de dados.....	26
Figura 3 – Modelo de <i>framework para gestão de dados</i>	29
Figura 4 – Modelo DAMA-DMBOK, segundo o guia.	31
Figura 5 - Demonstração de <i>Big Data</i> através dos 5 Vs.	36
Figura 6 – Governança de dados integrados com o ambiente corporativo	50
Figura 7 – Compreensão das etapas metodológicas.	60
Figura 8 – Etapas do Mapeamento Sistemático.....	61
Figura 9 – Síntese dos procedimentos de pesquisa bibliográfica.....	62
Figura 10 – A Concepção e a Função do Questionário.....	63
Figura 11 – Etapas da Pesquisa.	64
Figura 12 – Etapas da <i>Guideline</i> para implantação de Governança de Dados e <i>Big Data</i>	96
Figura 13 – Modelo para hierarquia A e B.....	97

ÍNDICE DE TABELAS

Tabela 1 - Critérios dos cenários de aplicação para a Administração Pública.	39
Tabela 2 – Avaliação de Prontidão para <i>Big Data</i> – Grupo Focal.	91

LISTA DE GRÁFICOS

Gráfico 1 – Cargos ocupados pelos entrevistados na Administração Pública.....	68
Gráfico 2 – Em qual poder trabalha o entrevistado	69
Gráfico 3 - Tempo de experiência com T.I do entrevistado	70
Gráfico 4 - Volumetria de armazenamento dos órgãos.	71
Gráfico 5 - Projeção de crescimento do banco de dados.....	72
Gráfico 6 - Possui projeto para ambiente próprio de <i>Big Data</i> ?	73
Gráfico 7 - Diferenciação entre os termos distintos de <i>Big Data</i>	74
Gráfico 8 - Considera que todos os termos acima possuem forte relação com o tema <i>Big Data</i> ?.....	75
Gráfico 9 - Tipos de armazenamento de dados do órgão.	75
Gráficos 10 – Possui dados não estruturados?	76
Gráficos 11 - Existe algum problema de negócio que poderia ser resolvido com <i>Big Data</i> ?	77
Gráficos 12 - Perspectiva para implantação de <i>Big Data</i> nos órgãos.	78
Gráficos 13 - O órgão já possui projetos de <i>Big Data</i> ?	79
Gráficos 14 - O que tem motivado o uso de <i>Big Data</i> no seu órgão público?	79
Gráficos 15 - Dificuldades encontradas na implementação do ambiente de <i>Big Data</i>	80
Gráficos 16 - Impactos com a implantação de <i>Big Data</i>	81
Gráficos 17 - Os impactos econômicos da adoção de <i>Big Data</i> nos órgãos.....	82
Gráficos 18 - As oportunidades de implementação do <i>Big Data</i> nos órgãos públicos.	82
Gráficos 19 - Os impactos que práticas de <i>Big Data</i> podem/poderiam trazer ao desenvolvimento de soluções nos órgãos públicos.....	83
Gráficos 20 - Os impactos citados na questão anterior são mais positivos que negativos nos órgãos públicos?	84
Gráficos 21 - A sua resposta à questão anterior envolveu impactos existentes e comprovados (Sim) ou impactos percebidos pelo conhecimento acerca de <i>Big Data</i> conhecido pelo respondente do questionário (Não).	85
Gráficos 22 - O órgão compartilha conhecimentos e aplicações de <i>Big Data</i> entre todos os setores do órgão?	86
Gráficos 23 - A equipe de TI e profissionais de dados compartilham conhecimentos entre todos os setores da organização?	86

Gráficos 24 - No ecossistema de <i>Big Data</i> é explorado ou adotado serviços de nuvem para processar ou conduzir o Analytics?	87
Gráficos 25 - Quantos profissionais de cientistas de dados o órgão possui servidores públicos?	88
Gráficos 26 - Quantos profissionais de cientistas de dados o órgão possui terceirizado?	88

ABREVIATURAS

Sigla	Significado
APF	Administração Pública Federal
API	Interface de Programação de Aplicativos
BI	Business Intelligence
CEO	Chief Executive Officer
COBIT	Control Objectives for Information and Related Technologies
COSO	Committee of Sponsoring Organization of the Treadway Commission
DGL	Data Governance Leaders
DW	Data Warehouse
GRC	Governança, Risco e Conformidade
IA	Inteligência Artificial
ITIL	Information Technology Infrastructure Library
OCDE	Organização para Cooperação e Desenvolvimento Econômico
PDF	Portable Document Format
PIB	Produto Interno Bruto
POSIX	Portable Operating System Interface
PSI	Informações do Setor Público
RFID	Radio Frequency Identification
RM	Ranking Médio
TCLE	Termo de Consentimento Livre e Esclarecido
TI	Tecnologia da Informação
TIC	Tecnologia da Informação e Comunicações

1. INTRODUÇÃO

Este trabalho analisa os impactos das boas práticas de Governança de Dados aliado a *Big Data* na Administração Pública Federal do Brasil, apontando desafios que os órgãos governamentais competentes encontram ao implantar tecnologias que tratam grandes volumes de dados. Tais tecnologias serão discutidas ao longo da dissertação com o intuito de desmistificar sua importância e aplicabilidade para a concepção de diretrizes sobre *Big Data*.

Os governos geram e coletam grande quantidade de dados por meio de suas atividades cotidianas (GONÇALVES & VARELLA, 2018). Embora os órgãos governamentais estejam cada vez mais reconhecendo o valor potencial derivado do uso da Governança de Dados e *Big Data*, há um déficit significativo no número de pesquisas publicadas sobre este tema quando comparamos o setor público ao setor privado.

No entanto, a análise de grandes volumes de dados, se constitui, sim, em um desafio para os órgãos componentes do setor público, assim como ocorre em organizações privadas. Uma razão para isso é que a legislação acerca de Governo Eletrônico ainda não foi implementada por completo (SOUZA, 2019). Em adição, há muito a ser feito em relação à segurança de TI.

Os benefícios das aplicações relacionadas à Governança de dados e do *Big Data* no setor público podem ser agrupados em três áreas principais, com base na classificação dos tipos de benefícios (REIS, 2015):

- análise avançada, por meio de algoritmos automatizados;
- melhorias na eficácia, proporcionando maior transparência interna;
- melhorias na eficiência, nas quais serviços podem ser oferecidos com base na personalização dos mesmos; e aprendizado com o desempenho de tais serviços.

Os apontamentos realizados por Reis (2015) enfatizam a importância da aplicação de tecnologias voltadas para grandes volumes de dados. No setor público são tratados vários segmentos de informações de diferentes setores, demandando uma análise apurada desses dados, a fim de auxiliar a tomada de decisões que tragam benefícios a curto e longo prazo.

Algumas ideias iniciais parecem evidentes com base na literatura correlata, como o fato de que é necessário desenvolvimento para os campos de escalabilidade da análise de dados, descoberta de padrões e aplicativos em tempo real. Ainda, melhorias na procedência do compartilhamento e integração de dados do setor público, e o fornecimento de mecanismos integrados de segurança e privacidade em aplicativos de *Big Data*.

1.1. Motivação

1.1.1 Motivação de mercado e barreiras

O ambiente de trabalho, foco da pesquisa, é composto por órgãos públicos ligados ao Governo Federal. O Brasil, por sua vez, é um comprador de produtos e serviços de Tecnologia da Informação - TI na América Latina.

Assim sendo, o ambiente privado de TI, especialmente na cidade de Brasília, é atingido pelas demandas do Governo Federal do Brasil. Deste modo, as tecnologias de Governança de Dados e *Big Data* passam a ser consideradas cada vez mais necessárias para busca de respostas que permitam uma melhor orientação de esforços nestes órgãos do Governo Federal ou mesmo de outras esferas governamentais no Brasil.

Algumas características que diferenciam o ambiente de trabalho de órgãos públicos do mercado privado é o fato de não existir concorrência direta entre estes órgãos. Este fato pode gerar ambiente de cooperação entre os aludidos, uma vez que o setor público não se trata de um setor produtivo, embora suas despesas representem uma “fatia” considerável do PIB do país.

A maior parte da receita do setor é arrecadada por meio de impostos e contribuições sociais. Assim, o impacto das tecnologias de *Big Data* está relacionado com aspectos de eficiência e eficácia: quanto mais eficiente for o setor público, melhor atendidos serão os cidadãos e menos recursos (impostos) serão necessários para fornecer o mesmo nível de serviço.

Outra característica particular do setor público é que este abrange a aplicação de Governança de Dados e *Big Data* para fornecer maior transparência, ou seja, dar acesso às informações dos governos aos cidadãos. Informações compiladas por sistemas de *Big Data* e Mineração de Dados tendem a fornecer insumos para que

cidadãos e empresas possam, baseados em dados concretos, participar do processo de tomada de decisão.

Em uma última análise, a qualidade dos serviços prestados pelo setor público também pode ser aprimorada com o uso de tecnologias de Governança de Dados e *Big Data*. Um bom exemplo é o direcionamento de recursos para serviços como educação, saúde, serviços sociais, políticas ativas e de segurança com base em informações geradas pelos sistemas sobre a eficiência e demanda de cada um desses setores.

1.1.2 Motivação técnica

Projetos de implantação de Governança de Dados e *Big Data*, assim como qualquer projeto, necessitam estar alinhados com os objetivos estratégicos da organização que os utilizam. Tal problemática será abordada ao longo do texto do presente documento. Em outras palavras, técnicas de *Big Data*, por si só, não são suficientes para fornecer subsídios de melhoria da atuação de qualquer organização.

No entanto, aspectos tecnológicos da Governança de Dados e *Big Data* são abordados, ainda que alguns órgãos não tenham encontrado o perfeito alinhamento entre técnicas e projetos de *Big Data* e sua missão institucional. Logo, como motivação técnica, temos que, em princípio, os benefícios da Governança de Dados e *Big Data* no setor público podem ser agrupados em três áreas principais, com base na classificação dos tipos de benefícios: a análise de *Big Data*, os seus processos e melhorias na eficácia e eficiência das rotinas e incrementos nos processos de tomada de decisão.

Para tanto, o uso de sistemas de suporte a tomada de decisão com uso de ferramentas de *Big Data* pode produzir, por exemplo, conjuntos de resultados que revelam padrões ou mecanismos mais eficientes de previsões de comportamento. Isso pode melhorar o processo decisório, desde que os resultados sejam trabalhados de forma correta e em alinhamento com as políticas públicas. Vale ressaltar que tais ferramentas também incluem a administração de poder computacional, aplicativos de Mineração de Dados, dentre outros.

Nesse contexto, essa área abrange os aplicativos que fornecem serviços de melhoria contínua, com base na personalização de serviços e aprendizado com o desempenho dos mesmos.

Tais apontamentos serão mensurados a partir da análise descritiva de um questionário que foi disponibilizado para órgãos públicos evidenciando um conjunto de fatores que podem beneficiar ou restringir as diretrizes de *Big Data*. Este questionário é parte importante deste objeto de estudo, e será abordado mais adiante neste trabalho.

1.2. Problemática e Justificativa

O problema foco da pesquisa é responder quais são as restrições e desafios relativos à implantação e uso das tecnologias de Governança de Dados e *Big Data* no setor público. Além disso, queremos conhecer onde os governos podem atuar como catalisadores no desenvolvimento de um ecossistema de dados por meio da abertura de seus próprios conjuntos de dados, gerenciando sua disseminação e uso (objetivos e oportunidades).

O setor público está se tornando cada vez mais consciente do valor potencial a ser obtido com o *Big Data*, à medida que os governos geram e coletam grandes quantidades de dados por meio de suas atividades cotidianas (YIU, 2012).

Os benefícios da Governança de Dados e o *Big Data* no setor público podem ser agrupados em três áreas principais, com base na classificação dos tipos de benefícios: análise avançada, por meio de algoritmos automatizados; melhorias na eficácia, proporcionando maior transparência interna; melhorias na eficiência, onde melhores serviços podem ser fornecidos com base na personalização dos serviços; e aprender com o desempenho de tais serviços.

Desta forma, examinam-se vários fatores e restrições que foram identificados, que podem impulsionar ou interromper o desenvolvimento de *Big Data* no setor, dependendo de como eles são abordados. As descobertas, após analisar os requisitos e as tecnologias atualmente disponíveis, mostram que há questões de pesquisa abertas a serem abordadas, ou seja, problemas para o desenvolvimento de tais tecnologias, para que soluções competitivas e eficazes possam ser construídas. Algumas dessas questões são:

- Por que tecnologias relacionadas ao *Big Data* não foram difundidas pelos órgãos públicos?
- Por que o governo negligencia uma base de dados que poderia ser trabalhada a seu favor, com as práticas de *Big Data*?
- Ou ainda, por que as empresas que fornecem soluções de governança de dados não são acionadas pelos governos, sabendo que, internamente, não existem projetos que deem continuidade na governança de dados e consequentemente na prática de *Big Data*?

Campos que lidam com a escalabilidade da análise de dados, descoberta de padrões e aplicativos em tempo real ainda precisam de um desenvolvimento relevante. Também são necessárias melhorias na procedência do compartilhamento e integração de dados do setor público. Além disso, é extremamente importante fornecer mecanismos integrados de segurança e privacidade em aplicativos de *Big Data*, pois o setor público coleta grandes volumes de dados confidenciais. Até mesmo porque respeitar a privacidade dos cidadãos é uma obrigação do setor público nacional.

Sabe-se que atividades governamentais como o gerenciamento de pensões e pagamentos de subsídios, cobrança de impostos, sistemas nacionais de saúde, registro de dados de tráfego e emissão de documentos oficiais geram um volume de dados absurdos diariamente. Além disso, são notórias as atuais tendências socioeconômicas e tecnológicas, com o aumento da produtividade em um ambiente com restrições orçamentárias significativas, a crescente demanda por serviços médicos e sociais e a padronização e interoperabilidade como requisitos importantes para as aplicações e tecnologias do setor público. Com isso em mente, o problema da pesquisa busca responder o seguinte:

Os órgãos públicos brasileiros estão devidamente preparados para a utilização das diretrizes de governança de dados e *Big Data*?

Como justificativa, destacam-se as de dados abertos, que são um ponto de partida para impulsionar um mercado de dados e das tecnologias de *Big Data*. Adicionalmente, as necessidades de melhoria de tomada de decisão, acrescida da possibilidade de maior participação do cidadão junto aos órgãos de governo.

As políticas ativas na área de dados abertos podem beneficiar o setor privado e, conseqüentemente, facilitar o seu crescimento no Brasil. Em contrapartida, isso também tende a beneficiar os orçamentos públicos com o melhor aproveitamento da receita tributária de uma crescente indústria de dados em âmbito nacional.

1.3. Objetivos

1.3.1 Objetivo Geral

Por meio da análise de boas práticas e aplicações da Governança de Dados e *Big Data* no setor público brasileiro, em âmbito federal, busca-se delimitar quais os objetivos e maiores desafios e dificuldades relacionadas com a adoção das tecnologias no aludido cenário. Para isso, lançaremos mão de uma pesquisa do tipo *survey* em instituições coligadas ao tema e de um grupo focal que ajudará na construção de um *guideline* como proposta para a adoção de *Big Data* para estas instituições.

1.3.2 Objetivos específicos

Os objetivos específicos consideram um escopo mais focado em aspectos relacionados com as necessidades e com os papéis dos órgãos públicos mediante a sociedade administrada, quais sejam, buscam responder às seguintes questões:

- Qual é o papel da governança de dados no setor público federal?
- Como governança de dados, pode melhorar o cenário problemático e econômico do Brasil?
- Quais são os impactos econômicos das soluções de *Big Data* no setor público?
- Quais as oportunidades e desafios da adoção de *Big Data* no setor público?
- Quais os impactos das práticas de *Big Data* no setor público?
- Quais problemas ligados ao setor público que dificultam as práticas e diretrizes de *Big Data*?

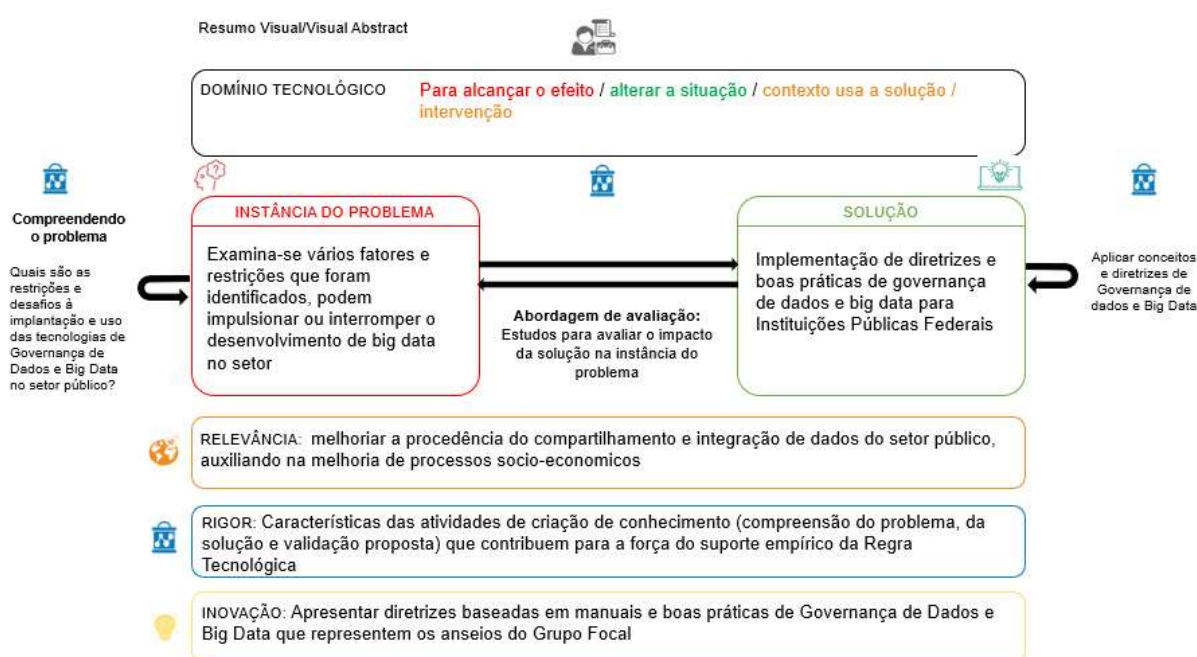
1.4. Contribuições

- Fornecer dados concisos que comprovem que a prática de *Big Data*, pode interferir de maneira positiva ou não na Administração Pública Federal;

- Contribuir com o estímulo de incremento de eficiência e eficácia dos processos e consequentemente na análise dos dados dos setores públicos;
- Dissertar sobre os processos de *Big Data* disseminando boas práticas na administração pública, contribuindo com literatura técnica e científica para o corpo acadêmico e futuros leitores que possam se interessar pelo tema.
- Dissertar sobre as diretrizes coligadas a Governança de Dados;
- Evidenciar a importância da governança de dados em conjunto com *Big Data*;
- Elaborar um *guideline* com o propósito de guiar os leitores na adoção de *Big Data* em conformidade com governança de dados, aplicadas a instituições públicas federais.

1.5. Ilustração metodológica

Figura 1 – Apresentação ilustrativa da metodologia



Fonte: próprio autor

1.6. Estrutura da dissertação

A estrutura desta dissertação pode ser compreendida na seguinte forma: a primeira seção diz respeito à introdução da temática, bem como sua importância e relevância para o meio acadêmico e profissional, relatando a problemática, justificativa, bem como o objetivo central e objetivos específicos que compõe este objeto de estudo. Posteriormente apresentamos as motivações técnicas e de mercado, que impulsionaram o autor para a realização desta dissertação.

A segunda seção destina-se ao desenvolvimento do referencial teórico desta dissertação, onde serão apresentados conceitos, diretrizes, aprofundando o conhecimento técnico e prático relacionado à temática deste objeto de estudo.

Na terceira seção apresenta-se a metodologia utilizada nesta dissertação. São apresentadas as práticas, métodos e aplicações para a exploração do estudo de caso que será apresentado na quarta seção.

A quarta seção traz as aplicações da metodologia, baseadas em três fases, onde: (i) explora-se a pesquisa *survey* através de um questionário aplicado a instituições públicas federais; (ii) aplica-se um questionário a um grupo focal; e (iii) apresentam-se as diretrizes propostas como solução para a aplicação das tecnologias para Governança de dados e *Big Data*.

Por fim, a quinta seção, destina-se a apresentação dos resultados, contribuições e conclusões desta dissertação.

2. REFERENCIAL TEÓRICO

Nesta seção são apresentados os tópicos correlatos ao tema tratado, por meio da revisão da literatura. Inicialmente, os tópicos são abordados de forma mais direta, considerando não apenas os conceitos, mas também sua importância e implicações diretas e indiretas no cenário proposto.

2.1. Governança de Dados

Nesta seção é apresentada uma concepção histórica sobre Governança de Dados, relatando sua real importância para o meio educacional, comercial e industrial.

A seguir, são apresentados ao leitor os riscos de se aplicar Governança de Dados, principalmente quando esta não é centralizada em setores que controlem

adequadamente suas diretrizes. Tais correlações são importantes para o entendimento de suas atribuições junto aos órgãos federais que será discutido posteriormente.

Por fim, traremos as aplicações de Governança de Dados. Tais evidências mostram requisitos benéficos para organizações que conseguem trabalhar suas diretrizes de forma eficiente, mostrando sua relevância ao tema proposto.

Dentre as várias situações, apresentaremos constantemente as relações entre os tópicos e a real situação governamental dentro das suas diretrizes.

2.1.1 Concepção histórica e sua relevância para o cenário

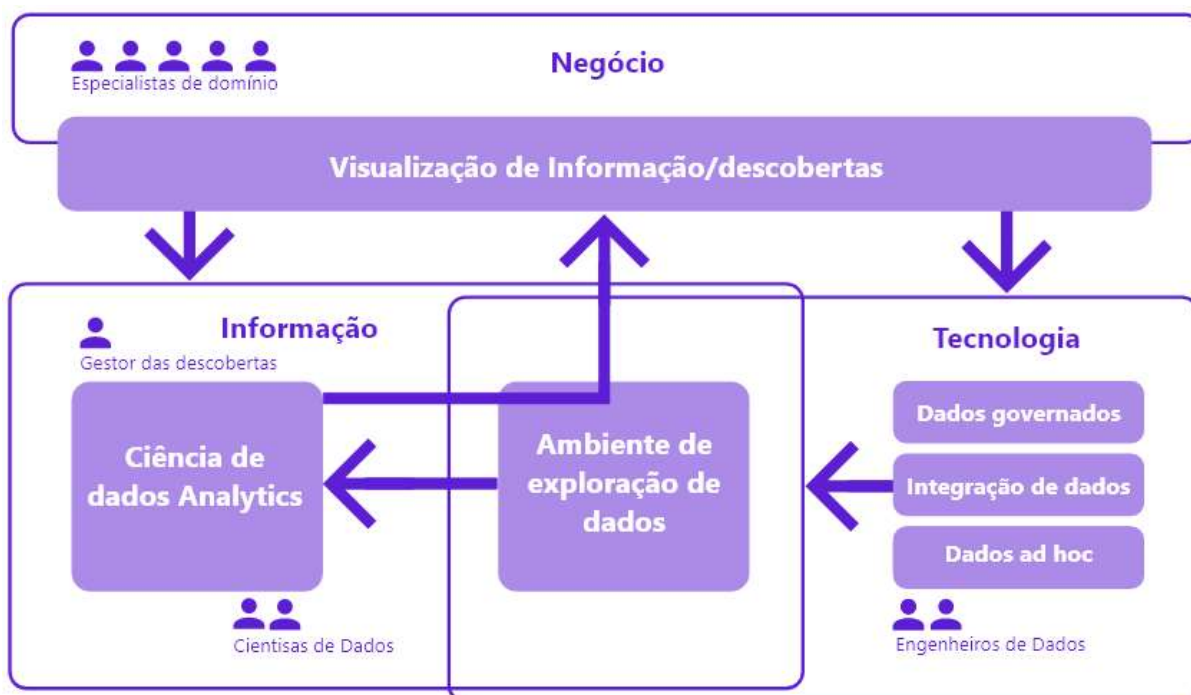
O *Data Governance* (Governança de dados), é uma estrutura de quatro vias que inclui disponibilidade, aplicabilidade, integridade e segurança. É um conjunto de processos, usado pelas partes interessadas que usam a tecnologia, para garantir que os dados importantes e críticos sejam gerenciados e protegidos (ESPINDOLA, et. al., 2017).

Envolve uma coordenação simplificada de indivíduos (pessoas), métodos (processos) e inovação (tecnologia) em uma ordem que resulta na realização do valor dos dados para qualquer organização (BARATA, 2015). Ela atua como uma ponte entre negócios e TI para decisões e iniciativas.

Embora pessoas, processos e tecnologia estejam no centro de uma boa governança de dados, este último é apenas um facilitador para a governança de dados. Ela deve ser implementada como um fluxo de trabalho disciplinado dentro da organização. Sem essa disciplina, os dados nunca seriam tratados como uma mercadoria valiosa.

Abaixo, na Figura 2, Stumpf (2017) mostra a importância de Governança de Dados nas organizações.

Figura 2 - Demonstrativo da utilização de governança de dados



Fonte: Stumpf (2017) – Modelo adaptado pelo autor

A Governança de Dados é necessária para garantir que os ativos de informações de uma organização sejam gerenciados de maneira formal, adequada, proativa e eficiente em toda a empresa para garantir confiabilidade e responsabilidade.

O *Data Governance* compreende a coleta de dados, a revisão e a padronização e a otimização do uso. Torna os dados consistentes. A Governança de Dados garante que os dados críticos estejam disponíveis no momento certo para a pessoa certa, de forma padronizada e confiável (CARVALHO, 2012). Isso infere uma melhor organização das operações de negócios. A adoção e implementação da Governança de Dados podem resultar em maior produtividade e eficiência de uma organização.

Contudo, o que não se sabe, não pode ser protegido. Esse ditado é ainda mais importante para a segurança cibernética. Pesquisas líderes mostram que 65% não têm a capacidade de detectar violações de dados com o *Data Governance*, que deve ser visto como um esforço da empresa (DGI, 2017).

Pode-se implementar em segmentos, mas deve-se sempre e necessariamente ter uma perspectiva corporativa. Para isso, primeiro deve ser estabelecido um órgão de governança que crie a estratégia e as políticas necessárias para a organização.

Este não é um exercício único, mas um esforço contínuo que requer monitoramento e manutenção.

O monitoramento, manutenção e revisão contínua dos dados são cruciais e importantes. O sucesso de qualquer programa de Governança de Dados pode ser obtido através da comunicação eficaz entre TI, Unidades de Negócios e partes interessadas em dados.

Outros fatores de sucesso incluem documentação e comunicação suficientes de alterações, problemas e tarefas em um plano de comunicação simplificado e predefinido. Um programa de governança de dados seguro e de alto desempenho ajudará qualquer organização a aproveitar seus dados genuínos e dignos dos negócios com a ajuda de soluções tecnológicas apropriadas, quando necessário.

O objetivo final de qualquer programa de Governança de Dados é permitir que as organizações reajam efetivamente aos desafios e oportunidades colocados pelo mercado, adotando a filosofia "*Data First*" (BARATA, 2015).

Como no caso de qualquer outro programa, as diferenças culturais e as abordagens isoladas para aquisição e gerenciamento de dados criam barreiras na formação e manutenção de um programa de Governança de Dados, além da falta de um patrocínio comercial sênior sustentado.

Uma base inadequada e a ausência de métricas são outros motivos para uma falha na Governança de Dados. Uma base adequada inclui: gerenciamento de dados, modelos de dados, metadados etc. Fundamentalmente, uma base sólida do que precisa ser governado é o requisito mais básico para a Governança de Dados (ESPINDOLA, et. al., 2017).

Uma organização não pode ser transformada da noite para o dia, especialmente em um programa de trabalho intensivo, como o *Data Governance*. Esperar uma mudança radical de um estágio "sem programa" para um programa "em toda a empresa" em alguns meses é uma expectativa irreal e pode levar à falha no controle de dados.

Subestimar a quantidade de trabalho a ser realizado, excesso de planejamento e menos execução, falta de compromisso de linha/negócio sustentado e ausência de equipe/escritório do programa também são alguns dos motivos que contribuem para a falha na governança de dados. Planejar, definir o escopo e executar um programa

de Governança de Dados pode ser desafiador, portanto pode ser feito em etapas para alcançar os resultados desejados.

2.1.2 Riscos aplicados à Governança de Dados

Em um ambiente de negócios volátil e de mudança dinâmica, o gerenciamento de riscos e a garantia da informação desempenham um papel crítico no crescimento e sustentação gerais de qualquer organização. Devido à crescente dependência de infraestrutura aprimorada digitalmente, os CEOs enfrentam o desafio assustador de encontrar o equilíbrio certo entre proteção e progresso, bem como privacidade e governança.

Além do dilema supracitado, há uma postura de segurança fragmentada e, às vezes desfocada, resultante de um mandato multirregulatório. Aí reside a necessidade de uma estrutura holística de risco e conformidade da governança.

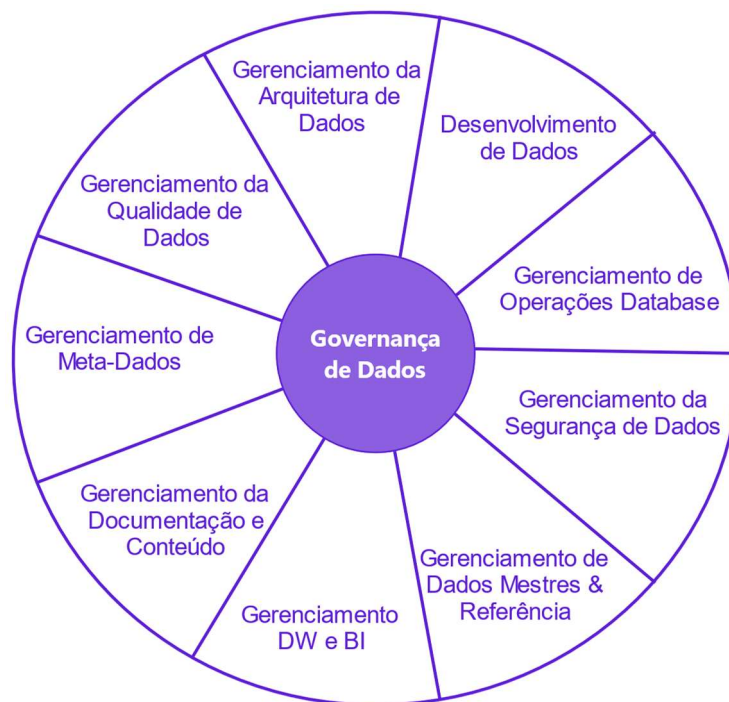
Santos (2010) destaca alguns riscos de governança e desafios de conformidade, como:

- Falta de visibilidade em dados críticos e ativos cruciais;
- Impacto e perdas devido a uma violação ou ameaça percebida;
- Cobertura incompleta de controles e ativos de segurança;
- Otimização de riscos de governança e ferramentas de gerenciamento de conformidade; e
- Visibilidade do risco atual de governança e postura de conformidade;

Governança, Risco e Conformidade (GRC) referem-se a uma estratégia para coordenar a governança geral de uma organização, o gerenciamento de riscos corporativos e a conformidade com os regulamentos. Devemos pensar no GRC como uma abordagem estruturada para alinhar a TI aos objetivos de negócios, enquanto gerencia efetivamente os riscos e atende aos requisitos de conformidade (SANTOS, 2010).

A Figura 03 ilustra um modelo sugerido por Stumpf (2017), com uma visão sistemática para um *framework* de gerência de dados, com base no modelo de DAMA-DMBOK.

Figura 3 – Modelo de *framework* para gestão de dados



Fonte: Stumpf (2017) – Modelo adaptado pelo autor

Gartner (2016) define que em um ambiente de TI o GRC possui três componentes principais:

- **Governança:** garante que as atividades organizacionais, como o gerenciamento de operações de TI, estejam alinhadas de maneira a apoiar os objetivos de negócios da organização.
- **Risco:** Certifica de que qualquer risco (ou oportunidade) associado às atividades organizacionais seja identificado e tratado de maneira a apoiar os objetivos de negócios da organização. No contexto de TI, isso significa ter um processo abrangente de gerenciamento de riscos, que entra na função de gerenciamento de riscos corporativos de uma empresa.
- **Conformidade:** garante que as atividades organizacionais sejam operadas de maneira a atender às leis e regulamentos que afetam esses sistemas. No contexto de TI, isso significa garantir que os sistemas de TI e os dados contidos nesses sistemas sejam usados e protegidos adequadamente.

O cumprimento da conformidade envolve controles de TI, bem como a auditoria desses controles para garantir que eles estejam funcionando conforme o planejado. As organizações também usam controles para gerenciar riscos identificados. De fato,

o termo "GRC" surgiu no início dos anos 2000, após muitos desastres financeiros corporativos amplamente divulgados, culminando em empresas lutando para melhorar seus processos de controle interno e governança (GARTNER, 2016).

As organizações desenvolvem uma estrutura de GRC para a liderança, organização e operação das áreas de TI da organização, para garantir que elas apoiem e permitam os seus objetivos estratégicos. Embora existam boas opções de software disponíveis para ajudar a simplificar as operações de GRC, o GRC é mais do que um conjunto de ferramentas de software.

Muitas organizações consultam uma estrutura para orientação no desenvolvimento e aprimoramento de suas funções GRC, em vez de criar uma a partir do zero. Estruturas e padrões fornecem blocos de construção que as organizações podem adaptar ao seu ambiente. Pode-se destacar, o COBIT, o COSO e o ITIL como os grandes players em diversos setores.

2.1.3 Entendendo sobre o DAMA-DMBOK: Conceitos e Diretrizes

Esta seção traz um resumo acerca da importância do guia DAMA-DMBOK (2018), uma grande fonte de conhecimento de diferentes áreas relacionadas ao gerenciamento de dados, reunidas e resumidas por uma forte equipe de profissionais.

O guia apresenta a 'Roda DAMA' (ver Figura 3), que explica as nove principais áreas de conhecimento. O modelo Hexágono de Fatores Ambientais é a base para a descrição de cada área de conhecimento e inclui, por exemplo, produtos, atividades, ferramentas etc. Mas ainda assim, o modelo de gerenciamento de dados DAMA tem suas próprias especificações, que consistem no seguinte:

- Pode ser aplicado a uma ou a várias áreas do conhecimento;
- Pode ser implementado, na prática, por meio de *frameworks* específicos.

Cada uma das áreas de conhecimento pertence a diferentes categorias. Por exemplo, Metadados e Referência e Dados Mestres são relacionados a dados, enquanto Armazenamento de Dados e Operações têm baixo aspecto tecnológico e operacional. Então, como devemos lidar com isso? Na verdade, a ideia é cada grupo de profissionais envolvidos nos processos de governança de dados, baseando-se na Roda DAMA, encontrar a "sua própria maneira" de realizar as aplicações práticas.

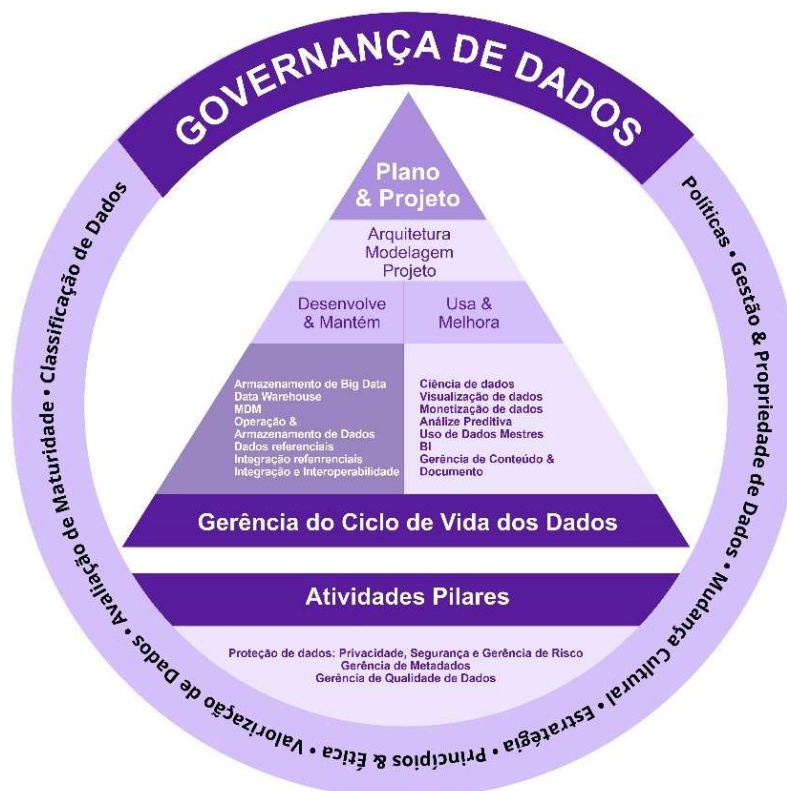
Ressalta-se que existem diretrizes que apoiam e norteiam tais procedimentos, porém o guia é claro com relação a sua prática.

Desta forma, o guia DAMA-DMBOK (2018), ressalta que existem três pilares principais que suportam o modelo de gerenciamento de dados DAMA:

- O gerenciamento de dados é uma função de negócios, assim como finanças, vendas, etc.;
- O gerenciamento de dados é considerado parte da organização de Tecnologia da Informação;
- O gerenciamento de dados é multifuncional; requer uma gama de habilidades e conhecimentos.

O assunto principal do gerenciamento de dados são os dados. O DAMA-DMBOK (ver Figura 4), do ponto de vista do conteúdo, classifica os dados em quatro categorias: referência, mestre, transacional e metadados. Desta forma, desperta a curiosidade de saber por que o DAMA-DMBOK considera os dados de referência e mestre como uma área de conhecimento separada e negligencia os dados transacionais.

Figura 4 – Modelo DAMA-DMBOK, segundo o guia.



