UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA – PATOS DE MINAS
ENGENHARIA ELETRÔNICA E DE TELECOMUNICAÇÕES

CAROLINE SILVÉRIO ROSA

Estudo sobre as técnicas e métodos de análise de dados no contexto de *Big Data*

Patos de Minas - MG
2018

CAROLINE SILVÉRIO ROSA

Estudo sobre as técnicas e métodos de análise de dados no contexto de *Big Data*

Trabalho de Conclusão de Curso, apresentado à banca examinadora como requisito parcial de avaliação da disciplina de TCC2 da graduação em Engenharia Eletrônica e de Telecomunicações, da Faculdade de Engenharia Elétrica, da Universidade Federal de Uberlândia, Campus Patos de Minas.

Orientador: Prof. Dr. Pedro Luiz Lima Bertarini

Patos de Minas - MG

2018

CAROLINE SILVÉRIO ROSA

Estudo sobre as técnicas e métodos de análise de dados no contexto de *Big Data*

Trabalho de Conclusão de Curso, apresentado à banca examinadora como requisito parcial de avaliação da disciplina de TCC2 da graduação em Engenharia Eletrônica e de Telecomunicações, da Faculdade de Engenharia Elétrica, da Universidade Federal de Uberlândia, Campus Patos de Minas.

Orientador: Prof. Dr. Pedro Luiz Lima Bertarini

Aprovado em: ___/___/____

Banca Examinadora:

Prof. Dr. Pedro Luiz Lima Bertarini

Prof. Dr. Daniel Costa Ramos

Prof. Dr. Davi Sabbag Roveri

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Eunice e Vicente, que nunca mediram esforços para que eu chegasse até aqui e que se sacrificaram em prol do meu sucesso e da minha felicidade.

Aos meus irmãos, Isabella e Gustavo, pelo carinho e cumplicidade.

À toda minha família e amigos, pelo incentivo e confiança.

Ao meu orientador Prof. Dr. Pedro Luiz Lima Bertarini pela paciência, pelos aconselhamentos e por sua imensa ajuda durante todos esses meses.

Aos colegas de UFU, pelo companheirismo nessa caminhada cheia de desafios e sacrifícios.

À toda comunidade da Universidade Federal de Uberlândia – Campus Patos de Minas, por todo conhecimento compartilhado.

"O mundo orientado por dados estará sempre ligado, sempre rastreando, sempre monitorando, sempre ouvindo e sempre observando - porque estará sempre aprendendo."

(David Reinsel)

RESUMO

O desenvolvimento da tecnologia de informação (TI), a ascensão da Internet da Coisas (IoT), o avanço da computação em nuvem e o uso massivo das mídias sociais possibilitou uma geração de grandes volumes dados, com perspectiva de alto crescimento deste número nos próximos anos. Esses dados possuem diversas fontes e características, sendo muitas vezes, de natureza não-estruturada. Nesse cenário, surge a análise de *Big Data* (BD), como um método de análise capaz de adquirir informações válidas de variados conjuntos de dados e assim auxiliar na tomada de decisões. As aplicações de BD são diversas, com destaque no marketing de empresas, no comércio online, na indústria, na saúde, na segurança, nas grandes companhias de tecnologia de informação, dentre outros. Este trabalho ressalta a importância do *Big Data* na sociedade atual, estudando seu histórico, suas perspectivas futuras e seus desafios (como a complexidade e a veracidade dos dados, a privacidade do usuário e a gestão dos resultados). Também é discutido sobre as ferramentas e técnicas dentro do contexto de análise *Big Data*, fazendo um levantamento de artigos e suas justificativas para utilização de cada aplicação.

Palavras-chave: *Big Data*, Análise de Dados, Mineração de Dados.

ABSTRACT

The development of Information Technology (IT), the rise of Internet of Things (IoT), the improvement of cloud computing, and the high use of social media caused the generation of a large volume of data, with a massive growth expected for the next years. This data has several sources and characteristics, being often unstructured. In this scenario, the Big Data (BD) analytics emerges as a method capable of acquiring valid information from various datasets, and improving decision making. There are a lot of applications of BD, emphasizing in marketing, e-commerce, industry, health-care, security, information technology companies, etc. This paper highlights the importance of Big Data in today's society, history, future perspectives, and challenges (like the complexity and the veracity of the data, the user privacy, and the management of the results). It is also discussed about tools and techniques for Big Data analysis, doing a survey of articles and their reasons to use each application.

Keywords*: Big Data, Big Data Analytics, Data Mining.*

LISTA DE FIGURAS

LISTA DE QUADRO

LISTA DE ABREVIATURAS E SIGLAS

ANN – *Artificial Neural Network* – Rede Neural Artificial

API – *Application Programming Interface* – Interface de Programação de Aplicativos

AR – *Augmented Reality* – Realidade Aumentada

BD – *Big Data*

BDA – *Business Analytics* – Análise de Negócios

BI – *Business Intelligence* – Inteligência de Negócios

BOT – Robô

GPS – *Global Positioning System* – Sistema de Posição Global

ETL – *Extract, Transform, Load* – Extrair, Transformar e Carregar

E/S – Entrada/Saída

HDFS – *Hadoop Distributed File System* – Sistema de arquivos distribuídos do Hadoop

IDC – *International Data Corporation* – Corporação Internacional de Dados

IoT – *Internet of Things* – Internet das Coisas

ML – *Machine Learning* – Aprendizagem de Máquina

PPDM – *Privacy Preserving Data Mining* – Mineração de Dados com Preservação da Privacidade

RFID – *Radio-Frequency Identification* – Identificação de Rádio Frequência

SNA – *Social Network Analysis* – Análise de Redes Sociais

SQL – *Structured Query Language* – Linguagem de Consulta Estruturada

TI – Tecnologia da Informação – *Information Technology*

TV – Televisão

VR – *Virtual Reality* - Realidade Virtual

WEKA – *Waikato Environment for Knowledge Analysis* – Ambiente Waikato para Análise de Conhecimento

LISTA DE SÍMBOLOS

EB *Exabytes*

PB *Petabytes*

ZB *Zettabytes*

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

Com o crescente avanço da tecnologia, a ascensão das mídias sociais (Facebook, Twitter, blogs e etc.) e o surgimento da Internet das Coisas (IoT), uma enorme quantidade de dados são gerados e registrados, em formato estruturado ou não estruturado [1]. Esse grande fluxo de informação tem despertado interesse de diversos estudiosos, corporações, indústrias e governo, que viram na análise de dados, diversas oportunidades no comércio, inovação e engenharia social [2].

Segundo REINSEL et al., estamos vivendo na chamada Era dos Dados. Os crescentes avanços tecnológicos levam a mudanças essenciais na maneira que a população mundial vive, trabalha e se diverte. Temos o surgimento de objetos inteligentes, o desenvolvimento da robótica, da impressão 3D, dentre as mais diversas inovações que fomentam a geração de dados no universo digital. Extrair e trazer simplicidade dos milhões de bytes gerados não se torna apenas uma oportunidade, mas uma necessidade no mundo atual e que está por vir [3].

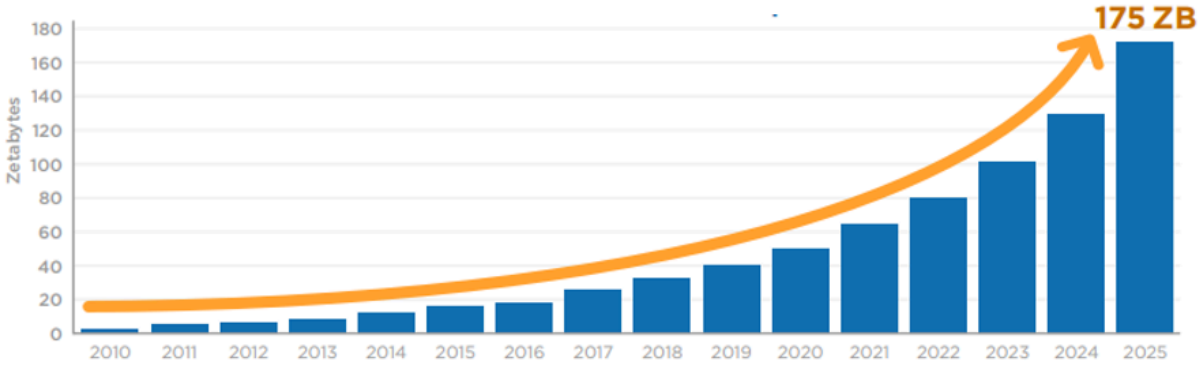
A rápida expansão no volume de dados e sua característica muitas vezes não estruturada, torna os meios tradicionais não suficientes para gerenciamento e análise. Não é mais possível endereçar toda a informação em linhas e colunas de bancos de dados convencionais [4]. Nesse cenário, surge então o conceito de *Big Data*, que tem como objetivo transformar dados imperfeitos e complexos em conhecimento útil [5].

Constituindo como um método de análise de dados mais inteligente e eficaz, *Big Data* tem sido usado pelas organizações como forma de adquirir informações válidas, que serão proveitosas para o negócio de alguma maneira [5]. As empresas utilizam os resultados obtidos para entender melhor e mais a fundo o mercado, criar novos produtos e serviços, e responder a mudança de padrões quase em tempo real [4].

Além do setor empresarial, governos também têm se interessado nas possibilidades de *Big Data*. É possível revelar percepções e comportamentos padrões de comunidades, por meio de análise de dados digitais e arquivos gerados sobre o que dizemos e fazemos usando diariamente dispositivos digitais [5]. Isso pode ser útil no gerenciamento de cidades, na área da saúde, na segurança pública e em diversos outros setores de interesse do cidadão, ou de políticos, como é o caso do uso de *Big Data* nas campanhas eleitorais [6].