Metadados sendo "dados sobre dados" descrevem todos os outros tipos. Cada tipo de dado, porém, tem sua qualidade. Portanto, "Qualidade de Dados", de acordo com DAMA-DMBOK, também é uma área de conhecimento. Arquitetura de dados e modelagem e design de dados são ferramentas que, de uma forma ou de outra, auxiliam na descrição da estrutura e do processamento de dados. A Arquitetura de Dados se concentra na documentação de fluxos de dados e cadeias de valor de dados. *Data Modeling and Design* é dedicado a definir requisitos de dados, fornecendo diferentes modelos de dados DAMA-DMBOK (2018).

Os dados também podem ser classificados pela forma como são armazenados. Assim, o DAMA-DMBOK apresenta uma classificação de dados estruturados e não estruturados. Os dados não estruturados ganham destaque na área de conhecimento de "Gerenciamento de Documentos e Conteúdo".

Os dados estruturados são mencionados em três áreas de conhecimento. "Armazenamento de Dados e Operações" fornece informações sobre design, implementação e operação de bancos de dados. A movimentação da organização e a consolidação de dados são o núcleo de "Integração e Interoperabilidade de Dados", e "DW e BI" se concentra em relatórios de dados.

Observando com atenção a Figura 04, essas três áreas de conhecimento estão relacionadas aos aspectos tecnológicos e operacionais do processamento de dados. No modelo, todos eles estão rodeados pelo retângulo "Tecnologia". A segurança de dados é projetada pela arquitetura e implementada por todos os blocos relacionados à tecnologia. A governança de dados é abrangente e diz respeito a processos, procedimentos e funções de negócios de gerenciamento de dados.

2.1.4 Aplicação da governança de dados

Por que a governança de dados é importante para as empresas? Isso pode parecer uma pergunta retórica. O poder dos dados em impulsionar as empresas hoje é bem reconhecido. Os dados são poderosos e é melhor que as ideias sejam orientadas por dados. Então, segue-se que a governança de dados também é importante. Uma empresa possui muitos dados recebidos externamente e mais dados são criados ou atualizados internamente, de modo que provavelmente os dados devem ser "governados" para que tenham mais qualidade.

Ladley (2012) define que a governança de dados é um conjunto de processos que garante que ativos de dados importantes sejam formalmente gerenciados em toda a empresa. Garante que informações confiáveis sejam usadas para processos críticos de negócios, tomada de decisão e contabilidade.

Desta forma, pode-se dizer que a falta de governança de dados efetiva dentro de uma empresa acarreta na existência de dados insuficientes. Esses dados de baixa qualidade se manifestam em definições inconsistentes, duplicadas, campos ausentes e outros tipos de dados clássicos.

A seguir, apresentamos a importância e os principais benefícios da governança de dados, segundo Ladley (2012):

1. Governança de Dados economiza dinheiro: simplesmente, a governança de dados aumenta a eficiência. Contas duplicadas, por exemplo, levam a esforços duplicados ou, no mínimo, à perda de tempo de rastreamento em seus setores de marketing, vendas, finanças ou análises. A Governança de Dados reduz erros no banco de dados da organização, fornecendo aos negócios uma base sólida para trabalhar. Tempo economizado na busca por inconsistências, é dinheiro economizado. Além disso, a governança de dados força a empresa a definir claramente seus dados principais e as regras que os regem. O início de um projeto de governança de dados é uma oportunidade de ouro para colocar todos na mesma página sobre as principais definições de dados. A aplicação disso garante maior eficiência operacional ao longo do tempo.

2. O risco da má governança dos dados e a falta de governança eficaz dos dados é uma preocupação de segurança por duas razões: riscos externos à segurança associados a dados sujos e não estruturados; e problemas de conformidade regulamentar. Dados inválidos e dados mal estruturados representam um risco à segurança pelo simples fato de que, se tivermos uma base inconsistente, não seremos capazes de identificar rapidamente quando algo errado acontece e nem como monitorar com eficiência quais dados estão em risco. Boas ferramentas e práticas de governança de dados facilitam o monitoramento do que está acontecendo na sua base e facilitam a visualização de quais áreas podem estar em risco. À medida que as pessoas continuam a entender a importância de seus dados pessoais, os governos começam a encarar a maneira que as empresas armazenam, protegem e usam os dados de seus clientes com muita seriedade.

3. Fornecendo clareza: reserve um tempo para refletir sobre o que a garantia de dados perfeitos significaria para as organizações e para o governo. A governança eficaz dos dados fornece a tranquilidade de que os dados geralmente são limpos, padronizados e precisos. Os efeitos disso repercutem em toda a empresa, garantindo que as métricas sejam precisas, conhecendo quais são as mais importantes e principalmente como e quando aplicá-las e ainda proporcionando uma maior confiança nas análises.

Atualmente, a governança de dados é fundamental para as empresas orientadas a dados. Agora se sabe por que a governança de dados é importante e, a partir desta afirmação, considere três fatores principais os quais as empresas estão se alinhando para se beneficiar no futuro próximo: Internet das Coisas, Inteligência Artificial e *Big Data*.

Todos esses três tópicos trazem grandes promessas para as empresas por meio de uma enorme quantidade de dados, através dos quais *insights* e inteligência podem ser obtidos; no entanto, o influxo de dados aumenta a necessidade de um programa eficaz de governança. Se uma empresa não se antecipar aos dados sujos provenientes da Internet das Coisas, da IA e do *Big Data*, o resultado principal pode ser apenas um grande pântano de dados, em vez do aumento da inteligência e dos lucros que os membros do conselho esperam.

Assim sendo, este objeto de estudo aborda conceitos fundamentais que auxiliam esta temática, relacionando a governança de dados aliada a uma destas três forças do *Big Data*, e aplica-se através de conceitos e diretrizes que podem e devem ser aproveitadas nas instituições federais, de modo que garanta efetividade, objetividade, redução de custos, entre outros.

2.2. *Big Data* e suas aplicações

Iniciamos esta seção apresentando os conceitos básicos sobre *Big Data*. É importante, neste primeiro momento, a familiarização com o termo para, posteriormente, entender suas aplicações.

Em um segundo momento, apresentamos as aplicações já destinadas ao contexto desta dissertação. Aqui é evidenciado não só o potencial de *Big Data* aplicado ao setor público federal, mas também suas restrições e barreiras governamentais que impedem sua aplicação com objetividade e efetividade.

Posteriormente, no tópico três desta seção, são citados alguns dos recursos de dados disponíveis para a Administração Pública, relatando conceitos, comparando com comissões especializadas (como por exemplo, a comissão europeia).

No quarto tópico são apresentados requisitos técnicos e não técnicos relacionados à aplicabilidade no setor público federal. As diretrizes apontadas são registros do Senado Federal, constituídos em 2012 que evidenciam e contribuem para esta temática.

No quinto e último tópico desta seção, apresentamos a importância de *Big Data* para a Administração Pública Federal. Questões relacionadas à sua contribuição direta e indireta, bem como os benefícios que traz ao governo são mensurados nesta seção.

2.2.1 Conceito histórico e atribuições

Big Data é um termo que descreve o grande volume de dados em velocidade de geração e transmissão elevadas e diferentes formatos - estruturados e não estruturados - que inunda uma empresa no dia-a-dia (IBM, 2020). Mas não é a quantidade de dados que é importante. É o que as organizações fazem com os dados que interessam. O *Big Data* pode ser analisado para obter informações que levam a melhores decisões e movimentos estratégicos de negócios (SILVA, 2015).

O uso do *Big Data* está se tornando comum pelas empresas para superar seus pares. Na maioria das indústrias, os concorrentes existentes e os novos entrantes usarão as estratégias resultantes dos dados analisados para competir, inovar e capturar valor (FEDERAL, 2012).

Big Data ajuda as organizações a criar novas oportunidades de crescimento e novas categorias de empresas que podem combinar e analisar dados do setor. Essas empresas possuem amplas informações sobre produtos e serviços, compradores e fornecedores, preferências do consumidor que podem ser capturadas e analisadas.

Embora o termo “*Big Data*” seja relativamente novo, o ato de coletar e armazenar grandes quantidades de informações para análise final tem muito tempo. O conceito ganhou força no início dos anos 2000, quando o analista do setor Doug Laney articulou a definição atual de *Big Data* (Figura 5) como os cinco Vs:

Figura 5 - Demonstração de *Big Data* através dos 5 Vs.

Volume	Variedade	Velocidade	Veracidade	Valor
 <p>Quantidade de dados que são gerados a cada segundo</p>	 <p>Tipos de dados que podem ser utilizados</p>	 <p>Velocidade a qual os dados são gerados e transmitidos</p>	 <p>Grau de organização e confiabilidade dos dados</p>	 <p>Necessidade de se transformar dados em valor</p>
Terabytes Registros Transações Arquivos	Estruturado Semiestruturado Não estruturado	Tempo real Streams Batch	Fidedignidade Autenticidade Origem Relevância Disponibilidade	Estatística Eventos Correlações Hipotética

Fonte: Tehseen (2016), modelo adaptado pelo autor.

De acordo com Dias (2018) e Oracle (2019), a Figura 5 acima pode ser definida da seguinte maneira:

- **Volume:** as organizações coletam dados de várias fontes, incluindo transações comerciais, mídias sociais e informações de dados de sensores ou de máquina para máquina. No passado, armazená-los seria um problema - mas as novas tecnologias (como o *Hadoop*¹) diminuíram o fardo. O termo '*Big Data*', em si, está relacionado a um tamanho enorme de dados. O tamanho desses dados desempenha um papel crucial na determinação do seu valor. Portanto, 'Volume' é uma característica que precisa ser considerada ao lidar com '*Big Data*'.
- **Variedade:** dados são fornecidos em todos os formatos - desde conjuntos de dados estruturados, dados numéricos em bancos de dados tradicionais, à documentos de texto não estruturados, e-mail, vídeo, áudio, dados de cotações de ações e transações financeiras. Variedade refere-se a fontes heterogêneas e à natureza dos dados, estruturados ou não estruturados.
- **Velocidade:** os dados são transmitidos em uma velocidade sem precedentes e devem ser tratados em tempo hábil. *Tags* RFID, sensores e medição inteligente estão aumentando a necessidade de lidar com *Torrents*

¹ <https://hadoop.apache.org/>

de dados em tempo real. O termo 'velocidade' refere-se à velocidade de geração de dados. A rapidez com que os dados são gerados e processados para atender às demandas determina o potencial real nos dados. O *Big Data Velocity* lida com a velocidade com que os dados fluem de fontes como processos de negócios, logs de aplicativos, redes e sites de mídia social, sensores, dispositivos móveis etc. O fluxo de dados é massivo e contínuo.

- Veracidade: um dos pontos mais importantes de qualquer informação é que ela seja verdadeira. Com *Big Data* não é possível controlar, por exemplo, cada *hashtag* do *Twitter* ou notícia falsa na internet, mas com análises e estatísticas de grandes volumes de dados é possível compensar as informações incorretas.
- Valor: O último V é o que torna *Big Data* relevante. Tudo bem ter acesso a uma quantidade massiva de informação a cada segundo, mas isso nada é irrelevante se não puder gerar valor. É importante que empresas entrem no negócio do *Big Data*, mas é sempre importante ter em mente o custo-benefício e tentar agregar valor ao que se está fazendo.

No passado, planilhas e bancos de dados eram as únicas fontes de dados consideradas pela maioria dos aplicativos. Os dados na forma de e-mails, fotos, vídeos, dispositivos de monitoramento, PDFs, áudio etc. também estão sendo considerados por aplicativos de análise. Essa variedade de dados não estruturados traz certos problemas para armazenamento, mineração e análise de dados (ORACLE, 2019).

Desta forma, a Oracle (2019) destaca os benefícios do *Big Data*, como:

- as empresas podem utilizar inteligência externa enquanto tomam decisões;
- melhor atendimento ao cliente;
- identificação precoce de risco para o produto/serviços se houver; e
- melhor eficiência operacional.

A importância do *Big Data* não gira em torno da quantidade de dados que uma empresa possui, mas de como ela utiliza os dados coletados. Toda empresa usa dados à sua maneira; quanto mais uma empresa usar seus dados de maneira eficiente, mais potencial de crescimento ela terá. A empresa pode pegar dados de

qualquer fonte e analisá-los para encontrar respostas que poderão permitir os itens de 1 a 7 descritos a seguir (ORACLE, 2019):

1. Economia de custos: algumas ferramentas de *Big Data*, como *Hadoop* e *Cloud-Based Analytics*, podem trazer vantagens de custo para os negócios quando grandes quantidades de dados devem ser armazenadas. Essas ferramentas também ajudam na identificação de maneiras mais eficientes de fazer negócios.

2. Redução de tempo: a alta velocidade de ferramentas como o *Hadoop* e a análise em memória podem identificar novas fontes de dados, ajudando as empresas a analisar dados e a tomar decisões rápidas com base em aprendizado.

3. Entender as condições do mercado: analisando *Big Data*, podemos entender melhor as condições atuais do mercado. Por exemplo, analisando o comportamento de compra dos clientes, uma empresa pode descobrir os produtos mais vendidos e produzi-los de acordo com essa tendência. Com isso, ela pode ficar à frente de seus concorrentes.

4. Controlar a reputação online: as ferramentas de *Big Data* podem fazer análises de sentimentos. Portanto, pode-se obter *feedback* sobre quem está dizendo o que diz respeito à sua empresa. Se desejar monitorar e melhorar a presença *on-line* dos negócios, as ferramentas de *Big Data* podem ajudar nisso.

5. Usar *Big Data Analytics* para aumentar a aquisição e retenção de clientes: o cliente é o ativo mais importante de que qualquer empresa depende. Não existe um único negócio que possa reivindicar sucesso sem primeiro ter que estabelecer uma base sólida de clientes. No entanto, mesmo com uma base de clientes, uma empresa não pode se dar ao luxo de desconsiderar a alta concorrência que enfrenta. Se uma empresa demora a aprender o que os clientes estão procurando, é muito fácil começar a oferecer produtos de baixa qualidade. No final, resultará em perda de clientela e isso cria um efeito geral adverso no sucesso dos negócios. O uso de *Big Data* permite que as empresas observem vários padrões e tendências relacionados ao cliente.

6. Usar *Big Data Analytics* para resolver o problema dos anunciantes e oferecer informações de marketing: a análise de *Big Data* pode ajudar a mudar todas as operações comerciais. Isso inclui a capacidade de corresponder às expectativas dos clientes, mudar a linha de produtos da empresa e garantir que as campanhas de marketing sejam poderosas.

7. Análise de *Big Data* como motor de inovações e desenvolvimento de produtos: outra vantagem do *Big Data* é a capacidade de ajudar as empresas a inovar e reconstruir seus produtos.

Os melhores exemplos de *Big Data* podem ser encontrados tanto no setor público quanto no privado. Desde publicidade direcionada, educação e indústrias já mencionadas (assistência médica, seguros, manufatura ou serviços bancários). Até o ano 2021, 1,7 *megabytes* de dados serão gerados a cada segundo para cada pessoa no planeta; o potencial de crescimento organizacional orientado a dados no setor de hospitalar é enorme (FAGUNDES et al., 2017).

Enfatizando a aplicação de *Big Data* no setor bancário, pode-se imaginar que a quantidade de dados neste setor está subindo de forma rápida a cada segundo. Desta forma, a aplicação de *Big Data*, pode ajudar a detectar:

- o uso indevido de cartões de crédito;
- uso indevido de cartões de débito;
- clareza de negócios;
- alteração nas estatísticas do cliente;
- lavagem de dinheiro;
- mitigação de riscos.

2.2.2 Aplicações potenciais de *Big Data* para a Administração Pública Federal e suas restrições

Quatro aplicações potenciais para o setor público foram descritas e desenvolvidas em Zillner et al. (2013) por demonstrar o uso de tecnologias de *Big Data* no setor público. A Tabela 1 apresenta um sumário sobre tais aplicações.

Tabela 1 - Critérios dos cenários de aplicação para a Administração Pública.

Resumo dos cenários de aplicação para a Administração Pública	
Nome	Monitoramento e supervisão de atividades regulamentadas para operadores de jogos de azar on-line
Sumário	Grandes volumes de dados disponíveis dificultam a regulação e a supervisão eficientes das atividades
Sinopse	Monitorar os operadores de jogos de azar on-line para o controle das atividades regulamentadas e a detecção de fraudes. O usuário deste aplicativo é o órgão público responsável pela atividade de supervisão. Este procedimento é uma obrigação regulatória da administração pública; os operadores de jogos de azar on-line devem fornecer as informações ao público regulador por meio de um canal de

comunicação específico. Os dados em tempo real são recebidos dos operadores de jogos a cada 5 minutos.

Objetivos de Negócios Garantir a conformidade com os regulamentos, prevenção e detecção de fraudes e investigação criminal.

Nome Eficiência operacional em agência de trabalho

Sumário Extrair valor de grandes volumes disponíveis de dados não utilizados

Sinopse Ative uma nova gama de serviços personalizados, melhore o atendimento ao cliente e reduza os custos de operação na agência federal do trabalho alemã. Todos os trabalhadores desempregados estavam recebendo os mesmos serviços padrão, apesar de terem perfis diferentes. Os dados históricos de seus clientes foram analisados, incluindo perfis, intervenções e o tempo necessário para encontrar um emprego.

Objetivos de Negócios Reduza o custo e melhore a qualidade do serviço: agora eles conseguem encontrar um novo emprego em um período mais curto.

Nome Segurança pública em cidades inteligentes

Sumário Grandes volumes de dados disponíveis a partir de sensores, mídias sociais e chamadas de emergência podem ser combinados para fornecer segurança pública eficaz.

Sinopse Cidades inteligentes equipadas com sensores e infraestruturas de comunicação ajudam o setor público a manter as cidades e seus cidadãos seguros. Ter informações precisas e atualizadas permite respostas melhores e mais rápidas durante emergências e resulta em menos danos e vítimas. As fontes típicas para obter essas informações podem vir de chamadas de resposta a emergências, câmeras de vigilância e forças móveis (como um carro-patrolha da polícia) que chegaram ao local. Nos últimos anos, as mídias sociais mostraram um potencial interessante para a coleta de informações que auxiliam na obtenção de uma imagem precisa da consciência situacional (Van Kasteren et al. 2014). Todas as informações coletadas são coletadas em um centro de comando e controle onde um operador pode decidir como orientar as forças móveis disponíveis.

Objetivos de Negócios Resposta rápida a emergências, prevenção de danos e menos vítimas.

Nome Policiamento preditivo usando dados aberto

Sumário Reutilização de dados públicos abertos para fornecer policiamento preditivo

Sinopse Governos de todo o mundo iniciaram iniciativas de dados abertos para disponibilizar dados do setor público em prol da transparência e permitir que terceiros ofereçam serviços com base nesses dados. Um desses serviços pode ser descrito como policiamento preditivo, onde dados históricos de crimes são usados para descobrir automaticamente tendências e padrões. Os padrões identificados ajudam a obter *insights* sobre os problemas relacionados ao crime que uma cidade está enfrentando e permitem um destacamento mais eficaz e eficiente das forças policiais (Wang et al. 2013; PredPol 2013).

Objetivos de negócios Diminuição significativa do crime, uso eficiente de forças móveis.

Wang et al. (2013) menciona que, para a *Big Data* na Administração Pública, as restrições podem ser resumidas como descrito a seguir:

- Falta de vontade política de fazer o setor público tirar proveito dessas tecnologias. Isso exigirá uma mudança de mentalidade dos funcionários de alto escalão do setor público.
- Falta de pessoas qualificadas e orientadas para os negócios, cientes de onde e como *Big Data* pode ajudar a solucionar os desafios do setor público e quem pode ajudar a preparar a estrutura regulatória para o desenvolvimento bem-sucedido dessas soluções.
- O novo Regulamento Geral de Proteção de Dados e as diretivas PSI trazem algumas incertezas sobre o impacto na implementação de iniciativas de *Big Data* e dados abertos no setor público. Assim, os dados abertos devem ser um catalisador do setor público para o setor privado para estabelecer uma poderosa indústria de dados.
- Ganhando impulso na adoção: hoje, há mais marketing sobre *Big Data* no setor público do que experiências reais para aprender quais aplicativos são mais rentáveis e como devem ser implantados. Isso requer o desenvolvimento de um conjunto padrão de soluções de *Big Data* para o setor.
- Numerosos órgãos da administração pública: (naqueles amplamente descentralizados), muita energia é perdida e permanecerá assim até que uma estratégia comum seja definida para a reutilização de plataformas entre tecnologias.

2.2.3 Recursos de dados disponíveis para a Administração Pública

De acordo com Zillner et al. (2013), a Informação do Setor Público (PSI) é definida da seguinte forma: abrange qualquer representação de atos, fatos ou informações – e qualquer compilação dos mesmos – qualquer que seja seu meio (escrito em papel ou armazenado em formato eletrônico ou como gravação sonora, visual ou audiovisual), realizado por órgãos públicos.

Um documento mantido por um órgão do setor público é um documento em que este órgão tem o direito de autorizar sua reutilização. Segundo Correia (2004), quanto à disponibilidade das informações produzidas por esses órgãos públicos, e

na ausência de diretrizes específicas, o organismo produtor é livre para decidir como disponibilizá-las: de forma direta aos usuários finais, estabelecendo um público/parceria privada ou terceirizar a exploração comercial dessas informações a operadores privados.

Desta forma, Correia (2004) esclarece que as atividades que não se enquadram na tarefa pública “normalmente incluem o fornecimento de documentos produzidos e cobrados em caráter comercial e em concorrência com outros no mercado”.

Sobre a natureza do PSI disponível, existem várias abordagens. O Livro Verde sobre o PSI (Comissão Europeia 1998) propõe algumas classificações, como:

- distinção PSI entre administrativa e não administrativa; e
- distinção do PSI em relação à sua relevância para o público.

Além disso, ele pode ser diferenciado de acordo com seu valor potencial de mercado e, em alguns casos, de acordo com o conteúdo dos dados pessoais:

- distinção PSI de acordo com seu anonimato.

Em comparação com outros setores, como a assistência médica, que produz uma quantidade de imagens eletrônicas, a quantidade mais importante de dados produzidos pelo setor público é textual ou numérica. Como resultado das iniciativas de governo eletrônico dos últimos 15 anos, 90% desses dados é criada em formato digital, de acordo com a McKinsey (MCKINSEY GLOBAL INSTITUTE, 2011).

Segundo uma pesquisa realizada para a formulação da Parceria Europeia de Valorização de *Big Data* para representantes do setor público (ZILLNER et al., 2013), o principal ativo de dados apontados pelos autores é todo o sistema de setor público, registros, Bancos de Dados e Sistemas de Informação, dos quais os mais significativos são:

- cidadãos, empresas e propriedades (por exemplo, registros de base, transações);
- dados fiscais;
- dados de segurança;
- gerenciamento de documentos, e as crescentes transações eletrônicas;
- compras e despesas públicas;
- órgãos e funcionários públicos;

- dados geográficos relacionados a cadastros;
- conteúdo relacionado à cultura, educação e turismo;
- documentos legislativos;
- dados estatísticos (dados socioeconômicos que poderiam ser usados pelo setor privado); e
- dados geoespaciais.

2.2.4 Requisitos da Administração Pública: Técnicos e não técnicos.

A agregação de dados por meio dos limites administrativos de maneira não baseada em solicitação é um desafio real. Isso se dá porque esses dados podem revelar informações pessoais e de segurança altamente confidenciais quando combinadas com várias outras fontes de dados, não apenas comprometendo a privacidade individual, mas também a segurança civil.

O direito de acesso aos conjuntos de dados necessários para uma operação deve ser justificado e obtido. Quando uma nova operação é realizada com dados existentes, é necessário obter uma notificação ou uma licença da Agência de Privacidade de Dados.

O anonimato deve ser preservado nesses casos para que a dissociação de dados seja garantida. Preocupações individuais com privacidade e segurança pública devem ser tratadas antes que os governos possam ser convencidos a compartilhar dados mais abertamente, não apenas publicamente, mas compartilhando de maneira restrita com outros governos ou entidades internacionais.

Outra dimensão é a regulamentação para o uso da computação em nuvem, de forma que o setor público possa confiar nos provedores desses serviços. Além disso, a falta de fornecedores de computação em nuvem de *Big Data* no mercado europeu também é uma barreira para adoção.

Faltam cientistas e tecnólogos especializados em dados que possam capturar e processar essas novas fontes de dados (WANG, et al., 2013). Enquanto as tecnologias de *Big Data* se tornam cada vez mais adotadas nos negócios, os profissionais qualificados da área se tornam cada vez mais difíceis de encontrar.

Os órgãos públicos podem se distanciar com as habilidades que já possuem, mas precisam garantir que estas avancem. Além das pessoas com orientação técnica, há falta de conhecimento nas pessoas com orientação comercial que estão cientes do que a *Big Data* pode fazer para ajudá-las a resolver os desafios do setor público.

Desta forma, o Senado Federal (2012) explica que os requerimentos não técnicos incluem:

- disposição de fornecer e adotar tecnologias de *Big Data* e também de saber usá-lo;
- necessidade de abordagens (políticas) nacionais ou europeias comuns - como as políticas nacionais de interoperabilidade de dados abertos. Falta de liderança neste campo; e
- uma incompatibilidade geral entre inteligência de negócios em geral e *Big Data* em particular no setor público.

Já em relação aos requisitos técnicos, o Senado Federal (2012) destaca os tópicos descritos nos parágrafos a seguir:

- Descoberta de padrões: identificação de padrões e semelhanças para detectar comportamentos criminais ou ilegais específicos no cenário de aplicação de monitoramento e supervisão de operadores de jogos de azar on-line (e também para cenários de monitoramento semelhantes no setor público). Esse requisito também é aplicável para melhorar a eficiência operacional na agência de trabalho e no cenário de policiamento preditivo.
- Compartilhamento de dados/integração de dados: necessário para superar a falta de padronização dos esquemas de dados e a fragmentação da propriedade dos mesmos. Integração de múltiplas e diversas fontes de dados em uma plataforma de *Big Data*.
- Informações em tempo real: ative a análise de dados novos/em tempo real para tomada de decisão instantânea.
- Segurança e Privacidade de Dados: procedimentos legais e meios técnicos que permitem o compartilhamento seguro de dados e a preservação da privacidade. As soluções para esse requisito podem desbloquear o amplo uso de *Big Data* no setor público. Os avanços na proteção e privacidade dos dados são fundamentais, pois podem permitir a análise de grandes

quantidades de dados pertencentes ao setor público sem divulgar informações sigilosas. Esses problemas de privacidade e segurança estão impedindo o uso de infraestruturas de nuvem (processamento, armazenamento) por muitos órgãos públicos que lidam com dados confidenciais.

- Transmissão de dados em tempo real: como a capacidade de instalar sensores está aumentando em cenários de aplicativos de cidades inteligentes, há uma alta demanda por transmissão de dados em tempo real. Será necessário fornecer recursos distribuídos de processamento e limpeza para sensores de imagem, a fim de não reduzir os canais de comunicação e fornecer apenas as informações necessárias para a análise em tempo real, provendo sistemas de conscientização situacional para os tomadores de decisão.
- Análise de linguagem natural: extraia informações de fontes on-line não estruturadas (por exemplo, mídias sociais) para permitir a mineração de sentimentos. Reconhecimento de dados de entradas de linguagem natural, como texto, áudio e vídeo.
- Análise preditiva: conforme descrito no cenário do aplicativo para policiamento preditivo, onde o objetivo é distribuir forças e recursos de segurança de acordo com a previsão de incidentes, forneça previsões com base no aprendizado de situações anteriores para prever a alocação ideal de recursos para serviços públicos.
- Modelagem e Simulação: ferramentas específicas do domínio para modelagem e simulação de eventos de acordo com dados de eventos passados para antecipar os resultados de decisões tomadas para influenciar as condições atuais em tempo real, por exemplo, em cenários de segurança pública.

2.2.5 A importância do *Big Data* na Administração Pública Federal

Por trás da política da montanha-russa que balança países como Estados Unidos, Alemanha, Japão, Coreia e também o Brasil, uma revolução muito mais silenciosa, mas igualmente importante, está em andamento. Os gerentes governamentais estão avançando no uso de “*Big Data*” e estão tendo impacto. É o centro de um esforço importante para transformar a saúde do governo federal e

melhorar os resultados dos programas federais, como argumentou um recente relatório da Academia Nacional de Administração Pública (NAOPA, 2018).

Desta forma, indaga-se: qual o problema relacionado ao *Big Data*? É tentador olhar para a revolução silenciosa dos dados como apenas o próximo passo pequeno e lógico na medição do desempenho do governo, uma continuação da evolução em curso que começou com a Revisão de Desempenho Nacional da Al Gore. Depois disso, vieram uma série de agendas de administração, cada uma delas focada (de maneiras muito diferentes) na produção de informações melhores para obter resultados superiores.

Desta forma Naopp (2018) retrata que o papel do *Big Data* na administração pública federal é relevante, onde o fornecimento de dados cresceu exponencialmente, com muito mais dados das mais diversas fontes. O governo não tem escolha sobre adotar a revolução da *Big Data* (VIDOR, 2019). Todas as atividades governamentais, desde o custo dos cuidados de saúde, até o estado do meio ambiente, têm gerado uma quantidade de dados sem precedentes.

O movimento é muito grande para ser ignorado. Não apenas o governo está produzindo seus próprios dados em um volume cada vez maior, como também surgem dados do setor privado, tanto através do investimento de empresas privadas quanto do crescimento explosivo das mídias sociais. Mesmo que o governo quisesse ignorar a revolução dos dados, não poderia - os dados estão chegando de todos os lugares.

Mais dados estão disponíveis em tempo real. O famoso programa CompStat, da cidade de Nova York, começou porque os comandantes da polícia perceberam que, coletando estatísticas, não precisavam esperar até o final do ano para obter os números dos crimes. Atualmente, contudo, o governo tem a seu dispor um enorme aparato tecnológico que captam informações em todos os lugares a qualquer momento, como câmeras de segurança, por exemplo.

Mais dados são granulares, conectados às operações reais dos programas. Os Centros de Controle e Prevenção de Doenças sabem como as vacinas contra a gripe estão afetando o surto de gripe e a Agência de Proteção Ambiental pode ver o impacto das regras do ar limpo na qualidade do ar - incluindo as emissões de instalações individuais. Os aplicativos *para smartphone* rastreiam todos os voos com destino a qualquer lugar, provendo a possibilidade de planejamento em tempo real. É

possível detalhar da política ampla aos resultados reais e acompanhar o que está acontecendo.

Pela primeira vez, funcionários do topo - no Gabinete de Administração e Orçamento e nos escritórios dos secretários de gabinete - e gerentes da linha de frente podem ver, simultaneamente os mesmos dados em tempo real. Esta é uma oportunidade emergente não totalmente desenvolvida. Mas a capacidade de funcionários do alto escalão do governo entenderem o que está acontecendo em escritórios individuais tende a, inevitavelmente, transformar a administração federal.

Mais desses dados estão vivendo separadamente dos mundos anteriores da análise de políticas (como análise de custo-benefício e avaliação de programas) e gerenciamento de desempenho (como avaliações em larga escala dos resultados do programa). Esse *tsunami* de dados está em uma realidade distinta - e crescente -. Ele se baseia em perguntas, dados, ferramentas e respostas diferentes.

Esses dados e análises de *Big Data* estão trazendo à tona perguntas antes inimagináveis. Quais regulamentos do governo estão nos livros há mais tempo? Que regulamentos estão mais conectados com os outros? Um estudo recente de *Big Data* da Deloitte (DELOITTE) descobriu que 12% de todas as seções do Código de Regulamentos Federais não foram tocadas desde pelo menos a década de 1970 - e que 67% não foram editadas desde que foram publicadas. Existem 17.800 seções que são correspondências extremamente próximas a outras seções, diferindo apenas em poucas palavras. Se estivermos interessados em simplificar as regulamentações federais, a análise de *Big Data* fornece uma maneira de saber onde procurar. Segundo Deloitte (2019), isso simplesmente não era possível antes.

A revolução dos dados não apenas fornece muito mais informações, mas também cria novas formas de comunicar o que está acontecendo. Há muito debate sobre os “desertos alimentares” nas cidades - áreas onde as pessoas têm pouco acesso a alimentos saudáveis e não têm um bom transporte para alcançar melhores lugares para comprá-los.

Um estudo de Baltimore produziu um mapa fascinante de desertos alimentares lá. Ele coloca quatro bancos de dados diferentes um sobre o outro para produzir uma análise clara e convincente de um problema muito complicado. Uma coisa é produzir sofisticadas equações de regressão múltiplas que exploram as interações de variáveis

importantes, mas de uma forma que ninguém, a não ser o conhecimento, pode digerir. Outra coisa é desenhar um mapa que capte questões importantes em bairros individuais, de uma forma que faça sentido para todos. Ver é acreditar e acreditar torna mais fácil agir sobre isso (BALTIMORE, 2019).

A própria natureza da revolução da *Big Data* pode ajudar a superar o problema fundamental de tantos esforços anteriores para levar uma melhor análise ao governo: elevados investimentos em diversos estudos que poucas pessoas leem e que produziram tão pouco resultado. Um bom mapa - sobre desertos alimentares, por exemplo - pode identificar problemas e ajudar a impulsionar soluções.

Essa revolução facilita o ataque ao problema ainda mais fundamental da oferta e demanda: produzir respostas para as perguntas dos formuladores de políticas, em vez de tentar convencê-los a fazer as perguntas sobre as quais temos dados.

Não é um trabalho fácil. De fato, uma das maiores limitações é o fato de não existirem muitos funcionários do governo especializados em mineração e análise de dados. Outra é que parte desse trabalho exige que as pessoas que sabem como detalhar, perguntem o que está acontecendo, por que isso importa e o que fazer a respeito. De fato, a revolução dos dados não é tanto sobre números, mas também sobre capital humano: levar as pessoas que sabem como impulsionar a revolução.

Mas não é como se o governo tivesse uma escolha. A sociedade, em geral, está evoluindo e os cidadãos não terão muita paciência com um governo que não acompanha. A boa notícia, porém, é que a revolução dos dados está proporcionando ao governo uma gama enorme de novas oportunidades. O mundo está constantemente procurando o novo e, no governo, não é diferente. A revolução dos dados tem o potencial de melhorar a saúde e a eficácia das organizações governamentais.

2.3. Governança de Dados aliado ao *Big Data*

Nesta seção, apresentamos tópicos sobre a relação entre Governança de Dados e *Big Data*. As razões pelas quais são cruciais para uma organização e como podem ser aproveitados dentro das Instituições Federais, serão alguns dos pontos explorados nas seções a seguir.

2.3.1 Por que a Governança de Dados é crucial para ambientes de *Big Data*?

Organizações em todo o mundo estão investindo em sistemas capazes de hospedar e processar dados de maneiras nunca antes imaginadas. Em alguns casos, as empresas estão até reformulando seus ambientes de TI existentes com base nesses novos sistemas. Esses sistemas de *Big Data* produziram resultados tangíveis: aumento de receita e custos mais baixos. No entanto, os resultados positivos estão longe de ser garantidos. Para realmente obter valor dos dados, essas novas plataformas devem ser governadas (HENRIQUES & COSTA, 2018).

O termo “governança” causa medo nos corações de muitos profissionais de dados. Como o termo geralmente é vagamente definido e mal compreendido, muitos simplesmente recorrem a uma abordagem apenas de tecnologia para resolver suas necessidades de governança. O que talvez seja menos conhecido é que as próprias tecnologias devem ser revisadas ao aperfeiçoar a governança de dados hoje (ROSSOUW, 2012).

A governança de dados não é a linhagem, administração ou gerenciamento de dados mestre. Cada um desses termos é frequentemente ouvido em conjunto com - e mesmo no lugar de - governança de dados. Na verdade, essas práticas são componentes dos programas de governança de dados de algumas organizações. São componentes importantes, mas são meramente componentes (STUMPF, 2017).

Segundo a IBM (2020), essencialmente, a governança de dados trata de gerenciar formalmente dados importantes em toda a empresa e, assim, garantir que o valor seja derivado. Embora os níveis de maturidade variem de acordo com a organização, a governança de dados geralmente é alcançada por meio de uma combinação de pessoas e processos, com a tecnologia usada para simplificar e automatizar aspectos desse processo.

Cita-se, por exemplo, segurança. Mesmo os níveis básicos de governança exigem que os ativos de dados importantes e sensíveis de uma empresa sejam protegidos. Os processos devem impedir o acesso não autorizado a dados confidenciais e expor todos ou parte desses dados a usuários com uma "necessidade de saber" legítima.

Tecnologias como sistemas de gerenciamento de identidades e recursos de gerenciamento de permissões simplificam e automatizam os principais aspectos dessas tarefas.

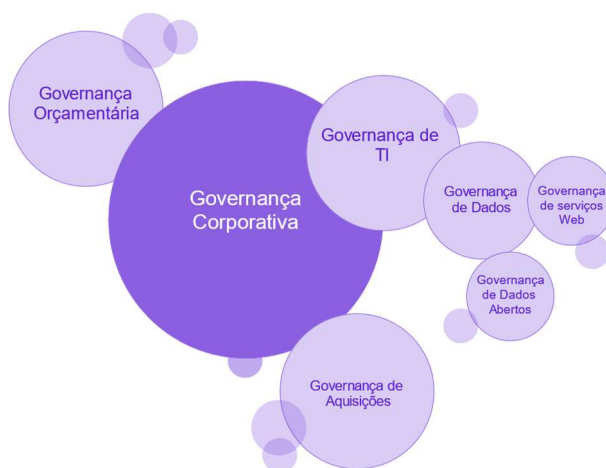
Algumas plataformas de dados simplificam ainda mais as tarefas, vinculando os registros existentes baseados em nome de usuário e senha, como o *Active Directory*. Isso permite maior expressividade ao atribuir permissões, além dos relativamente poucos graus de liberdade proporcionados pelos *bits* do modo POSIX.

Também se deve reconhecer que, à medida que a velocidade e o volume dos dados aumentarem, será quase impossível para humanos (por exemplo, administradores de dados ou analistas de segurança) classificá-los em tempo hábil. Às vezes, as organizações são forçadas a manter novos dados bloqueados em uma célula de retenção até que alguém os classifique e exponha adequadamente aos usuários finais (ROSSOUW, 2012).

Felizmente, os provedores de tecnologia estão desenvolvendo maneiras inovadoras para classificar dados automaticamente, diretamente quando recebidos ou imediatamente depois. Ao alavancar essas tecnologias, um pré-requisito essencial do processo de autorização é atendido, minimizando o tempo para o *insight* (STUMPF, 2017).

Stumpf (2017) enfatiza por meio da Figura 06 a integração entre as governanças (de Tecnologia da Informação, Orçamentária, Dados) e deixa um leque para aplicações de *Big Data*.

Figura 6 – Governança de Dados integrados com o ambiente corporativo



Fonte: Stumpf (2017) – modelo adaptado pelo autor

Assim sendo, questiona-se: Qual a diferença entre a governança de dados na era do *Big Data*? Fazendo referência aos 5 Vs do *Big Data*, governar esses sistemas pode ser uma tarefa complexa. As organizações geralmente são forçadas a agrupar *clusters* separados, cada um com seu próprio objetivo comercial ou armazenar e processar tipos de dados exclusivos, como arquivos, tabelas ou fluxos.

Arquiteturas convergentes simplificam bastante a governança (HENRIQUES; COSTA, 2018). Em sistemas convergentes, vários tipos de dados (por exemplo, arquivos, tabelas e fluxos) são integrados em um único repositório de dados que pode ser controlado e protegido de uma só vez.

Além dos 5 Vs, há outra diferença mais sutil. A maioria das distribuições de *Big Data*, se não todas, inclui uma amálgama de diferentes mecanismos de análise e aprendizado de máquina situados "em cima" do(s) repositório(s) de dados. *Spark*² e *Hive*³ são apenas dois dos mais populares em uso hoje. Essa flexibilidade é ótima para os usuários finais, porque eles podem simplesmente escolher a ferramenta mais adequada às suas necessidades específicas de análise.

O problema da perspectiva de governança é que essas ferramentas nem sempre respeitam os mesmos mecanismos ou protocolos de segurança, nem registram ações de forma completa, consistente ou em repositórios que podem ser dimensionados - pelo menos não "prontos para uso" (ROSSOUW, 2012).

Como resultado, os profissionais de *Big Data* podem ser flagrados ao tentar atender às demandas de conformidade ou de auditoria sobre, por exemplo, linhagem de dados - um componente de governança que visa responder perguntas como: "De onde esses dados vieram e o que aconteceu com o tempo?".

2.3.2 O papel integral da Governança de Dados em ambientes de *Big Data*

A Governança de Dados é um conceito multifacetado, mas fornece as ferramentas e os processos para promover o entendimento dos dados em toda a empresa. É um programa abrangente, não um projeto, e deve incluir um conjunto principal de soluções para fornecer uma base de governança adequada (YIU, 2012).

² <https://spark.apache.org/>

³ <https://hive.apache.org/>

Essas soluções incluem um glossário de negócios, dicionários de dados e linhagem de dados para definir dados, termos e atributos de negócios, bem como fontes de dados, uso, relacionamentos e interdependências. A governança de dados também deve atribuir claramente a responsabilidade e a propriedade entre os envolvidos, administradores e proprietários de dados, bem como um mecanismo para gerenciar consultas e resolver problemas.

Historicamente, a governança de dados tem sido intimamente associada à garantia da conformidade regulatória e, embora isso seja verdade, o papel da governança de dados é muito mais amplo na era do *Big Data*. Por exemplo, o gerenciamento de metadados é uma parte crucial da governança, uma vez que estes desempenham um papel importante para as organizações descobrirem *insights* analíticos (TEHSEEN, 2016).

A governança de dados também desempenha um papel crítico nos esforços de qualidade de dados, à medida que as organizações continuam lutando para avaliar, melhorar e relatar a qualidade de seus dados.

Os desafios que a governança de dados abrangente de hoje ajuda as organizações a superar - os de acessibilidade, usabilidade, significado e qualidade - todos aumentam exponencialmente no mundo do *Big Data*. Os ambientes de *Big Data* são potenciais tesouros para *insights*, mas, sem governança, responsabilidade, colaboração e suporte organizacionais adequados, eles podem ser buracos negros de dados não utilizados (SILVA, 2015).

A chave para governar esses ambientes é gerenciar e definir esses dados em toda a cadeia de suprimento de dados - um esforço que começa quando os dados são processados nas organizações e entram em qualquer ambiente interno, seja um *Data Warehouse* ou um *Data Lake*, e continua ao longo do seu ciclo da vida. Em toda a cadeia de fornecimento de dados, questões importantes precisam ser abordadas. Essas incluem:

Transparência e rastreabilidade são elementos-chave que podem ser rastreados por meio de metadados e linhagem de dados - de onde vieram os dados, por quais processos e sistemas foram movidos na organização e como eles foram transformados? A qualidade dos dados é uma preocupação constante, pois os dados são considerados para análise (SILVA, 2015).

Com possíveis transformações, o que se sabe sobre esses dados e em que se pode confiar? São precisos, consistentes e confiáveis? Os usuários de negócios podem depender disso para gerar análises e *insights* precisos? Acessibilidade e entendimento são fundamentais para usuários de negócios. Sem eles, é como ter um armazém cheio de equipamentos e materiais, mas nenhuma chave para a porta e nenhuma instrução sobre como usar o que está dentro - nada será construído ou o que será construído provavelmente desmoronará.

Desta forma, pode-se indagar: Quais dados estão disponíveis? Como é usado internamente? Como é definido e quais são os termos comerciais relacionados? Essas definições diferem entre as linhas de negócios ou departamentos? Os dados devem ser claramente categorizados, organizados e disponíveis para os usuários e bem definidos, para que os dados certos sejam selecionados para a tarefa correta.

Propriedade e colaboração são críticas. Não basta saber de onde vieram os dados e o que são; é preciso haver responsabilidade e prestação de contas contínuas por esses ativos. Quem possui os dados? Os proprietários e administradores de dados devem ser claramente definidos para que os usuários de negócios tenham recursos a quem recorrer, para perguntas sobre uso e aplicabilidade (SILVA, 2015).

Um programa abrangente de governança de dados responderá a todas essas perguntas e fornecerá uma estrutura sólida para que os dados organizacionais extraídos sejam confiáveis, compreensíveis e utilizáveis. Não fazer isso pode resultar em decisões de negócios baseadas em dados incompletos ou incorretos, que podem custar caro tanto em receita quanto em perda de reputação, produtividade e oportunidades perdidas.

A governança de dados está ganhando importância não apenas devido ao crescente volume e velocidade de dados e ao surgimento de ambientes de *Big Data*, mas também devido às crescentes complexidades regulatórias e ao desafio incansável de garantir a qualidade dos dados para produzir resultados de qualidade. E embora a governança tenha sido historicamente relegada a TI e à conformidade, são os líderes de negócios de hoje que precisam aproveitar dados e análises para obter vantagem competitiva e melhorar os resultados (SILVA, 2019).

Enfrenta-se uma democratização dos dados no sentido de que estes não são mais apenas uma preocupação da TI, pois as restrições de recursos e as demandas

de negócios direcionam uma crescente necessidade de empoderamento dos usuários de negócios por meio de recursos de autoatendimento. As empresas precisam ter dados na ponta dos dedos para resolver problemas de negócios, para que possam transformar rapidamente esses dados em *insights* acionáveis.

É por isso que a governança de dados é tão crítica na era do *Big Data*. A necessidade e o uso de dados estão cada vez mais nas mãos dos usuários de negócios, o que significa que precisa haver uma estrutura sólida que defina todos os aspectos dos dados e sua utilização (SILVA, 2019).

2.3.3 Como as diferenças de *Big Data* podem afetar a Governança de Dados

Embora o *Big Data* e a Governança de Dados já existam há algum tempo, o relacionamento entre eles ainda é uma novidade. A necessidade de governança de dados em *Big Data* surge da necessidade das empresas de obter informações precisas, confiáveis e acionáveis sobre os dados existentes. Na ausência de uma política adequada de governança de dados, os projetos de dados podem resultar em custos excessivos e em informações enganosas, o que acabaria por causar danos irrevogáveis aos negócios.

De acordo com (IBM, 2020), a governança de *Big Data* deve começar definindo uma política geral de governança de dados. Isso inclui políticas claras definidas para inventário, propriedade, qualidade e retenção de dados, além da segurança das informações.

A existência de políticas para cada uma delas permite que a empresa identifique as partes interessadas corretamente e aplique procedimentos para garantir a eficiência e a confiabilidade, enquanto reduz os riscos de falha. Em seguida, é necessário definir os padrões para os dados, começando pelos conjuntos de dados críticos que agregam mais valor para os negócios.

Isso inclui identificar elementos de dados críticos, definir metadados e relacionamentos e, finalmente, documentá-los para referência futura. Deve existir um processo de automação para lidar com grandes cargas de dados provenientes de diferentes fontes na organização, pois os processos manuais não podem atender a uma variedade tão grande de informações (IBM, 2020).

Hoje, os modelos analíticos são críticos para uma plataforma de *Big Data* e, portanto, andam de mãos dadas com uma política de governança de dados. O modelo analítico deve ser simples o suficiente para ser entendido e utilizado por usuários de negócios e analistas de dados. Além disso, deve existir um processo para atender aos novos requisitos analíticos dos usuários finais. Como pode ser observado, o advento de ferramentas mais novas como o *SlamData*⁴, os modelos analíticos se tornam menos críticos, no entanto, a necessidade de governança não diminui (IBM, 2020).

Finalmente, e mais importante, deve haver uma equipe dedicada para lidar com as responsabilidades da governança de *Big Data*. A equipe deve consistir de um patrocinador, diretor de dados e vários especialistas no assunto para monitorar, revisar e aprimorar constantemente a qualidade do processo e as informações.

Em resumo, não se espera que o *Big Data* dê seu verdadeiro valor a uma empresa sem uma política de governança de dados estabelecida. A governança de dados é uma novidade no que diz respeito ao *Big Data*, no entanto, está cada vez mais chamando a atenção das equipes *Big Data* envolvidas com o tema. Em pouco tempo, ela se tornará obrigatória para qualquer projeto de *Big Data*, permitindo que as empresas aproveitem totalmente seu poder.

2.4. Aplicações de Governança de Dados e *Big Data* em Instituições Públicas Federais: Uma visão globalizada

Uma boa governança de dados pode contribuir para definir uma visão comum; melhorar a implementação e coordenação coerentes; e fortalecer as bases institucionais, regulatórias, de capacidade e técnicas para controlar e gerenciar melhor o ciclo de valor dos dados, ou seja, coletar, gerar, armazenar, proteger, processar, compartilhar e reutilizar dados, como forma de aumentar a confiança e agregar valor (SILVA, 2019).

A boa governança de dados é imprescindível para governos que pretendem se tornar mais orientados a dados como parte de sua estratégia digital (IBM, 2020). Pode ajudar a extrair valor dos ativos de dados, permitindo maior acesso, compartilhamento e integração a nível organizacional, além de aumentar a eficiência e a

⁴ <https://www.selecthub.com/big-data-analytics-tools/slamdata/>

responsabilidade geral. No entanto, embora o conceito não seja novo, a maioria dos governos está lutando para colocá-lo em prática.

Barc (2019) menciona que a governança de dados é cada vez mais relevante para as práticas de proteção de dados em escala global, de maneira mais exclusiva e explícita. No entanto, uma abordagem forte e desequilibrada da superproteção de dados pode reduzir o valor do seu compartilhamento, como na prestação de serviços públicos transfronteiriços.

Recentemente, o uso indevido de dados por empresas privadas e a crescente preocupação dos cidadãos sobre o gerenciamento de dados no setor público provocaram uma intervenção do governo para melhorar a proteção de dados pessoais (OCDE, 2019). Como resultado, o uso ético e transparente dos dados está agora no topo da agenda política.

Os fluxos de dados aumentaram nas organizações, setores e fronteiras, adicionando outro nível de complexidade à governança de dados em um mundo globalizado e interconectado. Ela não é mais uma questão dentro dos limites organizacionais, mas uma preocupação multinacional resultante do compartilhamento de dados além-fronteiras.

Nesse contexto, instrumentos internacionais, como o Regulamento Geral de Proteção de Dados, procuraram “devolver aos cidadãos o controle sobre seus próprios dados” (OCDE, 2019) e tomar medidas internacionais para impedir o uso indevido dos mesmos. O regulamento geral sobre proteção de dados avançou agenda, subjacente à necessidade de estruturas comuns para garantir a proteção desses dados.

No entanto, a superproteção de dados pode resultar do mal-entendido dos regulamentos nacionais e internacionais e conduzir a mudanças em termos de abordagens políticas. Por exemplo, de abertura “por padrão” para “abrir se possível, protegido se necessário”. O desafio global neste estágio é garantir o equilíbrio entre fluxos de dados gratuitos e proteção de dados (OCDE, 2019).

Embora os países da OCDE venham frequentemente definindo elementos relevantes para a governança de dados do setor público no contexto do governo digital, dados abertos, gerenciamento de dados e/ou estratégias e/ou políticas de IA, esses elementos são frequentemente fragmentados. Em alguns cenários, essa desconexão está profundamente enraizada nos intrincados arranjos de governança

que apoiam essas políticas ou na falta de clareza em termos de liderança e responsabilidades. Este fato coloca, portanto, barreiras importantes para a integração e compartilhamento de dados (BARC, 2019).

Uma governança holística de dados também pode ajudar a habilitar o governo como plataforma (uma das principais dimensões de um governo digital). Por exemplo, o desenvolvimento de ferramentas de dados comuns, mas flexíveis (como plataformas de compartilhamento de dados) fornece soluções que podem ser reutilizadas em todo o setor público (ALGMIN & ZAID, 2018).

Em um nível mais técnico, a fragmentação também resulta de desafios herdados em termos de qual organização gera e controla os dados e a impossibilidade de compartilhá-los e acessá-los à luz de acordos legais específicos. Isso leva a políticas e soluções técnicas isoladas que aumentam a impossibilidade de construir um governo integrado e conectado.

A falta de um modelo abrangente de governança de dados pode levar à proliferação ou duplicação de padrões e soluções técnicas para compartilhamento de dados. Isso dificulta a interoperabilidade entre diferentes organizações e setores, afetando diretamente a possibilidade de integração desses dados, processos e organizações. Também leva a repetidas solicitações para os cidadãos fornecerem seus dados pessoais várias vezes ao setor público desnecessariamente (ALGMIN & ZAID, 2018).

Uma estrutura de governança de dados deve garantir o gerenciamento adequado dos dados durante todo o seu ciclo de vida (Ghavami, 2015). Por exemplo, nos últimos anos, o movimento aberto de dados do governo permitiu uma discussão mais aprofundada da necessidade de fortalecer a liderança e administração de dados no setor público.

Isso também abriu uma discussão mais técnica sobre práticas aprimoradas de gerenciamento de dados, por exemplo, em torno da produção, armazenamento, processamento e compartilhamento para uma maior abertura de dados. No entanto, esses elementos não foram entendidos como parte de esforços mais amplos de dados do setor público que conectam todos os estágios do ciclo de valor dos dados.

Os países repentinamente perceberam o valor da catalogação de dados para fins de abertura e descoberta, mas falharam em reconhecer como essas iniciativas

também traziam benefícios políticos relevantes para a produtividade no setor público. Por outro lado, em alguns países da OCDE, uma cultura bem estabelecida de eficiência do setor público levou ao desenvolvimento de registros de dados como um meio de melhorar o compartilhamento interinstitucional de dados.

Essa mentalidade ofuscou o crescente valor da abertura de dados do governo e do envolvimento e colaboração com atores externos para encontrar soluções para os desafios das políticas. Como resultado, os países que outrora lideraram o antigo movimento de governo eletrônico (com forte foco na eficiência) ficaram muito atrás daqueles que duplicaram os esforços para compartilhar e abrir dados aos usuários como forma de promover negócios e inovação social (GHAVAMI, 2015).

Países da OCDE, como Canadá, Irlanda, Holanda, Reino Unido e Estados Unidos, avançaram ou estão caminhando para a definição de estratégias abrangentes de dados como meio de aumentar a coesão do setor público e promover a integração de políticas e ferramentas (OCDE, 2019).

Essas estratégias abrangem a maioria, se não todos, os estágios do ciclo de valor dos dados do governo (desde a produção de dados, passando pela sua proteção até a abertura e reutilização dos mesmos). Ainda assim, cada estágio requer arranjos específicos, pois eles produzem benefícios políticos específicos. Por exemplo, dados abertos permitem o uso como uma plataforma para maior envolvimento e colaboração do usuário, e melhores práticas de produção de coleta de dados podem ajudar a reduzir o viés político.

A OCDE observou que um forte foco em questões técnicas, como principal resultado da governança de dados, pode desencaminhar decisões políticas relacionadas a este tema. Por exemplo, concentrando-se principalmente na adoção de soluções tecnológicas, como interfaces de programação de aplicativos (APIs) e padrões de dados, em vez de somente permitir o aspecto organizacional, governança e cultura adequadas a contextos para enfrentar os desafios políticos. Todos esses são elementos-chave para uma boa governança de dados (OCDE, 2019).

Em alguns casos, os países da OCDE investiram recursos para definir papéis estratégicos (por exemplo, administradores e diretores de dados) e apoiar a governança de dados através da definição de um tecido institucional mais forte. O estabelecimento dessas funções estratégicas pode ajudar no dimensionamento e na

sustentação da implementação de políticas e na construção de uma maior maturidade de dados no setor público (OCDE, 2018).

Isso ocorreu no contexto de estratégias ou políticas de dados abertos na Coreia do Sul e Estados Unidos, por exemplo. No entanto, na maioria dos países a liderança e/ou administração de dados ainda são incompreendidos, limitando a governança de dados às atividades do departamento de TI e não como um fator que pode ajudar a alcançar as metas da política por meio de melhores práticas de gerenciamento e compartilhamento de dados.

3. METODOLOGIA E APLICAÇÃO DA PESQUISA

3.1. Compreendendo as etapas metodológicas

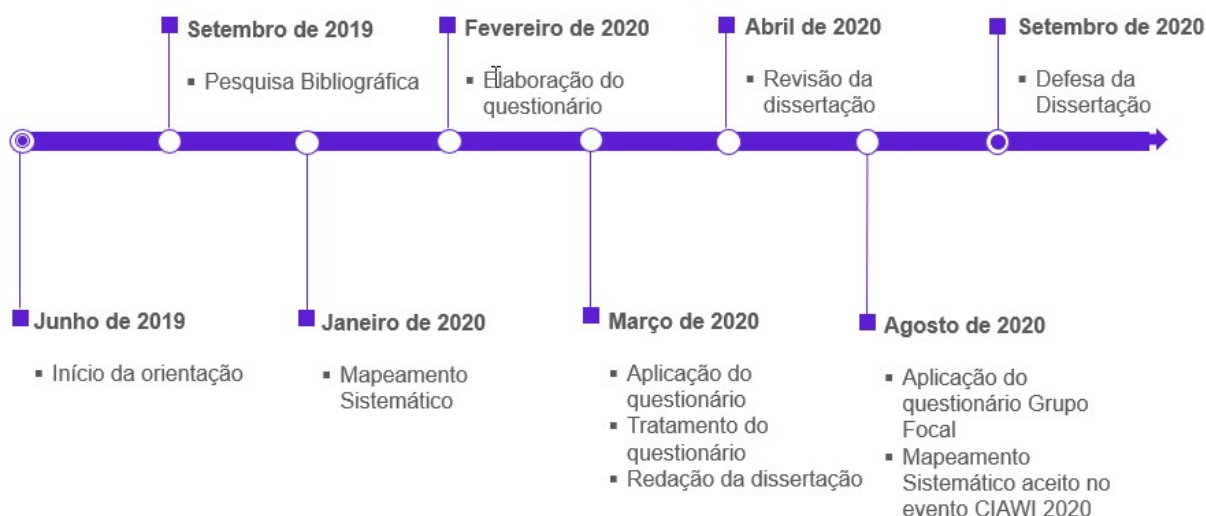
Esta dissertação foi desenvolvida em várias etapas metodológicas. A primeira pode-se dizer que é o embasamento do capítulo destinado à revisão da literatura, onde foi explorado através de pesquisas diretas (livros e obras autorais) e indiretas (teses de doutorado, mestrado, artigos, revistas, jornais, etc.) para compor o entendimento temático em questão.

Em um segundo momento, apresenta-se o mapeamento sistemático (APÊNDICE A), cujo estudo antecede esta dissertação, que também é de cunho acadêmico. O artigo foi aprovado para publicação na 17ª Conferência Ibero-Americana, Lisboa – Portugal, em dezembro de 2020.

Na etapa seguinte, foi desenvolvida uma pesquisa do tipo *survey* (APÊNDICE B). Esta pesquisa foi fundamental para a compreensão e entendimento das aplicações de *Big Data* nas Instituições Públicas Federais. Um questionário foi disponibilizado pelo *Google Forms* durante um período e enviado aos principais órgãos brasileiros. Os relatos dos participantes foram registrados e podem ser conferidos na seção “Resultados e Discussão”.

Abaixo, na Figura 7, é possível verificar todas as etapas desta dissertação, compreendendo desde o início das atividades relacionadas ao período de orientação científica, pesquisa bibliográfica, mapeamento sistemático, pesquisa *survey*, pesquisa com grupo focal, confecção do *guideline* e conclusão dos estudos.

Figura 7 – Compreensão das etapas metodológicas.



Fonte: Elaborado pelo autor.

Posteriormente, foi realizada uma pesquisa de grupo focal (APÊNDICE C). Tal pesquisa foi destinada a um grupo de especialistas em governança de dados e *Big Data*. Os procedimentos foram embasados na literatura de Devenport (2014), que serviu de guia e fundamentação para a formulação do conjunto de perguntas conforme Apêndice C.

Por fim, baseando-se na pesquisa de grupo focal, foi proposto um *guideline* como parte colaborativa deste objeto de estudo. Tal contribuição, traz relatos entre diretrizes e regras para que futuras instituições possam utilizar, para implantação das boas práticas de Governança de Dados e *Big Data*.

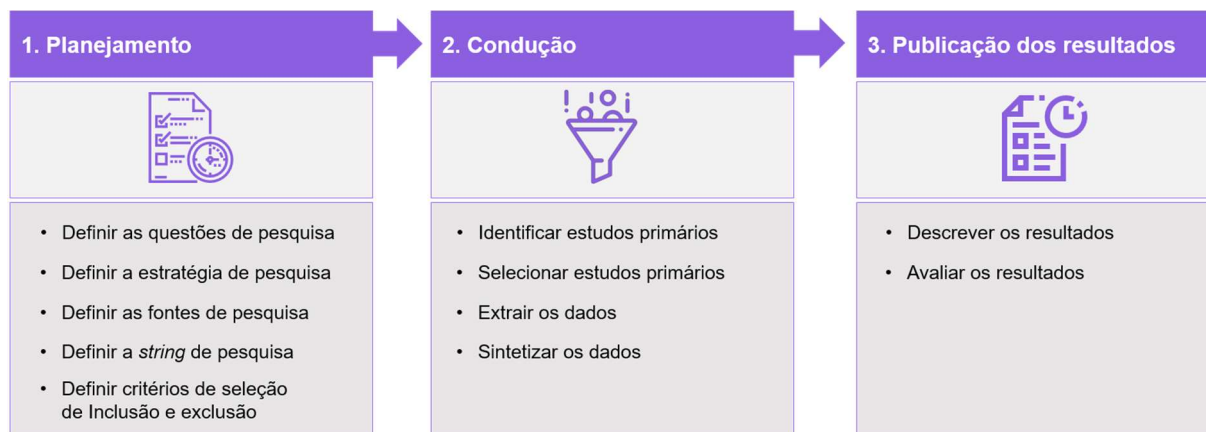
3.2. Apresentação do Mapeamento Sistemático

O artigo (APÊNDICE A) apresenta o Mapeamento Sistemático (SANTOS; BARRETO, 2015) da literatura referente uma análise do impacto das práticas de *Big Data* na Administração Pública Federal - APF, a qual busca obter subsídios para responder quais os objetivos e maiores desafios quando da implantação de *Big Data* e/ou Ecossistemas de *Big Data*.

O artigo é antecessor a esta dissertação, e contribuiu diretamente para o engajamento e aprofundamento desta pesquisa. O Mapeamento Sistemático está fortemente relacionado com o conhecimento da literatura correlata que pode auxiliar no entendimento da questão em epígrafe.

A Figura 08 ajuda na compreensão das etapas que constituiu o Mapeamento Sistemático que antecedeu a esta dissertação.

Figura 8 – Etapas do Mapeamento Sistemático.



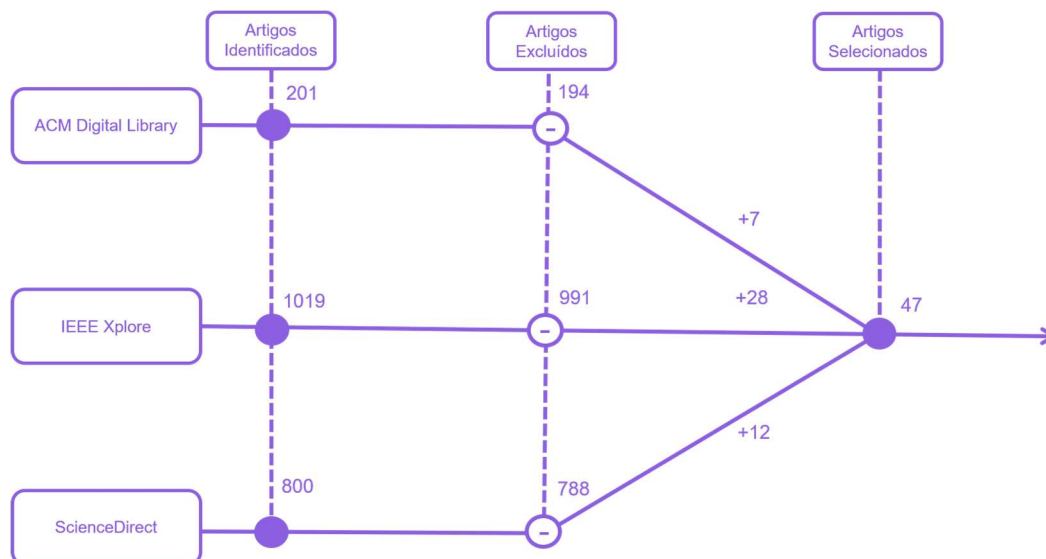
Fonte: Elaborado pelo autor.

As pesquisas iniciais realizadas no artigo contribuem diretamente para a compreensão das diretrizes de implantação de *Big Data* no setor público. Tais diretrizes precisam estar perfeitamente alinhadas com a missão institucional do órgão que se propõe a implantá-las, mas os gestores públicos nem sempre percebem suas instituições aptas a implantarem projetos de *Big Data*.

Desta forma surgiu a necessidade de aplicar o questionário referente ao *survey*, pois se entendia que por mais que se tenham investimentos públicos para a tecnologia em questão, a prática estava longe de ser ideal em várias instituições que participaram do processo.

Na Figura 09 está registrada a síntese dos resultados aplicados junto aos procedimentos referentes à pesquisa de cunho bibliográfico.

Figura 9 – Síntese dos procedimentos de pesquisa bibliográfica.



Fonte: Elaborado pelo autor.

3.3. Apresentação da pesquisa *Survey*

Para este estudo foi escolhida a metodologia de pesquisa bibliográfica para embasamento do *survey*.

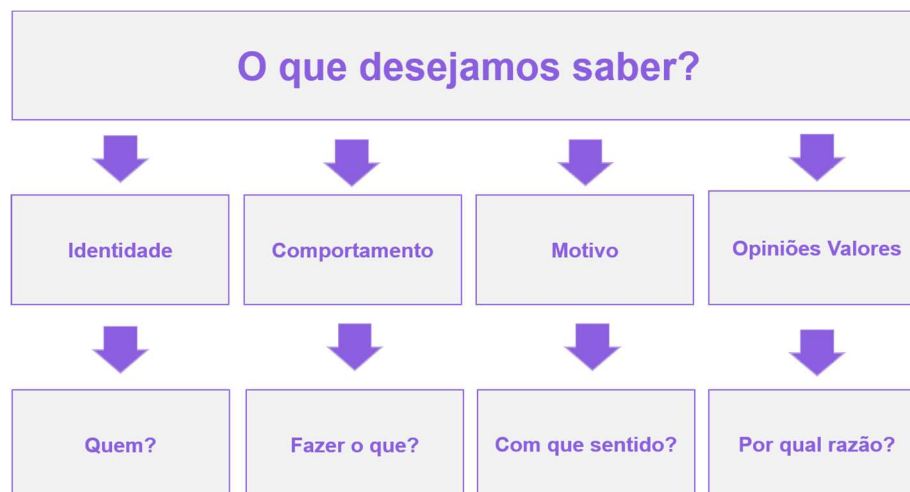
A pesquisa bibliográfica foi fundamental para o entendimento não só dos conceitos básicos referentes ao *Big Data*, como também das motivações para seu uso e das tecnologias relacionadas em órgãos públicos. Como evidenciado ao longo de todo o presente texto, a maioria dos artigos encontrados, em especial aqueles que se relacionam com a utilização de *Big Data* por organizações governamentais, estão ligados a casos fora do Brasil.

A discussão detalhada da bibliografia foi objeto da seção “Referencial Teórico” e se encontram sintetizadas, ainda, no mapeamento sistemático que antecedeu a presente dissertação, da qual a cópia pode ser visualizada no Apêndice A.

Neste sentido, identificou-se a necessidade de uma pesquisa de campo. A pesquisa de tipo *survey* consiste no método de coleta de informações diretamente de pessoas a respeito de seus conhecimentos e vivência profissional no foco da pesquisa (BOCCATO et al., 2006).

A partir dos objetivos traçados e do público alvo que deseja atingir elaboram-se as perguntas. Com isso deve ser elaborada uma estratégia de ação que deve permear a elaboração da maioria das questões. Estas indagações estão ilustradas na Figura 10.

Figura 10 – A Concepção e a Função do Questionário



Fonte: Moscarola (1990), modelo adaptado pelo autor.

Para a realização de tal pesquisa, e com respostas das quais se pudessem tomar alguma validade de conclusões ou de direcionamentos, definiu-se que o perfil do respondente ou entrevistado deveria obedecer a um dos seguintes critérios:

- ter alguma relação de trabalho com a área de TI;
- ter conhecimento mínimo sobre *Big Data* e termos relacionados;
- trabalhar prestando serviços dentro de órgãos públicos, ainda que se tratem de trabalhadores contratados em regime de terceirização;
- ter algum conhecimento acerca do órgão público para o qual prestassem serviço, bem como dos resultados ou projetos relacionados com o uso do *Big Data*.

Logo, qualquer questionário que fosse utilizado na pesquisa deveria coletar estas informações juntamente com as demais, para que se aferisse algum grau de pertinência do conjunto de respostas coletadas sobre *Big Data* no serviço público.

Algumas das perguntas aludiram a questões de volumetria de dados guardados, uma vez que parte do público que trabalha com *Big Data* pode entender apenas como operações sobre conjuntos de dados a partir de determinado volume.

As demais perguntas procuram aludir às questões principais que esta dissertação procura auxiliar a responder. Cumpre lembrar ao leitor que o objetivo principal da pesquisa é que, pela análise de práticas e aplicações de *Big Data* no setor público brasileiro, em âmbito federal, se entenda quais os objetivos e maiores desafios e dificuldades relacionados com a adoção da *Big Data*.

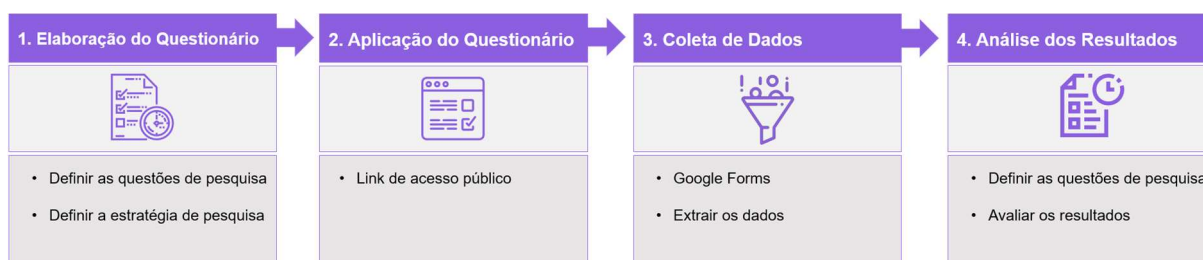
É necessário, ainda, destacar os objetivos específicos, os quais consideram um escopo mais focado em aspectos relacionados com as necessidades e com os papéis dos órgãos públicos mediante a sociedade administrada, o que se busca inferir com base nas questões elencadas a seguir.

- Quais são os impactos econômicos das soluções de *Big Data* no setor público?
- Quais as oportunidades e desafios da adoção de *Big Data* no setor público?
- Quais os impactos das práticas de *Big Data* no setor público?

Com base em questões que envolvessem estas perguntas ligadas aos objetivos geral e específicos da dissertação, seria possível o alinhamento do caso da Administração Pública Brasileira com os casos aferidos em outros países, bem como a adequação dos projetos de *Big Data* nos órgãos públicos às suas respectivas missões institucionais, algo que é considerado condição fundamental de casos de sucesso, conforme visualizado na literatura correlata.

Em adição, seria possível visualizar como está sendo motivada a implantação e utilização do *Big Data* e que resultados podem já ser compreendidos, sempre com foco nos resultados principais que se espera de organizações ligadas ao governo, em especial questões de eficiência e eficácia. A Figura 11 demonstra as etapas básicas do método utilizado.

Figura 11 – Etapas da Pesquisa.



Fonte: Elaborado pelo autor.

A seguir se apresenta as etapas acerca da pesquisa de campo:

1. **Elaboração do Questionário:** além das premissas apresentadas em epígrafe, as questões foram elaboradas com base no conhecimento e experiência profissional de casos observados no dia a dia de uma área de tecnologia relacionada a aplicações para *Big Data*, focando no setor público federal. O autor deste trabalho exerceu a função de construtor do formulário bem como sua divulgação por meio da plataforma *Google Forms*, atendendo todos os requisitos de segurança e normalização institucional. As perguntas do questionário foram elaboradas abordando questões como: o cargo no setor público, qual o poder de atuação (Executivo, Legislativo ou Judiciário), além de questões relevantes sobre TI, experiências, conhecimento sobre *Big Data*, diretrizes, políticas de segurança para o armazenamento da informação, entre outras.
2. **Aplicação do Questionário:** a aplicação do questionário foi feita por meio de *link* de acesso para o público alvo escolhido para realização da pesquisa. O formulário foi enviado para profissionais ligados à tecnologia *Big Data* (sempre focando em Setores Públicos Federais) de diferentes regiões e portes, proporcionando uma maior abrangência e alcance na coleta dos dados.
3. **Coleta de Dados:** foi utilizada a ferramenta on-line de formulário do *Google Forms* com perguntas descritivas com opções de respostas para a escolha (Apêndice B).
4. **Análise dos Resultados:** a elaboração do questionário faz-se necessário para a compreensão dos processos e diretrizes que norteiam a gestão de políticas de armazenamento e segurança para *Big Data*, buscando por meio das perguntas o foco na utilização, políticas, sistemas, segurança, armazenamento, entre outros.

3.3.1 Etapas da aplicação do questionário

Buscando uma compreensão na literatura, May (2004) relata que o questionário é somente um conjunto de questões realizado para gerar dados com o intuito de atingir objetivos específicos de um projeto. Chagas (2000) menciona que a aplicação de questionários é de suma importância para a obtenção de conhecimentos específicos

em áreas diferentes e confeccionar um questionário não é uma tarefa das mais fáceis, já que, além do esforço, envolve tempo e bastante conhecimento sobre a proposta.

Diante de tais concepções este objeto de estudo limitou-se na formulação de um questionário com 33 questões, conforme Apêndice B, cumprindo as seguintes etapas:

- 1ª etapa: disponibilização do Termo de Consentimento Livre e Esclarecido (TCLE), que busca a conscientização do participante de maneira livre e espontânea, informando sobre os objetivos do questionário, levantamento de dados e informações relevantes sobre os processos;
- 2ª etapa: um campo obrigatório que disponibiliza ou não o formulário para ser preenchido pelo participante, uma pergunta direta se deseja ou não participar do mesmo.
- 3ª etapa: nesta etapa, exploraram-se as informações básicas do participante como cargo que ocupa na empresa (Administração Pública), em qual sistema de poder trabalha, qual órgão, etc.
- 4ª etapa: compreendeu um conjunto de informações básicas relativas ao emprego do *Big Data* em si. Informações de familiarização com *Big Data*, experiência na área, estrutura dos órgãos públicos para aplicar *Big Data*, entre outros.
- 5ª etapa: compreende-se em um conjunto de questões mais abrangentes relacionadas ao *Big Data*, enfatizando políticas, normas, regimentos, sistemas de segurança e possíveis tecnologias que poderão ser acopladas ao sistema.

Conforme explanado anteriormente, o meio utilizado para a realização desta pesquisa foi através do *Google Forms*. O fato de a plataforma ser uma das mais utilizadas atualmente, os serviços da Google são acessados e conhecidos pela área acadêmica, técnica e científica, permitindo assim uma melhor disseminação dos formulários, estatísticas geradas por meio dos questionamentos (perguntas diretas de múltipla escolha) e o compartilhamento dos resultados com todos aqueles que participaram diretamente da pesquisa.

O *Google Forms* passou a ser utilizado por acadêmicos, cientistas, etc., por possuir uma interface intuitiva, simples e de fácil compreensão. Para quem realiza o questionamento (formatação das perguntas), a plataforma oferece várias opções de

controle, além de permitir a geração de gráficos que facilitam a compreensão dos resultados.

A pesquisa foi disponibilizada durante 14 dias, contendo um número de 83 participantes, porém 11 registros foram excluídos por inconsistência nos dados, distribuídos em 45 órgãos públicos, dos quais todos aceitaram participar da mesma. Deste modo, avaliou-se a participação dos envolvidos mediante os questionamentos obtendo dados precisos sobre as diretrizes de gestão de tecnologias, políticas e diretrizes de segurança voltadas para *Big Data*.

3.4. Apresentação da pesquisa referente ao Grupo Focal

Através da pesquisa *survey*, descrita na seção anterior, foi possível perceber a necessidade de se realizar uma nova pesquisa, mais direcionada e com especialistas que estão vivenciando na prática os conceitos desta dissertação.