A chamada *Global DataSphere*, ou Esfera de Dados Global, que engloba todos os dados criados, capturados e replicados no mundo, em qualquer ano, crescerá de 33 zettabytes em 2018 para 175 zettabytes até 2025 [7]. Isso significa um aumento significativo de informações e conteúdos disponíveis na rede. Segundo GANTZ, o investimento em infraestrutura no domínio digital e das telecomunicações crescerá 40% entre 2012 e 2020, sendo *Big Data* uma das áreas chave de investimento [8]. A Figura 1 mostra a previsão do crescimento da Esfera de Dados Global, de 2010 a 2025.

Figura 1 - Tamanho Anual do *Global Datasphere*, de 2010 até 2025.



Fonte: [7]

O conceito de *Big Data* está cada dia mais tendo espaço e despertando interesse de companhias e profissionais da área de tecnologia. Surgiram então, diversas técnicas de análise de dados diferentes, o que pode ser, em um primeiro momento, problemático. O grande volume de informação e plataformas de estudo e análise de *Big Data* deixa o processo de escolha confusa, tornando um grande desafio saber qual a técnica e ferramenta ideal para cada objetivo ou contexto. A análise de dados é utilizada por diversos setores da sociedade, que possuem objetivos totalmente distintos entre si. Por isso, não existe apenas uma fórmula ou solução para a obtenção de resultados no contexto de *Big Data*. Cada situação e conjunto de dados requer uma técnica diferente de análise, o que influencia diretamente no êxito do estudo.

Neste contexto, o estudo das ferramentas e plataformas de *Big Data* se torna essencial para o segmento tecnológico. Este trabalho busca apresentar e explorar o conceito de *Big Data*, as ferramentas e métodos disponíveis para estudo e as plataformas utilizadas para o desenvolvimento de análise *Big Data*. Assim, o trabalho a ser elaborado tem como tema o estudo sobre técnicas de análise de dados no contexto de *Big Data*.

1.1. OBJETIVOS

Este trabalho tem como objetivo estudar conceitos de *Big Data*, as técnicas disponíveis e analisar as melhores opções de ferramentas para diferentes finalidades.

1.1.1. Objetivos Gerais

Apresentar o tema Big Data e suas técnicas de análise de dados.

1.1.2. Objetivos Específicos

- Introduzir conceitos de *Big Data*, histórico e sua importância na sociedade, fornecendo uma base de estudo sobre o tema.
- Apresentar ferramentas e plataformas disponíveis para análise de dados, que foram utilizados e descritos em artigos.
- Discutir motivos pelos quais diferentes ferramentas de *Big Data* são utilizadas em situações diversas.

1.2. JUSTIFICATIVAS

Com o avanço da tecnologia nas telecomunicações, em especial da telefonia móvel e da Internet, estamos cada vez mais conectados e gerando mais informações aptas para análise, no universo digital. Segundo GANTZ, a maioria dessas informações é produzida assistindo TV digital, interagindo com mídias sociais, enviando imagens e vídeos de celulares para a Internet, e assim por diante [8].

Contudo, de acordo com LUVIZAN, o volume de pesquisas acadêmicas em *Big Data* no Brasil ainda é consideravelmente pequeno [6]. Levando em consideração a importância do

Big Data para a sociedade atual e seu papel essencial no futuro das telecomunicações, este trabalho procura fomentar a base de estudo acerca do tema, passando por sua importância e apresentando as principais técnicas disponíveis atualmente sobre o assunto.

1.3. CONSIDERAÇÕES FINAIS

Foi apresentado no Capítulo 1 uma visão geral sobre o que foi pesquisado e estudado no Trabalho de Conclusão de Curso 2, passando pelos motivos da escolha do tema e objetivos a serem alcançados. No Capítulo 2 será apresentado um referencial teórico sobre o assunto do projeto, passando pelo histórico de *Big Data* e seus conceitos principais. No Capítulo 3 será abordado as diversas técnicas e métodos de análise no contexto de *Big Data*, fazendo uma descrição de cada uma e mencionando suas vantagens e desvantagens. No Capítulo 4, o trabalho será concluído, trazendo uma visão geral dos assuntos abordados durante a pesquisa.

CAPÍTULO 2

DESENVOLVIMENTO TEÓRICO

Big Data é uma expressão relativamente nova, porém, não representa uma novidade [9]. De acordo com LUVIZAN, apesar de existir registro de pesquisas sobre grandes volumes de dados datados ainda na década de 70, foi nos anos 2000 que a modalidade se expandiu drasticamente. Isso se deve a um grande desenvolvimento nas técnicas de processamento, armazenamento e transmissão de dados e, principalmente, no aumento na geração de dados, que chegou a volumes extremamente maiores dos vistos anteriormente [6].

REINSEL [3] descreve a evolução da computação em relação ao processamento de dados em três grandes períodos:

- Antes de 1980, quase todas as informações eram armazenadas em *datacenters* dedicados. Mesmo existindo o acesso a dados por meio de um dispositivo terminal localizado remotamente, a capacidade de computação desse dispositivo era insignificante. Os dados e o poder de processamento estavam concentrados no *mainframe*[1] e seu uso era quase que exclusivamente para fins comerciais.

- Entre 1980 a 2000, a ascensão dos computadores pessoais e a Lei de Moore[2] democratizaram o poder da computação e da transferência de dados. O *datacenter* não apenas armazenava dados em um hub, mas os gerenciava e transmitia para dispositivos de terminal. Isso impulsionou o surgimento das indústrias de entretenimento digital, como músicas, filmes e jogos.

- A partir dos anos 2000, com a popularização e a evolução das redes banda larga *wireless* e por meio da Amazon, Google, Microsoft e outros provedores de serviços populares, o *datacenter* se estendeu à arquitetura em nuvem. O poder computacional continuou a evoluir, principalmente com surgimento de novos dispositivos, como telefones

[1] Computador de grande porte, usado para hospedar bancos de dados comerciais, servidores de transações e aplicativos que exigem um maior grau de segurança e disponibilidade. [59]
[2] A Lei de Moore diz que o número de transistores incorporados em um chip será aproximadamente o dobro a cada 24 meses [60].

celulares, consoles de jogos e objetos inteligentes, que por sua vez, geram uma quantidade enorme de dados [3].

A Figura 2 ilustra e resume os três períodos mencionados acima:

Figura 2 - Evolução da computação.

Fonte: Adaptado de [3].

Quase 30% dos dados do mundo precisarão de processamento em tempo real [7]. Padrões em mídias sociais, cruzamento de informações, correlação de dados, informações médicas, tudo isso pode gerar informação válida. Nessas circunstâncias, surge o *Big Data*, que torna possível extrair informações de grandes bases de dados ainda inexplorados no universo digital [8].

Apesar de inicialmente ter sido usada para referir a apenas “um grande volume de dados”, o termo *Big Data* atualmente vem recebendo outras características. Segundo MAÇADA, “passou também a se referir ao grande volume de informações estruturadas e não estruturadas originadas de diversas fontes, com os quais as organizações deveriam fazer uso, visando à melhoria do processo decisório” [9].

Nem todos os dados são necessariamente úteis para *Big Data*. Alguns tipos de dados são melhores para esse tipo de análise, como por exemplo filmagens de câmera de segurança, dispositivos médicos, mídias sociais e imagens postadas na internet [8]. Na Figura 3, temos a porcentagem da quantidade de dados que podem ser úteis quando caracterizados e analisados, nos anos de 2013 e 2020, de acordo com a IDC (*International Data Corporation*).

Figura 3 - Oportunidade de Big Data.

Fonte: Adaptado de [10]

Sobre as oportunidades no contexto de *Big Data*, GANTZ diz:

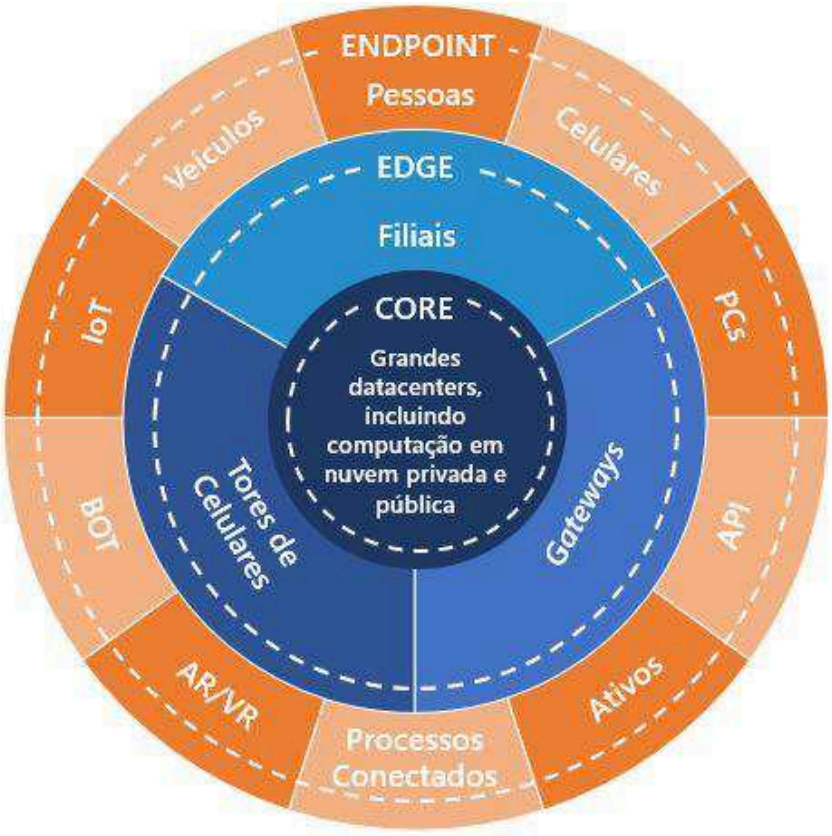
> A grande maioria dos novos dados gerados não é estruturada. Isso significa que, na maioria das vezes, sabemos pouco sobre os dados, a menos que sejam de alguma forma caracterizados ou "etiquetados" - uma prática que resulta em metadados. Os metadados são um dos subsegmentos que mais crescem no universo digital, embora ainda constituindo uma pequena parte do universo digital em geral. [8].

"A IDC definiu três locais principais onde o conteúdo digital é criado: *Core*, que são os tradicionais *datacenters* e em nuvem; *Edge*, a infraestrutura empresarial reforçada, como torres de celular e filiais; e os *Endpoints*, PCs, *smartphones*, e dispositivos IoT" [7]. Esses três locais estão ilustrados na Figura 4.

O *Core* consiste de *datacenters* corporativos e de nuvem, privados e públicos. Um dos principais impulsionadores do crescimento do *Core* é o desenvolvimento da computação em nuvem, que passa a ser cada vez mais uma opção para empresas, em detrimento aos *datacenters*

tradicionais. No *Edge*, temos servidores e dispositivos que são protegidos por corporações, porém não estão em *datacenters* centrais. Temos neste local *datacenters* menores para tempos de resposta mais rápidos, torres de celular e *Gateways*. Os *Endpoint* são os dispositivos na extremidade da rede. Dentre eles, temos PCs, celulares, API (Interface de Programação de Aplicativos), ativos (bens, valores), processos conectados, realidade aumentada e realidade virtual, BOT (robôs), IoT (Internet das Coisas), veículos e pessoas [7].

Figura 4 - Propagação dos dados do *Core* até *Endpoint*, e o inverso.



Fonte: Adaptado de [7].

JIN et al [11] descreve a significância de Big Data em várias perspectivas:

- No desenvolvimento nacional, se tornará um novo marco da força de um país, esperando-se que competições econômicas e políticas entre os países se baseiem, além de aspectos tradicionais, na exploração do potencial do *Big Data*;

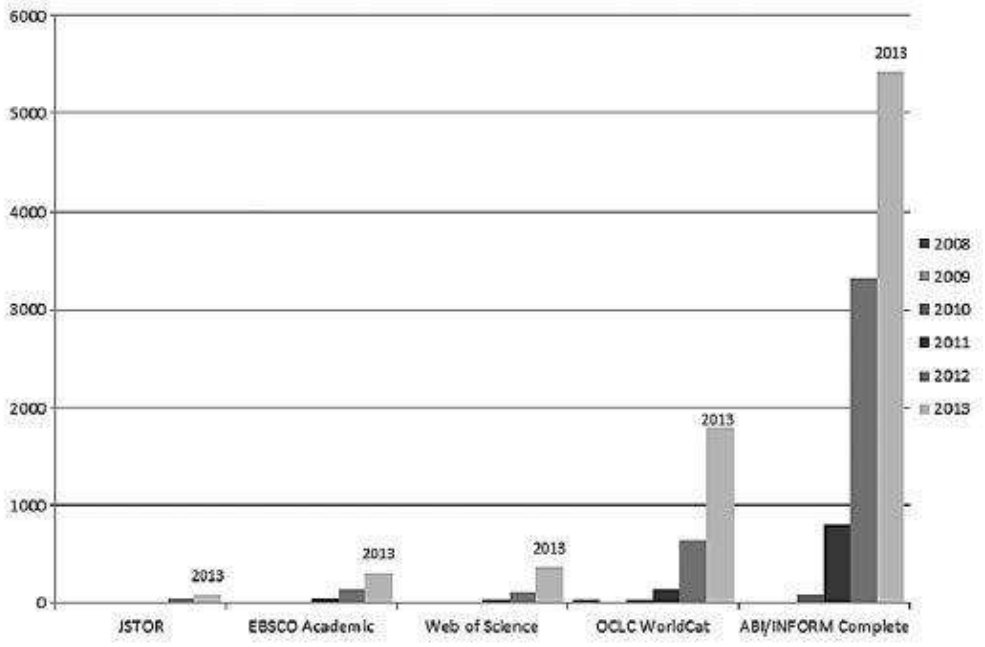
- Nas indústrias, é uma ferramenta chave para melhorar a competitividade das empresas, que buscam extrair informações, conhecimento e até inteligência com *Big Data*. É o foco da nova geração de TI e suas aplicações;
- Na pesquisa científica, pesquisadores podem usar *Big Data* para encontrar ou extrair informações necessárias, sem necessidade de acessar diretamente os objetos a serem estudados;
- Na pesquisa interdisciplinar, trazendo o *Big Data* como o objeto de pesquisa. Visa generalizar a extração do conhecimento dos dados e constitui uma disciplina interdisciplinar chamada *Data Science*;
- Ajudando pessoas a perceber melhor o presente, uma vez que fornece uma riqueza de informações sobre a sociedade. Isto pode nos fazer entender melhor as recentes relações em comunidade, além de ajudar na tomada de decisões;
- Ajudando pessoas a perceber e prever melhor tendências para o futuro. Essa análise tem sido aplicada para abordar questões sociais, incluindo na saúde pública e no desenvolvimento econômico [11].

No meio acadêmico, o número de publicações, eventos e iniciativas dedicadas ao tema tem crescido constantemente. A Figura 5, desenvolvida por EKBIA et al., apresenta número de publicações dentro de cinco banco de dados acadêmicos diferentes, nos anos de 2008 a 2013, com Big Data no título ou como palavra-chave [2].

LUVIZAN detectou em sua pesquisa, um início significativo do interesse sobre o tema *Big Data* a partir de 2011, tanto nas categorias acadêmicas como nas não acadêmicas, se mantendo a partir de então numa crescente, denominando o que se conhece como “onda de interesse” pelo assunto [6].

A IDC, em seu Guia Mundial de Gastos em Análise de Dados e *Big Data*, prevê que as receitas mundiais para soluções de *Big Data* e *Business Analytics* (BDA) atingirão US$260 bilhões em 2022. Isso significa uma taxa de crescimento anual de 11,9%, entre os anos de 2017 e 2022. As receitas do BDA deverão totalizar US$166 bilhões em 2018, um aumento de 11,7% em relação a 2017. Mais da metade de todas as receitas do BDA serão destinadas a serviços de TI. Os investimentos em software crescerão para mais de US$90 bilhões em 2022. As compras de servidores para armazenamento para BDA crescerão em 7,3%, atingindo quase US $27 bilhões em 2022 [12].

Figura 5 - Número de publicações com a frase "Big Data", em cinco bancos de dados acadêmicos entre 2008 e 2013.



Fonte: [2]

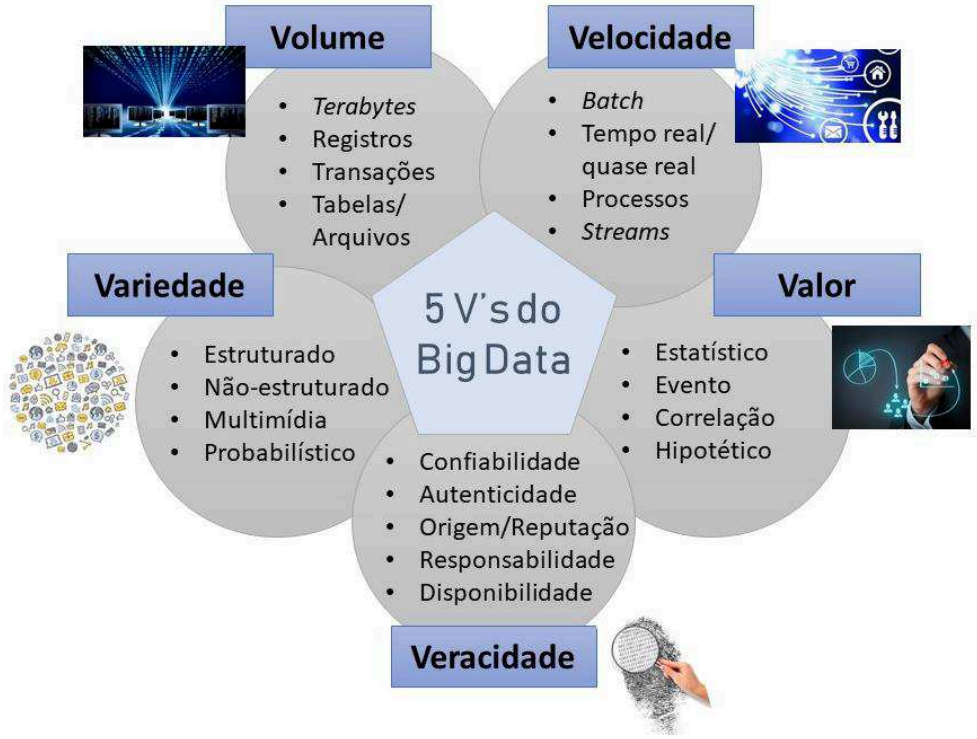
2.1. OS 5 V'S DO BIG DATA

Os dados a serem analisados em Big Data, são caracterizados por algumas dimensões. A magnitude dos dados gerados, a rapidez na qual os dados são gerados constantemente e a diversidade deles formam o que é chamado de três Vs: Volume, Velocidade e Variedade [13]. Posteriormente, foram adicionados os conceitos de Veracidade e Valor, passando a formar os cincos V's do Big Data, como pode ser visto na Figura 6 [14].

O Volume de Big Data, que se refere à dimensão dos dados, é medido atualmente em *petabytes* (PB), *exabytes* (EB) ou *zettabytes* (ZB). A Velocidade, ou ainda rapidez na criação de dados, dá ao processo de Big Data a capacidade de tomada de decisão em tempo quase real. A Variedade informa se os dados têm características estruturadas e não-estruturadas, dando a possibilidade de trabalhar, além de dados tradicionais, dados comportamentais, dados probabilísticos e multimídia [13] [14]. A Veracidade é a preocupação constante em manter os dados confiáveis. Deve-se verificar a origem dos dados e sua autenticidade [14]. Uma vez que o Volume, a Velocidade e a Variedade de dados aumentam constantemente, é importante estar

atento a qualidade da informação [15]. Já Valor, refere-se ao processo de descobrir, destacar e dar valores a informações, que podem estar ocultas ou não, em meio a grandes conjuntos de dados distintos [16].

Figura 6 - Os 5 V's do Big Data.