Para modificar comportamentos é necessário primeiro que se entenda o contexto no qual ocorrem os significados e a importância a eles atribuídos pelos seus agentes. Enquanto uma pequena parcela do significado do contexto comportamental é de cunho pessoal e particular, grande parte deste é culturalmente moldada e socialmente construída, (GILBERT, 1991).

Mediante esta necessidade, surgiu então o interesse pela pesquisa de grupo focal. Desta forma, apesar das análises preliminares realizadas nas instituições públicas federais, percebeu-se a necessidade de se buscar algo mais específico para o entendimento das ações e diretrizes para aplicação de governança de dados aliadas a *Big Data*, e, deste modo, optou-se por uma pesquisa de Grupo Focal.

A pesquisa foi realizada com um grupo específico de profissionais, contendo sete integrantes, onde foram debatidas várias questões relacionadas à temática durante um pouco mais de 1 hora. A seleção levou em conta requisitos sobre o tema, tempo de experiência, prática, vivência, entre outros, na tentativa de refinar os dados obtidos para gerar como proposta de implantação das boas práticas de governança de dados aliados ao *Big Data* e uma *guideline*, cumprindo as seguintes etapas:

- 1ª etapa: seleção da equipe com experiência relevante acerca dos tópicos em discussão, com conhecimentos específicos de Governança de Dados e *Big Data*;

- 2ª etapa: disponibilização do Termo de Consentimento Livre e Esclarecido (TCLE), que busca a conscientização do participante de maneira livre e espontânea, informando sobre os objetivos do Grupo Focal, levantamento de dados e informações relevantes sobre os processos;
- 3ª etapa: elaboração do roteiro de discussão, dividido em seis tópicos: dados, negócio do órgão, liderança, metas, tecnologia, equipe de TI e analistas de dados, com 35 questões, conforme Apêndice C.
- 4ª etapa: condução da entrevista em um ambiente on-line e interativo;
- 5ª etapa: análise dos resultados.

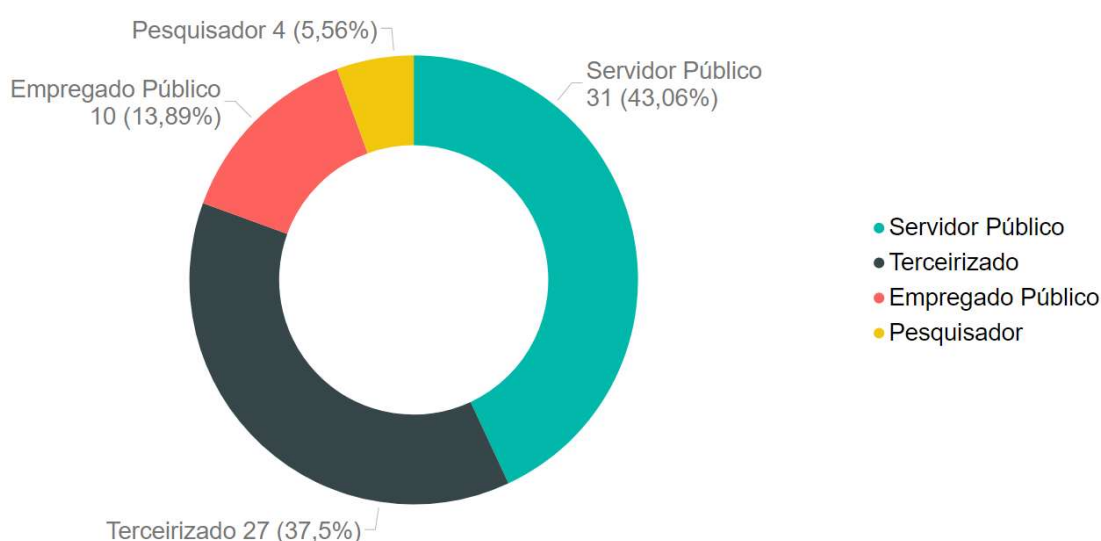
4. RESULTADOS E DISCUSSÕES

4.1. Análise da aplicação da pesquisa *Survey*

Inicialmente, convém destacar que a pesquisa de campo foi utilizada como ferramenta acessória na busca por respostas às perguntas principais e auxiliares da pesquisa em relato. O intuito do questionário era dimensionar o envolvimento dos órgãos públicos federais com as técnicas e diretrizes empregadas ao *Big Data*.

Entre as questões de 1 a 5 empregadas no questionário em apêndice, se buscaram meios de identificação do perfil do público respondente.

Gráfico 1 – Cargos ocupados pelos entrevistados na Administração Pública

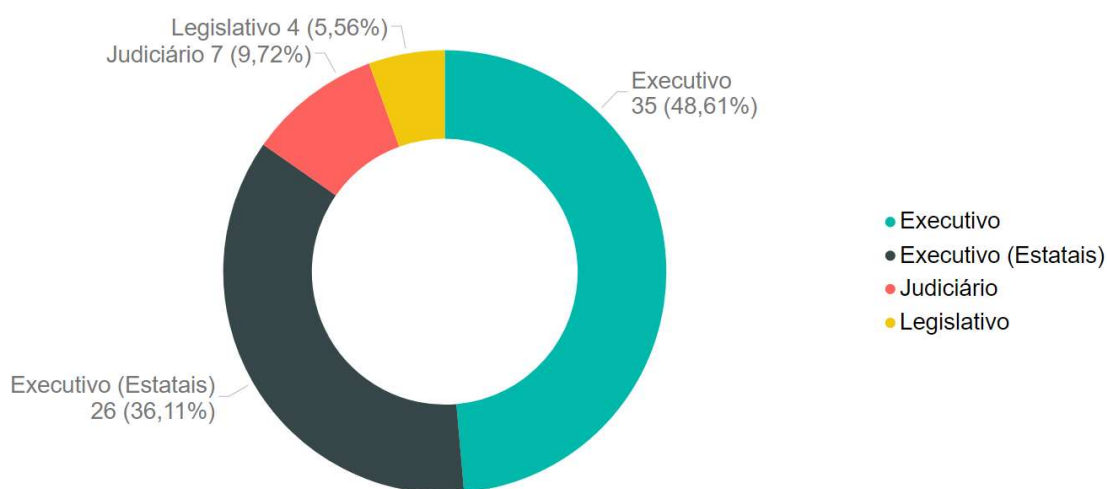


Fonte: próprio autor.

No Gráfico 1 acima, com relação ao cargo ocupado pelos entrevistados na administração pública federal, nota-se que 31 pessoas, ou seja, 43,06% pertencem à classe dos servidores públicos. Já 27 pessoas (37,5%) pertencem ao grupo dos profissionais terceirizados. Empregado público aparece em 3º lugar com 10 pessoas ou 13,89% e como pesquisadores apenas 4 pessoas ou 5,56%. Logo, a maioria dos entrevistados são servidores e empregados públicos, o que pode representar um quórum que, embora ligado à área de TI, teria condições de fazer a conexão, em teoria, dos trabalhos da área com as atividades institucionais de cada instituição pública representada.

Analisando a segunda questão, no Gráfico 2, ligada ao poder que os entrevistados trabalham, a maioria trabalha no poder executivo, representando 35 pessoas ou 48,61%. Enquanto 26 pessoas, ou 36,11%, trabalham no setor Executivo (estatais), 7 pessoas no Judiciário e 4 pessoas no Legislativo, representando apenas 9,72% e 5,56% respectivamente.

Gráfico 2 – Em qual poder trabalha o entrevistado

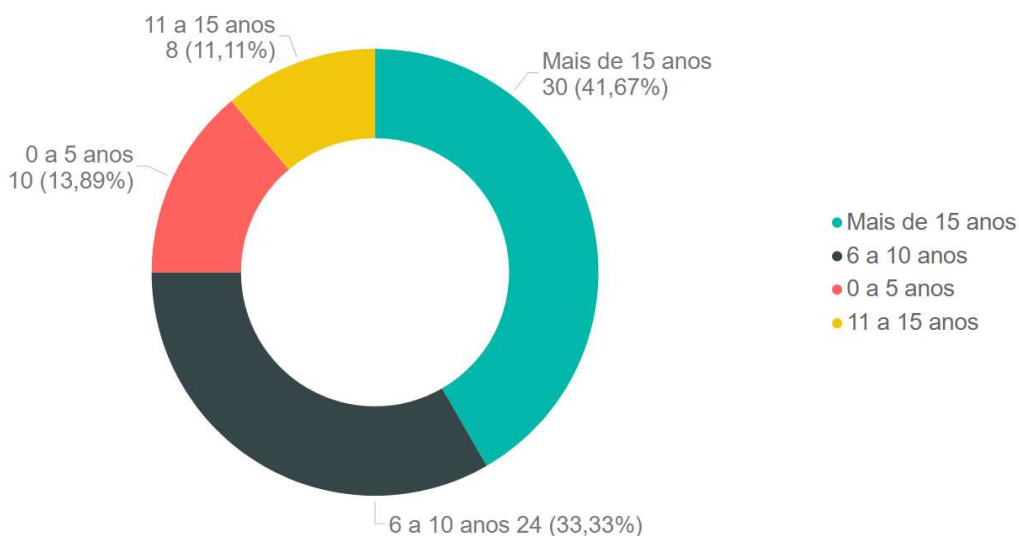


Fonte: próprio autor.

Na terceira questão explorou-se em qual órgão o entrevistado trabalha. A quantidade total foi de 45 órgãos.

A quarta questão, da qual o resultado é compilado no Gráfico 3, explora o tempo de experiência dos entrevistados em relação ao emprego em TI.

Gráfico 3 - Tempo de experiência com T.I do entrevistado.



Fonte: próprio autor.

A maioria (cerca de 52%) possui mais de 10 anos de experiência em seus cargos. Estas perguntas iniciais trataram de traçar um pouco do perfil demográfico dos respondentes e seus locais de trabalho. O perfil indicou que se trata de um conjunto de profissionais com experiência em TI e que trabalham em órgãos majoritariamente do Poder Executivo. Portanto, órgãos mais ligados à execução de políticas e serviços públicos, e que, assim, guardariam, a princípio, certas similaridades e resultados de utilização de tecnologias de *Big Data*.

A partir desta etapa, se iniciam os questionamentos que buscam compreender informações com as tecnologias relacionadas ao *Big Data*. Desta forma, a questão cinco retrata apenas os estados dos quais os órgãos estão lotados e, na próxima questão, iniciam-se assuntos correlacionados com a aludida tecnologia.

Já a questão 6, buscou compreender qual era o grau de familiaridade com os termos concernentes ao *Big Data*. Noventa e quatro por cento (94,44%) afirmaram estar familiarizados com a terminologia correlata ao tema.

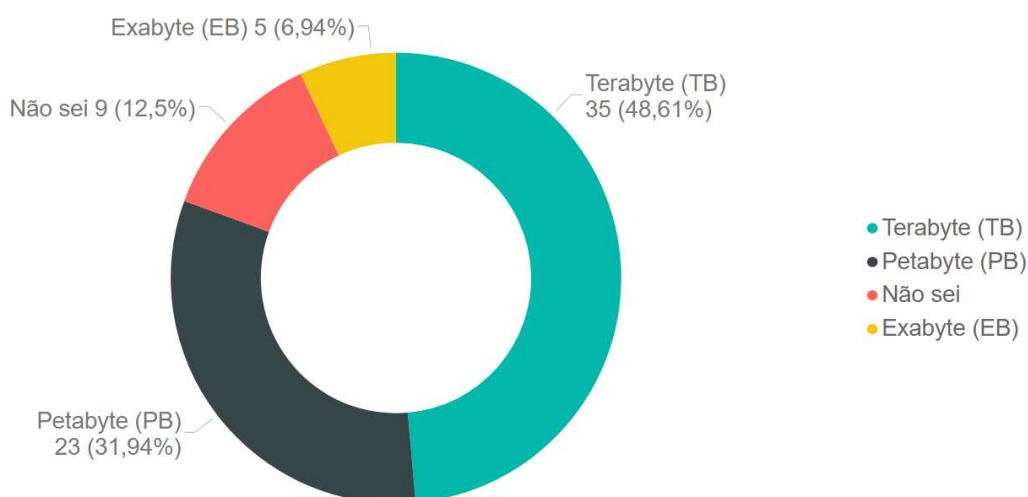
Na sétima questão, inicialmente buscou-se compreender qual era o nível de experiência relacionada ao *Big Data*. Porém, mesmo com as respostas, concluiu-se que tais questionamentos não significam de fato que o entrevistado possui experiência com o tema, até porque esta experiência não pode ser medida de forma direta neste objeto de estudo.

Na oitava questão, buscou-se compreender, na visão do entrevistado, a relação entre a volumetria de dados e o *Big Data*. A totalidade das respostas indicou que há a relação, indicando que todos os entrevistados estavam cientes que o ambiente de *Big Data* estaria relacionado com grandes volumes de dados. Assim teriam condições de, mesmo que não concordassem acerca do tamanho exato da massa dados que poderia ser considerada “*Big Data*”, estariam cientes que este volume não pode ser pequeno.

Posteriormente, na nona questão, busca-se saber se no órgão em que o entrevistado trabalha existe ambiente para *Big Data*. Dos respondentes, 32% afirmaram que não e 68% que sim. Esta questão pode contrastar com as anteriores, principalmente com relação à experiência dos entrevistados ligados ao *Big Data*. Como analisado anteriormente, o fato de trabalhar na área por um determinado período, não é suficiente para aferir o grau de experiência na área. Ainda, não se pode afirmar que o fato de 32% dos órgãos não dispor deste ambiente significa que seus funcionários não tenham experiência em *Big Data*.

Na décima questão buscou-se identificar qual era a volumetria de armazenamento de dados do órgão em que trabalha o respondente. A compilação deste resultado está representada no Gráfico 4.

Gráfico 4 - Volumetria de armazenamento dos órgãos.



Fonte: próprio autor.

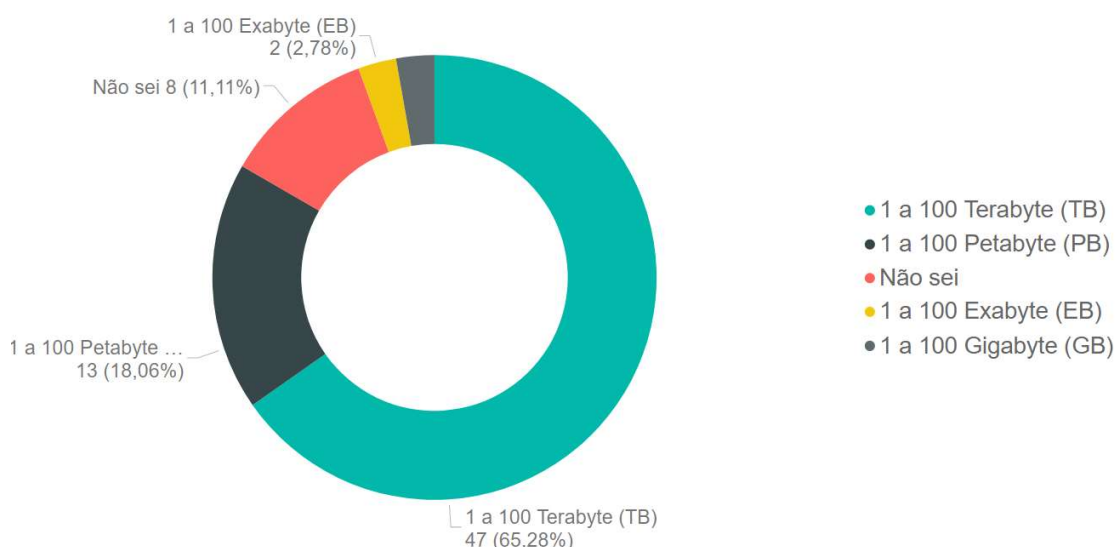
Verifica-se que boa parte dos órgãos trabalha com *terabytes* de informação, 48,61%, na visão dos entrevistados. Outro grupo indicou volumes na casa de

petabytes, representando 31,94% do total. Doze por cento (12,5%) não souberam responder e 6,94% trabalham com *exabytes* de informações. *terabytes*, contudo, pode não representar, necessariamente, um ambiente de *Big Data*.

Assim, a não ser que as tecnologias utilizadas fossem indiscutivelmente relacionadas ao *Big Data*, poderia haver questionamento acerca das respostas anteriores quanto à existência de ambientes de *Big Data*, pois com base na volumetria de *terabytes* não há segurança na afirmação da existência de tal ambiente.

Continuando, na décima primeira questão perguntou-se qual é a projeção do crescimento do volume de dados para o ano de 2020 (Gráfico 5).

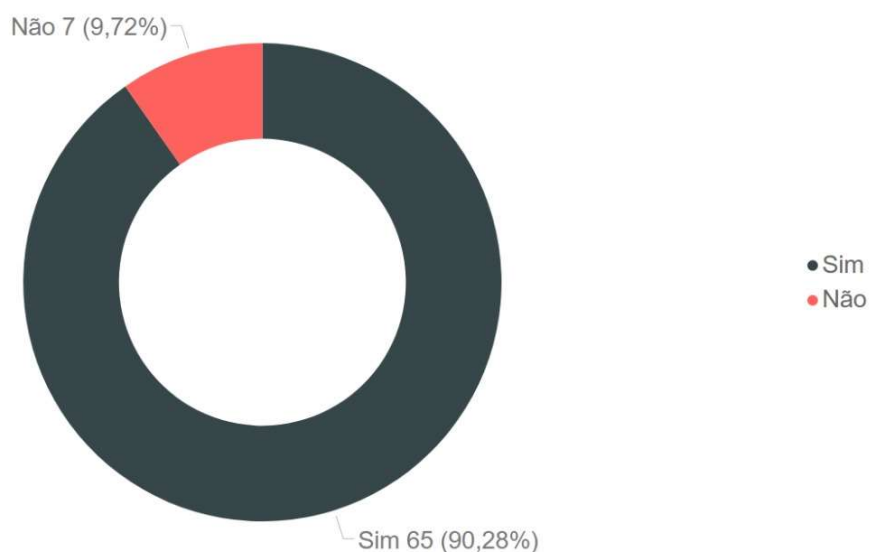
Gráfico 5 - Projeção de crescimento do banco de dados.



Fonte: próprio autor.

Sessenta e cinco por cento 65,28% disseram que era esperado entre 1 a 100 *terabytes* de volume de dados, 18,06% entre 1 a 100 *petabytes* e menos de 2,78% esperam de 1 a 100 *gigabytes* e *exabytes*, respectivamente. Por fim, 11,11% não souberam responder.

Já na décima segunda questão questionou-se se, com o crescimento esperado, o órgão onde trabalha o respondente passaria a ter um ambiente de *Big Data*. O resultado está compilado no Gráfico 6.

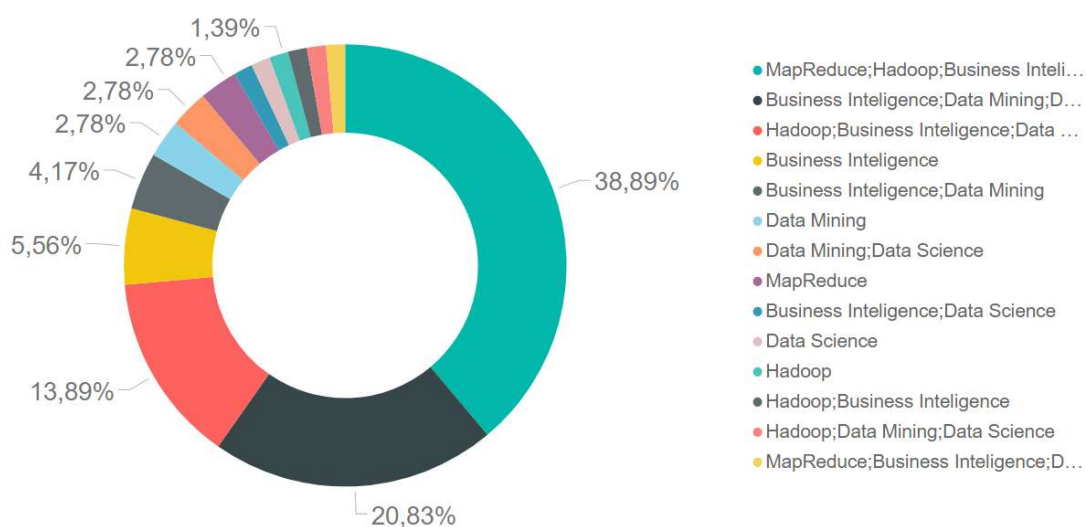
Gráfico 6 - Possui projeto para ambiente próprio de *Big Data*?

Fonte: próprio autor.

Mais de noventa por cento (90,28%) afirmaram que possuem projetos para aderirem ao próprio ambiente de *Big Data* e 9,72% não possuem.

Ou seja, se confirmou que a maioria dos profissionais respondentes, de fato, compreende que um volume de dados na casa de *terabytes* é suficiente para caracterizar um ambiente de *Big Data*, e que, portanto, para estes, a influência das tecnologias componentes do *Big Data* para a definição da existência de um ambiente é mais importante que volumes que cheguem a *petabytes*, por exemplo.

Na questão 13 (Gráfico 7) e 14 (Gráfico 8) buscou-se verificar se os usuários estariam familiarizados com algumas tecnologias relacionadas com o *Big Data*, e se sabiam diferenciá-las de outras também comuns nestes ambientes.

Gráfico 7 - Diferenciação entre os termos distintos de *Big Data*.

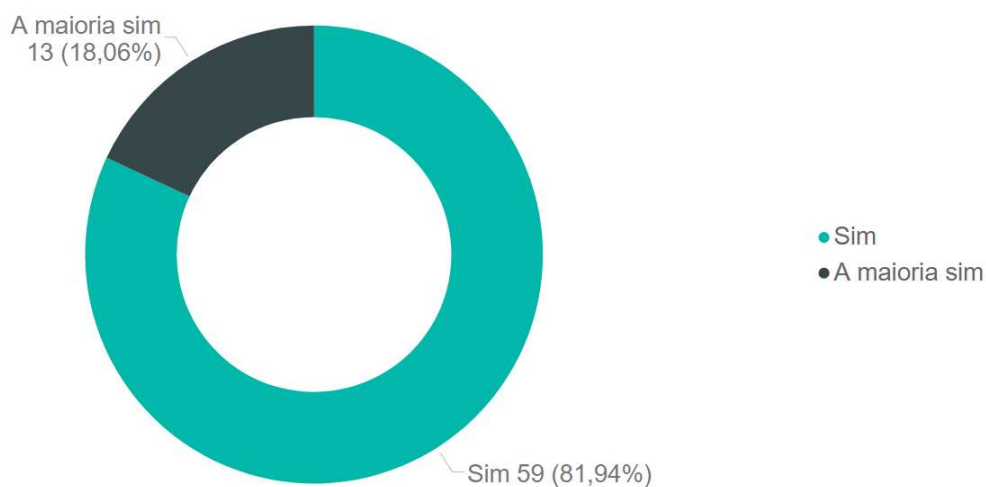
Fonte: próprio autor.

Os três maiores grupos de respondentes a esta questão indicaram que a maior parte dos participantes sabia diferir pelo menos três dos conceitos. Estes três maiores grupos representam cerca de 72% das respostas. O resultado, ainda, demonstrou que mais de um terço deles sabia diferir todos os conceitos, conforme vemos no Gráfico 7.

O resultado demonstra que ao menos uma parte dos que afirmaram existir um ambiente de *Big Data* em suas repartições conhecia inclusive acerca de algumas das tecnologias do *Big Data*. Portanto, é fato que a composição deste ambiente estaria ligada às tecnologias, além do volume de dados propriamente dito. Este dado não apenas corrobora que eles conheciam o conceito, mas que de alguma forma o aplicavam em seu próprio ambiente profissional.

A décima quarta questão serviu de complemento à décima terceira, verificando se os respondentes sabiam acerca da profundidade de relação entre os conceitos apresentados e o ambiente de *Big Data*.

Gráfico 8 - Considera que todos os termos acima possuem forte relação com o tema *Big Data*?

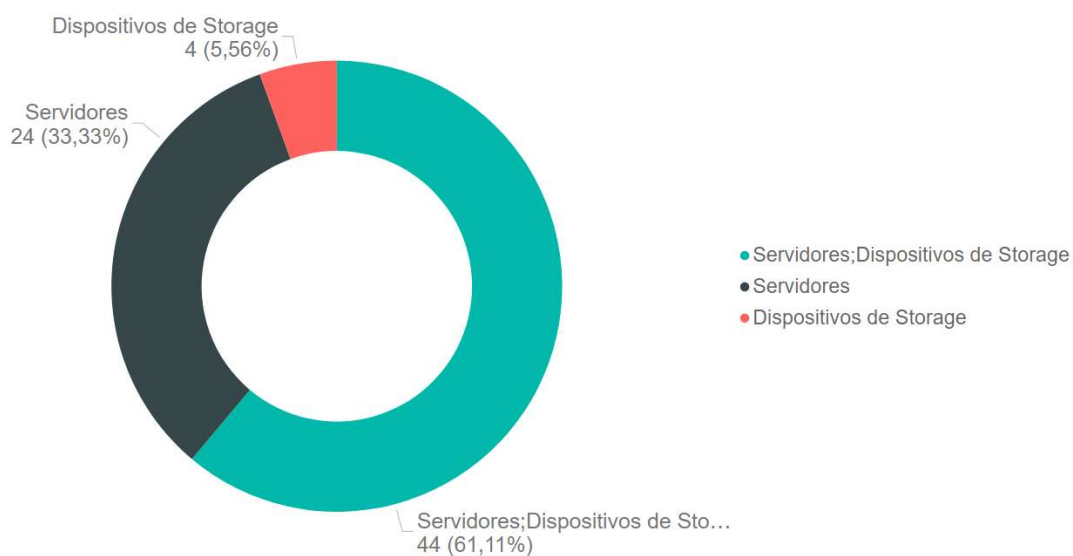


Fonte: próprio autor.

Oitenta e um por cento (81,94%) tinha em mente que os conceitos/tecnologias estariam fortemente relacionados com o *Big Data*, conforme demonstrado no Gráfico 8 acima.

O resultado ilustrado no Gráfico 9 é relativo à décima quinta questão, que se referia à quais os tipos de armazenamento de dados do órgão.

Gráfico 9 - Tipos de armazenamento de dados do órgão.

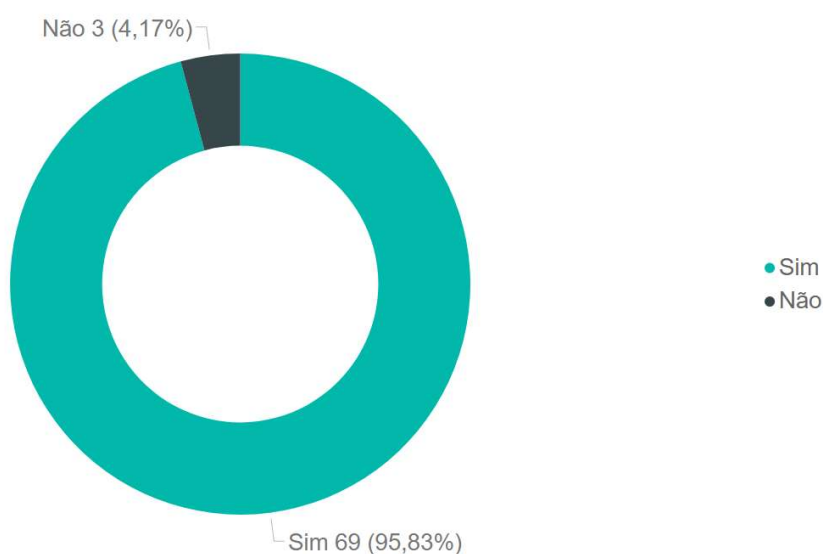


Fonte: próprio autor.

Sessenta e um por cento (61,11%) disseram Servidores Dispositivos de *Storage*; 33,33% marcaram Servidores; e 5,56% responderam Dispositivos de *Storage*.

A questão seguinte, a décima sexta, buscava conhecer sobre a infraestrutura de armazenamento de dados estruturados e não estruturados. O resultado está apresentado no Gráfico 10.

Gráficos 10 – Possui dados não estruturados?



Fonte: próprio autor.

É possível notar que noventa e cinco por cento (95,83%) disseram que sim, haviam dados não estruturados enquanto que 4,17% disseram que não. Ou seja, os dados nos sugerem que há apenas uma parcela muito pequena de órgãos que trabalha exclusivamente com dados estruturados.

Já na décima sétima questão, questionava acerca do acesso dos órgãos a um grande volume de dados, sejam estes estruturados ou não. Deve-se ressaltar que em um ambiente de *Big Data*, devem ser considerados os 5 Vs (Volume, Velocidade, Variedade, Veracidade e Valor), porém, especificadamente nesta questão, buscou-se apenas a compreensão sobre o acesso dos órgãos a estes dados.

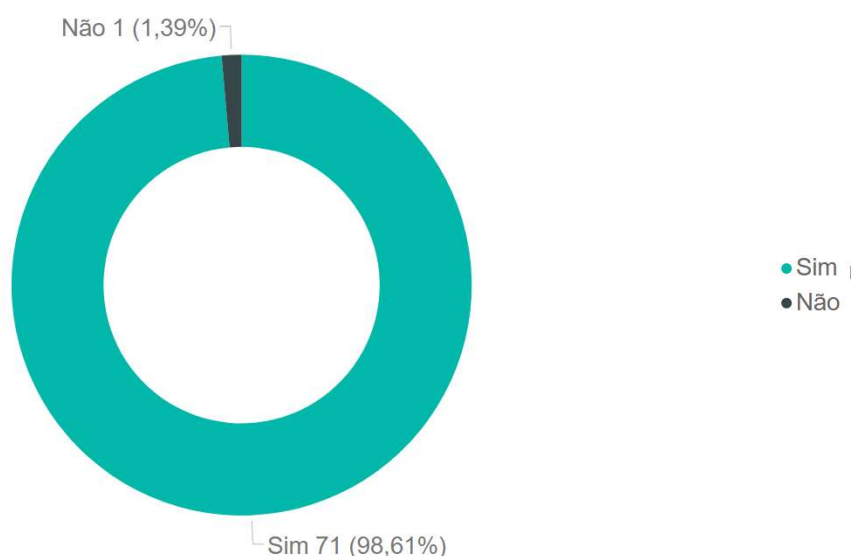
Por acesso, podemos entender que, embora o órgão não possua um alto volume de dados, ele tem permissão de coletar grande quantidade de informações, muitas vezes não estruturadas, demandando um aparato ferramental e ambiente necessário para o processamento que caracteriza o *Big Data*.

Até esta questão, em resumo, o cenário indicava que os respondentes tinham noção suficiente do que era *Big Data*, tinham um relativo conhecimento sobre as tecnologias relacionadas, seus órgãos tinham pessoas com algum nível de conhecimento nestas tecnologias, tinham acesso a um grande volume de dados e projetos de *Big Data* em seus ambientes profissionais. Deste primeiro conjunto de informações coletadas, relativas à metade das perguntas do questionário objetivou, em suma, aferir se as respostas seguintes do questionário (aquelas que se alinham com as perguntas principais e secundárias da presente dissertação), poderiam, de fato, contribuir para construir o entendimento foco do trabalho.

Verificou-se que sim, haviam ambientes de *Big Data* na Administração Pública Brasileira, segundo a amostra de respostas do questionário, que está contando com profissionais que tem as noções mínimas da tecnologia, e que estão, em sua maioria, com compromissos profissionais de longa data com suas respectivas repartições. Com este cenário, foi possível, então, na continuidade da análise das respostas do questionário averiguar aspectos do alinhamento da TI de *Big Data* descrita com os negócios das repartições públicas representadas, permitindo a aferição, ao menos dentro da amostra obtida, quanto aos questionamentos do trabalho em tela.

Na décima oitava questão (Gráfico 11), os entrevistados foram questionados se existiria algum problema de negócio que poderia ser resolvido com as tecnologias voltadas para *Big Data*.

Gráficos 11 - Existe algum problema de negócio que poderia ser resolvido com *Big Data*?

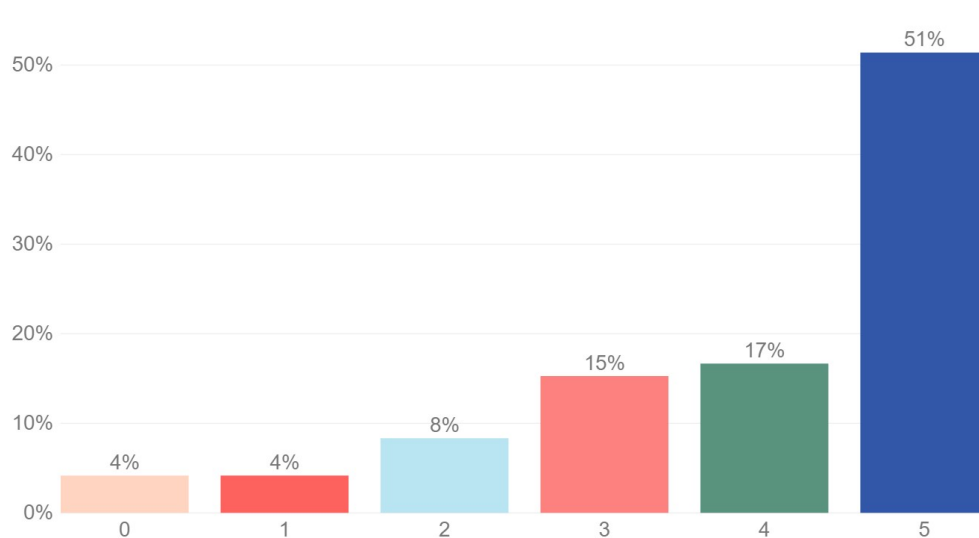


Fonte: próprio autor.

Noventa e oito por cento (98,61%) afirmaram que seus órgãos tinham problemas que poderiam ser resolvidos com *Big Data*. Os resultados mostram que os entrevistados tinham ciência sobre esses problemas, demonstrando uma janela de oportunidades para tais aplicações.

Na décima nona questão, buscou-se saber se havia prospecção de implantação de *Big Data* nos órgãos que ainda não tinham a aludida estrutura; o resultado está compilado no Gráfico 12.

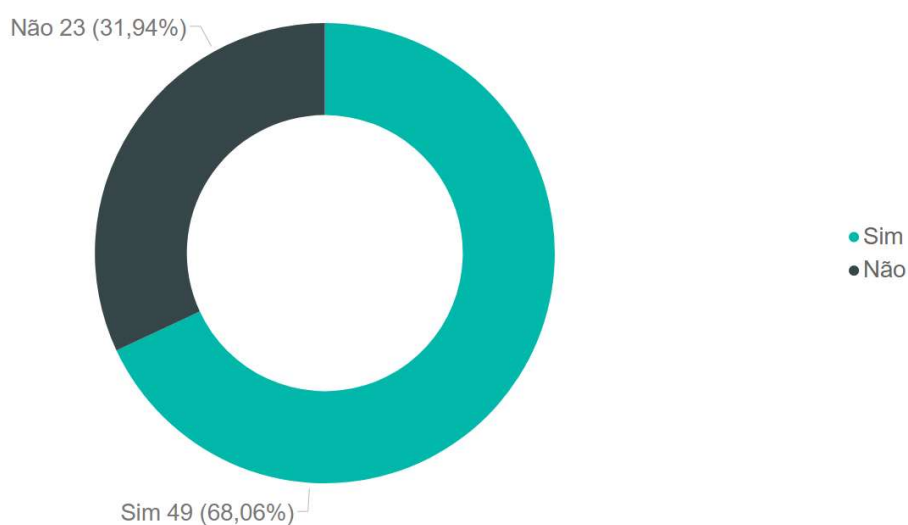
Gráficos 12 - Perspectiva para implantação de *Big Data* nos órgãos.



Fonte: próprio autor.

Esta pergunta demonstra que além de muitos órgãos já possuírem o *Big Data*, estão com novos projetos de utilização.

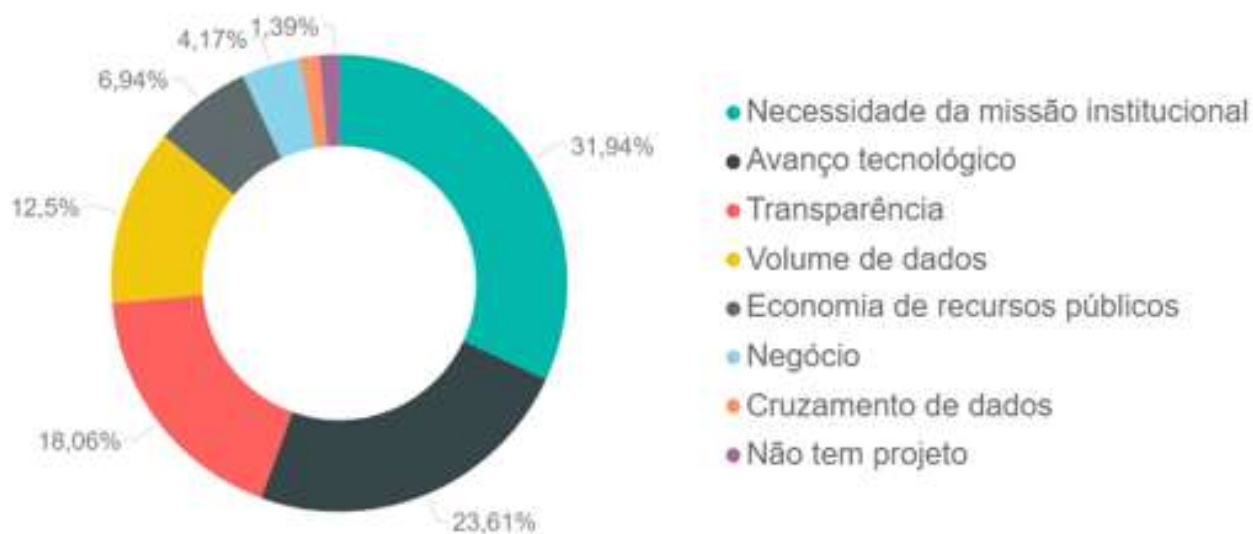
Isso ficou ainda mais claro quando analisamos a resposta da pergunta seguinte, acerca da existência de projetos de *Big Data* de uma maneira mais direta, sendo que houve resposta positiva de 68,06% dos entrevistados.