Fonte: Adaptado de [14]

ZHOUA et al organizou essas cinco dimensões em uma pilha, como pode ser visto na Figura 7. A camada "*Big*" é a mais fundamental e a camada "*Data*" é a central para *Big Data*. A camada inferior, que contém "Volume" e "Velocidade", depende mais fortemente dos avanços tecnológicos. A camada superior, "Valor", é orientada para as aplicações e seus impactos no mundo real, aproveitando o poder estratégico do *Big Data* [17].

Figura 7 - Pilha de Big Data

Fonte: Adaptado de [17]

2.2. *BIG DATA* E A INTERNET DAS COISAS (IoT)

Uma taxa enorme de objetos está sendo conectada à Internet, compreendendo o que é conhecido hoje como "Internet das Coisas" (*Internet of Things* ou IoT) [18]. Dentre esses objetos temos sensores, bancos de dados e outros dispositivos ou *software*. Há muitos domínios nos quais IoT facilita a vida humana de maneira notável, incluindo assistência médica, automação, transporte e respostas emergenciais a desastres naturais [18]. Os diversos sensores produzem diferentes tipos de características, como por exemplo as *tags* de identificação por radiofrequência (RFID), que informam localização e tempo, os GPS's, que identificam localização e marcapassos que capturam informações sobre o coração [19].

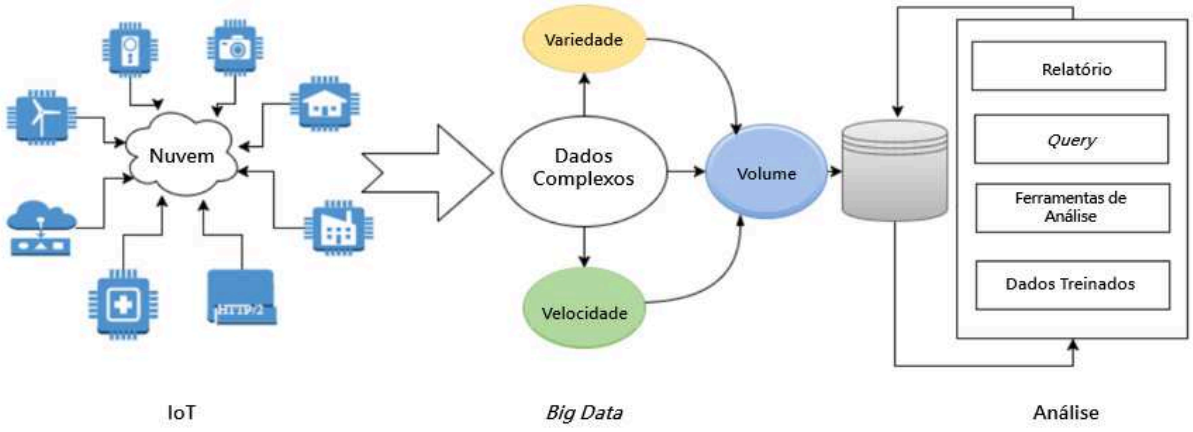
Os objetos de IoT tem incorporado a si uma interface de rede, permitindo que as comunicações entre eles forneçam diversos serviços para os usuários [20]. Muitas oportunidades são apresentadas pela capacidade de analisar e utilizar grandes quantidades de dados de IoT, incluindo aplicativos em cidades inteligentes, sistemas inteligentes de transporte e de rede, medidores inteligentes de energia e dispositivos remotos de monitoramento de saúde do paciente [21].

> Hoje, mais de 5 bilhões de consumidores interagem com dados todos os dias. Até 2025, esse número será 6 bilhões, ou 75% da população mundial. Em 2025, cada pessoa conectada terá pelo menos uma interação com dados a cada 18 segundos. Muitas dessas interações ocorrerão graças aos bilhões de dispositivos IoT conectados em todo o mundo, que deverão criar mais de 90 ZB de dados em 2025 [7].

O crescimento de dados produzidos via IoT desempenhou um papel importante no cenário de Big Data [21]. Segundo O'LEARY, uma enorme parcela dos dados tem origem atribuída aos objetos do universo da Internet das Coisas, visto que os mesmos geram grandes Volumes de dados. A Velocidade dos dados associados a IoT é muito maior do que com o processamento tradicional, pois os sensores podem capturar dados continuamente. Esses dados também possuem grande Variedade, uma vez que temos cada vez mais diferentes tipos de sensores e diferentes fontes de dados. Por último, temos que a Veracidade dos dados está cada dia mais confiável, à medida que a qualidade do sensor e outros dados melhoram com o tempo. Esses fatores associados impulsionam a geração de *Big Data* pela Internet da Coisas [19].

Na Figura 8, temos ilustrado a relação de IoT com *Big Data.*

Figura 8 - Relação entre IoT e análise de *Big Data*.



Fonte: [21]

MARJANI divide a figura em três etapas. A primeira consiste em gerenciar as fontes de dados de IoT, que podem ser as mais diversas, como câmeras, objetos domésticos inteligentes, sensores hospitalares, etc. Esses dados podem ser armazenados na nuvem. Na segunda etapa, os dados gerados são chamados de *Big Data* e são baseados em Volume, Velocidade e Variedade. A última etapa aplica ferramentas para analisar os grandes conjuntos de dados de IoT armazenados. Os quatro níveis de análise começam com os Dados de Treinamento (conjunto de dados rotulados utilizados para a análise), passam para as Ferramentas de Análise, *Query* (consulta ao banco de dados) e por fim, a geração de Relatórios [21].

2.3. COMPUTAÇÃO EM NUVEM

Computação em nuvem (*cloud computing*) é uma arquitetura poderosa e extremamente bem-sucedida, que revolucionou a infraestrutura da computação e permitiu a realização de tarefas complexas e de larga escala [1] [22]. Elimina-se nessa modalidade, a necessidade de se manter *hardware* de computação, espaço dedicado e *software* [1].

Os serviços em nuvem abrangem várias funções de TI, desde armazenamento e computação até bancos de dados e serviços de aplicativos. A necessidade de armazenar, processar e analisar grandes quantidades de conjuntos de dados fez que com organizações e indivíduos adotassem a computação em nuvem [1]. Um grande número de aplicações está sendo atualmente implementado na nuvem. Esse número crescerá devido à falta de recursos da computação em servidores locais, aos custos reduzidos e ao aumento do volume de dados produzidos e consumidos [23]. Em 2025, a IDC prevê que 49% dos dados armazenados no mundo residirão em ambientes de nuvem pública [7].

A Figura 9 é um gráfico da IDC, em parceria com a Seagate, que mostra a relação em porcentagem do uso de *Datacenters* Corporativos comparado ao uso da Nuvem para armazenamento de dados. Percebe-se que a Nuvem teve um grande e rápido crescimento nos últimos anos, com perspectiva de ultrapassar os *Datacenters* Corporativos nos próximos anos. Isso acontece porque as empresas passaram a buscar cada vez mais utilizar a Nuvem (pública e privada) para as necessidades de processamento de dados, se tornando o novo repositório de dados corporativos [7].

Empresas e consumidores estão encontrando na Nuvem uma opção cada vez mais atraente, uma vez que ela permite o acesso rápido e sempre disponível aos dados, mesmo à medida que um maior número de dispositivos com maiores níveis de inteligência esteja conectado a várias redes. O consumidor deixa de preocupar cada vez mais com a capacidade de armazenamento disponível nos seus dispositivos *Endpoint*, e mais em fazer uso da Nuvem [7].

Segundo HASHEM, computação em nuvem e *Big Data* são intimamente correlacionados. O *Big Data* fornece aos usuários a capacidade de usar computação para processar e analisar conjuntos de dados em tempo hábil e a infraestrutura de computação em nuvem pode servir como uma plataforma eficaz para lidar com o armazenamento de dados necessário para realizar análise de *Big Data* [1]. Se por um lado o desenvolvimento da

computação em nuvem fornece soluções para o armazenamento e processamento de *Big Data*, o surgimento de *Big Data* também acelera o desenvolvimento da computação em nuvem [16].

Figura 9 - Dados armazenados em Nuvens públicas vs *Datacenters* Tradicionais.

Fonte: [7]

2.4. APLICAÇÕES IMPORTANTES DE BIG DATA

Como dito anteriormente, a análise *Big Data* tem se destacado mundialmente em diversos seguimentos, uma vez que o estudo de dados traz diversos benefícios para toda a sociedade. É atualmente uma área de interesse de diversas companhias, que cada vez mais investem no processamento e análise de dados. Nesta seção, serão abordados dois importantes conceitos de aplicações de *Big Data* no ambiente de negócios.

2.4.1. *Business Intelligence* (BI)

Business Intelligence (BI) é um termo em TI, surgido no final dos anos 80, usado para referir a uma ampla gama de processos e *softwares* usados para coletar, analisar e disseminar dados, ajudando empresas a entender melhor seus negócios e tomar decisões mais oportunas [24] [25]. As empresas estão aproveitando dados para melhorar as experiências dos clientes, abrir novos mercados, tornar funcionários e processos mais produtivos e criar novas fontes de vantagem competitiva [7].

As ferramentas de BI permitem extrair, transformar e carregar (ETL – *Extract, Transform, Load)* dados para análise e disponibilizar em relatórios, alertas e cartões de pontuação [24]. As oportunidades associadas a dados e análises em diferentes organizações ajudaram a gerar um interesse significativo em BI [25]. Apesar de possuir grandes quantidades de dados, as companhias encontram dificuldade na análise, devido ao caráter muitas vezes redundante e inconsistente dessas informações, tornando-as complexas de administrar [26]. A análise de *Big Data*, além de ajudar a lidar com esses dados complexos, amplia o escopo do BI, que geralmente foca nos bancos de dados internos da empresa, agora buscando extrair valor de dados externos [27].

2.4.2. Big Data aplicado ao *Marketing*

Com o avanço da tecnologia, o consumidor comum é atualmente um gerador de dados, tanto os tradicionais e estruturados, quanto comportamentais, contemporâneos, não estruturados. EREVELLES em seu artigo, propõe o estudo de marketing em *Big Data* em três parâmetros: capital físico, humano e organizacional. Os recursos de capital físico incluem *software* ou uma plataforma de coletar, armazenar ou analisar no contexto de *Big Data*. Os recursos de capital humano incluem a percepção de cientistas e estrategistas que possuem conhecimentos de dados, para capturarem informações das atividades do consumidor, e assim gerenciarem e extraírem percepções importantes e úteis na análise *Big Data*. Por último, os recursos de capital organizacional incluem uma estrutura organizacional que permite à empresa transformar essas percepções adquiridas em ações, podendo ser necessário alterar processos de negócios para agir de acordo com o conhecimento conquistado no processo [13].

2.5. SEGURANÇA E PRIVACIDADE

O desenvolvimento da mineração de dados no contexto de *Big Data* exige uma crescente coleta e processamento de grandes quantidades de dados pessoais, como por exemplo localização, registros criminais, hábitos de compra, histórico médico e de crédito, acarretando cada vez mais na preocupação do público com a privacidade e segurança de dados [1] [28]. A

falta de normas e padrões entre os sites de comércio eletrônico e a sofisticação e a dedicação dos hackers, por exemplo, colocam em risco informações privadas consideráveis [8].

A privacidade é comumente vista como o direito dos indivíduos de controlar informações sobre si mesmos [28]. Considera-se violação deste direito o acesso não autorizado a dados pessoais, a descoberta indesejada de informações embaraçosas, o uso de dados pessoais para fins diferentes daquele para o qual os dados foram coletados, etc [29]. Um campo da mineração de dados, chamado de "mineração de dados com preservação da privacidade" (PPDM – *privacy preserving data mining*), ganhou um grande desenvolvimento nos últimos anos. Tem como objetivo proteger informações confidenciais contra divulgação não solicitada ou não autorizada e, enquanto isso, preservar a utilidade dos dados [29].

REINSEL diz que independentemente de onde os dados vêm, as empresas devem enfrentar o desafio de gerenciar. Apesar dos usuários gerarem seu próprio conteúdo nas mídias sociais, a plataforma da mídia social deverá armazenar os dados em sua infraestrutura. As empresas têm autoridade para acessar e gerenciar esses dados pessoais em crescimento e, portanto, devem assumir maior responsabilidade pelos riscos de privacidade e segurança [3].

Além disso, temos um aumento do número de sensores embutidos, que capturam dados muitas vezes sem o usuário perceber. Isso abre espaço para vazamento de informações pessoais, tornando mais urgentes as exigências de proteção de segurança dos dados. Embora alguns tipos de dados não tenham requisitos de segurança fortes, como fotos de telefones celulares, conteúdo de sites públicos e materiais de código aberto, a maioria das informações possui esse requisito, como informações financeiras corporativas, informações de identificação pessoal e prontuários médicos [3]. REINSEL divide os dados em cinco níveis:

1. Apenas privacidade: por exemplo um endereço de e-mail em um upload do YouTube;
2. *Compliance*[3]: e-mails que podem ser descobertos em litígios ou sujeitos a regras de retenção;
3. Custódia: informação da conta, uma violação que poderia levar ou ajudar no roubo de identidade;

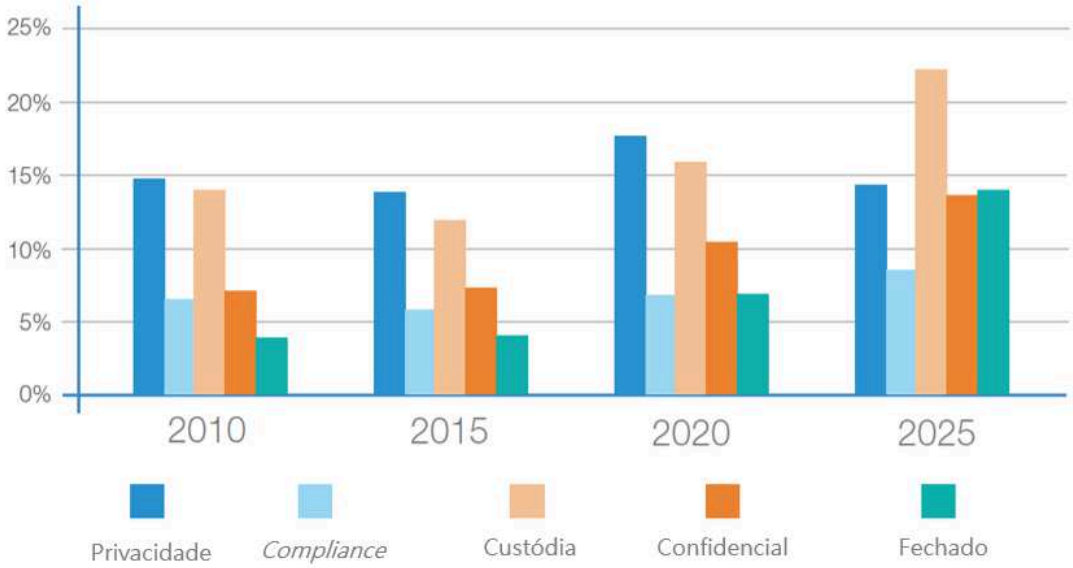
[3] Termo que significa "estar em conformidade com", obedecer, satisfazer o que foi imposto, comprometer-se com a integridade [61].

4. Confidencial: informações que o originador deseja proteger, como segredos comerciais, listas de clientes, memorandos confidenciais, etc;

5. Fechado: informações que exigem a mais alta segurança, como transações financeiras, arquivos pessoais, registros médicos, inteligência militar, etc [3].

A Figura 10 ilustra esses cinco níveis em porcentagens de dados, entre os anos de 2010, 2015, 2020 e 2025. É possível notar observando o gráfico, que o número percentual de dados que precisam de proteção tem um significativo aumento para 2025. Como já mencionado no referente trabalho, a predição feita para 2025 é a *Global Datasphere* chegar a 173ZB de tamanho. Isso significa que será cada vez mais necessário a busca por técnicas e normas relacionadas à segurança de dados.

Figura 10 - Porcentagem de dados que precisam de proteção.



Fonte: [3]

Em termos de Brasil, a IDC diz que em 2014, mais de 40% das informações do universo digital brasileiro que precisava de proteção, não estava sendo protegido, como ilustrado na Figura 11 [30].

Figura 11 - Dados que precisam de proteção, mas não possuem, no Brasil.

Fonte: Adaptado de [30]

Quando se utiliza serviços gratuitos (e-mail, redes sociais, etc), o usuário se torna automaticamente a fonte de dados da empresa, que pode analisar esses dados para melhorar a satisfação do cliente, ou ainda vender a terceiros para análise posterior. PERERA et al diz que provavelmente no futuro, os provedores de serviços oferecerão as seguintes opções ao cliente: o consumidor paga pelo serviço, com a garantia de sua privacidade protegida, ou tem serviço gratuito, porém em troca autoriza a coleta e uso de seus dados [31].

Para GANTZ, essa questão é mais sociológica e organizacional que tecnológica, uma vez que mesmo com o desenvolvimento das soluções, elas serão ineficientes sem a mudança comportamental das empresas. “Existe também uma grande necessidade da padronização entre os sites que salvam, coletam e reúnem informações privadas, para que as informações pessoais dos indivíduos sejam mantidas dessa maneira” [8].

2.6. DESAFIOS DA ANÁLISE *BIG DATA*

A análise de dados, especialmente no ambiente de *Big Data*, conta com vários desafios, como redundância, inconsistência, ruído, heterogeneidade, discretização e categorização de dados, o alto custo com infraestrutura e a necessidade de métodos de análises específica para trabalhar com grandes volumes de dados. [17]. Entre esses desafios, alguns são causados pelas

características do *Big Data*, alguns por seus atuais modelos e métodos de análise, e alguns, pelas limitações dos atuais sistemas de processamento de dados [11].

JIN et al considera como os maiores desafios em *Big Data*: a complexidade dos dados, a complexidade computacional e a complexidade do sistema. Os diferentes tipos de dados, seu grande volume e sua constante e rápida mudança torna sua percepção, representação, compreensão e processamento muito mais complicados, e resulta em aumento na necessidade de maior complexidade computacional. Sistemas de processamento *Big Data*, adequados para lidar com uma diversidade de tipos de dados e aplicativos, são a chave para o sucesso da análise de grandes volumes de dados [11].

ZICARI [32] agrupa os desafios de *Big Data* em três dimensões: dados, processos e gerenciamento:

- Desafios de dados: refere-se às características dos dados, como volume, variedade, velocidade, veracidade, qualidade e disponibilidade dos dados, como encontrar e coletar dados, relevância e etc;
- Os desafios do processo: relacionado às técnicas de como capturar, alinhar e transformar dados, como modelar matematicamente ou por simulação e como entender, visualizar e compartilhar os resultados;
- Desafios de gestão: aspectos de privacidade, segurança, política e ética [32].

Chen et al [16] define os desafios chave para *Big Data*:

- Representação de dados: muitos conjuntos de dados têm determinados níveis de heterogeneidade. A representação de dados tem como objetivo tornar os dados mais significativos para análise computacional e interpretação do usuário.
- Redução da redundância e compactação de dados: geralmente, há um alto nível de redundância nos conjuntos de dados. A redução de redundância e a compactação de dados são eficazes para reduzir o sistema sem que os valores potenciais dos dados sejam afetados.
- Gerenciamento do ciclo de vida dos dados: os valores em *Big Data* dependem da atualização dos dados. Deve ser desenvolvido um princípio de importância dos dados para decidir quais devem ser armazenados e quais devem ser descartados.

- Confidencialidade de dados: a análise de *Big Data* pode ser entregue a um terceiro para processamento somente quando forem tomadas medidas preventivas adequadas para proteger esses dados confidenciais, para garantir sua segurança.
- Gestão de energia: com o aumento do volume de dados e demandas analíticas, o processamento, o armazenamento e a transmissão de *Big Data* consumirão inevitavelmente mais e mais energia elétrica.
- Escalabilidade: o algoritmo analítico deve ser capaz de processar conjuntos de dados cada vez mais complexos e em expansão.
- Cooperação: a análise de *Big Data* é uma pesquisa interdisciplinar, que requer que especialistas em diferentes campos cooperem.