Gráficos 13 - O órgão já possui projetos de *Big Data*?

Fonte: próprio autor.

O que estaria a motivar estes projetos? Essa seria a pergunta mais adequada no sentido de apoiar os questionamentos principais da dissertação. A que oportunidades de negócio o *Big Data* estaria a ser uma resposta ou solução?

A pergunta 21 abordou esse assunto. Nas respostas, compiladas no Gráfico 14, podemos observar que a maior motivação era a missão institucional do órgão representado.

Gráficos 14 - O que tem motivado o uso de *Big Data* no seu órgão público?

Fonte: próprio autor.

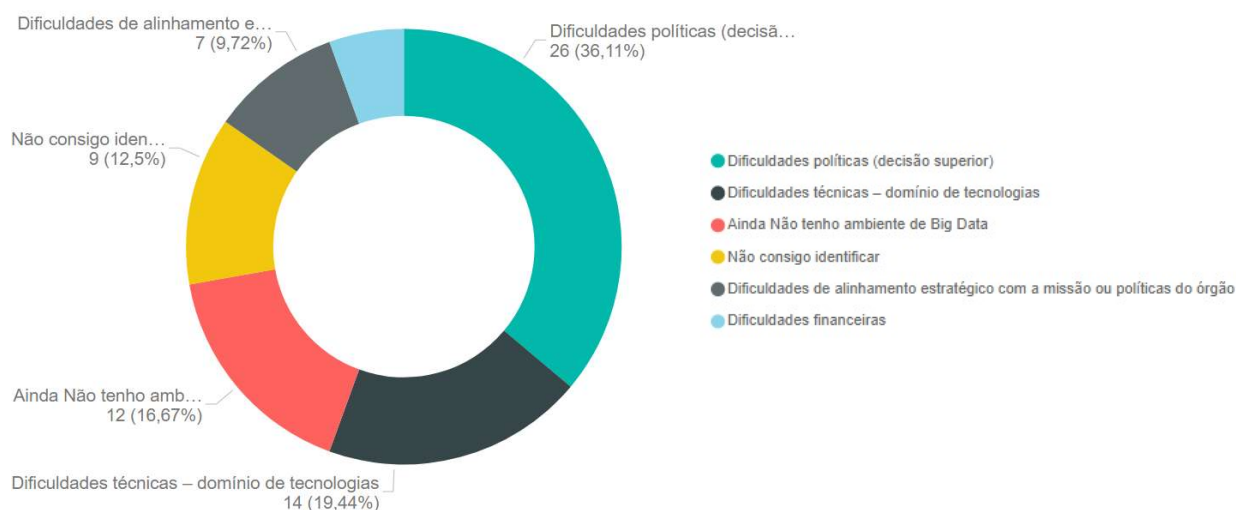
Isso demonstrou concordância com o fato descrito nos capítulos anteriores, ou seja, o *Big Data* precisa estar alinhado ao atendimento da missão institucional pra que seus projetos sejam utilizáveis. O maior motivador, a maior oportunidade de implantação do *Big Data* deve depender de forma direta desta característica.

Um grupo respondeu que a motivação tinha relação com avanços tecnológicos. O fato é preocupante, pois demonstra que há pessoas que não tem encontrado no *Big Data* uma solução de problemas públicos, mas o inverso, precisa de motivações no órgão para justificar investimentos em *Big Data*.

Outro grupo ainda relacionou os projetos com a eficiência de gastos públicos. No entanto, se tratou de um grupo representativo. Há ainda de se destacar que uma parcela considerável dos respondentes relaciona os projetos de *Big Data* com questões de transparência. Dessa forma, esse aspecto alinha as respostas com as expectativas iniciais da pesquisa em relato.

Quanto às dificuldades encontradas, objetos da pergunta vigésima segunda, apresenta-se o Gráfico 15.

Gráficos 15 - Dificuldades encontradas na implementação do ambiente de *Big Data*.



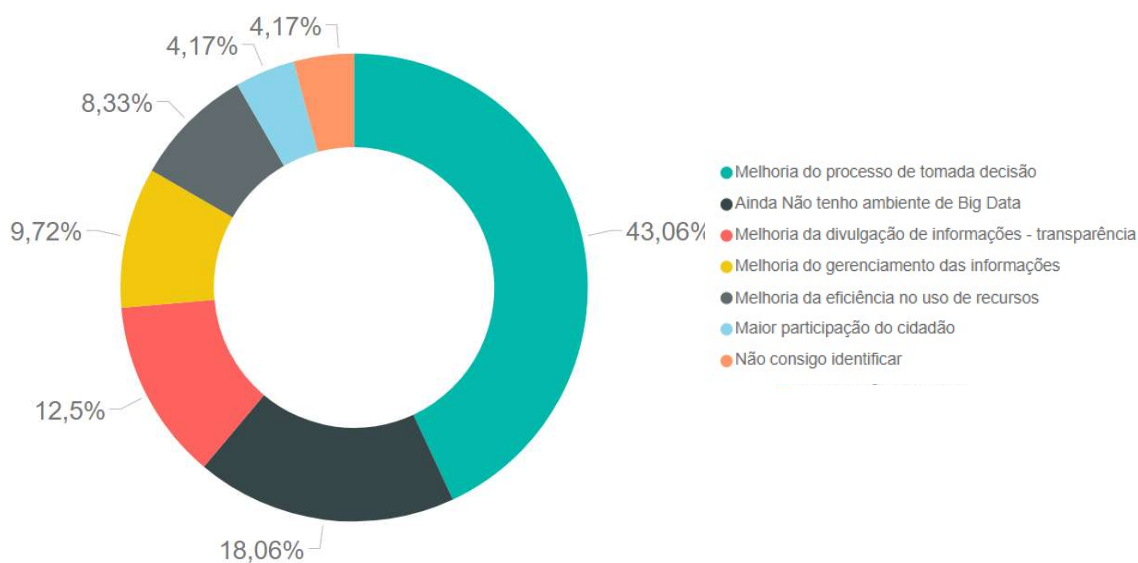
Fonte: próprio autor.

Os grupos de destaque elencados como barreiras na implantação de projetos de *Big Data* foram os das dificuldades políticas e técnicas. Há, portanto, um enorme desafio ligado a estas duas questões. As dificuldades políticas também podem estar relacionadas com questões burocráticas. Determinados projetos da Administração

Pública Brasileira sofrem com a descontinuidade causada pelas trocas políticas frequentes nos escalões superiores dos órgãos públicos. Estas respostas também se mostraram muito úteis quando da revisão das perguntas norteadoras da pesquisa.

A pergunta seguinte buscou aferir quais os principais resultados e impactos com a implantação do *Big Data*, acompanhada de uma pergunta mais relacionada com questões de eficiência no gasto de recursos públicos (Gráfico 16).

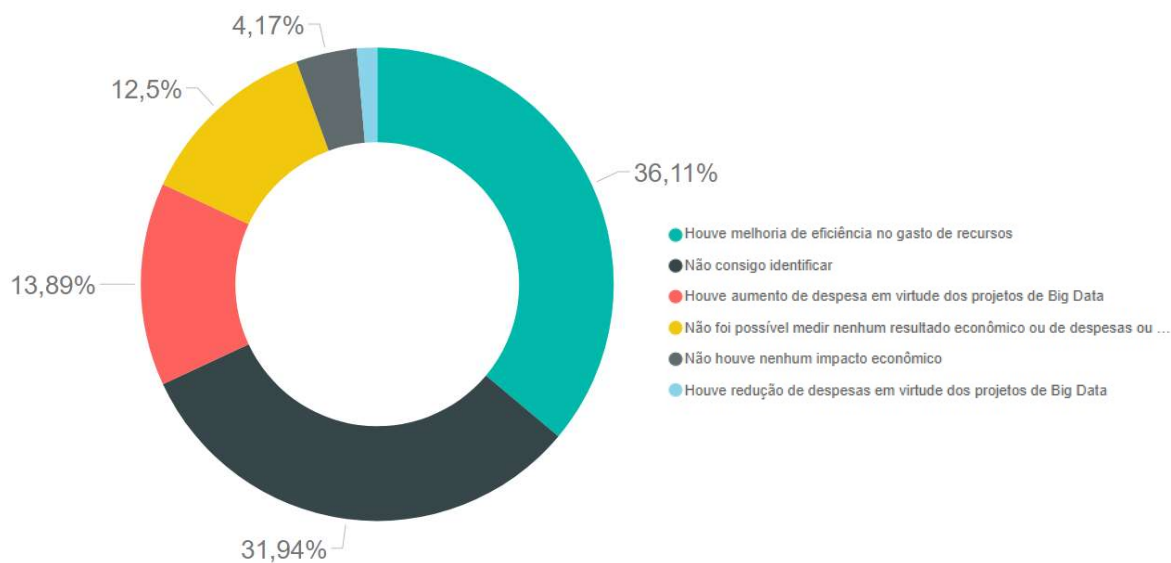
Gráficos 16 - Impactos com a implantação de *Big Data*.



Fonte: próprio autor.

As respostas para estas questões destacaram que o *Big Data* tem sido um promotor de melhoria de tomada de decisão e que tem contribuído significativamente quanto a questões de transparência.

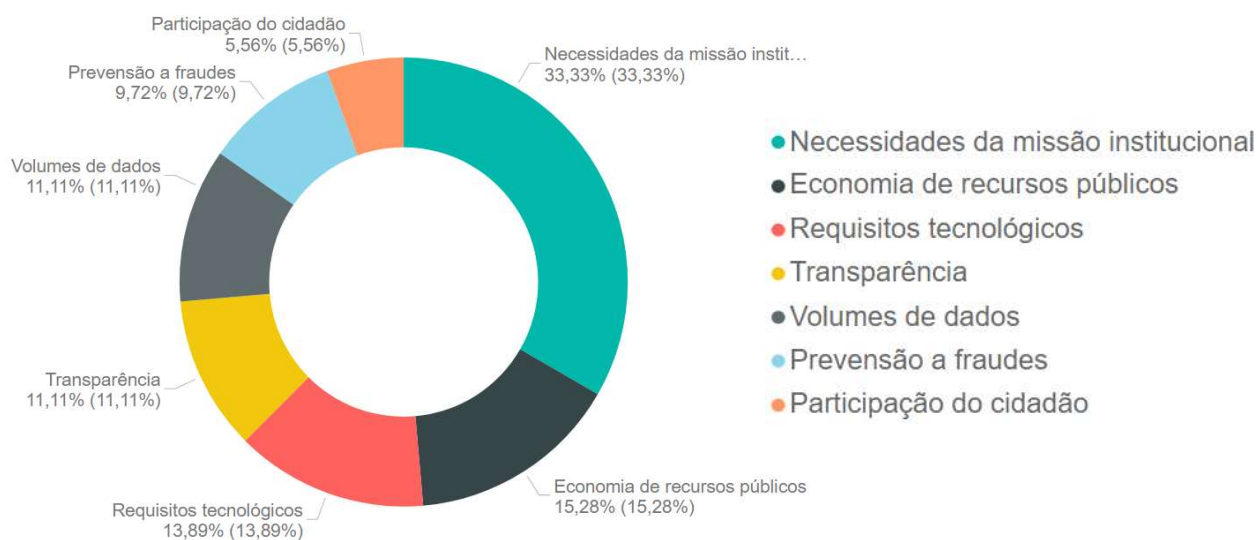
Parte dos respondentes informou que houve melhoria quanto à eficiência de gastos públicos, embora uma parte considerável não tenha conseguido identificar os impactos econômicos. Esperava-se que uma parte maior dos entrevistados conseguisse identificar melhorias de eficiência econômica.

Gráficos 17 - Os impactos econômicos da adoção de *Big Data* nos órgãos.

Fonte: próprio autor.

Um resultado considerado interessante foi o de que parte do grupo identificou um aumento de gastos. Não ficou claro, contudo, se este aumento estaria ligado apenas ao setor de TI.

A pergunta seguinte questionava os participantes acerca das oportunidades de utilização do *Big Data*, ou seja, não apenas quanto a projetos que já utilizam, mas quanto a motivadores de novos projetos acerca do tema no setor público (Gráfico 18).

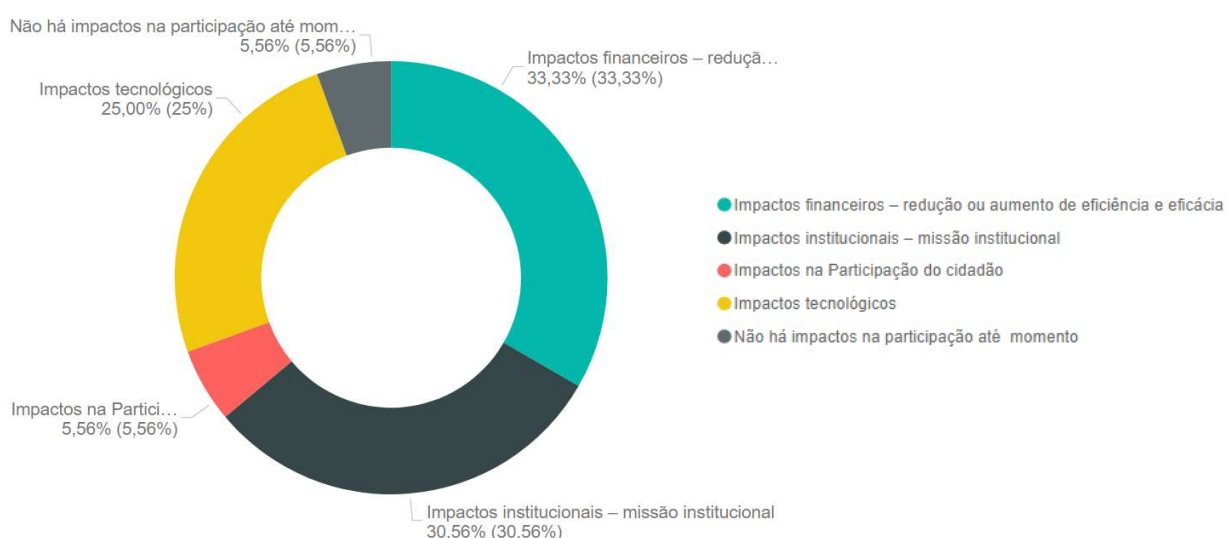
Gráficos 18 - As oportunidades de implementação do *Big Data* nos órgãos públicos.

Fonte: próprio autor.

As respostas indicaram os dois maiores motivadores, primeiro, as necessidades relacionadas com a missão institucional e, depois, a economia de recursos públicos.

A vigésima sexta questão buscou analisar quais os impactos que os projetos de *Big Data* podem trazer às instituições representadas, conforme demonstradas no Gráfico 19 abaixo:

Gráficos 19 - Os impactos que práticas de *Big Data* podem/poderiam trazer ao desenvolvimento de soluções nos órgãos públicos.

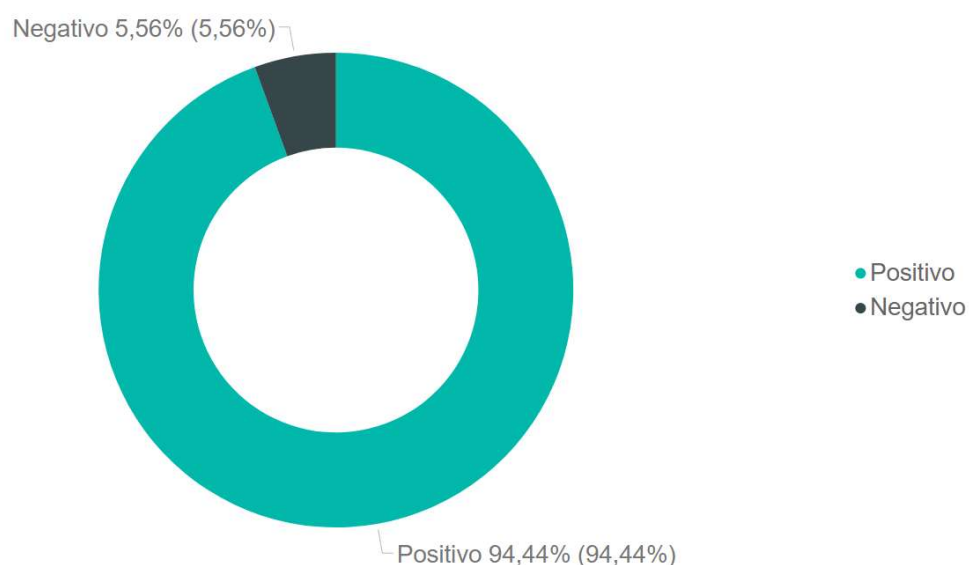


Fonte: próprio autor.

Três grupos de resposta se destacam: impactos financeiros (eficiência nos gastos públicos), impactos na missão institucional (o que também se traduz em eficiência e eficácia) e impactos tecnológicos (como a TI impacta e é impactada com os referidos projetos).

A mesma linha de questionamento segue na questão 27, representada pelo Gráfico 20 abaixo, onde verifica se os impactos são mais positivos ou negativos na visão dos respondentes.

Gráficos 20 - Os impactos citados na questão anterior são mais positivos que negativos nos órgãos públicos?

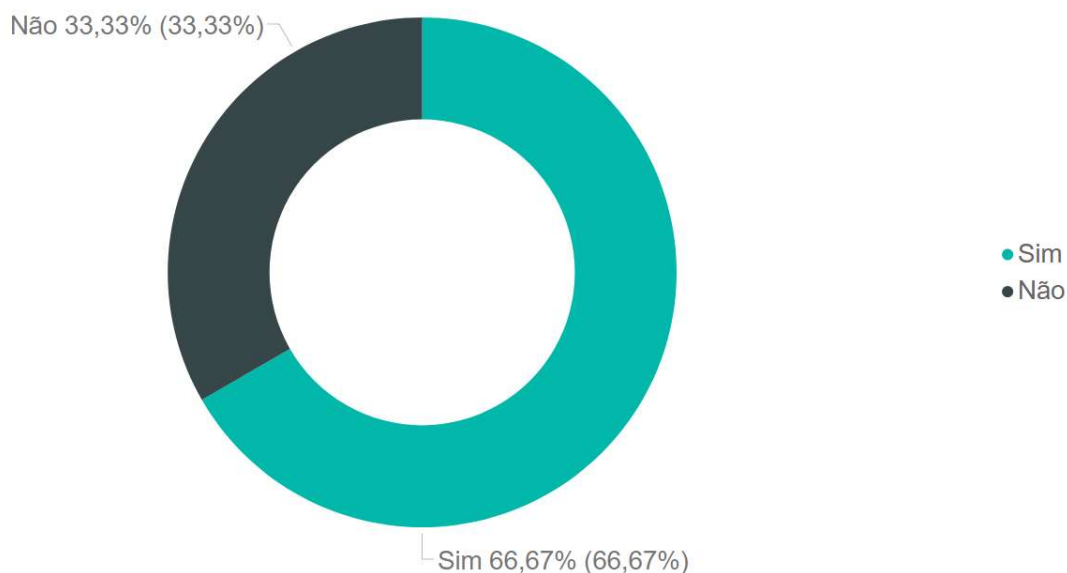


Fonte: próprio autor.

A questão inicial pode parecer muito óbvia. No entanto, como exemplo, se tomarmos que um grupo se referiu a impactos tecnológicos, estes poderiam ser negativos caso houvesse a necessidade de implantação de itens de tecnologia para os quais uma instituição não estivesse preparada. Igualmente, uma implantação com condução incorreta poderia afetar o cumprimento da missão institucional. No entanto, a maioria dos entrevistados (94,44%) identificou que os impactos seriam positivos.

A pergunta 28 também estava relacionada com estes impactos. Perguntou-se se os impactos estariam ligados a casos ou situações reais ou conhecidas pelos entrevistados (Gráfico 21).

Gráficos 21 - A sua resposta à questão anterior envolveu impactos existentes e comprovados (Sim) ou impactos percebidos pelo conhecimento acerca de *Big Data* conhecido pelo respondente do questionário (Não).



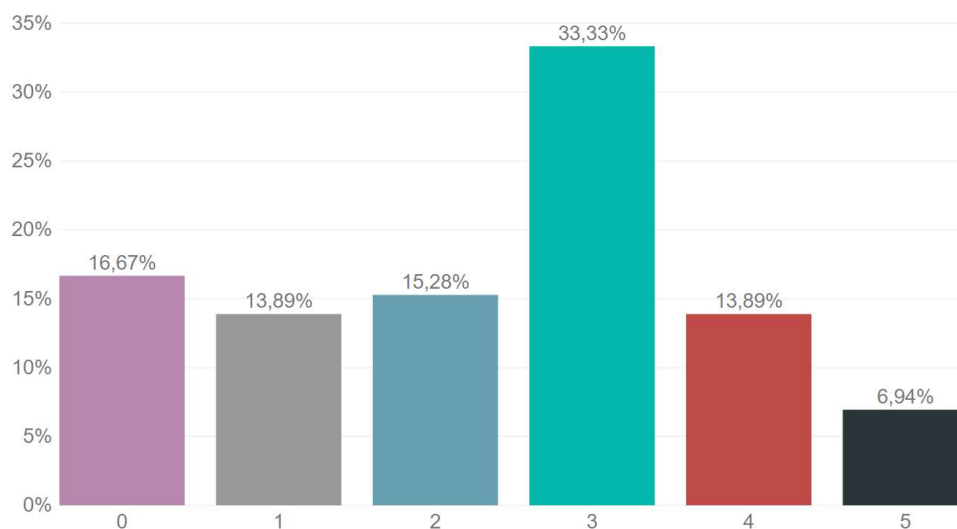
Fonte: próprio autor.

Sessenta e seis por cento (66,67%) relacionam que a sua medição de impactos que podem ocorrer com projetos de *Big Data* é positiva de acordo com o conhecimento e experiências anteriores.

Da vigésima nona à trigésima terceira questões se buscou conhecer um pouco o quadro de distribuição de conhecimento e de presença de profissionais de ciência de dados, bem como algumas informações de utilização de tecnologia de nuvem.

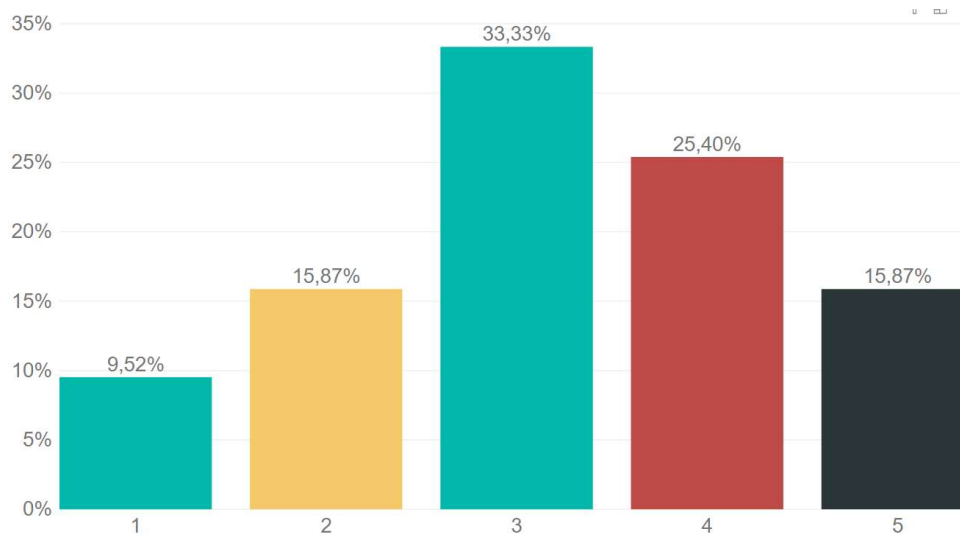
O compartilhamento de conhecimento se mostrou mediano e os resultados estão compilados nos Gráficos 22 e 23, relacionados entre os profissionais de TI e os outros setores das instituições.

Gráficos 22 - O órgão compartilha conhecimentos e aplicações de *Big Data* entre todos os setores do órgão?



Fonte: Próprio autor

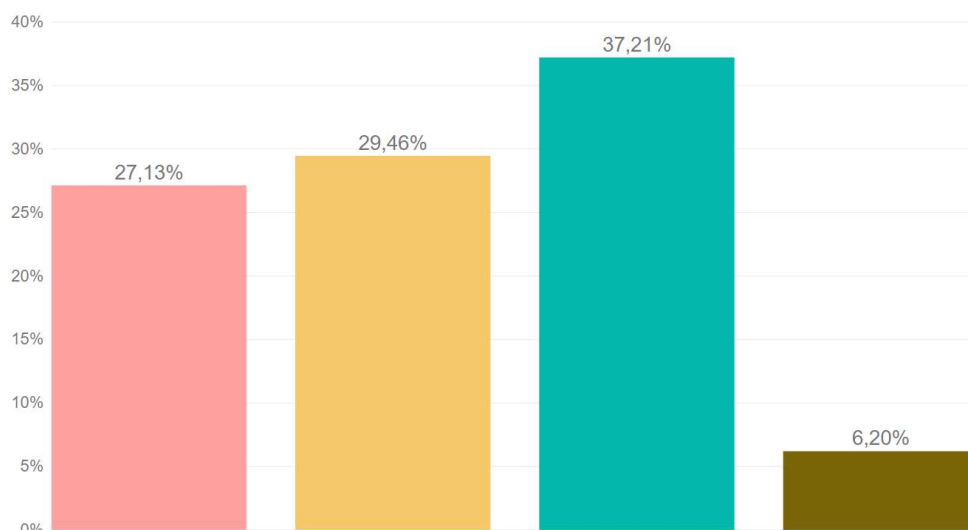
Gráficos 23 - A equipe de TI e profissionais de dados compartilham conhecimentos entre todos os setores da organização?



Fonte: próprio autor.

A questão 31 buscou verificar o uso de repositórios e serviços em nuvem para processamento dos dados (haveria o uso em que intensidade). A intensidade de uso destes repositórios foi considerada média, o que remete à ideia de que parte pode estar ocorrendo dentro das infraestruturas das instituições representadas. O resultado está compilado no Gráfico 24.

Gráficos 24 - No ecossistema de *Big Data* é explorado ou adotado serviços de nuvem para processar ou conduzir o *Analytics*?

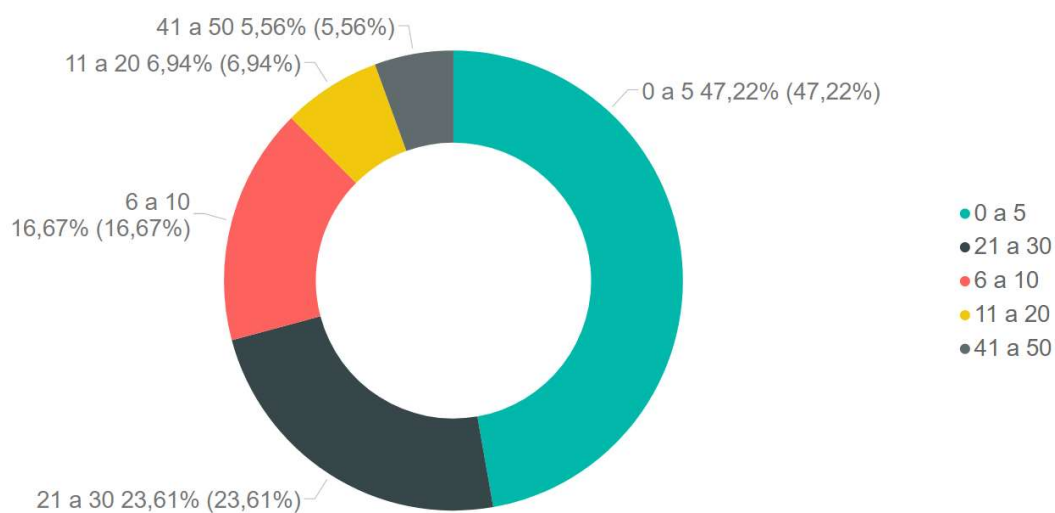


Fonte: próprio autor.

Por fim, as duas últimas perguntas buscaram analisar a quantidade de servidores cientistas de dados, tanto profissionais concursados (pergunta 32) quanto terceirizados (pergunta 33). Os resultados estão compilados nos Gráficos 25 e 26, logo abaixo.

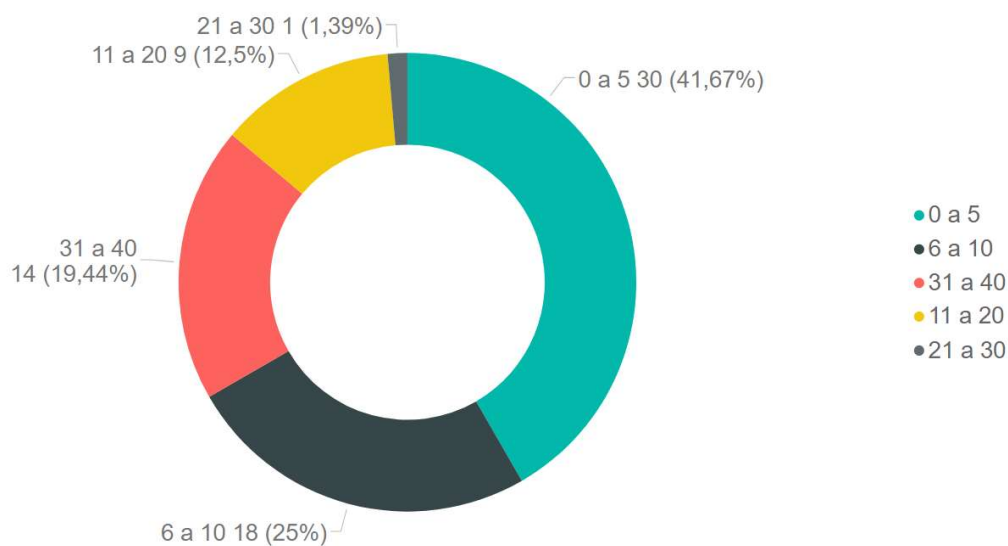
Apesar disso, os resultados mostram que as questões podem não ter ficado claras o suficiente para os participantes. Isso porque seria improvável, dado o tamanho do corpo de servidores em TI nos órgãos públicos, que algum órgão possa ter uma quantidade de servidores dedicados à ciência de dados maior que 30 ou 40 pessoas.

Gráficos 25 - Quantos profissionais cientistas de dados o órgão possui de servidores públicos?



Fonte: próprio autor.

Gráficos 26 - Quantos profissionais cientistas de dados o órgão possui de terceirizados?



Fonte: próprio autor.

Realizando uma análise geral acerca dos dados demográficos dos participantes, podemos constatar que uma parcela significativa dos respondentes é oriunda do setor público, conforme o cenário desejado pela pesquisa. Além disso, boa parte dos participantes trabalha no setor executivo em seus órgãos públicos e também possui experiência no tocante a aplicação da tecnologia da informação em suas tarefas.

Apesar de o questionário ter sido disponibilizado por meio do Google *Forms*, a fim de garantir uma abrangência nacional, boa parte dos entrevistados estão concentrados no Distrito Federal, mas também há registros de participação de outros estados como Acre, Piauí, Paraná, entre outros. Apesar de a tecnologia avançar em passos largos e muitas vezes os trabalhadores não conseguirem acompanhar essa evolução na mesma velocidade, o questionário denota um ponto interessante sobre a experiência dos entrevistados em relação às diretrizes e boas práticas relacionadas ao *Big Data*, ou seja, há uma aceitação e conhecimento sobre os termos empregados neste objeto de estudo.

É comum nos depararmos com estruturas precárias em órgãos públicos, ou até mesmo a falta de recursos para investimentos em tecnologia, capital humano, capacitação, entre outros. Porém, para este objeto de estudo, denota-se que boa parte dos órgãos públicos que foram representados pelos participantes da pesquisa possui ambiente propício para a realização e empregabilidade de diretrizes para *Big Data*. Além disso, há investimentos próprios para volumes de dados, que, de acordo com os resultados, tendem a continuar crescendo.

Ouro ponto que merece destaque, é a questão sobre a diferenciação entre termos distintos relacionados ao *Big Data*, que confundem até mesmo os profissionais ligados à área. Neste caso, boa parte dos entrevistados, ou seja, 38% dizem ter conhecimento sobre tais termos, sabem suas diferenças e real empregabilidade. Podemos notar, a partir disso, que estes profissionais possuem conhecimento suficiente para empregar as diretrizes e boas práticas relacionadas à temática.

Outros pontos explorados, como problemas relacionados a negócios que poderiam ser resolvidos com a aplicação de *Big Data* trouxeram achados importantes, mostrando que a maioria dos respondentes acredita que a prática diária das diretrizes de *Big Data* pode auxiliar significativamente nas decisões e nas tarefas diárias dos órgãos que participaram. Para os órgãos que não possuíam ambiente próprio para

direcionar requisitos para as aplicações de *Big Data*, vários deles possuem projetos que viabilizam a empregabilidade e implantação de *Big Data*, facilitando assim análises e trabalhos futuros.

Apesar das dificuldades encontradas na aplicação, conforme demonstrado nas análises anteriores, vários órgãos encontram-se motivados em realizar modificações estruturais para implantação de *Big Data*, além de demonstrarem interesse na capacitação de profissionais e alocação de recursos para o crescimento da volumetria de dados.

No geral, foi percebido um esforço dos órgãos públicos que ainda não possuem um ambiente próprio para *Big Data*. É válido citar que tais órgãos parecem comprometidos com projetos futuros, fortalecendo as diretrizes, aumentando o leque de possibilidades. Isso corrobora com o que foi explorado na revisão da literatura deste objeto de estudo, acerca da importância de *Big Data* para o setor público federal não só no Brasil, mas em uma escala global.

4.2. Análise da aplicação da pesquisa de grupo focal

Para análise e interpretação dos dados coletados, é importante destacar que foi utilizada a técnica de pesquisa qualitativa de Grupo Focal para obtê-los a partir de reuniões em grupo e compreender as técnicas e diretrizes empregadas na Governança de Dados e ao *Big Data*.

4.2.1 Perfil dos participantes

Os participantes foram selecionados a partir de sua experiência com iniciativas de *Big Data* e Governança de Dados e suas posições estratégicas nos Órgãos Públicos.

Compuseram o grupo focal três Engenheiros de Dados e quatro Cientistas de Dados. A média de experiência com Governança de Dados e *Big Data* foi de 4 anos. Ainda, quatro participantes com Mestrado e três com especialização.

4.2.2 Avaliação de Prontidão para *Big Data* – Grupo Focal

Devenport (2014) Avaliação de Prontidão para *Big Data*, Modelo DELTTA adaptado para a Governança de Dados e *Big Data* (Dados, Negócio do Órgão, Liderança, Metas, Tecnologia e Equipe de TI e Analistas de Dados).

Para uma melhor análise dos resultados, foi utilizada uma abordagem qualitativa para estabelecer o *Ranking* Médio (RM). Para realizar este cálculo, lançamos mão do método de análise da Escala Likert (MALHORTA, 2001; OLIVEIRA, 2005). A escala utilizada era composta por 5 pontos para mensurar o grau de concordância ou discordância das questões avaliadas.

As intensidades dos níveis da escala definida foram distribuídas da seguinte forma: (1) discordo totalmente; (2) discordo um pouco; (3) não concordo nem discordo; (4) concordo um pouco; e (5) concordo totalmente. Assim, os valores menores que 3 são considerados como discordantes e, maiores que 3, como concordantes, considerando uma escala de 5 pontos. O valor 3 foi considerado “indiferente” ou “sem opinião”.

Tabela 2 – Avaliação de Prontidão para *Big Data* – Grupo Focal.