Os artigos possuem alguns pontos em comum. A questões de privacidade e segurança de dados são um grande desafio para *Big Data*. A crescente onda de escândalos relacionadas a venda e uso de dados traz insegurança aos usuários. Dados são usados não mais para apenas o uso comercial, mas para uso antropológico e político. Com estudo comportamental, é possível influenciar e manipular a população para seus interesses. As leis e normas de privacidade devem ser adequadas para trabalhar agora com esse novo mundo de dados. Discussões sobre os limites do uso de dados pessoais estão em alta, se tornando não apenas um problema de falha computacional, mas sim social e ético.

Em relação aos dados utilizados em *Big Data,* os principais desafios estão na sua complexidade e no processo de análise. Novas tecnologias, métodos de análise, ferramentas, *software*, soluções em geral devem ser desenvolvidas a medida que o número de dados aumenta.

2.7. CONSIDERAÇÕES FINAIS

Nesse capítulo, foi apresentado uma base teórica sobre o tema *Big Data*, descrevendo conceitos, histórico e sua importância para a sociedade atual. A análise BD é desafiadora, pois foge da análise tradicional e compreende uma nova maneira de lidar com dados cada vez mais variados e volumosos. No Capítulo a seguir, será discutido sobre técnicas e ferramentas no contexto de Big Data.

CAPÍTULO 3

ESTUDO DAS FERRAMENTAS E TÉCNICAS NO CONTEXTO DO *BIG DATA*

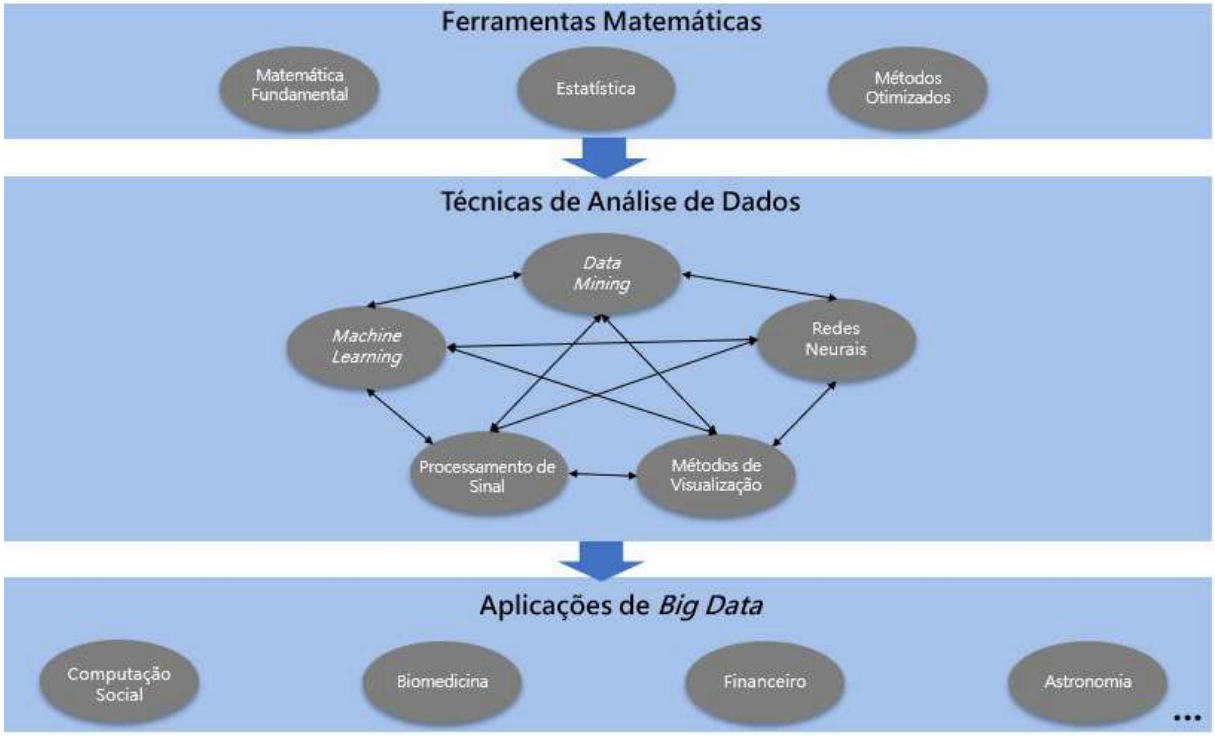
No Capítulo 2, foi apresentado o conceito de *Big Data*, sua relevância e importância na sociedade atual e seus principais desafios. Dentro dos desafios, vimos que a complexidade dos dados e dos processos pode ser um problema. Devido à complexidade dos dados e seu grande volume, a análise *Big Data* tem como desafio a escolha da metodologia a ser seguida para chegar aos objetivos previstos [33]. Neste Capítulo, serão discutidos os métodos e técnicas de análise disponíveis na literatura. O objetivo será entender qual método pode ser mais eficiente de acordo com a finalidade desejada.

A estruturação deste Capítulo será baseada na estruturação seguida proposta por YAQOOB et al [34], que fala em sua pesquisa sobre as ferramentas de processamento disponíveis, as técnicas de análise *Big Data* e as ferramentas de *data mining*. Durante a apresentação das ferramentas e técnicas, serão discutidas suas vantagens e desvantagens.

Segundo PAI, as organizações usam várias técnicas e tecnologias para agregar, manipular, analisar e visualizar *Big Data*. Essas técnicas podem ter origem na estatística, ciência da computação, matemática aplicada e economia. Algumas foram desenvolvidos para uso específico de *Big Data* e algumas foram adaptadas para trabalhar nesse cenário [35]. Na Figura 12 ilustra essas disciplinas nas quais as técnicas Big Data estão inseridas. A análise *Big Data* necessita de técnicas multidisciplinares para processar com eficiência o grande volume de dados em tempos de execução limitados [35]. Algumas dessas disciplinas estão mencionadas abaixo:

- Matemática Fundamental: Uso de técnicas e fórmulas matemáticas clássicas, objetivando a resolução de problemas relacionados a correlação de dados.
- Estatística: coleção de técnicas matemáticas que ajudam a analisar e apresentar dados [36]. A análise estatística e suas decisões são baseadas nas noções de probabilidade, que mensura como o acaso afeta certos eventos ou resultados [37].

Figura 12 - Técnicas de *Big Data*.



Fonte: [35]

- Métodos Otimizados: tem por objetivo minimizar o custo de produção ou maximizar a eficiência da produção. Um algoritmo de otimização é um procedimento que é executado iterativamente comparando várias soluções até que uma solução ótima ou satisfatória seja encontrada [38]. São aplicados para resolver problemas quantitativos em muitas áreas, como física, biologia, engenharia e economia [35].

- *Data Mining*: é um conjunto de técnicas para extrair informações (encontrar padrões) de dados, incluindo análise de agrupamento, classificação, regressão e aprendizado de regras de associação [35].

- *Machine Learning*: conjunto de métodos que podem detectar automaticamente padrões em dados e, em seguida, usar os padrões descobertos para prever dados futuros ou executar outros tipos de tomadas de decisão sob incerteza [39].

- Redes Neurais: *Artificial Neural Network* (ANN) é paradigma de programação de inspiração biológica que permite que um computador aprenda a partir de dados observacionais [40].

- Processamento de Sinal: operar, analisar, sintetizar um sinal de alguma forma, para extrair alguma informação útil [41].

- Métodos de Visualização: são as técnicas usadas para criar tabelas, imagens, diagramas e outras formas de exibição intuitivas para entender os dados [35].

Essas técnicas são usadas para as mais diversas aplicações no contexto de *Big Data*, com menção na Figura 12 para a Computação Social (interações em mídias sociais na internet), Biomedicina, na área Financeira e na Astronomia.

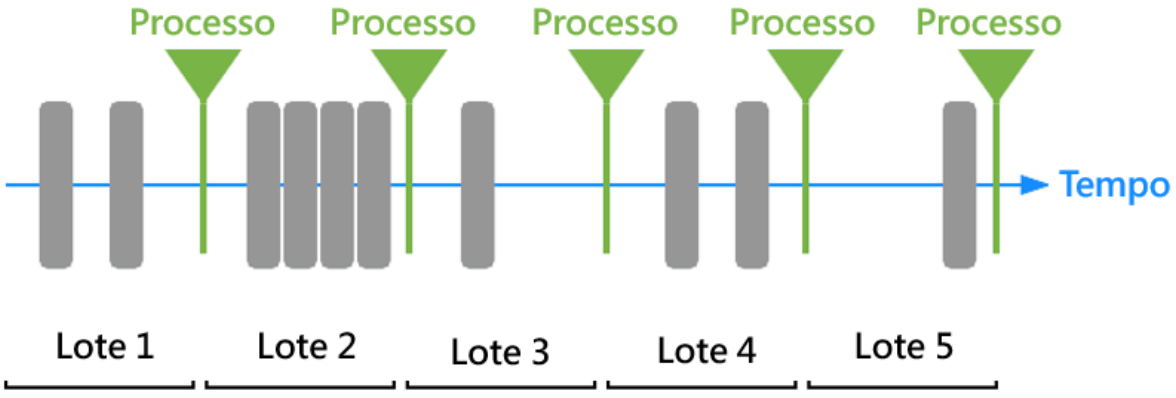
As seções seguintes examinam métodos de processamento de dados importantes, assim como técnicas de análise. Assim, apresenta uma visão geral de como Big Data pode ser utilizado na prática.

3.1. PROCESSAMENTO *BATCH*

O processamento *Batch*, ou em lote, é o processamento de um grande volume de dados de uma só vez. Os trabalhos são normalmente concluídos simultaneamente em ordem sequencial contínua. É uma maneira extremamente eficiente de processar grandes quantidades de dados coletados ao longo de um período de tempo [42]. Historicamente, a grande maioria das tecnologias de processamento de dados foi projetada para Processamento *Batch* [43].

Na Figura 13, temos uma representação de como funciona o Processamento *Batch*. Novos elementos são coletados e agrupados em lotes. Só quando o agrupamento tiver terminado, o lote então é processado [43].

Figura 13 - Processamento *Batch*.



Fonte: [43]

Uma breve comparação de ferramentas de Processamento *Batch*, levando em consideração seus pontos fortes e fracos é apresentada no Quadro 1. Essas técnicas serão descritas nas próximas seções deste capítulo.

Quadro 1 - Comparação de técnicas de Processamento Batch.

Comparação de técnicas de processamento Batch			
Tecnicas de Processamento Batch	Descrição	Pontos Fortes / Vantagens	Pontos Fracos / Desvantagens
Hadoop	Para executar processamento de aplicativos com uso intensivo de dados	- Dados distribuídos - Processamento - Tarefas independentes - Fácil manuseio de falha parcial - Escalonamento linear em casos ideais - Modelo de programação simples	- Modelo de programação restritiva - Junções de vários conjuntos de dados, o que o deixa complicado e lento - Gerenciamento de Cluster Difícil - Nó mestre único - Configuração não óbvia dos nós
Skytree Server	Para processar grandes quantidades de dados em alta velocidade	- Processamento rápido e preciso de quantidades volumosas de conjunto de dados - Análise avançada - Aprendizado de máquina de alto desempenho	- Alta complexidade
Dryad	Para melhorar os programas paralelos e distribuídos e ampliar a capacidade de processamento de um pequeno a um grande número de nós	- Programação mais fácil - Comparado com o MapReduce, é mais flexível - Permite múltiplas entradas e saídas	- Inadequado para programa iterativo e de aninhamento - A conversão de computação irregular em gráfico de fluxo de dados é muito difícil.
Pentaho	Para gerar relatórios a partir de um grande volume de dados estruturados e não estruturados	- Fácil acesso aos dados - Relatórios rápidos devido a técnicas de cache em memória - Visualização detalhada - Integração perfeita	Inconsistente na maneira em que eles trabalham Análise menos avançada em comparação com o Tableau
Tableau	Para processar grandes quantidades de conjuntos de dados	- Ótima visualização de dados - Solução de baixo custo para atualizar - Excelente suporte móvel	- Falta de capacidade de previsão - Segurança arriscada - Problemas de gerenciamento de mudança

Fonte: Adaptado de [34].

3.1.1. HADOOP

O *Hadoop* é uma plataforma de *software* da *Apache Software Foundation*, *open source*, escrito em Java, que permite o processamento de grandes conjuntos de dados em *clusters* de computadores. Ele possui dois componentes principais: a estrutura de programação HDFS e *MapReduce*, que estão intimamente relacionados entre si [44].

- HDFS (*Hadoop Distributed File System*): é um sistema de arquivos distribuído, projetado para armazenar conjuntos de dados muito grandes de forma confiável e transmitir para os usuários. O HDFS é altamente tolerante a falhas e pode ser expandido de um único servidor para milhares de máquinas, cada uma oferecendo armazenamento local. O HDFS consiste em dois tipos de nós, o "Mestre" – gerencia a hierarquia de sistemas, e vários "Escravos" – nós de dados [45].

- *MapReduce*: é um modelo de programação para processar e gerar grandes conjuntos de dados, úteis para atividades no mundo real [46]. Ele funciona nos termos das funções *map* (mapeamento) e *reduce* (reduzir). A função *map* considera o par chave/valor (chave sendo identificador do registro e valor o seu conteúdo) como entrada, e gera pares chave/valor intermediários. A função *reduce* mescla todos os pares associados à mesma chave (intermediários) e gera uma saída [1]. DEAN diz que "mais de dez mil programas *MapReduce* distintos foram implementados internamente no Google nos últimos quatro anos e uma média de cem mil trabalhos *MapReduce* são executados nos *clusters* do Google todos os dias, processando um total de mais de vinte petabytes de dados/dia" [46].

Apesar das muitas vantagens do Hadoop, como processamento de dados distribuídos, tarefas independentes, falhas parciais fáceis de lidar, escalonamento linear e modelo de programação simples, há muitas desvantagens do Hadoop, como o modelo de programação restritiva, junções de vários conjuntos de dados que fazem um difícil e lento gerenciamento de *clusters*, nó mestre único, configuração não óbvia dos nós, dentre outros [34].

Entre os artigos pesquisados para este trabalho, a utilização do Hadoop teve grande destaque, com um grande número de aplicações usando a plataforma.

O artigo "*Big data framework for analytics in smart grids*", de MUNSHI, A. A. e MOHAMEDA, Y. A. I., apresenta uma implementação, utilizando Hadoop, para *smart grids*. Foram comparados dois cenários: para uma única casa e uma *smart grid* que contém mais de

6000 medidores inteligentes. Os autores explicam a escolha do Hadoop para análise de *Big Data* em *smart grids* destacando algumas de suas vantagens e recursos [47]:

- Escalabilidade: o Hadoop permite que a infraestrutura de hardware seja ampliada e reduzida sem a necessidade de alterar os formatos de dados. Se novos bairros ou utilitários de geração forem adicionados à rede elétrica, nós e dispositivos de armazenamento adicionais poderão ser adicionados ao *cluster* existente sem afetar a funcionalidade dos nós existentes.

- Computação eficiente em tempo real: computação maciçamente paralela para servidores *commodity*, levando a uma redução considerável no custo por *terabyte* de armazenamento, o que torna a computação paralela acessível com o crescente volume de dados de *smart grid*, além de permitir tarefas de mineração e previsão de baixa latência em dados de fluxo.

- Flexibilidade: O Hadoop é capaz de absorver vários tipos de dados de inúmeras fontes. Diferentes tipos de dados de várias fontes podem ser agregados para análise posterior. Assim, muitos desafios dos vários tipos de dados da *smart grid* podem ser abordados.

- Tolerância a falhas: dados perdidos e falhas de computação são comuns em dados de *smart grid*. O Hadoop pode recuperar os dados e as falhas de cálculo causadas pela falha no nó ou pelo congestionamento da rede, armazenando os dados em muitos nós e distribuindo o trabalho de computação para outros nós saudáveis no *cluster* [47].

No artigo "*A Big Data Framework for Mining Sensor Data Using Hadoop*", de EL-SHAFEIY, E. A., EL-DESOUKY A. I., é utilizado Hadoop para mineração de *nodesets* frequentes para grandes quantidades de dados de sensores. Os autores utilizaram o Hadoop porque alguns pesquisadores propuseram o uso de MapReduce para lidar com grandes fluxos de dados de detecção, pois ele consegue minerar o espaço de busca de maneira distribuída. O MapReduce usa um sistema de arquivos distribuídos, que é particularmente otimizado para melhorar o desempenho de E/S (entrada/saída) [48].

O artigo "*E-Commerce Trends and Future Analytics Tools*", de KUMAR, P. e CHANDRASEKAR, S., tem como objetivo avaliar as tendências em constante mudança e explorar as tecnologias de informação e ferramentas futuristas do comércio eletrônico [49]. Os autores justificaram o uso de Hadoop para análise em e-commerce dizendo:

> Na crescente complexidade atual do Big Data (textos, comentários, dados de sensores, *emojies*, vídeos, imagens, áudio, etc.), o Hadoop oferece a melhor escolha possível. (...) Extração, processamento, indexação e análise *ad hoc* são recursos exclusivos do Hadoop. Oracle, IBM e Microsoft adotam o Hadoop, incluindo o *open source* Apache Spark [49].

No artigo "*Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics*", de STRANG, K. D. e SUN, Z., o Hadoop foi usado no Google *News* para coletar informações complexas sobre terrorismo. Os autores têm experiência com Hadoop, bem como com as aplicações relacionadas de Ambari, Avro, Cassandra, Chukwa, Flume, Hbase, Hive, Mahout, Pig, Soir, Faísca, Sqoop, Tez, YARN e ZooKeeper. Segundo o artigo, nenhum desses aplicativos relacionados fornece técnicas de análise estatística sem programação considerável. A proposta de pesquisa foi usar o Hadoop para coleta de *Big Data* e os resultados serem integrados com software de análise de texto e estatística [50].

3.1.2. Skytree Server

O Skytree Server é utilizado para processar grandes quantidades de dados em alta velocidade. É fácil de usar e fornece uma interface de linha de comando na qual os usuários podem inserir comandos. O Skytree Server tem cinco usos: sistema de recomendação, identificação de anomalias, *clustering,* segmentação de mercado e análise preditiva. O foco principal do Skytree Server é a análise de dados em tempo real. Ele é otimizado para a implementação de algoritmos de *machine learning* em *Big Data*, usando mecanismos que são notavelmente mais rápidos do que os de outras plataformas. Ele pode manipular bancos de dados relacionais, arquivos planos e dados estruturados e não estruturados. Apesar das muitas vantagens do Skytree Server, a alta complexidade é uma das limitações [34].

3.1.3. Dryad

O Dryad é um mecanismo de execução distribuída, de propósito geral. A estrutura operacional do Dryad é um gráfico acíclico direcionado, no qual vértices representam programas e arestas representam canais de dados. O Dryad executa operações nos vértices em *clusters* e transmite dados por meio de canais de dados. A estrutura de operação do Dryad é coordenada por um programa central chamado gerenciador de tarefas, que pode ser executado em *clusters* ou estações de trabalho através da rede. O Dryad permite que os vértices usem qualquer quantidade de dados de entrada e saída, enquanto o MapReduce suporta apenas um conjunto de entrada e saída [16]. O Dryad realiza várias funções, incluindo geração de gráficos, métricas de desempenho, processo de agendamento de processos, visualização, tratamento de falhas, tolerância a falhas e reexecução [34].