QUESTÕES		FREQUÊNCIA DE SUJEITOS					
		1	2	3	4	5	RM
DADOS							
1	No órgão, temos acesso a grandes volumes de dados não estruturados ou os dados chegam em tempo real e em grande velocidade para análise.	2			2	3	3,6
2	No órgão, agrupamos dados de várias fontes internas em um <i>Data Warehouse</i> ou <i>Datamart</i> para facilitar o acesso.	2			4	1	3,3
3	No órgão, mantemos definições e padrões consistentes nos dados que usamos para análise.		5	1	1		2,4
4	No órgão, usuários, tomadores de decisão e desenvolvedores de soluções para áreas do órgão confiam na qualidade de nossos dados.		2	2	3		3,1
DADOS: Quais as percepções, desafios e oportunidades?							
Análise do Grupo Focal:							
5	É muito comum encontrar muitos dados nos órgãos, o problema não são os dados, o problema vai mais na linha de extrair valor dos dados, muito se fala que o dado é o novo petróleo. Não faz muito sentido ter uma volumetria muita alta e não saber o que extrair dos dados e produzir valor.						
NEGÓCIO DO ÓRGÃO							
6	Empregamos institucionalmente uma combinação de abordagem de <i>Big Data</i> e Governança de Dados para atingir as metas do órgão.		3	1	2	1	3,1
7	A nossa gestão assegura que as unidades de negócio e as áreas do órgão trabalhem para estabelecer as prioridades da organização relativas ao <i>Big Data</i> e Governança de Dados.	2	5				1,7
8	Capacitamos nossa equipe de TI e profissionais de análise de dados para fomentar a aprendizagem e compartilhar conhecimentos entre todos os setores da organização.	1	2	2	2		2,7
9	As nossas iniciativas e infraestrutura de <i>Big Data</i> e Governança de Dados recebem aportes e recursos adequados para desenvolver as competências e recursos dos quais precisamos.		1	2	4		3,4
10	Colaboramos com parceiros de áreas, com Órgãos para os quais prestamos suporte e outros integrantes do nosso	4	2			1	1,9

	ecossistema de negócios para compartilhar conteúdo e aplicações de <i>Big Data</i> .						
	NEGÓCIO DO ÓRGÃO: Quais as percepções, desafios e oportunidades? Análise do Grupo Focal: Existe um foco em capacitação de prestadores de serviços terceirizados para os órgãos públicos. E um direcionamento de projetos de treinamento e investimento de capacitação na implementação de <i>Big Data</i> de Governança de Dados. Muitos órgãos optam pela contratação de profissionais do mercado. Os servidores públicos e prestadores de serviços terceirizados de <i>Big Data</i> do Tribunal de Contas da União – TCU e Câmara dos Deputados, são incentivados a realizarem capacitações nas áreas.						
11							
LIDERANÇA							
12	Os dirigentes e líderes que estão na instituição há mais tempo ponderam regularmente sobre oportunidades que o <i>Big Data</i> e a Governança de Dados oferecem ao nosso negócio.	1	2	4			2,4
13	Os dirigentes e chefes encorajam as unidades e seus líderes a incorporar <i>Big Data</i> e Governança de Dados a seus processos decisórios e de negócios.		4	1	2		2,7
14	Os dirigentes e líderes que estão na instituição há mais tempo utilizam o <i>Big Data</i> e a Governança de Dados para orientar tanto suas decisões estratégicas quanto táticas.	1	4	2			2,1
15	Os gestores das áreas do Órgão utilizam o <i>Big Data</i> e o <i>Analytics</i> para orientar suas decisões.		4	3			2,4
16	O processo de priorização e alocação de nossos ativos de <i>Big Data</i> (dados, pessoal, <i>software</i> e <i>hardware</i>) é orientado e examinado pelo núcleo de altos dirigentes.	1	3	1	2		2,6
	LIDERANÇA: Quais as percepções, desafios e oportunidades? Análise do Grupo Focal:						
17	Os processos são frágeis, falta priorização e descontinuação de iniciativas após a troca da gestão e liderança, sem pesar em longo prazo. Iniciativas de <i>Big Data</i> realizadas pelas equipes técnicas e sem a participação da alta gestão e liderança.						
METAS							
18	Priorizamos nossas iniciativas de <i>Big Data</i> e Governança de Dados de acordo com as oportunidades que geram mais valor para nos destacarmos em termos de resultados institucionais.	1	3	2		1	2,6
19	Consideramos novos serviços e funcionalidades baseados em <i>Big Data</i> como um aspecto do nosso processo de inovação.	1		4	2		3,0
20	Identificamos oportunidades internas para <i>Big Data</i> e Governança de Dados avaliando nossos processos, estratégias e os ambientes onde a instituição atua.		2	3	2		3,0
21	Conduzimos regularmente experimentos baseados em dados para avaliar o que está – ou não – se revelando eficaz no nosso negócio.	2	2	3			2,1
22	Avaliamos as decisões que tomamos com apoio do <i>Analytics</i> e dos dados para verificar se novas fontes de dados não estruturados poderiam proporcionar modelos melhores.	5	1			1	1,7
	METAS: Quais as percepções, desafios e oportunidades? Análise do Grupo Focal:						
23	Os Órgãos Públicos não tem uma cultura de difundir iniciativas integradas de <i>Big Data</i> . Na prática, alguns órgãos já estão pensando em novos processos de inovação e no entendimento de <i>Data Driven</i> para extrair conhecimento e solucionar problemas. As oportunidades e iniciativas de <i>Big Data</i> existem, experimento e mais iniciativa das equipes técnicas do desenvolvimento do que metas efetivas das instituições, ainda precisam evoluir e promover um novo <i>mindset</i> .						
TECNOLOGIA							
24	Temos explorado ou adotamos abordagens de computação paralela (por exemplo: Hadoop, MPI, Cuda, OpenCL e técnicas de Dataflow) para o processamento de <i>Big Data</i> .		1	3	2	1	3,4

25	Temos um grau elevado de maturidade no uso de visualização de dados para esclarecer questões ou decisões de negócio.	2	1	3	1		2,4
26	Temos explorado ou adotado serviços em nuvem para processar dados ou conduzir a Governança de Dados.	2	4			1	2,1
27	Temos explorado ou adotado <i>software</i> de código aberto para o <i>Big Data</i> e o <i>Analytics</i> .		1	1	3	2	3,9
28	Temos explorado ou adotado ferramentas para processar dados não estruturados, como texto, vídeo ou imagens.		1	1	3	2	3,9
TECNOLOGIA: Quais as percepções, desafios e oportunidades?							
Análise do Grupo Focal:							
29	Os Órgãos estão utilizando soluções de tecnologias híbridas para atender necessidades específicas. E também ferramentas ideais para lidar com ativos de <i>Big Data</i> . Falta expertise para utilizar as ferramentas para resolver o problema.						
EQUIPE DE TI E ANALISTAS DE DADOS							
30	Temos um número suficiente de potenciais cientistas de dados e profissionais de <i>Analytics</i> competentes para atingir nossos objetivos analíticos.	7					1,0
31	Nossos profissionais de tecnologias e de <i>Analytics</i> atuam como consultores de confiança aos nossos gestores, ajudando-os em tomadas de decisão importantes e na inovação orientada a dados.		3	3	1		2,7
32	Nossos profissionais de tecnologias e de Governança de Dados conhecem as disciplinas e os processos de negócio aos quais o <i>Big Data</i> e <i>Analytics</i> estão sendo ou podem ser aplicados.		3	2	2		2,9
33	Nossos profissionais de tecnologias e profissionais de gerenciamento de dados trabalham com eficácia em equipe na condução de projetos de <i>Big Data</i> , <i>Analytics</i> e Governança de Dados.		3	3	1		2,7
34	Temos programas (internos ou em parceria com organizações externas) para desenvolver ciência de dados e habilidades analíticas em nossos profissionais.	1		2	3	1	3,4
EQUIPE DE TI E ANALISTAS DE DADOS: Quais as percepções, desafios e oportunidade?							
Análise do Grupo Focal:							
35	Carência de profissionais qualificados ainda é uma das maiores barreiras para a implantação de <i>Big Data</i> na gestão de TI entre os Órgãos Públicos. Embora esses desafios apresentem impasses, está mais claro que o investimento de <i>Big Data</i> e Gestão de Dados proporcionará maior eficiência e benefícios no serviço Público Federal.						
Ranking Médio (RM)							

4.2.3 Entendendo o cenário atual

A pesquisa supracitada busca melhorar a compreensão das dificuldades de praticar governança de dados e *Big Data*, bem como as carências, processos de melhorias citadas nos tópicos subjacentes.

Pode-se dizer que o *Big Data* está redefinindo o panorama do gerenciamento de dados, de processos de extração, transformação e carregamento de tecnologias (como *Hadoop*) para limpar e organizar dados não estruturados em aplicativos de *Big Data*.

Embora o setor de negócios seja líder no desenvolvimento de aplicativos de *Big Data*, o setor público começou a obter *insights* para apoiar a tomada de decisões

em tempo real a partir de dados em movimento de rápido crescimento de várias fontes, incluindo a *Web*, sensores biológicos e industriais, vídeo, e-mail e comunicações sociais.

Percebe-se que a prática de governança de dados aliada ao *Big Data* em governos pode ajudá-los a servir seus cidadãos e superar os desafios nacionais (como controle de gastos em áreas como a saúde, criação de empregos, desastres naturais e terrorismo). Contudo, também há algum ceticismo quanto à possibilidade de realmente melhorar as operações do governo, visto que estes devem desenvolver iniciativas e adotar tecnologias (como *Hadoop* e *NoSQL*) para transformar dados em informação útil e organizada.

Aqui, pergunta-se acerca da capacidade dos governos de implementar alguns dos aplicativos de *Big Data* associados ao setor de carência, como os supracitados acima. Primeiro, se compartilham os setores em termos de objetivos, missões, processos de tomada de decisão, atores de decisão, estrutura organizacional e estratégias, depois se voltam para várias aplicações atuais em tecnologias avançadas.

Desta forma, salienta-se sobre a real necessidade de se aplicar Governança de dados aliadas a boas práticas de *Big Data*, não só visando um conjunto de dados estruturados, válidos, com riqueza de informações governamentais de interesse público e que podem (e devem) ajudar a sanar vários dos problemas que o Brasil vem enfrentando. É válido dizer ainda que:

- empresas, governos e a comunidade de pesquisa podem derivar valor da enorme quantidade de dados digitais que coletam;
- os governos dos principais países de Tecnologia da Informação e Comunicações (TIC) iniciaram projetos de aplicação de *Big Data* para aumentar a eficiência operacional, transparência, bem-estar dos cidadãos, envolvimento em assuntos públicos, crescimento econômico e segurança nacional; e
- a análise de projetos de aplicativos de *Big Data* por governos oferece orientação aos países seguidores para suas próprias iniciativas de *Big Data* futuras nesta área.

Essas observações são interessantes justamente porque expressam direitos e deveres enquanto órgãos, governos, comunidade, entre outros. A riqueza de informações que são coletadas hoje através dos grandes centros de redes distribuídas, principalmente na internet, ajudam na compreensão de vários fatores socioeconômicos que, se forem bem tratados, extraídos e compreendidos, podem poupar vários recursos, abrindo espaço para novos investimentos direcionados justamente a suprir a carência de cada setor.

4.2.4 Guideline para implantação de Governança de Dados e *Big Data* em Instituições Públicas Federais

Segundo Barata (2015), a qualidade dos dados usados em negócios atualmente, é importante para geração de históricos, práticas, vantagens e desvantagens, entre clientes e organizações. Consequentemente, para que as organizações forneçam bons resultados de negócios, seus dados devem ser precisos e o uso desses dados deve ser controlado por meio de políticas e monitoramento (BARATA, 2015).

Desta forma, diretrizes para implantação e prática de governança de dados aliadas ao Big Bata evitam erros e garantem uma gestão de qualidade dos dados. A etapa mais crítica desde processo, contudo, é estabelecer a estrutura dessa governança, uma vez que ela é a base para o sucesso dessas iniciativas não apenas para as organizações comerciais e industriais, mas também governamentais.

As equipes de governança de dados têm o dever de garantir que os dados sejam tratados de maneira uniforme e eficaz, desde modelos comuns para a estrutura de governança de dados à aplicação de políticas. Desta forma, explora-se como estabelecer uma estrutura de governança de dados, baseando-se em um *guideline*, desmistificando os processos e etapas essenciais (BARC 2019).

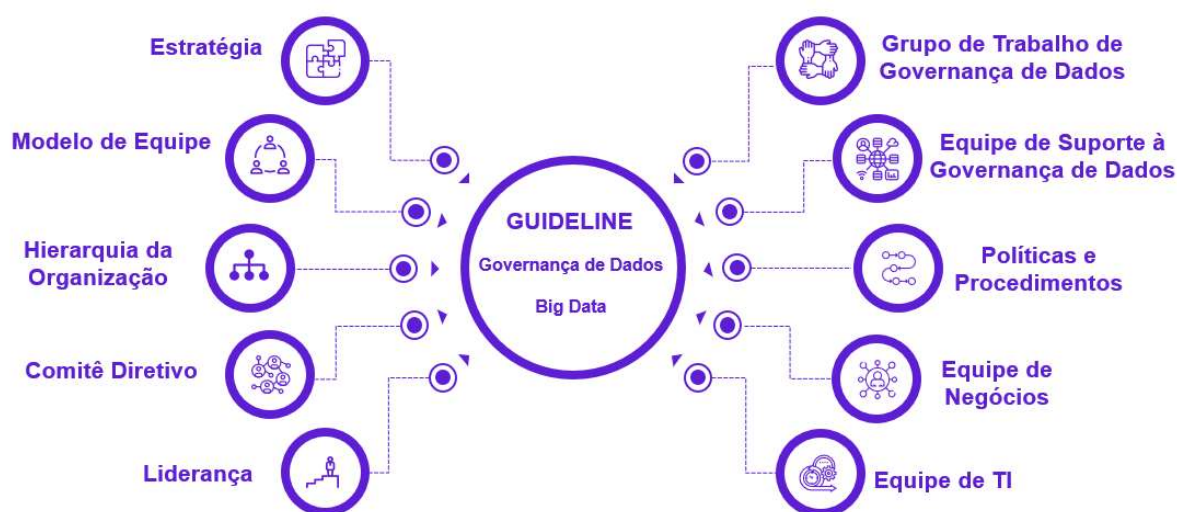
Neste guia, estão incluídos os principais papéis e funções necessários para melhorar de forma mais eficaz a qualidade dos dados de uma organização. Seguindo as etapas descritas, espera-se que líderes de negócios e de TI possam melhorar a qualidade dos dados, alcançar as metas de negócios e, por fim, reduzir despesas associadas a dados ruins.

A governança de dados envolve a tomada de decisões, gerenciamento e prestação de contas relacionadas aos dados em uma organização. Frequentemente,

uma equipe de governança de dados é construída para garantir que os dados sejam tratados de maneira uniforme e eficaz, sendo responsável direta pela qualidade dos dados. Os programas de governança de dados são projetados para preparar regras e regulamentos para uma organização e para lidar com quaisquer problemas que possam surgir em relação aos dados. Eles também garantem a conformidade com as políticas. Eles informam a uma empresa quem é o proprietário dos dados e quem pode executar certas funções com eles (BARATA, 2015).

Na Figura 12 ajuda na compreensão das etapas que constitui a *Guideline* para implantação de Governança de Dados e *Big Data* em Instituições Públicas Federais.

Figura 12 – Etapas da *Guideline* para implantação de Governança de Dados e *Big Data*



Fonte: Elaborado pelo autor.

Baseando-se nas concepções de Barc (2019), propõe-se as etapas (diretrizes) descritas a seguir, como processos de implantação e prática de governança de dados e *Big Data* para organizações de uma forma geral, onde:

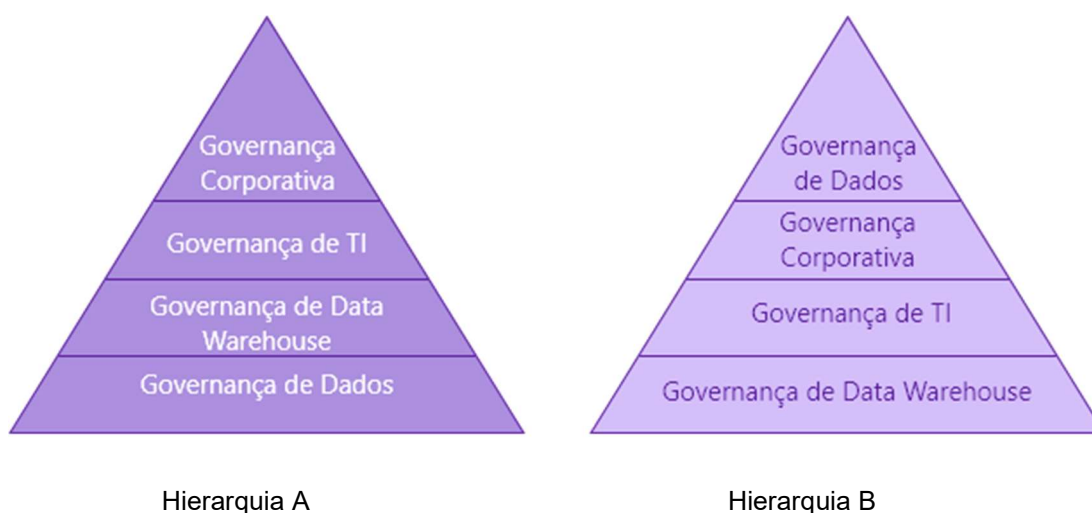
1ª etapa – determinar a estratégia: essa estratégia pode ser iniciada escrevendo um estatuto de governança de dados com a ajuda das partes interessadas e dos envolvidos diretamente no projeto. O regulamento incluirá uma declaração de missão e uma explicação dos objetivos gerais do programa. Os executivos do alto escalão provavelmente terão que assinar o estatuto antes que a implementação possa começar oficialmente. O modelo escolhido para a estrutura de governança de dados deve estar incluído neste regulamento e explicar quem tem autoridade e em que nível

com relação aos dados. Este documento também deve deixar explícito a amplitude do alcance dos padrões de governança de dados.

2ª etapa – escolher um modelo para uma equipe de governança de dados: há muitos modelos de governança de dados disponíveis, mas nem todos serão adequados para todas as organizações. Um modelo comum que adota uma abordagem em três camadas inclui um grupo de executivos de nível sênior, geralmente chamado de Comitê Diretor, que toma decisões de alto nível; um grupo de gerenciamento intermediário, o *Data Governance Office*, que oferece orientação; e outro grupo chamado Grupo de Trabalho de Governança de Dados, que responsável por grande parte do esforço administrativo. O Grupo de Trabalho contém especialistas no assunto em negócios e Tecnologia da Informação (TI).

3ª etapa – escolher a hierarquia certa para a organização: a governança corporativa é comumente configurada conforme mostrado nos diagramas da Figura 12. A "Hierarquia A" tem políticas começando no ponto de governança corporativa da pirâmide que então passam pelos canais abaixo dela, primeiro para governança de TI, depois para governança de *Data Warehouse* e, finalmente, para governança de dados e tudo e todos relacionados à governança de dados.

Figura 13 – Modelo para hierarquia A e B



Fonte: Elaborado pelo autor.

Na “Hierarquia B”, a governança de dados orienta todas as decisões. Nesse tipo de *framework*, a governança dos dados recebe tratamento preferencial. A estrutura dessa governança cobre regras e políticas de negócios, a qualidade, integridade e segurança dos dados, além da conformidade com regras, regulamentos,

auditorias, controles e muito mais. Aqueles que estão selecionando a melhor hierarquia para sua organização devem decidir se desejam se concentrar em uma abordagem de governança corporativa ou governança de dados.

Na função de governança de TI em ambas as hierarquias, um grupo de governança de TI garante que as estratégias e políticas de dados de toda a empresa sejam executadas em seus sistemas e bancos de dados. É esta equipe a responsável por decidir quem tomará decisões em vários domínios, como infraestrutura, arquitetura e priorização.

A governança do *Data Warehouse* determina o tipo de modelagem que melhor garantirá o uso eficaz desde recurso e monitora a conformidade com todas as políticas relacionadas. *Data Warehouses* são investimentos de longo prazo que precisam ser monitorados e financiados continuamente ao longo de sua vida.

4ª etapa – selecionar o Comitê Diretivo: o Comitê Diretivo é composto por executivos de alto nível e partes interessadas que, normalmente, são os vice-presidentes sêniores e das linhas de negócios cujos dados estão sendo controlados, como Marketing, Finanças, Compras, etc. O chefe do Comitê Diretivo deve ser um executivo de nível sênior que se reporta ao CEO, que é a pessoa que tem autoridade para aprovar orçamentos de projetos, bem como colocá-los em uma lista de prioridades. Vale ressaltar que os executivos do Comitê Diretivo são responsáveis por suas respectivas linhas de negócios.

O Comitê Diretivo também inclui patrocinadores executivos em negócios e TI, bem como o presidente do Escritório de Governança de Dados. Um patrocinador executivo é alguém do nível executivo sênior da empresa que será capaz de promover a aceitação da governança de dados em toda a organização. Os patrocinadores devem ser líderes comprovados que conhecem bem seus negócios e são capazes de lidar com várias funções relacionadas à governança.

O Comitê Diretivo dirige o processo de governança de dados em toda a organização e garante que as políticas e procedimentos desenvolvidos para os dados e para a qualidade destes, sejam seguidos. Também apóia e autoriza empreendimentos, patrocina e financia programas de governança e os prioriza. O Comitê também aprova cartas e estratégias para projetos de dados, aprova e faz alterações em sugestões, aprova financiamento e recomenda projetos. Além disso, o

Comitê orienta sobre a estratégia de negócios e trata da resolução de conflitos. Esse grupo deve estar disposto a participar de todas as iniciativas relacionadas à governança de dados na organização e relatar as informações ao *Data Governance Office* e ao *Data Governance Working Group*.

5ª etapa – configurar o *Data Governance Office*: o *Data Governance Office*, também chamado de *Master Data Council*, aplica a governança de dados. As funções incluem o Líder de Governança de Dados (DGL), representantes de TI e coordenador. O DGL geralmente é subordinado direto do vice-presidente e preenche a função como um compromisso de trabalho em tempo integral. Eles trabalham com todas as unidades de negócios e TI para que a governança de dados e as estratégias sejam comuns em toda a organização. O DGL também impõe políticas, propõe em quais projetos de governança de dados investir, coordena os grupos de negócios e tecnologia, estabelece métricas de sucesso, monitora e relata qualidade de dados e métricas de governança e trabalha com líderes de negócios e recursos de TI para priorizar e resolver problemas. O DGL deve ser um líder, ter perspicácia política e saber quem são os principais influenciadores. Eles também devem ser qualificados em governança de dados e práticas do setor. O DGL deve ser alguém que está na organização há vários anos e que já conhece os objetivos da empresa e o panorama político.

6ª etapa – escolher o grupo de trabalho de governança de dados: os membros do Grupo de Trabalho de Governança de Dados ocupam cargos de gerência ou superiores, responsáveis pela ligação entre as áreas de negócios e TI. Eles impulsionam o gerenciamento e a qualidade dos dados para áreas específicas e têm experiência no assunto para negócios e questões de TI. O Grupo de Trabalho pode recomendar projetos com base no uso de dados.

Os membros devem ser líderes e capazes de tomar decisões em equipe. As funções nesse grupo podem incluir líder de qualidade de dados, administrador de dados, líder de metadados e arquiteto de dados. Os profissionais de TI do Grupo de Trabalho precisam ter experiência tecnológica e conhecimento sobre modelagem, análise e migração de dados, bem como redação técnica e apresentação.

7ª etapa – selecionar a equipe de suporte à governança de dados: funções mais específicas devem ser selecionadas para suportar a estrutura de governança de dados. Isso inclui o seguinte:

- Proprietário dos dados: cria políticas e é o proprietário da qualidade dos dados para domínios de dados mestres; o proprietário dos dados é a pessoa que é o contato para quaisquer problemas com os dados.
- *Data Steward*: cria políticas, as coloca em prática e as aplica, bem como corrige problemas de qualidade de dados. Existe um *Data Steward* para cada unidade de negócios. Eles trabalham dentro de suas próprias organizações e conduzem treinamento para usuários sobre políticas e procedimentos. Uma das principais funções do *Data Steward* é garantir que os mesmos dados não existam em vários departamentos, criando duplicações desnecessárias. O *Data Steward* também coordena a coleta de dados e gerencia sua acessibilidade.
- Arquiteto de dados: projeta, estrutura, organiza e mantém dados. O arquiteto de dados mapeia a estrutura e a organização dos dados relevantes para funções de negócios. Essa função costuma ser uma espécie de "intermediário" que entende as necessidades de dados dos grupos de negócios e traduz essas necessidades em estruturas de dados funcionais dentro das políticas de dados da organização.
- Modelador de dados: obtém e documenta regras de negócios para qualidade de dados;
- Analista de dados: localiza e pesquisa problemas para os proprietários dos dados e analisa a qualidade dos dados minuciosamente, por registro e caso.

8ª etapa – desenvolver e aplicar políticas e procedimentos: a maneira mais eficaz de melhorar a qualidade dos dados e evitar os custos associados a dados ruins é instituir melhores práticas de gerenciamento de dados. A definição de uma equipe de qualidade de dados vinculada ao patrocínio de nível executivo pode resultar na criação de políticas, procedimentos e padrões aplicáveis que irão aprimorar a qualidade dos dados em toda a empresa.

Desta forma, o grupo de trabalho de Governança de Dados cria políticas com base nas metas do programa no regulamento. Essas políticas podem ser associadas

aos dados de diferentes maneiras, como regras de nomenclatura, resolução de conflitos, segurança, entre outras. As políticas devem ser aprovadas pelo Comitê Diretivo e devem ser aplicadas.

Garantir que todos na organização estejam aderindo às políticas é importante, especialmente no que diz respeito à qualidade dos dados. O não cumprimento das políticas e procedimentos pode levar aos mesmos erros de qualidade de dados que as políticas foram projetadas para ajudar a prevenir.

As políticas podem ser aplicadas tornando-as compreensíveis e documentadas para que todos tenham fácil acesso a elas. Uma política que não está prontamente disponível raramente será lida. Elas devem conter métricas de valor tangível para mostrar como a adesão às políticas pode ter um impacto significativo nos negócios. As políticas permanecem um documento vivo e são atualizadas conforme as estratégias e metas da empresa mudam.

9ª etapa – estabelecer uma equipe de negócios qualificada: a governança de dados precisa de especialistas tanto na área de negócios quanto da área de TI. Os especialistas em negócios comunicam a outras pessoas quais são as necessidades do negócio em relação aos dados. Eles devem saber como os dados são coletados e como serão usados pelos tomadores de decisão, como os CEOs. A equipe de negócios – não a equipe de TI – lida com a propriedade dos dados e seus membros são os especialistas no assunto. Eles sabem o que a empresa está buscando e como os dados podem ajudar a cumprir essas metas. Eles também precisam ter conhecimento técnico suficiente para explicar as necessidades de TI para a respectiva equipe de forma eficaz.

Os executivos devem ter autoridade da alta administração para garantir o cumprimento dos padrões e políticas relacionadas à qualidade dos dados. Eles também devem compreender como os dados ruins entram em um sistema e desenvolver maneiras de corrigir o problema. Eles têm autoridade para fazer alterações em dados de baixa qualidade nos níveis corporativo e departamental e podem auditar e medir o sucesso das soluções de qualidade de dados.

10ª etapa – estabelecer uma equipe de TI qualificada: os membros da equipe de TI precisam ser altamente qualificados em tecnologia. Eles fornecem ferramentas para que os integrantes da equipe de negócios trabalhem com os dados. Eles suportam os

aplicativos que são usados pelos dados da empresa e devem ser capazes de entender as políticas de negócios e incorporá-las à TI. Eles devem manter o número de ferramentas no mínimo; quanto mais simples, melhor.

Os membros da equipe de TI devem ser capazes de codificar para movimentação e armazenamento de dados e ter conhecimento sobre modelos de dados e bancos de dados. Eles precisam entender como traduzir modelos de negócios em modelos de dados e determinar regras para limpar os dados se houver erros predominantes. Os especialistas em TI devem trabalhar com departamentos em toda a organização e entender as necessidades de dados de cada um, bem como os aplicativos requeridos.

Enquanto o lado comercial lida com o desenvolvimento de políticas, a TI lidará com os recursos reais de conformidade e segurança de dados. A TI também pode fornecer métricas para ajudar os especialistas de negócios a avaliar a qualidade dos dados e a aderência às políticas e determinar se alguma alteração precisa ser feita.

5. CONSIDERAÇÕES FINAIS

A Governança de Dados é fundamental para lidar com eles de maneira eficaz e inculir qualidade de dados em toda a organização, independente da sua área de atuação. Escolher o melhor modelo para uma organização, selecionar os membros certos da equipe e garantir a conformidade contínua com as políticas e procedimentos são a base para a construção de um programa eficaz e bem-sucedido que aprimorará os dados de uma empresa nos próximos anos.

Esta dissertação buscou fornecer uma visão geral de um plano de implantação e prática de Governança de Dados aliada ao *Big Data*. Por intermédio de um *guideline* foi possível descrever etapas que darão suporte à criação e manutenção de uma Governança de Dados eficaz.

Primeiramente a organização deve determinar sua estratégia de uso e governança de dados. Em seguida, um modelo e uma hierarquia para governar o uso de dados devem ser escolhidos. O terceiro passo é a implementação de toda uma estrutura organizacional de governança de dados.

Um modelo comum que adota uma abordagem em três camadas inclui um grupo de executivos de nível sênior, geralmente chamado de Comitê Diretor, que toma decisões de alto nível; um grupo de gerenciamento intermediário, o *Data Governance Office*, que oferece orientação; e outro grupo denominado Grupo de Trabalho de Governança de Dados, que realiza grande parte do trabalho administrativo.

Finalmente, as políticas devem ser criadas, documentadas, comunicadas e aplicadas. Este planejamento contribui para uma estrutura de governança de dados eficaz e fornecerá às organizações uma estrutura sobre a qual construir sua abordagem de governança. Cada organização, contudo, deve lidar com as características e restrições do seu próprio contexto, mas espera-se que as etapas apresentadas nesse trabalho sirvam de guia para uma base sólida na direção de uma governança de dados efetiva.

Organizações em todo o mundo estão investindo em sistemas capazes de armazenar e processar dados de maneiras antes inimagináveis. Em alguns casos, as empresas estão até mesmo reformulando seus ambientes de TI existentes com base nesses novos sistemas. Esses sistemas de *Big Data* produziram resultados tangíveis:

aumento das receitas e redução de custos. No entanto, os resultados positivos estão longe de serem garantidos. Para realmente obter valor dos dados de alguém, essas novas plataformas devem ser controladas.

O termo governança de dados amedronta muitos profissionais de dados. Como muitas vezes é vagamente definido e mal compreendido, muitos simplesmente optam por uma abordagem exclusivamente tecnológica para resolver suas necessidades de governança. A complexidade que acompanha muitos sistemas de *Big Data* torna essa abordagem baseada em tecnologia especialmente atraente, embora seja bem conhecido que a tecnologia sozinha raramente será suficiente. O que talvez seja menos conhecido é que as próprias tecnologias devem ser revisitadas ao se otimizar para governança de dados hoje.

Felizmente, é possível resolver a linhagem de dados usando uma abordagem mais prescritiva e em sistemas que podem ser dimensionados proporcionalmente às demandas de *Big Data*. Em particular, uma arquitetura baseada em fluxos permite que as organizações "publiquem" dados (ou informações sobre os dados) que são ingeridos e transformados dentro do *cluster*. Os consumidores podem então "assinar" esses dados e povoar os sistemas *Downstream* da maneira que for considerada necessária.

Esta dissertação buscou, por meio de uma revisão da literatura, abordar conceitos que ajudassem na compreensão da temática explorada. Entender sobre diretrizes de *Big Data* e Governança de dados, de maneira geral, é uma tarefa complexa e exige uma série de análises e etapas que precisam ser seguidas rigorosamente para que possam dar bons frutos. A prática de Governança aliada ao *Big Data*, não é trivial para qualquer organização e, no caso explorado pelo presente trabalho, para órgãos públicos federais ainda são necessários investimentos direcionados e compreensão das diretrizes para se chegar ao fator comum de sucesso.

5.1. Contribuições e Trabalhos Futuros

Esta dissertação teve como objetivo principal, relacionar as diretrizes sobre boas práticas de Governança de Dados aliada ao *Big Data* no setor público, buscando o entendimento sobre os objetivos e maiores desafios relacionados com a adoção da *Big Data* no contexto estudado.

As contribuições estão indiretamente ligadas a perceber os motivadores e os benefícios obtidos com as práticas de Governança de Dados e *Big Data* nas organizações públicas, em suma, em conformidade com os objetivos da dissertação. A revisão da literatura trouxe que como principais motivadores à utilização de Governança de Dados e *Big Data* por órgãos governamentais estão, entre outros:

- a capacidade de melhoria de decisões;
- o incremento da participação dos cidadãos;
- o melhor processamento de dados;
- a economia de recursos públicos; e
- questões de transparência.

Como desafios se destacam a questão do alinhamento de projetos de Governança de Dados e *Big Data* com a missão institucional. As oportunidades estariam fortemente relacionadas com motivações.

Diante da escassez de casos relacionados ao contexto disponíveis na literatura, foi realizada uma pesquisa de campo para verificar o alinhamento destas questões, oportunidades, desafios e motivadores de projetos de Governança de Dados e *Big Data* nos órgãos públicos do Governo Federal do Brasil.

Com base nos resultados apresentados no Capítulo 4 pudemos avaliar, primeiramente, se os entrevistados teriam conhecimento que os permitissem responder com segurança as questões centrais da dissertação quanto ao que ocorre nos órgãos públicos que representavam. Uma vez verificado que a maioria dos respondentes era composta por pessoas com conhecimento teórico e aptas a aplicar técnicas de Governança de Dados e *Big Data* em suas instituições, lançamos mão de outras questões que visavam responder as questões centrais deste trabalho.

Quanto aos impactos econômicos e Governança de Dados e *Big Data*, se verificou que existem questões econômicas, majoritariamente positivas, como redução de gastos de recursos públicos, embora parte considerável das respostas não tenha passado segurança quanto a este fato. Uma possível razão para isso pode estar relacionada com a área de atuação dos participantes nos seus respectivos órgãos, uma vez que nenhum deles trabalha envolvido diretamente com os investimentos no setor de TI. No entanto, com base nas respostas recebidas, o caso

brasileiro foge da ideia constante dos reportados na literatura, onde o uso do *Big Data* tende a promover a redução dos gastos públicos.

As oportunidades e desafios na adoção das boas práticas Governança de Dados e *Big Data* mostraram-se muito concentrados em aspectos da missão institucional, porém com parte dos desafios estando relacionados a questões políticas e administrativas dos órgãos. Os impactos, por sua vez, estariam ligados a uma melhoria na tomada de decisão, além, é claro, do impacto econômico já comentado.

Em todos os momentos os respondentes deixaram claro que o aspecto da missão institucional não vem sendo esquecido pelos órgãos que possuem projetos e ambientes Governança de Dados e *Big Data*. Isso mostra-se alinhado com o que encontramos na literatura. No entanto, pareceu existir pouco destaque na maioria das respostas aos aspectos de transparência e participação do cidadão, embora a literatura tenha sido rica em citar estas questões quando estamos falando deste assunto.

Uma resposta possível ao que o objetivo principal da pesquisa buscava analisar é que um dos principais motivadores com relação à implementação de projetos de Governança de Dados e *Big Data* é a melhoria no processo de tomada de decisão, objetivando a melhoria de eficiência no gasto de recursos públicos. Outra motivação notável está relacionada a questões de transparência, embora sem o mesmo destaque se comparado às questões econômicas em si. Entretanto, também foi possível perceber que projetos de Governança de Dados e *Big Data* ainda têm enfrentado muitas barreiras burocráticas e políticas.

Outra descoberta relevante foi que a maioria dos órgãos que possuem ambientes e projetos Governança de Dados e *Big Data* já observaram que o alinhamento à missão institucional é o primeiro passo. Por sua vez, se entende que este alinhamento é capaz de ajudar na superação das barreiras levantadas por questões políticas.

Quanto à contribuição deste trabalho no setor público, se verificou que, ao menos para novos projetos em órgãos que ainda não possuem projetos de Governança de Dados e *Big Data* ou que os possuam de maneira muito incipiente, pode-se verificar melhor o alinhamento com a missão institucional. Isso também tende

a melhorar os estudos de impactos, inclusive os econômicos, uma vez que estudos mais detalhados podem auxiliar a derrubar as barreiras políticas.

Os novos projetos podem também considerar melhor os aspectos de transparência e participação do cidadão, pois verificou-se pouco destaque de tais fatores nos projetos apurados. Estes tópicos também podem ser elencados de maneira a superar os desafios políticos ou, ao menos, reduzi-los. A disseminação desse conhecimento de estruturação de projetos com maior detalhamento de informações entre os prestadores de serviço de TI pode gerar uma melhoria do entendimento do escopo dos mesmos, não apenas do ponto de vista tecnológico, mas também do ponto de vista da contribuição social que os servidores de TI podem trazer para a comunidade. Isso pode melhorar, inclusive, o comprometimento com cada projeto, ou seja, trata-se de mais um potencial impacto positivo na Governança de Dados e *Big Data*.

Como trabalho futuro, podemos ampliar o escopo da pesquisa a mais órgãos e colaboradores, bem como focar questões como desafios e oportunidades de forma mais detalhada. Ainda, como trabalho derivado, se poderia estudar melhor projetos que trabalhem questões de transparência e de participação do cidadão de forma mais direta e ativa na forma de estudos de caso. A própria compreensão do porquê os profissionais de TI desses órgãos não darem ênfase nesses tópicos pode se constituir em um bom escopo de pesquisa.

REFERENCIAL BIBLIOGRÁFICO

ALGMIN, A. and J. ZAINO (2018), ***Trends in Data Governance and Data Stewardship***: A 2018 DATAVERSITY Report, DATAVERSITY Education, LLC, http://content.dataversity.net/rs/656-WMW-918/images/Trends%20in%20Data%20Governance%20and%20Stewardship_FinalR-P-Graphs.pdf Acessado em: 06 Ago. 2020.

BALTIMORE (2019). **Mapa dos desertos alimentares de Baltimore City**. Disponível em: https://translate.googleusercontent.com/translate_c?depth=1&hl=pt-BR&prev=search&rurl=translate.google.com.br&sl=en&sp=nmt4&u=https://www.baltimoresun.com/maryland/baltimore-city/bal-bmorefoodmap-graphic-20150610-htmlstory.html&usq=ALkJrhiGA5F5zvCldW-OTPLTdbHxtgZhAQ. Acesso em: 12 mai. 2020.

BARATA, André Montoia. **Governança de dados em organizações brasileiras: uma avaliação comparativa entre os benefícios previstos na literatura e os obtidos pelas organizações**. 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/100/100131/tde-28072015-215618/pt-br.php> . Acesso em: 06 Ago. 2020.

BARC (2019), ***Data Governance: Definition, Challenges & Best Practices***, Bi-Survey.com, <https://bi-survey.com/data-governance> Acessado em: 06 Ago. 2020.

BOCCATO, V. R. C. **Metodologia da pesquisa bibliográfica na área odontológica e o artigo científico como forma de comunicação**. Rev. Odontol. Univ. Cidade São Paulo, São Paulo, v. 18, n. 3, p. 265-274, 2006.

CARVALHO, Mónica Isabel Machado. **Data Governance: estudo e aplicação na EDP distribuição**. 2012. Faculdade de Economia, Universidade de Évora, Évora, 2012. 57 p. Disponível em: <http://hdl.handle.net/10316/21346> . Acesso em: 06 Ago. 2020

COMISSÃO EUROPEIA (1998). COM (1998). ***Public sector information: A key resource for Europe. Green paper on public sector information in the information society***. European Commission. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01972240500388180> Acesso em: 09 mar. 2020.

CORDEIRO, Daniel. **Apache Hadoop: Conceitos técnicos e práticos, evolução e novas possibilidades**. 2012. Departamento de Ciências e Computação, Instituto de Matemática e Estatística – Universidade de São Paulo. Disponível em: <http://www.each.usp.br/dc/papers/erad-hadoop-DanielCordeiro.pdf> Acesso em: 01 mar. 2020.

CORREIA, Z. P. (2004). ***Toward a stakeholder model for the co-production of the public-sector information system***. *Information Research*, 10(3), paper 228. Retrieved February 27, 2013. Disponível em: <http://InformationR.net/ir/10-3/paper228.html>. Acesso em: 09 mar. 2020.

DANA-PMBOK (2018) *Navigating the Labyrinth is a concise guide to data management, providing an executive's overview to the DMBOK2*. Utgitt 2018-06-01. Disponível em: https://doc.difi.no/nasjonal-arkitektur/kunnskap_bok-navigating-the-labyrinth/ Acesso em: 12 ago. 2020.

DATA GOVERNANCE INSTITUTE (DGI). *Definitions of Data Governance*. [S.l.]: DGI, c2017a. Disponível em: http://www.datagovernance.com/adg_data_governance_definition/. Acesso em: 06 Ago. 2020.

DAVENPORT, Thomas H., 2014 - *Big Data at work: dispelling the myths, uncovering the opportunities*. pages cm: ISBN 9781422168165.

DELOITTE (2019). *Using advanced analytics to drive regulatory reform*. Disponível em: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-ps-using-advanced-analytics-to-drive-regulatory-reform.pdf> Acesso em: 12 mai. 2020.

DIAS, Jhenny Kelly do C. *Big Data: Conceitos Básicos*. Disponível em: Acesso em: <http://www.cogef.ms.gov.br/wp-content/uploads/2018/04/Apresentac%CC%A7a%CC%83o-conceitual-do-Big-Data.pdf> 29 fev. 2020.

DRUID (2019). *Apache Druid is a high performance real-time analytics database*. Disponível em: <https://druid.apache.org/technology> Acesso em: 10 mar. 2020.

ESPINDOLA, Priscilla Lüdtke. *Governança de dados aplicada à ciência da informação: análise de um sistema de dados científicos para a área da saúde*. Universidade do Estado de Santa Catarina. 2017. Disponível em: <file:///C:/Users/wesle/AppData/Local/Temp/8651080-Texto%20do%20artigo-42059-7-10-20180831.pdf> Acesso em: 06 Ago. 2020.

FAGUNDES, Priscila Basto., MACEDO, Dyllon Douglas J., FREUND, Gislaine Parra. *A produção científica sobre qualidade de dados em Big Data: Um estudo na base de dados WEB of Science*. Universidade Federal de Santa Catarina, 2017. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/download/8650412/pdf/> Acesso em: 01 mar. 2020.

FILHO, Osilmar, M. C., CARDOSO, Marcelo de Castro. *Arquitetura em Big Data para organização de milhões de ligações em uma empresa de Telecom*. Curso de Engenharia de Computação Centro Universitário de Anápolis (UniEvangélica) – Anápolis, GO – Brasil. 2019. Disponível em: http://repositorio.aee.edu.br/bitstream/aee/319/1/TCC2_2016_2_OsilmarMendonca.pdf Acesso em: 01 mar. 2020.

GAMA, João & RODRIGUES, Pedro. (2007). *Data Stream Processing*. 10.1007/3-540-73679-4_3. Disponível em: https://www.researchgate.net/publication/236007656_Data_Stream_Processing/citation/download Acesso em: 10 mar. 2020.

GARTNER (2016) **GRC Redefined: Gartner Adopts an Integrated Risk Management View**. Disponível em: <https://blogs.gartner.com/john-wheeler/grc-redefined-gartner-adopts-an-integrated-risk-management-view/>. Acesso em: 29 julho 2020.

GHAVAMI, P. (2015), **Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics**, CreateSpace Independent Publishing Platform.

GILBERT, MJ. The anthropologist as alcoholologist: qualitative perspectives and methods in alcohol research. *Int J Addict*, v.25, n.2A, p.127-48, 1990-91.

GONÇALVES, Tânia Carolina Nunes Machado., VARELLA, Marcelo D. **Os desafios da Administração Pública na disponibilização de dados sensíveis**. Escola De Direito De São Paulo Da Fundação Getulio Vargas Revista Direito Gv | São Paulo | V. 14 N. 2 | 513-536 | Maio-Ago 2018. Disponível em: <http://www.scielo.br/pdf/rdgv/v14n2/1808-2432-rdgv-14-02-0513.pdf> Acesso em: 09 mar. 2020.

GONDINE, Sônia M. G. **Grupos focais como técnica de investigação qualitativa: desafios metodológicos**. 2002. Disponível em: <https://www.scielo.br/pdf/paideia/v12n24/04.pdf> Acesso em: 13 ag. 2020.

GUPTA, Anish., GUPTA, Manish K., **HIVE- Processing Structured Data in HADOOP**. International Journal of Scientific & Engineering Research Volume 8, Issue 6, June-2017 45 ISSN 2229-5518. Disponível em: https://www.researchgate.net/publication/318338588_HIVE-Processing_Structured_Data_in_Hadoop Acesso em: 10 mar. 2020.

HENRIQUES, Daniela A. COSTA, Helena R. **Big Data – como utilizar a extraordinária quantidade de informações coletadas por novas tecnologias para obter vantagens competitivas**. UNA – 2018. Disponível em: http://revistapensar.com.br/tecnologia/pasta_upload/artigos/a69.pdf. Acesso em: 06 Ago. 2020.

IBM (2020) **Governança de dados**. Disponível em: <https://www.ibm.com/br-pt/analytics/data-governance>. Acesso em: 18 ago. 2020.

KAFKA, Apache. **O que é o Apache Kafka?** 2019. Disponível em: <http://kafka.apache.org/>. Acesso em: 18 mar. 2020.

KLEPPMANN, Martin. Making Sense of Stream Processing: The Philosophy Behind Apache Kafka and Scalable Stream Data Platforms. 2016. Disponível em: https://assets.confluent.io/m/2a60fabledb2dfbb1/original/20190307-EB-Making_Sense_of_Stream_Processing_Confluent.pdf Acesso em: 10 mar. 2020.

LADLEY J. **Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program**. The Morgan Kaufmann Series on Business Intelligence. Morgan Kaufmann. 2012.

MAGHAM, Ravi. **Pro Apache Phoenix: na SQL driver for Hbase**. 2017. Disponível em: https://www.academia.edu/40508955/PRO_APACHE_PHOENIX_AN_SQL_DRIVER_FOR_HBASE Acesso em: 18 mar. 2020.

MAY, T. **Pesquisa social: questões, métodos e processo**. Porto Alegre, Artmed, 2004.

MCKINSEY GLOBAL INSTITUTE. (2011, June). **Big Data: The next frontier for innovation, competition, and productivity**. McKinsey & Company. Disponível em: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>. Acesso em: 09 mar. 2020.

MOSCAROLA J., Enquête et analyse de données en marketing avec le Sphinx Vuibert 1991.,

NAOPA (2018). National Academy of Public Administration. **Strengthening Organizational Health and Performance in Government**. Disponível em: [https://www.napawash.org/uploads/Strengthening Organizational Health and Performance in Government.pdf](https://www.napawash.org/uploads/Strengthening_Organizational_Health_and_Performance_in_Government.pdf) Acesso em: 12 mai. 2020.

NARKHEDE, Neha., SHAPIRA, Gwen., PALINO, Todd. **Kafka: The Definitive Guide**. 2017. Disponível em: <https://book.huihoo.com/pdf/confluent-kafka-definitive-guide-complete.pdf> Acesso em: 18 mar. 2020.

NETFLIX (2019). Spark and Spark Streaming at Netflix. Disponível em: <https://www.slideshare.net/SparkSummit/spark-and-spark-streaming-at-netflix-sedakar-daxini> Acesso em: 18 mar. 2020,

OECD (2018), **Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact**, OECD Digital Government Studies, OECD Publishing, Paris, <https://dx.doi.org/10.1787/9789264305847-en>. Acesso em: 06 Ago. 2020.

OECD (2019), **Digital Government Review of Argentina: Accelerating the Digitalisation of the Public Sector**, OECD Publishing, Paris, <https://doi.org/10.1787/354732cc-en> . Acesso em: 06 Ago. 2020.

ORACLE (2019) **O que é Big Data?** Disponível em: <https://www.oracle.com/br/big-data/guide/what-is-big-data.html> Acesso em: 29 fev. 2020.

PETRY, Marcio. (2016). **Big Data e a implementação de um Sistema Distribuído com Apache Hadoop: um estudo exploratório**. Disponível em: [https://www.researchgate.net/publication/318600685 Big Data e a implementacao de um Sistema Distribuido com Apache Hadoop um estudo exploratorio/citation/download](https://www.researchgate.net/publication/318600685_Big_Data_e_a_implementacao_de_um_Sistema_Distribuido_com_Apache_Hadoop_um_estudo_exploratorio/citation/download) Acesso em: 01 mar. 2020.

PHOENIX, Apache. **OLTP e análise operacional para Apache Hadoop**. 2019. Disponível em: <https://phoenix.apache.org/> Acesso em: 18 mar. 2020.

PREDPOL (n.d.). **Retrieved September 08, 2013, from PredPol** Web site: <https://www.predpol.com/>. Acesso em: 09 mar. 2020.

QBAL, Muhammad & SOOMRO, Tariq. (2015). **Big Data Analysis: Apache Storm Perspective**. International Journal of Computer Trends and Technology. 19. 9-14.

10.14445/22312803/IJCTT-V19P103. Disponível em:
https://www.researchgate.net/publication/271196175_Big_Data_Analysis_Apache_Storm_Perspective/citation/download Acesso em: 18 mar. 2020.

REIS, Rodrigo Xavier. **A utilização do fenômeno *Big Data* na Administração Pública A experiência do PENSA na Prefeitura Municipal do Rio de Janeiro**, 2015. Disponível em:
<https://www.tce.rj.gov.br/documents/454798/49753202/Rodrio%20Xavier%20dos%20Reis.pdf> Acesso em: 09 mar. 2020.

ROSSOUW, L. **Big Data – grandes oportunidades**. Gen Re – Risk Insights, vol. 16, nº 2, 2012.

SANTANA, Lucas Vasconcelos. **Introduzindo o Apache Storm MAC5742 - Computação Paralela e Distribuída** Instituto de Matemática e Estatística Universidade de São Paulo. 2015. Disponível em:
<https://www.ime.usp.br/~gold/cursos/2015/MAC5742/reports/ApacheStorm.pdf> Acesso em: 18 mar. 2020.

SANTOS, Isabel Maria Francisca dos. **Uma proposta de governança de dados baseada em um método de desenvolvimento de arquitetura empresarial**. 2010. 140 p. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010. Disponível em:
<http://www2.uniriotec.br/ppgi/banco-de-dissertacoes-ppgi-unirio/ano-2010/uma-proposta-de-governanca-de-dados-baseada-em-um-metodo-de-desenvolvimento-de-arquitetura-empresarial/view>. Acesso em: 06 Ago. 2020.

SANTOS, Marco A., BARRETO, Raimundo. **Mapeamento Sistemático**. Universidade Federal do Amazonas –UFAM Instituto de Computação -IComp. 2015. Disponível em:
<https://arxiv.org/ftp/arxiv/papers/1504/1504.01027.pdf> Acesso em: 20 ago. 2020.

SENADO FEDERAL (2012). **Manual de Padronização de Atos Administrativos Normativos**. Disponível em:
<https://www2.senado.leg.br/bdsf/bitstream/handle/id/496338/000960587.pdf?sequence=1&isAllowed=y> Acesso em: 09 mar. 2020;

SILVA, Bruno A. (2019). **Apache Hadoop**. Universidade Federal de São Carlos. Disponível em:
<https://dcomp.sor.ufscar.br/verdi/topicosCloud/Hadoop.pdf> Acesso em: 01 mar. 2020.

SILVA, Petrônio. (2015). **Big Data e Data Science: Admirável Mundo Novo**. 10.13140/RG.2.1.2665.4806. Disponível em:
https://www.researchgate.net/publication/289253933_Big_Data_e_Data_Science_Admiravel_Mundo_Novo Acesso em: 29 fev. 2020.

SOUZA, Milton Pereira. **Mecanismos para viabilizar transparência ativa e abertura de dados adequados a governança de informações digitais da ANAC**. Universidade Federal de Minas Gerais. 2019. Disponível em:
<https://repositorio.ufmg.br/bitstream/1843/30472/1/MiltonPereiradeSouza.pdf> Acesso em: 09 mar. 2020.

STORM, Apache. **Porque utilizar o Apache Storm**. 2019. Disponível em:
<https://storm.apache.org/index.html> Acesso em: 18 mar. 2020.

STUMPF, Ricardo D. **O porquê de governança de dados em organizações de controle.** TCU – 2017. Disponível em: <https://webcache.googleusercontent.com/search?q=cache:vJld4o5ftqYJ:https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/1383/1529+&cd=1&hl=pt-BR&ct=clnk&gl=br&client=firefox-b-d> Acesso em: 06 Ago. 2020.

TALHAOUI, Mohamed Amine. (2018). **Real-time Data Stream Processing - Challenges and perspectives.** International Journal of Computer Science Issues. 14. 10.20943/01201705.612. Disponível em: [https://www.researchgate.net/publication/326985824_Real-time_Data_Stream_Processing - Challenges and Perspectives/citation/download](https://www.researchgate.net/publication/326985824_Real-time_Data_Stream_Processing_-_Challenges_and_Perspectives/citation/download) Acesso em: 10 mar. 2020.

TEHSEEN (2016). **Big Data e os 3 Vs.** Disponível em: <http://tehseen.dbsdatapoints.com/wp-content/uploads/sites/107/2016/04/3vs1.jpg>, Acesso em: 28 fev. 2020.

VAN KASTEREN, T., ULRICH, B., SRINIVASAN, V., & NIESSEN, M. (2014). **Analyzing tweets to aid situational awareness.** 36th European Conference on Information Retrieval. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-06028-6_79. Acesso em: 09 mar. 2020.

VIDOR, D. M. (2019). **Big Data, Revolução Digital e o Direito.** Disponível em: <https://jus.com.br/artigos/75132/bigdata-revolucao-digital-e-o-direito>. Acesso em: 20/08/2020.

WANG, T., RUDIN, C., WAGNER, D., & SEVIERI, R. (2013). **Detecting patterns of crime with series finder.** Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Disponível em: <https://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/viewPaper/7018> Acesso em: 09 mar. 2020.

YANG, Fangin., TSCHETTE, Eric., LÉAUTÉ, Xavier., **Druid: A real time analytical data store.** Disponível em: <http://static.druid.io/docs/druid.pdf> Acesso em: 10 mar. 2020.

YIU, C. (2012). **The Big Data Opportunity. Making government faster, smarter and more personal.** London: Policy Exchange. Disponível em: <http://ict-industry-reports.com.au/wp-content/uploads/sites/4/2013/05/2012-The-Big-Data-opportunity-in-Government-UK-Policy-Exchange-July-2012.pdf> Acesso em: 21 julho 2020.

ZILLNER, S., RUSITSCHKA, S., MUNNÉ, R., LIPPELL, H., LOBILLO, F., HUSSAIN, K. et al. (2013): **D2.3.1. First draft of the sectorial requisites.** Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4). Disponível em: http://big-project.eu/sites/default/files/D2.4.1_FINAL_v0_8.pdf Acesso em: 09 mar. 2020.

ZIMMERMANN, Marlene Harger., MARTINS, Pura Lúcia Oliver. **Grupo focal na pesquisa qualitativa: relato de experiencia.** 2008. Disponível em: https://educere.bruc.com.br/arquivo/pdf2008/211_86.pdf. Acesso em: 13 ago. 2020.

APÊNDICE A – MAPEAMENTO SISTEMÁTICO

MAPEAMENTO SISTEMÁTICO SOBRE UMA ANÁLISE DO IMPACTO DAS PRÁTICAS DE BIG DATA NA ADMINISTRAÇÃO PÚBLICA FEDERAL

Fernando Gonçalves de Oliveira¹, Washington Henrique Carvalho Almeida¹,
Luciano de Aguiar Monteiro¹, Fernando Escobar²,
Rodrigo Otávio Ribeiro Hagstrom² e Daniel Alves da Silva²

¹CESAR School, Cais do Apolo, 77, Bairro Recife – Recife-PE, CEP: 50030-390, Brasil

²Universidade de Brasília – UNB, Campus Universitário Darcy Ribeiro, Brasília -DF, CEP: 70910-900, Brasil
fernando.infinite@gmail.com, whca@cesar.school, lam@cesar.school,
fernando.escobar.br@gmail.com, rodrigo.hagstrom@gmail.com, daniel.alves@redes.unb.br

RESUMO

Este artigo apresenta um mapeamento sistemático da literatura referente à análise do impacto das práticas de Big Data na Administração Pública Federal, com a premissa de que as instituições públicas guardam semelhanças em missões institucionais, ainda que em locais diferentes e em setores diferentes. O principal objetivo dessa pesquisa é reunir subsídios para responder quais os objetivos e os maiores desafios acerca da implantação de Big Data e/ou Ecossistemas de Big Data, nos órgãos públicos federais brasileiros. O levantamento dos artigos foi realizado em três bases de dados (*ACM Digital Library*, *IEEE Xplore* e *Science Direct - Elsevier*), utilizando *strings* de busca, a fim filtrar um conjunto de artigos que estejam alinhados com os tópicos e respostas para as questões de pesquisa. Este mapeamento sistemático é uma etapa balizadora de uma pesquisa mais ampla que, em suma, pode direcionar estruturas de projetos de Big Data no setor público, em especial na Administração Pública Federal brasileira. Compreendeu-se que a implantação de Big Data no setor público precisa estar perfeitamente alinhada com a missão institucional do órgão que se propõe a implantá-lo, mas que os gestores públicos nem sempre percebem suas instituições como aptas a implantarem projetos desta natureza. Tal problema pode demandar que seja apresentado todo um conjunto de diferentes projetos de Big Data em diferentes esferas e missões institucionais. Finalmente, se percebe que há pouca literatura relacionada com o caso brasileiro, o que evidencia a pertinência, tanto do mapeamento sistemático quanto de etapas futuras da pesquisa.

PALAVRAS-CHAVE

Big Data. Administração Pública Federal. Ecossistema de Big Data. Mapeamento Sistemático.

ABSTRACT

This article presents a systematic mapping of the literature regarding the analysis of the impact of Big Data practices in the Federal Public Administration, with the premise that public institutions have similarities in institutional missions, even though in different locations and in different sectors. The main objective of this research is to gather information to answer what are the objectives and the biggest challenges regarding the implementation of Big Data and/or Big Data Ecosystems, in the Brazilian federal public agencies. The survey of the articles was carried out in three databases (ACM Digital Library, IEEE Xplore and Science Direct - Elsevier), using search strings, in order to filter a set of articles that were related to the results and answers to the research questions. This systematic mapping is a steppingstone for broader research that, in short, can direct structures of Big Data projects in the public sector, especially in the Brazilian Federal Public Administration. It was understood that a Big Data implementation in the public sector needs to be perfectly aligned with an institutional mission of the body that proposes to implement it, but that public managers do not always perceive their institutions as able to implement projects of this nature. Such a problem may require that a whole set of different Big Data projects be presented in different spheres and institutional missions. Finally, it is noticed that there is little quantity of literature related to the Brazilian case, which highlights the relevance, both of the systematic mapping and of future research stages.

KEYWORDS

Big data. Federal Public Administration. Big Data ecosystem. Systematic Mapping.

1. INTRODUÇÃO

Atualmente, indivíduos e organizações, sejam privadas, públicas ou do terceiro setor, vêm gerando grandes volumes de dados, em uma proporção que cresce constantemente (Akoka *et al.*, 2017). Complementarmente, a popularização dos dispositivos conectados (*Internet of Things – IoT*), sensores e aplicações que coletam e geram dados em forma de imagem, áudio e vídeo, tem contribuído significativamente para esse volume e sua curva ascendente. Esse contexto ganha impulso, relevância e complexidade com o advento da quarta revolução industrial, conhecida como Indústria 4.0 (Akoka *et al.*, 2017; Almeida *et al.*, 2019).

Big Data é um termo genérico para a enorme quantidade de dados digitais não estruturados, normalmente coletados de múltiplas fontes; o termo também representa os novos desafios que envolvem a complexidade, a segurança e os riscos à privacidade, bem como a necessidade de novas tecnologias e habilidades humanas, redefinindo o cenário do gerenciamento de dados e dos processos de extração, transformação e carregamento (Kim *et al.*, 2014).

Akoka *et al.* (2017) realizaram um trabalho de enfoque meta analítico com artigos sobre Big Data, o qual demonstra que a quantidade de estudos sobre o termo vem crescendo, o que pode estar relacionado com o aumento da percepção acerca da importância de tal conjunto de tecnologias e de sua implantação. Para responder à questão sobre quais os assuntos produzidos a partir das pesquisas de Big Data, os autores classificaram os trabalhos pesquisados em domínios de conhecimento, abrangendo, entre outros, energia, medicina, ecologia, finanças e governo. A presença do domínio “governo” demonstra a relevância do tema em relação a pesquisas com Big Data. Entretanto, de um total de 441 trabalhos analisados, apenas 35 pesquisas reportaram casos governamentais, posicionando esse tópico na oitava posição de um universo de doze domínios. Isso evidencia a carência de trabalhos relacionados ao domínio “governo” ou “setor público” e uma série de oportunidades que podem ser exploradas.

No caso do Brasil, há evolução no entendimento de que os órgãos do governo federal devem estar envolvidos em um ambiente que permita a adequada transparência das atividades públicas, bem como a necessidade de controle e conhecimento de informações, em larga escala, para melhor gestão das atividades relacionadas com os serviços públicos e governamentais (Victorino *et al.*, 2017). Os autores complementaram seu trabalho apresentando uma proposta de Ecossistema de Big Data, para dar suporte a análise de dados abertos governamentais, como resposta ao contexto de transparência. Estas características sugerem que o caso brasileiro guarda semelhanças com os casos internacionais mais ricamente presentes na literatura correlata, bem como semelhanças entre os órgãos brasileiros de diferentes setores, o que se tornou uma premissa no trabalho relatado neste artigo.

Este artigo tem como tema o mapeamento sistemático referente à análise do impacto das práticas de Big Data na Administração Pública Federal do Brasil, ocorrido ou a ocorrer, abordando o tema por meio de pensamentos de autores que delineiam as bases para o estabelecimento do ecossistema de Big Data no setor público e as boas práticas relacionadas. Assim, o objetivo principal é entender quais os propósitos e os maiores desafios e dificuldades relacionados com a adoção do Big Data na Administração Pública (AP). A pesquisa permite identificar iniciativas e a visão dos ecossistemas de Big Data que possam auxiliar ambientes profissionais ou organizacionais com aplicabilidade prática. O estudo em tela atinge, inicialmente, esse objetivo, sendo o ambiente profissional, no caso, a Administração Pública.

O restante deste artigo está estruturado da seguinte forma: a Seção 2 aborda alguns conceitos centrais para o entendimento da terminologia que compõe o cenário de Big Data e a Seção 3 apresenta a Metodologia da pesquisa. Na Seção 4, são apresentadas as conclusões deste mapeamento sistemático.

2. CONCEITOS

Segundo Victorino *et al.* (2017), citando Schimdt (2010), a humanidade criou, até o ano de 2003, cerca de cinco exabytes de dados. O volume, desde então, dobra a cada dois anos em média (Gantz and Reinsel, 2011), o que fez com que no ano de 2019 fossem criados cerca de 160 exabytes de dados. Consolidando o conceito de Big Data, nas palavras de Araújo Júnior and Souza (2018), fundamentados em Ramos (2015), Big Data é um conjunto de tecnologias utilizadas para tratar grandes volumes de dados. É importante destacar que, não necessariamente, os dados tomados para tratamento estão em formatos similares, compatíveis. Em um cenário como este, conforme destacam os mesmos autores, tornam-se necessárias soluções computacionais mais efetivas em termos de custos e mais inovadoras no tratamento da informação.

Victorino *et al.* (2017) ainda destacam, como uma destas soluções computacionais, o Big Data, o qual possibilita a gestão de dados em grande velocidade de processamento, mas que não se refere especificamente à sua organização. Araújo Júnior and Souza (2018) destacam que o conceito de Big Data tem sua origem no *Business Intelligence* (BI), que é, por sua vez, resultado do processo de mineração de dados analíticos, voltado para a combinação de dados, formatação de metadados e de conteúdo a serem usados no processo decisório. O BI é o uso

de variadas fontes de dados e informações, utilizadas para definir estratégias e subsidiar processos organizacionais. Logo, os conceitos de Big Data e BI convergem e, por isso, alguns autores consideram as duas concepções como apenas uma (Araújo Júnior and Souza, 2018; Barbieri, 2001).

Araújo Júnior and Souza (2018), Davenport (2014), Márquez and Lev (2017) são outros que também consideram os conceitos similares. Entretanto, não se verifica que os autores se preocupem em citar a organização dos dados utilizados nos processos, nem mesmo acerca do armazenamento. Ainda, conforme Inmon and Linstedt (2001), os conceitos são similares quando em relação a modelos que envolvem a captura, apresentação e geração de relatórios para a tomada de decisão.

Davenport (2014), citado por Victorino *et al.* (2017), afirma que o termo Big Data é muito genérico, utilizado para conjuntos de dados que não são comportados em repositórios habituais. Logo, o termo estaria relacionado a conjuntos de dados massivos e volumosos. Não caberiam, portanto, em servidores tradicionais e bancos de dados relacionais, demandando novas tecnologias de armazenagem e tratamento.

É comum, nos artigos analisados, que os autores abordem a definição de “3Vs” (Akoka *et al.*, 2017; Araújo Júnior and Souza, 2018; Davenport, 2014; Laney, 2001), caracterizando o Big Data, quaisquer que sejam, conforme definições presentes em Araújo Júnior and Souza (2018):

- volume: permite a modelagem e a análise de extensos volumes de dados e o que será extraído e armazenado;
- velocidade: fornece resultados de pesquisa com agilidade, em tempo compatível com a demanda, na qual os dados serão gerados, a rapidez na ação tomada sob os dados ou a taxa de mudança nos dados;
- variedade: processa e fornece conjuntos de dados em diferentes tipos e formatos.

A estes “3Vs”, Davenport (2014) acrescentou, posteriormente, como destacam Victorino *et al.* (2017), outros dois “Vs”, totalizando, então, os “5Vs” da definição do Big Data. São os outros dois “Vs”, também na definição de Araújo Júnior and Souza (Araújo Júnior and Souza, 2018; Hashem *et al.*, 2015):

- veracidade: processa e fornece conjuntos de dados fiéis aos dados processados e armazenados nos bancos, diminuindo incertezas e aumentando a confiabilidade dos dados; e
- valor: agrega valor informacional; informação de valor claro para as organizações que as utilizam.

A compilação dos “5Vs” é graficamente representada na Figura 1.

Volume	Variedade	Velocidade	Veracidade	Valor
				
Quantidade de dados que são gerados a cada segundo	Tipos de dados que podem ser utilizados	Velocidade a qual os dados são gerados e transmitidos	Grau de organização e confiabilidade dos dados	Necessidade de se transformar dados em valor
Terabytes Registros Transações Arquivos	Estruturado Semiestruturado Não estruturado	Tempo real Streams Batch	Fidedignidade Autenticidade Origem Relevância Disponibilidade	Estatística Eventos Correlações Hipotética

Figura 14. Demonstração de Big Data por meio dos 5Vs.

O processamento de grandes massas de dados é um dos principais motivos pelos quais gestores têm buscado projetos de Big Data (Almeida *et al.*, 2019). Portanto, o provimento de capacidade analítica a partir da extração de novas visões derivadas de grandes volumes de dados, o que em bases de dados tradicionais pode ser considerado complexo, bem como as demais estruturas que compõem o Ecossistema de Big Data.

2.1 ECOSSISTEMA DE BIG DATA

Segundo Victorino *et al.* (2017), revisitando os conceitos elencados em Shin and Choi (2015), assim como em outras ciências, o conceito de ecossistema estaria relacionado com as interações entre os diversos componentes de um sistema. No contexto do Big Data, portanto, estaria relacionado com as interações entre tecnologia, governo, indústria, mercados, usuários e sociedade, sendo examinados ou investigados os efeitos do ambiente de Big Data em todos os setores envolvidos. Victorino *et al.* (2017) citam, ainda, a similaridade desta definição com a encontrada em Demchenko *et al.* (2014), na qual a abordagem biológica é mantida, mas com a diferença na definição do ecossistema, como um complexo de facilidades técnicas e componentes construídos acerca de certa origem de dados específica e de suas aplicações, sendo que o complexo de componentes inter-relacionados é

voltado para as atividades de armazenamento, processamento, visualização e entrega dos resultados. O ecossistema compreenderia, ainda, alguns componentes arquiteturais, tecnologias e outros componentes para o processamento e a análise do Big Data, devendo compreender o gerenciamento do seu ciclo de vida, ou seja, o fluxo de transformação dos dados e toda sua infraestrutura de segurança, bem como todo o conjunto necessário de componentes e políticas para prover controle no acesso aos dados (Almeida *et al.*, 2019).

É possível observar, se tomadas as explicações em epígrafe, que os dados não se encontram em repositórios tradicionais, que são oriundos de diversas fontes, que o volume de dados é massivo (Almeida *et al.*, 2019) e que Big Data e *Cloud Computing* (ou Computação em Nuvem, em língua portuguesa) estão relacionados. Esta relação é parte do trabalho de Hashem *et al.* (2015), no qual os autores detalham que Big Data e *Cloud Computing* se complementam, fornecendo aos usuários a capacidade de utilizar poder computacional em larga escala para processar consultas distribuídas em vários conjuntos de dados e devolver as resultantes em tempo hábil. De forma subjacente, a Computação em Nuvem fornece outros mecanismos como, por exemplo, o do *framework* Hadoop e do MapReduce.

O MapReduce (Dean and Ghemawat, 2008) é um exemplo de aplicação de Big Data em um ambiente de Computação em Nuvem, o qual permite o processamento de grandes quantidades de conjuntos de dados, armazenados em paralelo no *cluster* Hadoop (Almeida *et al.*, 2019; Kazemi, 2018), habilitando a Computação em *Cluster*. A Computação em *Cluster* apresenta bom desempenho em ambiente de sistema distribuído, com características tais como: baixo consumo de energia, bom aproveitamento de armazenamento e boa taxa de utilização e adaptação do uso da rede de computadores e de comunicações. Bollier *et al.* (2010) destacam a capacidade da Computação em *Cluster* de fornecer um contexto favorável ao crescimento de dados. Já Miller *et al.* (2013) argumentam que a falta de disponibilidade de dados se torna cara, porque os usuários transferem mais decisões para métodos analíticos; o uso incorreto dos métodos e/ou pontos fracos inerentes podem produzir decisões erradas com alto custo.

O Hadoop é um *framework* de computação distribuída, desenvolvido em Java, voltado para *clusters* e processamento de grandes massas de dados. Após mapear os dados distribuídos e aplicar as transformações ou regras de negócio (processo MapReduce), o *framework* consolida o resultado e apresenta a resposta ao usuário final (Almeida *et al.*, 2019; O'Driscoll *et al.*, 2013; Kazemi, 2018). De acordo com Jadhav *et al.* (2018), a estrutura do Hadoop consiste em um sistema de arquivos distribuídos (*Hadoop Distributed File System* – HDFS) e outros projetos como Map Reduce, Yarn, Hive, Ping, Mahout, Avro, Sqoop, Oozie, Chukwa, Flume e Zookeeper chamados de ecossistema de componentes, que ajudam a resolver problemas de processamento de Big Data, como podemos observar na Figura 2.

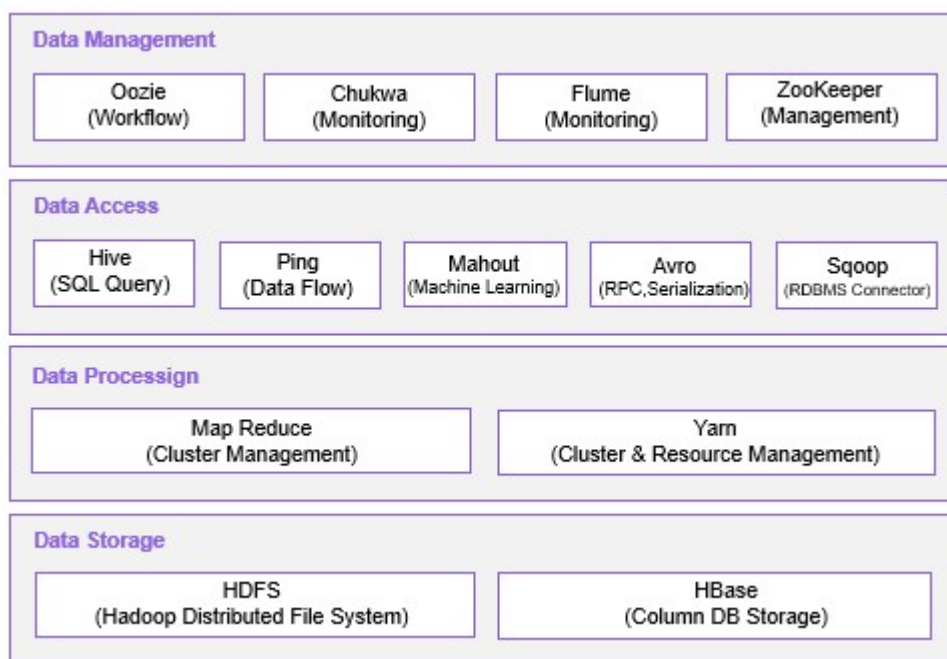


Figura 15. Jadhav et al. (2018) Componentes do ecossistema do Hadoop. Adaptado pelo autor

Estas relações entre os componentes do Ecossistema de Big Data, incluindo os processos e ferramentas para o trabalho com as massas de dados, são destacadas por Almeida *et al.* (2019). No referido trabalho, se verifica a relação com a Computação em Nuvem/distribuída, bem como com ferramentas e métodos de processamento de dados não utilizados na computação tradicional. Mais uma vez, até os sistemas de guarda de dados e a relação

destes volumes com a era da informação e da demanda por alto poder computacional para decisões negociais são compreensíveis, além de ressaltar a complexidade do Ecossistema de Big Data.

Estas decisões e os processos para se obtê-las, por meio da análise de grandes volumes de dados, em uma análise preliminar, são a razão da existência do Big Data (Almeida *et al.*, 2019). Razão esta que é reforçada no contexto da Administração Pública Federal, constituindo-se em ferramenta fundamental para apoiar os processos de tomada de decisões e de cumprimento dos normativos de transparência governamental.

2.2 BIG DATA NO SETOR PÚBLICO

Conforme citado na Seção 1, o setor público carece da utilização de ferramentas de Big Data, seja por questões de melhor controle das informações para tomada de decisões, seja pela necessidade de transparência das ações estatais junto à coletividade administrada.

Mesmo com estas características, o setor público encontra diferenças com relação a implantação do Big Data e seus ecossistemas. Segundo Klievink *et al.* (2017), o Big Data está sendo implementado com sucesso no setor privado e em pesquisas científicas. No entanto, o setor público parece estar ficando para trás. As organizações governamentais muitas vezes reconhecem as oportunidades relacionadas com Big Data, mas não se avaliam como aptas para a utilização de grandes volumes de dados, nem se consideram adequadamente equipadas para o Big Data. Os autores tratam exatamente acerca desta insegurança, apresentando uma estrutura para avaliar a prontidão das organizações públicas. O método utilizado busca desmistificar o conceito de Big Data para organizações do setor público holandeses (Klievink *et al.* 2017), conforme expresso em termos de características organizacionais específicas e mensuráveis. Os resultados obtidos sugerem que as organizações podem ser tecnicamente capazes de utilizar as tecnologias de Big Data, mas não obterão ganhos significativos com essas atividades se os aplicativos não se ajustarem às suas organizações e às principais tarefas estatutárias.

De acordo com Janssen and Van Den Hoven (2015), considerando que além de serem massivos, os dados podem ser abertos, compondo o que é literariamente chamado de *BOLD* (do inglês, *Big and Open Linked Data*), a utilização de Big Data por entes governamentais resulta em novas oportunidades e tem o potencial de transformar o governo e suas interações com o público. Como efeito colateral, o *BOLD* oferece a oportunidade de analisar até mesmo o comportamento dos indivíduos, aumentar o controle e reduzir a privacidade. Ao mesmo tempo, em linha com as melhores práticas, o *BOLD* pode ser usado para criar um governo transparente, reforçando a abordagem dos autores mencionados anteriormente, e com os objetivos de implantação do Big Data em entes governamentais. Ainda segundo Janssen and Van Den Hoven (2015), transparência e privacidade são consideradas importantes valores sociais e democráticos, que estão mudando a forma como os cidadãos informados participam de processos democráticos. Os conceitos de transparência e privacidade têm muitas interpretações e são considerados difíceis de serem compreendidos, o que dificulta, por consequência, as formas de implementação dos sistemas relacionados com as informações detidas por entes estatais.

Considerando as premissas dos parágrafos anteriores, a motivação deste artigo é o estudo da implementação do Big Data no setor público brasileiro, buscando por trabalhos relacionados com os entes governamentais de outros países e brasileiros, com a premissa inicial de que as semelhanças entre os objetivos institucionais de órgãos governamentais guardam semelhanças embora em diferentes regiões. Passa-se, agora, à Seção 3 para entendimento dos procedimentos adotados.

3. METODOLOGIA

3.1 SELEÇÃO DOS ARTIGOS

O mapeamento sistemático está relacionado com o conhecimento da literatura correlata, que pode auxiliar no entendimento da questão em epígrafe. Desta forma, para atingir o objetivo proposto, o processo do Mapeamento Sistemático foi conduzido, conforme disposto na Figura 3.

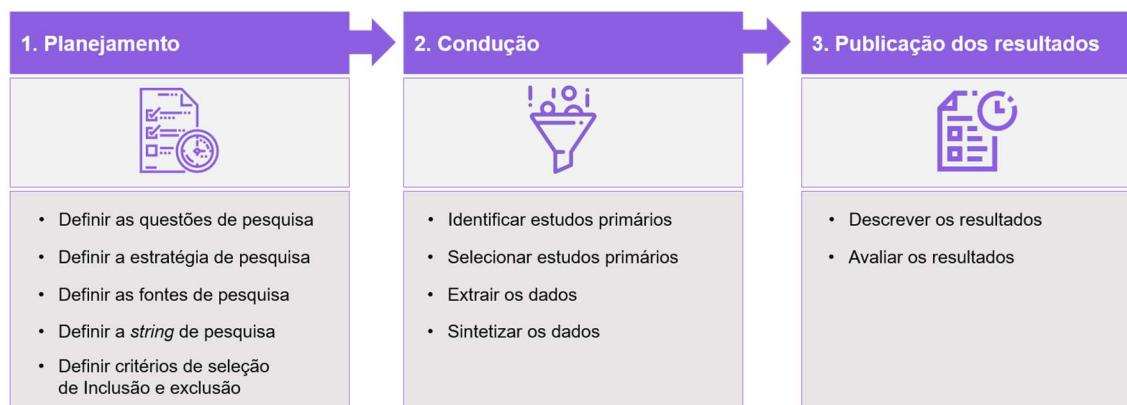


Figura 16. Condução do mapeamento sistemático. Autoria própria.

Considerando os objetivos e os desafios da pesquisa citados anteriormente, chegamos à seguinte questão principal (QP):

(QP) Quais os objetivos e maiores desafios ou dificuldades de implantação de Big Data no setor público?

Complementarmente, como perguntas secundárias (PS), que, integradas, constroem o conhecimento necessário para responder à questão principal, listam-se:

(PS1) Quais os impactos econômicos nas soluções de Big Data no setor público?

(PS2) Quais as oportunidades e desafios da adoção de Big Data no setor público?

(PS3) Quais os impactos que práticas de Big Data trazem ao desenvolvimento de soluções?

Trata-se sobre mapeamento sistemático da literatura o processo pautado em:

- 1) Elaboração de “strings” de pesquisa, orientadas pela estratégia de busca;
- 2) Variedade de fontes de pesquisa para localização dos estudos;
- 3) Definição de critérios de inclusão e exclusão dos estudos selecionados;
- 4) Avaliação da qualidade metodológica das produções recuperadas.
- 5) Bases de dados consultadas e estratégias de busca.

O levantamento dos artigos foi realizado em três plataformas de pesquisa, conforme exposto na Tabela 1.

Tabela 3. Repositórios selecionados. Elaboração própria.

Repositório	Endereço
ACM Digital Library	https://dl.acm.org/
IEEE Xplore	https://ieeexplore.ieee.org/
ScienceDirect – Elsevier	https://www.sciencedirect.com/

Considerando tais bases de pesquisa, algumas combinações de termos foram fundamentais para obtenção de artigos capazes de auxiliar no mapeamento sistemático e na obtenção do estado da arte. Procedeu-se ao cruzamento das principais palavras-chave relacionadas aos temas investigados, quais sejam: “Big Data”, “Setor Público” e “Governo”, além de outros eventualmente necessários para o enriquecimento das fontes de pesquisa, com objetivo de abarcar uma maior quantidade de produções, evitando que algum estudo muito importante ou de muito destaque fosse excluído no levantamento. Assim, algumas *strings* de pesquisa foram montadas e todas as publicações selecionadas foram referentes ao período de 2016 a 2020, conforme apresentadas na Tabela 2. Devido às particularidades dos motores de busca de cada plataforma utilizada, a *string* de busca principal precisou sofrer algumas adaptações, a fim de assegurar um resultado que englobasse um volume maior de estudos relevantes.

Tabela 4. Strings de busca. Elaboração própria.

ID	Base de dados	Query utilizada
1	ACM Digital Library	<i>“query”: AllField: (“public sector”) AND AllField: (“Big Data”) AND AllField: (“difficulties” OR “opportunities” OR “economic impacts” OR “economic advantages”) “filter”: Publication Date: (01/01/2016 TO 12/31/2020)</i>
2	IEEE Explore	<i>(“Big Data” AND “Government” OR “Big Data” AND “Public Sector”)</i>
3	ScienceDirect - Elsevier	<i>“Big Data” AND “Government” AND “Public Sector”</i>

Foram considerados alguns critérios de seleção para que um determinado estudo fosse incluído como passível de análise para a pesquisa, conforme apresentado na Tabela 3.

Tabela 5. Critérios de inclusão de artigos. Elaboração própria.

CI	Critério de inclusão de artigos
1	Apresenta os objetivos e maiores desafios/dificuldades de implantação de Big Data
2	Aborda os impactos econômicos nas soluções de Big Data no Setor Público
3	Aborda as oportunidades e desafios da adoção de Big Data

De maneira análoga, tomaram-se alguns critérios para exclusão de artigos que apareceram entre os resultados das *queries* apresentadas na Subseção 3.4, os quais estão descritos na Tabela 4.

Tabela 6. Critérios de exclusão de artigos. Elaboração própria.

CE	Critério de exclusão de artigos
1	Repetidos
2	Muito semelhantes
3	Não acessível
4	Publicado como “artigo curto” ou somente “poster”
5	Sem resumo ou abstract
6	Publicado apenas resumo
7	Versão mais antiga de outro artigo já considerado ou incluído
8	Não auxilia na resposta de nenhuma das questões da pesquisa

3.2 ESTUDOS DOS RESULTADOS

Após a aplicação das *strings* de busca descritas na Tabela 2, em cada base dados, chegamos a um conjunto de estudos que estão sumarizados na Tabela 5. Para uma melhor visualização, os artigos foram agrupados por ano e por base de dados. Em seguida, as Tabelas 6 e 7 trazem o resultante de estudos após a aplicação dos critérios de inclusão e exclusão, respectivamente, também agrupados por base/ano.

Tabela 7. Artigos por base/ano. Elaboração própria.

Base de dados	2016	2017	2018	2019	2020	TOTAL
ACM Digital Library	35	41	55	53	17	201
IEEE Explore	193	233	254	257	082	1019
Science Direct – Elsevier	108	138	154	242	158	800
TOTAL GERAL	336	412	463	552	257	2020

Tabela 8. Artigos excluídos por base/ano. Elaboração própria.

Base de dados	2016	2017	2018	2019	2020	TOTAL
ACM Digital Library	35	40	51	51	17	194
IEEE Explore	188	222	243	256	82	991
Science Direct – Elsevier	106	135	152	241	154	788
TOTAL GERAL	329	397	446	548	253	1973

Tabela 9. Artigos selecionados por base/ano. Elaboração própria.

Base de dados	2016	2017	2018	2019	2020	TOTAL
ACM Digital Library	0	1	4	2	0	7
IEEE Explore	5	11	11	1	0	28
Science Direct – Elsevier	2	3	2	1	4	12
TOTAL GERAL	7	15	17	4	4	47

Para a priorização dos artigos, bem como das respostas das perguntas, foi aplicada a metodologia de enfoque meta-analítico, conforme descrito nos trabalhos de Mariano and Rocha (2017), Mariano *et al.* (2011) e Sánchez-Meca and López-Pina (2008). Essa abordagem tende a apresentar os trabalhos de maior relevância acadêmica ou com maior quantidade de citações acadêmicas em destaque dentro de um conjunto previamente obtido. Gráficamente, as atividades de tal metodologia estão esquematizadas na Figura 4. Os detalhes de cada uma das três etapas representadas na aludida Figura 4 estão descritas em Mariano and Rocha (2017).

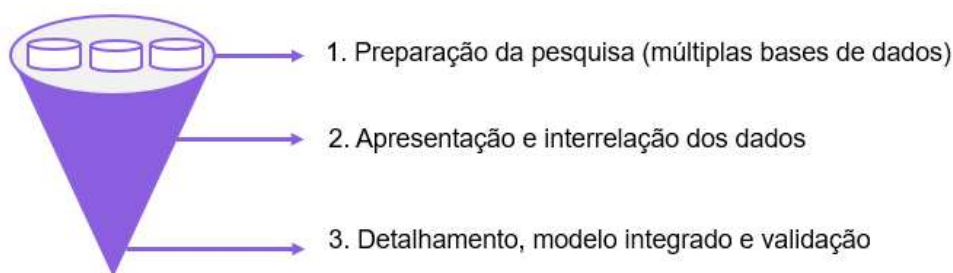


Figura 17. Diagrama Enfoque Meta Analítico. Extraído de Mariano and Rocha (2017).

No entanto, destaca-se que alguns artigos, não resultantes da priorização com o enfoque meta-analítico, bem como do procedimento principal descrito ao longo do presente trabalho, foram mantidos ou incluídos como base literária da pesquisa. Isso se deve ao fato de haver relação direta destes trabalhos com o objetivo desta pesquisa, como, por exemplo, o artigo de Victorino *et al.* (2017), que trata especificamente de Big Data no setor público brasileiro, embora ainda encontre poucas citações ou relevância dentro das bases pesquisadas.

Com fundamento nos estudos selecionados, algumas contribuições podem ser identificadas, possibilitando inferir respostas às perguntas secundárias PS1 e PS2 relatadas na Subseção 3.1. Tal resultado pode ser obtido por meio de enfoque em certos termos componentes da busca principal. O quantitativo de estudos selecionados, que apoiam as respostas às perguntas secundárias, está sintetizado na Tabela 8.

Tabela 10. Bases selecionadas. Elaboração própria.

Base de Dados	PS1	PS2
ACM Digital Library	7	4
IEEE Xplore	8	10
ScienceDirect – Elsevier	4	5

Gráficamente, considerando os procedimentos descritos entre as Subseções 3.1 a 3.5, sintetiza-se os resultados conforme apresentados na Figura 5.

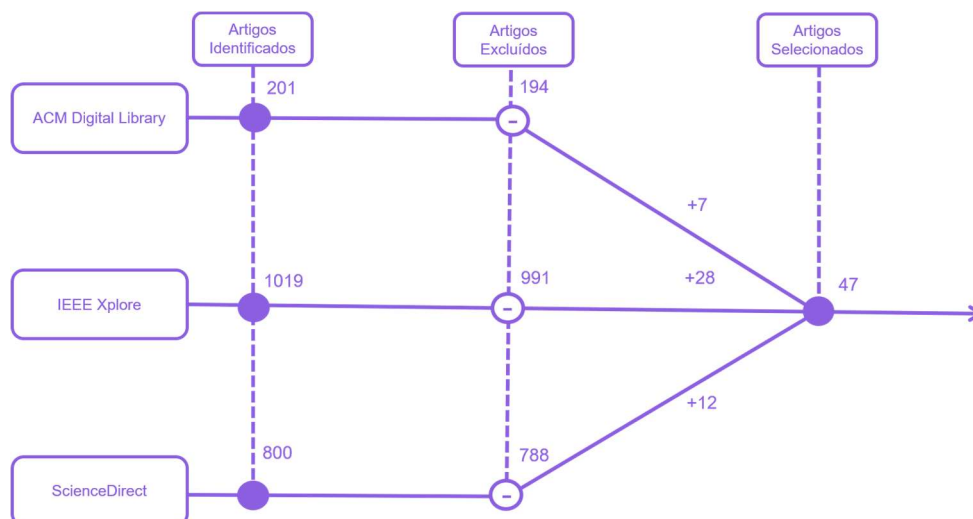


Figura 18. Síntese dos resultados dos procedimentos da pesquisa bibliográfica. Elaboração própria.

Passa-se agora para a Seção 4, na qual se apresentam as discussões.

4. DISCUSSÕES

Para selecionar artigos capazes de auxiliar na resposta da PS1 (impactos econômicos), pode-se focar nos artigos que citam impactos econômicos e criação de valor para organizações, por meio da adoção de Big Data, no caso, no setor público. Com este diapasão, estão sendo identificadas contribuições como as de Khtira *et al.* (2017) ou mesmo de Klievink *et al.* (2017), as quais mencionam as possibilidades de criação de valor em diferentes áreas de atuação dos órgãos governamentais, cada qual em seu segmento. Igualmente, citam uma melhor eficiência quanto à análise e coleta de informações e na interação com a comunidade administrada, permitindo, em última análise, melhor emprego dos recursos públicos ou mesmo sua economia.

Para a PS2 (oportunidades e desafios), foram encontradas contribuições no artigo de Klievink *et al.* (2017), o qual relata entre as dificuldades de implementação do Big Data no setor público, inicialmente, a dúvida sobre a capacidade, incluindo aspectos de infraestrutura computacional e da adoção de ferramentas de Big Data. Em adição, é informado no aludido trabalho que possuir a infraestrutura e as ferramentas pouco têm a acrescentar se o trabalho de Big Data não estiver muito correlacionado com a missão institucional de um determinado órgão público ou governamental. Ou seja, não acrescenta valor se não houver um projeto de Big Data conectado à missão institucional.

Ainda referente à PS2, Khtira *et al.* (2017) citam, ainda, que o Big Data no setor público se torna uma necessidade também em virtude das novas maneiras de a sociedade interagir com os órgãos governamentais. Essa participação popular é capaz de gerar uma enorme quantidade de dados, as quais as organizações públicas podem e devem saber manusear e tratar de forma a obter informação capaz de conduzir políticas públicas e orientar a tomada de decisão. As oportunidades de aplicações do Big Data para o setor público vão desde grandes aplicações em bases de dados de governos nacionais, passando pela aplicação de serviços públicos em comunidades locais, como exemplificado pelo trabalho de Malomo and Sena (2017).

Finalmente, considerando a PS3 (impactos), as dificuldades na implementação do Big Data estão muito ligadas aos desafios citados na PS2, em especial quando se consideram as organizações públicas. Inicialmente, o já citado artigo de Klievink *et al.* (2017) aponta a questão da compreensão da capacidade de implementação, considerando aspectos de infraestrutura computacional. Igualmente, a dificuldade de construir um projeto de Big Data realmente capaz de agregar valor significativo à missão institucional do órgão em questão. O artigo de Malomo and Sena (2017) também enfatiza que as organizações públicas, neste caso representadas por seu corpo técnico-funcional, muitas vezes não dominam as questões sobre como o Big Data pode ser útil no crescimento de eficiência das atividades públicas. Os artigos não se furtam, porém, a citar também dificuldades técnicas, relacionadas, por exemplo, com a multiplicidade de fontes de dados, muitas vezes derivadas de dados não estruturados, principalmente em virtude da forma como são coletados.

A pesquisa, conforme mencionado em epígrafe, encontra considerável relevância, uma vez que as dificuldades de implementação do Big Data podem ser comuns a diversas entidades públicas e até mesmo privadas. As eventuais soluções, por sua vez, em virtude de muitas dificuldades serem comuns, podem se constituir em soluções que auxiliem, portanto, diversas organizações.

Tem-se, portanto, que a principal questão a ser tratada no trabalho ora proposto é: Quais os objetivos e maiores desafios e dificuldades de implantação do Big Data no setor público? Segundo Victorino *et al.* (2017), os objetivos das organizações governamentais, ao participar de um ambiente de Big Data, estão diretamente relacionados com as suas funções de fornecimento de serviços públicos e com a necessidade de provimento de informações aos cidadãos, incluindo questões relacionadas com a transparência de governo, o que corrobora, mesmo que este autor tenha escrito seus trabalhos com vistas ao caso brasileiro, o explicitado pelos citados artigos utilizados na análise das questões descritas.

Na revisão foram encontrados desafios para a implantação de Big Data no setor público (Al-Sai and Abualigah, 2017). Como forma de responder à questão principal deste trabalho, os objetivos e desafios estão classificados em perspectiva, conforme Tabela 9.

Tabela 11. Desafios e possíveis soluções propostas. Adaptado de Al-Sai and Abualigah (2017).

Desafios	Soluções Possíveis
Perspectiva de Tecnologia	<ul style="list-style-type: none"> • Aumentar a capacidade e utilização do armazenamento disponível, como recursos em nuvem para armazenamento e processamento de Big Data. • Implementar questões de segurança questões de política, legislação e regulamentação. • Usar ferramentas eficazes de gerenciamento de dados para analisar dados de estrutura, desestruturação e estrutura simulada de várias fontes. • Fornecer e integrar recursos eficientes de Big Data
Perspectiva de Pessoas	<ul style="list-style-type: none"> • Aprimorar as habilidades do capital humano disponível em Big Data, como análise de dados e segurança de dados; e importar especialistas que podem apoiar a implementação de Big Data. • Implementar uma iniciativa nacional que apoia a colaboração entre Big Data e governo eletrônico. • Aumentar a conscientização sobre Big Data e o impacto no governo eletrônico. • Permitir que os cidadãos sejam mais criativos e se expressem efetivamente nas redes sociais.
Perspectiva de Processos de Negócio	<ul style="list-style-type: none"> • Desenhar uma estratégia eficaz que suporte o roteiro para a implementação de Big Data no governo eletrônico. • Estabelecer parcerias públicas e privadas. • Criar uma rede entre comunidade e governo sob a égide de Big Data. • Formular estratégias eficazes e um roteiro que leve a um ambiente de Big Data. • A liderança deve motivar, envolver, influenciar o suporte à implementação de Big Data no governo eletrônico. • Integrar um grande volume de dados valiosos no governo eletrônico aos processos de tomada de decisão. • Capacitar o papel da e-Liderança e Gerenciamento da informação na formulação do plano estratégico e decisões.

Além dos desafios supracitados, existem ainda outros obstáculos para o tema virar realidade dentro do setor público. Estudos de outros países, como no caso da Turquia, demonstram que o custo, infraestrutura insuficiente, questões de privacidade e segurança, burocracia governamental, falta de legislação sobre o tema, adaptação dos servidores ao processo e a integração dos sistemas de TI existentes são grandes obstáculos para a efetiva adoção de soluções de Big Data (Dogdu *et al.*, 2018).

5. CONCLUSÃO

Este artigo apresenta um mapeamento sistemático sobre o tema de implantação do Big Data no setor público, como pesquisa componente de esforços na verificação de como e porque os órgãos do Governo Brasileiro implantam as aludidas soluções de Big Data.

A consulta de publicações recentes sobre o tema tem fornecido subsídios para formulação das respostas sobre as perguntas de pesquisa, bem como sobre ferramentas e processos para adoção de tais soluções, as quais podem maximizar a produtividade no desenvolvimento de software e implantação de Big Data.

É importante destacar que foram pesquisados e analisados principalmente estudos internacionais, de acordo com a proposta inicial de bases adotadas, assim como dos critérios, tomando como premissa que o caso brasileiro

guarda semelhanças com os casos internacionais já documentados. A semelhança é sugerida nas poucas publicações brasileiras sobre o tema.

As dificuldades levantadas em diversas organizações públicas, ainda que não brasileiras, podem ser comuns a diversos tipos de organização que pretendam implantar Ecossistemas de Big Data no Brasil, tanto públicas, nas diversas esferas de governo, quanto privadas.

A revisão sistemática permitiu que se formulassem direcionadores das futuras pesquisas sobre implantação de Big Data nas organizações públicas brasileiras, organizados pelas perspectivas de tecnologia, de pessoas e de processos de negócio, abordando questões de infraestrutura e conhecimento tecnológico, objetivos estratégicos das organizações públicas e suas missões institucionais, bem como sua forte relação com questões de publicidade e transparência e da eficiência do gasto público.

Devido ao baixo número de artigos publicados sobre o contexto brasileiro, espera-se que este estudo venha a atenuar esse cenário, orientando estudos futuros, que poderão focar em documentar a realidade brasileira e descrever casos junto às organizações governamentais (nos diferentes poderes e nas diferentes esferas governamentais), levantando particularidades relevantes e intrínsecas ao contexto público brasileiro, algo que não foi possível identificar com o mapeamento sistemático reportado neste estudo.

AGRADECIMENTOS

Os autores agradecem o apoio das Agências Brasileiras de Pesquisa, Desenvolvimento e Inovação CAPES (concessão 23038.007604/2014-69 FORTE), CNPq (concessão 465741/2014-2 INCT em Segurança Cibernética) e a FAP-DF (concessões 0193.001366/2016 UIoT e 0193.001365/2016 SSDDC), bem como as cooperações com o Ministério da Economia (TED DIPLA 005/2016), Advocacia Geral da União (TED AGU 697.935/2019) e a CESAR School.

REFERÊNCIAS

- Akoka, J., Comyn-Wattiau, I. and Laoufi, N., 2017. Research on Big Data—A systematic mapping study. *Computer Standards & Interfaces*, 54, 105-115. [Online]. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0920548917300211>.
- Al-Sai, Z. A. and Abualigah, L. M., 2017. Big data and E-government: A review. In 2017 8th *International Conference on Information Technology (ICIT)*, pp. 580-587, IEEE. doi: 10.1109/ICITECH.2017.8080062.
- Almeida, W. H. C., de Aguiar Monteiro, L., de Lima, A. C., Hazin, R. R. and Escobar, F., 2019. Survey on Trends in Big Data: Data Management, Integration and Cloud Computing Environment. *SOFTENG 2019: The Fifth International Conference on Advances and Trends in Software Engineering*, p. 7.
- Araújo Júnior, R. H. and de Sousa, R. T. B., 2016. Estudo do ecossistema de Big Data para conciliação das demandas de acesso, por meio da representação e organização da informação. *Ciência da Informação*, 45(3). [Online]. Disponível em: <http://www.periodicos.ufpb.br/index.php/pbcib/article/view/39276>.
- Barbieri, C., 2001. BI-business intelligence: Modelagem e Tecnologia. Axcel Books, 2001. [On-line]. Disponível em: <https://books.google.com.br/books?id=HD1KAAAACAAJ>.
- Bollier, D., Firestone, C. M. and others, 2010. The promise and peril of big data. Aspen Institute, Communications and Society Program Washington, DC.
- Davenport, T., 2014. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.
- Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113. [Online]. Disponível em: <http://portal.acm.org/citation.cfm?doid=1327452.1327492>.
- Demchenko, Y., De Laat, C. and Membrey, P., 2014. Defining architecture components of the Big Data Ecosystem. In 2014 *International Conference on Collaboration Technologies and Systems (CTS)* (pp. 104-112). IEEE. [Online]. Disponível em: <http://ieeexplore.ieee.org/document/6867550/>.
- Dogdu, E., Ozbayoglu, M., Yazici, A. and Karakaya, Z., 2018. Perceptions, Expectations and Implementations of Big Data in Public Sector, 2018 3Rd *Int. Conf. Comput. Sci. Eng.*, pp. 615-620.
- Schmidt, E., 2010, "Every 2 days we create as much information as we did up to 2003," [Online]. Disponível em: <http://techcrunch.com/2010/08/04/schmidt-data/>.
- Kim, G. H., Trimi, S. and Chung, J. H., 2014. Big-data applications in the government sector, *Communications of the ACM*, vol. 57, no. 3, pp. 78-85. [Online]. Disponível em: <http://dl.acm.org/citation.cfm?doid=2566590.2500873>.
- Gantz, J. and Reinsel, D., 2011. Extracting value from chaos. IDC iVIEW, 1142(2011), 1-12.
- Miller, H. E. and others, 2013. Big-data in cloud computing: A taxonomy of risks, *Information Research*, vol. 18, no. 1, pp. 18-1.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U., 2015. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98-115. [Online]. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0306437914001288>.

- Inmon, W. H. and Linstedt, D., 2014. Data architecture: a primer for the data scientist: big data, data warehouse and data vault. Morgan Kaufmann.
- Jadhav, B., Patankar, A. B. and Jadhav, S. B., 2018. A Practical approach for integrating Big data Analytics into E-governance using hadoop. 2018 *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. doi:10.1109/icicct.2018.8473353.
- Janssen, M. and Van Den Hoven, J., 2015. Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? [Online]. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0740624X15001069>
- Kazemi, U., 2018. A Survey of Big Data: Challenges and Specifications. *CiiT International Journal of Software Engineering and Technology*, Vol 10, No 5.
- Klievink, B., Romijn, B. J., Cunningham, S. and de Bruijn, H., 2017. Big data in the public sector: Uncertainties and readiness. *Information systems frontiers*, 19(2), 267-283. [Online]. Disponível em: <http://link.springer.com/10.1007/s10796-016-9686-2>.
- Laney, D., 2001. Application delivery strategies. META Group, Stamford.
- Malomo, F. and Sena, V., 2017. Data intelligence for local government? Assessing the benefits and barriers to use of in the public sector. *Policy & Internet*, 9(1), 7-27. [Online]. Disponível em: <http://doi.wiley.com/10.1002/poi3.141>
- Mariano, A. M. and Rocha, M. S., 2017. Revisão da literatura: apresentação de uma abordagem integradora. In *AEDEM International Conference* (Vol. 18).
- Mariano, A. M., García Cruz, R. and Arenas Gaitán, J., 2011. Meta análises como instrumento de pesquisa: Uma revisão sistemática da bibliografia aplicada ao estudo das alianças estratégicas internacionais. *Congresso Internacional de Administração* [Online]. Disponível em: <http://rgdoi.net/10.13140/RG.2.1.1150.1522>.
- Márquez, F. P. G. and Lev, B., 2017. Big data management. Springer.
- O'Driscoll, A., Daugelaite, J. and Sleator, R. D., 2013. 'Big data', Hadoop and cloud computing in genomics. *Journal of biomedical informatics*, 46(5), 774-781.
- Khtira, R., Elasri, B. and Rhanoui, M., 2017. From Data to Big Data: Moroccan Public Sector, in *Proceedings of the 2nd International Conference on Cloud and Applications - BDCA'17*. Tetouan, Morocco: ACM Press, pp. 1-6. [Online]. Disponível em: <http://dl.acm.org/citation.cfm?doid=3090354.3090401>.
- Ramos, A., 2015. Infraestrutura Big Data Com Opensource. *Ciencia Moderna*. [On-line]. Available: <https://books.google.com.br/books?id=Y7HjswEACAAJ>.
- Sánchez-Meca, J. and López-Pina, J. A., 2008. El enfoque meta-analítico de generalización de la fiabilidad. [The Meta-analytic approach of reliability generalization], *Acción Psicológica*, vol. 5, no. 2, pp. 37-64. [Online]. Disponível em: <http://revistas.uned.es/index.php/accionpsicologica/article/view/457>.
- Shin, D. H. and Choi, M. J., 2015. Ecological views of big data: Perspectives and issues. *Telematics and Informatics*, 32(2), 311-320. [Online]. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0736585314000665>.
- Victorino, M. C., Shiesl, M., Oliveira, E. C., Ishikawa, E., de Holanda, M. T. and de Lima Hokama, M., 2017. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais concetados. *Informação & sociedade*, 27(1), p. 19.

APÊNDICE B – QUESTIONÁRIO APLICADO

ANÁLISE DO IMPACTO DAS BOAS PRÁTICAS DE GOVERNANÇA DE DADOS ALIADO A *BIG DATA* NA ADMINISTRAÇÃO PÚBLICA FEDERAL

Meu nome é Fernando Gonçalves de Oliveira, sou aluno do Mestrado Profissional em Engenharia de Software no CESAR SCHOOL sob a Orientação do Professor Felipe Silva Ferraz.

A pesquisa tem como objetivo analisar o impacto das boas práticas de Governança de Dados aliado a *Big Data* na Administração Pública Federal.

Sua participação é muito importante e em última análise contribui com o aprimoramento das aplicações de tecnologias da informação no âmbito da Administração Pública Brasileira.

Os resultados serão divulgados sempre em conjunto, sendo assegurado o sigilo concernente a organizações e acerca de respostas individuais.

Caso tenha interesse nos resultados ou qualquer dúvida sobre o trabalho, entre em contato pelo e-mail: fernando.infinite@gmail.com

Agradecemos pela sua importante participação nesta pesquisa!

-
1. Seu cargo na Administração Pública Federal?
 - a) Servidor Público
 - b) Terceirizado
 - c) Outros:
 2. Em qual poder trabalha?
 - a) Executivo
 - b) Executivo (Estatais)
 - c) Judiciário
 - d) Legislativo
 3. Em qual órgão você trabalha?
Sua resposta
 4. Qual o seu tempo de experiência em TI?
 - a) 0 a 5 anos
 - b) 6 a 10 anos
 - c) 11 a 15 anos
 - d) Mais de 15
 5. Atuando em qual Estado?
(lista dos estados)
 6. Você acredita estar familiarizado com as definições e termos correlatos ao *Big Data*?
 - a) Sim
 - b) Não
 7. Qual sua experiência com o assunto *Big Data*?
 - a) 0 a 5 anos
 - b) 6 a 10 anos
 - c) 11 a 15 anos
 - d) Mais de 15
 8. Em seu entendimento o termo *Big Data* depende da volumetria de dados que órgão guarda?
 - a) Sim
 - b) Não
 9. Seu órgão possui ambiente de *Big Data*?
 - a) Sim
 - b) Não
 10. Qual a volumetria de armazenamento do órgão?

- a) Gigabyte (GB)
 - b) Terabyte (TB)
 - c) Petabyte (PB)
 - d) Exabyte (EB)
 - e) Não sei
11. Qual a projeção de crescimento de banco de dados para 2020?
- a) 1 a 100 Gigabyte (GB)
 - b) 1 a 100 Terabyte (TB)
 - c) 1 a 100 Petabyte (PB)
 - d) 1 a 100 Exabyte (EB)
 - e) Não sei
12. Você considera que após o crescimento projetado o seu órgão vai passar a ter um ambiente de *Big Data*?
- a) Sim
 - b) Não
13. Entre os termos a seguir, quais você conhece e sabe diferenciar dos demais
- a) MapReduce
 - b) Hadoop
 - c) Business Intelligence
 - d) Data Mining
 - e) Data Science
14. Você considera que todos os termos acima possuem forte relação com o tema de *Big Data*?
- a) Sim
 - b) Não
 - c) A maioria SIM
 - d) A maioria não
 - e) Não tenho como opinar com segurança
15. Quais os tipos de armazenamento de dados do órgão?
- a) Servidores
 - b) Dispositivos de Storage
 - c) Cloud
 - d) Híbridos
 - e) Outros
16. Possui dados não estruturados?
- a) Sim
 - b) Não
17. Órgão possui acesso a um grande volume de dados para análise, tanto estruturado como não estruturado, numa escala de 0 a 5?
- 0 a 5
18. Existe algum problema de negócio que poderia ser resolvido com *Big Data*?
- a) Sim
 - b) Não
19. Existe a perspectiva de implantar *Big Data* no órgão, numa escala de 0 a 5?
- 0 a 5
20. Sua organização já possui projetos de *Big Data*?
- a) Sim
 - b) Não
21. O que mais tem motivado o uso de *Big Data* no seu órgão público?
- a) Necessidade da missão institucional

- b) Transparência
 - c) Economia de recursos públicos
 - d) Volume de dados
 - e) Avanço tecnológico
 - f) Outros
22. Quais foram as dificuldades encontradas na implementação do ambiente de *Big Data*:
- a) Dificuldades políticas (decisão superior)
 - b) Dificuldades técnicas – domínio de tecnologias
 - c) Dificuldades financeiras
 - d) Dificuldades de alinhamento estratégico com a missão ou políticas do órgão
 - e) Ainda não tenho ambiente de *Big Data*
 - f) Não consigo identificar
 - g) Outros
23. Quais foram os impactos com a implantação de *Big Data*:
- a) Melhoria do processo de tomada decisão
 - b) Melhoria da divulgação de informações – transparência
 - c) Melhoria da eficiência no uso de recursos
 - d) Maior participação do cidadão
 - e) Melhoria do gerenciamento das informações
 - f) Ainda não tenho ambiente de *Big Data*
 - g) Não consigo identificar
24. Quais os impactos econômicos da adoção do *Big Data* em seu órgão?
- a) Não houve nenhum impacto econômico
 - b) Houve aumento de despesa em virtude dos projetos de *Big Data*
 - c) Houve redução de despesas em virtude dos projetos de *Big Data*
 - d) Houve melhoria de eficiência no gasto de recursos
 - e) Houve piora da eficiência no gasto de recursos por dificuldades de uso das tecnologias de *Big Data*
 - f) Não foi possível medir nenhum resultado econômico ou de despesas ou não há projetos de *Big Data* implementados
 - g) Não consigo identificar
 - h) outros
25. Quais as oportunidades de implementação do *Big Data* no seu órgão?
- a) Necessidades da missão institucional
 - b) Transparência
 - c) Economia de recursos públicos
 - d) Participação do cidadão
 - e) Requisitos tecnológicos
 - f) Volumes de dados
 - g) Não há oportunidades identificadas
 - h) outras
26. Quais os impactos que práticas de *Big Data* podem/poderiam trazer ao desenvolvimento de soluções em seu órgão
- a) Impactos financeiros – redução ou aumento de eficiência e eficácia
 - b) Impactos tecnológicos
 - c) Impactos institucionais – missão institucional
 - d) Impactos na participação do cidadão
 - e) Não há impactos identificados até o momento
 - f) outras
27. Os impactos citados na questão anterior são mais positivos que negativos em seu órgão?
- a) Positivo
 - b) Negativo

28. A sua resposta à questão anterior envolveu impactos existentes e comprovados (Sim) ou impactos percebidos pelo conhecimento acerca de *Big Data* conhecido pelo respondente do questionário (Não)?
- a) Sim
 - b) Não
29. O órgão compartilha conhecimentos e aplicações de *Big Data* entre todos os setores do órgão, numa escala de 0 a 5?
- 0 a 5
30. A equipe de TI e profissionais de dados compartilha conhecimentos entre todos os setores da organização, numa escala de 0 a 5?
- 0 a 5
31. No ecossistema de *Big Data* é explorado ou adotado serviços em nuvem para processar dados ou conduzir o *analytics*, numa escala de 0 a 5?
- 0 a 5
32. Quantos profissionais de cientistas de dados o órgão possui?
- Terceirizados: 0 a 5 | 6 a 10 | 11 a 20 | 21 a 30 | 31 a 40 | 41 a 50
- Concursados: 0 a 5 | 6 a 10 | 11 a 20 | 21 a 30 | 31 a 40 | 41 a 50
33. Deseja se identificar, deixe seu nome e contato:
- Sua resposta

APÊNDICE C – QUESTIONÁRIO APLICADO – GRUPO FOCAL

AVALIAÇÃO DE PRONTIDÃO PARA *BIG DATA*

Indique seu grau de concordância em relação às seguintes afirmações referente a Avaliação de Prontidão para *Big Data*.

Escala Likert

1. Discordo totalmente;
2. Discordo um Pouco
3. Não concordo nem discordo
4. Concordo um pouco
5. Concordo totalmente

DADOS

1. No órgão, temos acesso a grandes volumes de dados não estruturados ou os dados chegam em tempo real e em grande velocidade para análise.
2. No órgão, agrupamos dados de várias fontes internas em um *Data Warehouse* ou *Datamart* para facilitar o acesso.
3. No órgão, mantemos definições e padrões consistentes nos dados que usamos para análise.
4. No órgão, usuários, tomadores de decisão e desenvolvedores de soluções para áreas do órgão confiam na qualidade de nossos dados.
5. DADOS: Quais as percepções, desafios e oportunidades?

NEGÓCIO DO ÓRGÃO

6. Empregamos institucionalmente uma combinação de abordagem de *Big Data* e Governança de Dados para atingir as metas do órgão.
7. A nossa gestão assegura que as unidades de negócio e as áreas do órgão trabalhem para estabelecer as prioridades da organização relativas ao *Big Data* e Governança de Dados.
8. Capacitamos nossa equipe de TI e profissionais de análise de dados para fomentar a aprendizagem e compartilhar conhecimentos entre todos os setores da organização.
9. As nossas iniciativas e infraestrutura de *Big Data* e Governança de Dados recebem aportes e recursos adequados para desenvolver as competências e recursos dos quais precisamos.
10. Colaboramos com parceiros de áreas, com Órgãos para os quais prestamos suporte e outros integrantes do nosso ecossistema de negócios para compartilhar conteúdo e aplicações de *Big Data*.
11. NEGÓCIO DO ÓRGÃO: Quais as percepções, desafios e oportunidades?

LIDERANÇA

12. Os dirigentes e líderes que estão na instituição há mais tempo ponderam regularmente sobre oportunidades que o *Big Data* e a Governança de Dados oferecem ao nosso negócio.
13. Os dirigentes e chefes encorajam as unidades e seus líderes a incorporar *Big Data* e Governança de Dados a seus processos decisórios e de negócios.
14. Os dirigentes e líderes que estão na instituição há mais tempo utilizam o *Big Data* e a Governança de Dados para orientar tanto suas decisões estratégicas quanto táticas.
15. Os gestores das áreas do Órgão utilizam o *Big Data* e o *Analytics* para orientar suas decisões.
16. O processo de priorização e alocação de nossos ativos de *Big Data* (dados, pessoal, software e hardware) é orientado e examinado pelo núcleo de altos dirigentes.
17. LIDERANÇA: Quais as percepções, desafios e oportunidades?

METAS

18. Priorizamos nossas iniciativas de *Big Data* e Governança de Dados de acordo com as oportunidades que geram mais valor para nos destacarmos em termos de resultados institucionais.
19. Consideramos novos serviços e funcionalidades baseados em *Big Data* como um aspecto do nosso processo de inovação.
20. Identificamos oportunidades internas para *Big Data* e Governança de Dados avaliando nossos processos, estratégias e os ambientes onde a instituição atua.

21. Conduzimos regularmente experimentos baseados em dados para coletar dados sobre o que está – ou não – se revelando eficaz no nosso negócio.
22. Avaliamos as decisões que tomamos com apoio do *Analytics* e dos dados para verificar se novas fontes de dados não estruturados poderiam proporcionar modelos melhores.
23. METAS: Quais as percepções, desafios e oportunidades?

TECNOLOGIA

24. Temos explorado ou adotamos abordagens de computação paralela (por exemplo: *Hadoop*, *MPI*, *Cuda*, *OpenCL* e técnicas de *Dataflow*) para o processamento de *Big Data*.
25. Temos um grau elevado de maturidade no uso de visualização de dados para esclarecer questões ou decisões de negócio.
26. Temos explorado ou adotado serviços em nuvem para processar dados ou conduzir a Governança de Dados.
27. Temos explorado ou adotado software de código aberto para o *Big Data* e o *Analytics*.
28. Temos explorado ou adotado ferramentas para processar dados não estruturados, como texto, vídeo ou imagens.
29. TECNOLOGIA: Quais as percepções, desafios e oportunidades?

EQUIPE DE TI E ANALISTAS DE DADOS

30. Temos um número suficiente de potenciais cientistas de dados e profissionais de *Analytics* competentes para atingir nossos objetivos analíticos.
31. Nossos profissionais de tecnologias e de *Analytics* atuam como consultores de confiança aos nossos gestores, ajudando-os em tomadas de decisão importantes e na inovação orientada a dados.
32. Nossos profissionais de tecnologias e de Governança de Dados conhecem as disciplinas e os processos de negócio aos quais o *Big Data* e *Analytics* estão sendo ou podem ser aplicados.
33. Nossos profissionais de tecnologias e profissionais de gerenciamento de dados trabalham com eficácia em equipe na condução de projetos de *Big Data*, *Analytics* e Governança de Dados.
34. Temos programas (internos ou em parceria com organizações externas) para desenvolver ciência de dados e habilidades analíticas em nossos profissionais.
35. EQUIPE DE TI E ANALISTAS DE DADOS: Quais as percepções, desafios e oportunidades?