3.1.4. Weka/Pentaho

Weka, abreviação de *Waikato Environment for Knowledge Analysis* (Ambiente Waikato para Análise de Conhecimento), é um software gratuito de *machine learing* e de mineração de dados, *open-source*, escrito em Java. O Weka fornece funções como processamento de dados, seleção de recursos, classificação, regressão, *clustering*, regra de associação e visualização, etc. O Pentaho é um dos softwares de BI *open-source* mais populares. Ele inclui uma plataforma de servidor web e várias ferramentas para suportar relatórios, análises, gráficos, integração de dados e mineração de dados, etc., todos os aspectos do BI. Os algoritmos de processamento de dados da Weka também estão integrados no Pentaho e podem ser correlacionados [16].

Apesar das muitas vantagens do Pentaho, como fácil acesso a dados, relatórios rápidos devido a técnicas de *caching* na memória, visualização detalhada e integração perfeita, há muitas desvantagens do Pentaho, como o Pentaho *suíte*, que são inconsistentes na maneira em que funcionam e análises menos avançadas em comparação com o Tableau [34].

3.1.5. Tableau

O Tableau é uma das ferramentas de BI mais rápidas. É rápido de implantar, fácil de aprender e muito útil para o cliente. O Tableau possui cinco produtos principais que facilitam as diversas necessidades de profissionais e organizações. Eles são:

- Tableau desktop: para uso individual;
- Servidor tableau: colaboração para qualquer organização;
- Tableau online: BI em caso de nuvem;
- Tableau *reader*: usado para ler arquivos salvos no desktop tableau;
- Tableau *public*: para publicar dados interativos online.

Apesar das muitas vantagens do Tableau, como a ótima visualização de dados, soluções de baixo custo para atualização e excelente suporte móvel, há muitas desvantagens, como falta de capacidade de previsão, segurança arriscada e problemas de gerenciamento de mudanças [34].

RAJESWARI et al [51] define algumas vantagens e desvantagens do uso do Tableau:

Vantagens:

- Tem uma excelente interface de usuário.
- Bom recurso de integração é. Pode se integrar a outras plataformas de *Big Data*, como o Hadoop;
- É compatível com dispositivos móveis;
- Software de baixo custo, fácil de atualizar e consome menos espaço de memória.

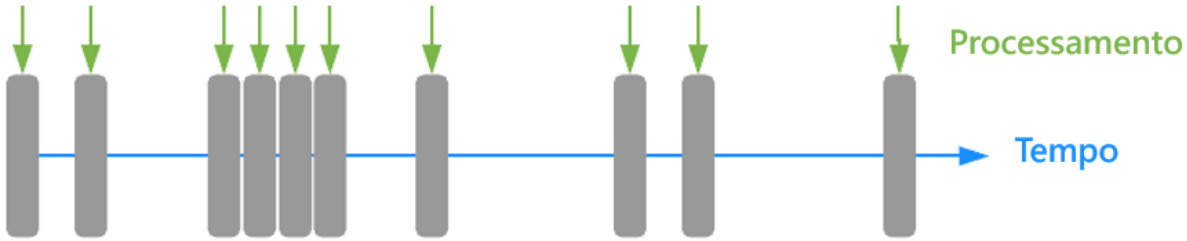
Desvantagens:

- É necessário processamento inicial de dados. E isso deve ser feito por especialista;
- Não suporta todas as estatísticas [51].

3.2. PROCESSAMENTO *STREAM*

No processamento *Stream*, ou em fluxo, cada novo dado é processado quando chega. Ao contrário do Processamento Batch, não há espera até que o próximo intervalo de processamento. Os dados são processados como peças individuais, em vez de serem processados como um lote de cada vez [43]. A Figura 14 ilustra o Processamento *Stream*.

Figura 14 - Processamento *Stream*.



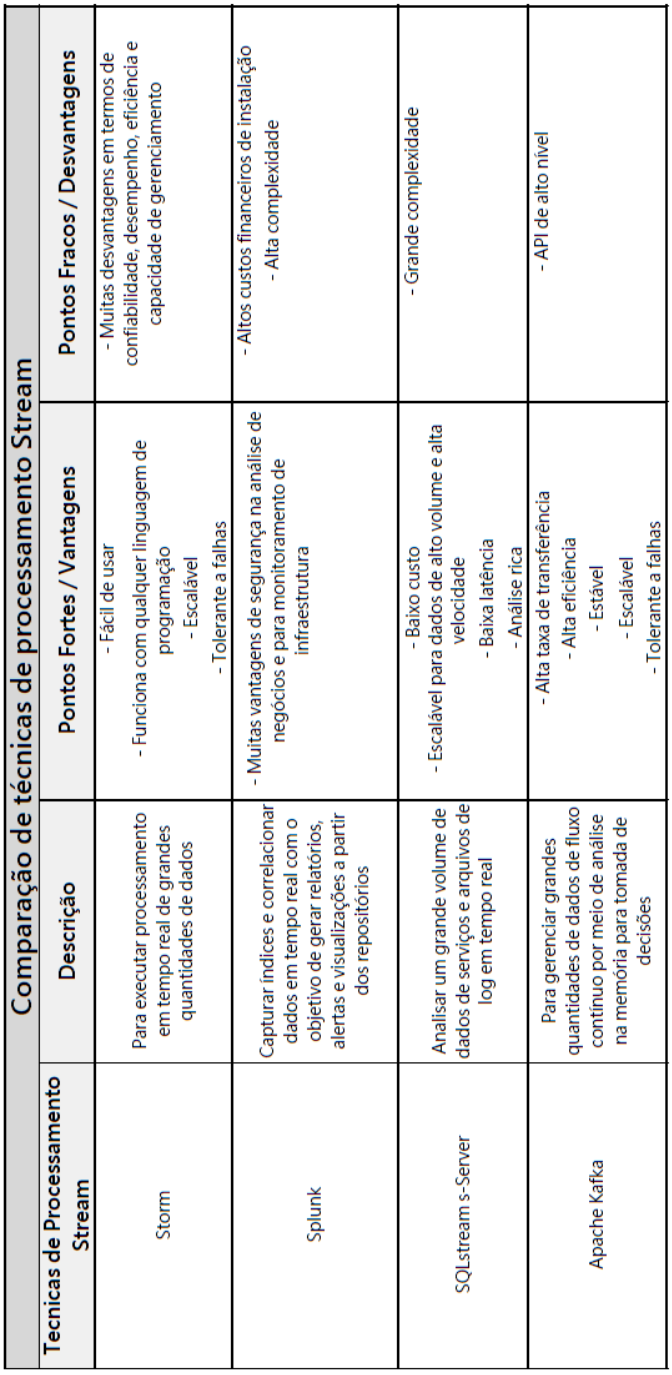
Fonte: [43].

Segundo SHIFF, o processamento *Stream* é altamente benéfico se os eventos que você deseja acompanhar estiverem ocorrendo com frequência e juntos no tempo. Se utiliza também quando precisa ser detecção imediata e responder rápida. É, portanto, útil para tarefas como detecção de fraudes e segurança cibernética. Se os dados da transação forem processados por *Stream*, as transações fraudulentas poderão ser identificadas e interrompidas antes de serem concluídas [42]. O Quadro 2 apresenta uma comparação de algumas técnicas de Processamento *Stream*, com os pontos fortes e fracos de cada uma.

3.2.1. Storm

O Apache Storm é uma plataforma *open-source*, originalmente desenvolvida pela Backtype e depois adquirida pelo Twitter [52]. Apesar de muitas vantagens do Storm, como ser fácil de usar, funcionar com qualquer linguagem de programação, ser escalável e tolerante a falhas, há muitas desvantagens do Storm em termos de confiabilidade, desempenho, eficiência e gerenciamento [34].

Quadro 2 - Comparação de técnicas de Processamento *Stream*.



Comparação de técnicas de processamento Stream			
Tecnicas de Processamento Stream	**Descrição**	**Pontos Fortes / Vantagens**	**Pontos Fracos / Desvantagens**
Storm	Para executar processamento em tempo real de grandes quantidades de dados	- Fácil de usar - Funciona com qualquer linguagem de programação - Escalável - Tolerante a falhas	- Muitas desvantagens em termos de confiabilidade, desempenho, eficiência e capacidade de gerenciamento
Splunk	Capturar índices e correlacionar dados em tempo real com o objetivo de gerar relatórios, alertas e visualizações a partir dos repositórios	- Muitas vantagens de segurança na análise de negócios e para monitoramento de infraestrutura	- Altos custos financeiros de instalação - Alta complexidade
SQLstream s-Server	Analisar um grande volume de dados de serviços e arquivos de log em tempo real	- Baixo custo - Escalável para dados de alto volume e alta velocidade - Baixa latência - Análise rica	- Grande complexidade
Apache Kafka	Para gerenciar grandes quantidades de dados de fluxo continuo por meio de análise na memória para tomada de decisões	- Alta taxa de transferência - Alta eficiência - Estável - Escalável - Tolerante a falhas	- API de alto nível

Fonte: Adaptado de [34]

3.2.2. Splunk

O Splunk captura índices e correlaciona dados em tempo real, com o objetivo de gerar relatórios, alertas e visualizações dos repositórios. Ele é projetado para diagnosticar problemas

de infraestrutura de TI e fornecer inteligência para operações comerciais. Muitas empresas renomadas, como Amazon, Senthub e Heroku, utilizam o Splunk. O Splunk apresenta os resultados de várias maneiras (por exemplo, gráficos e alertas). Apesar das muitas vantagens do Splunk, como segurança na análise de negócios e o monitoramento de infraestrutura, há algumas desvantagens do Splunk, como alto custo e alta complexidade [34] .

3.2.3. SQLstream s-Server

SQLstream s-Server é uma plataforma para analisar um grande volume de serviços e dados não estruturados em tempo real. Ele realiza coleta, agregação, integração e enriquecimento em tempo real nos dados de *streaming*. A plataforma emprega a linguagem SQL para suas operações subjacentes. SQLstream s-Server trabalha rápido porque não usa tecnologia de banco de dados. Os dados não são armazenados nos discos, mas são processados na memória por meio de consultas SQL contínuas. Apesar de muitas vantagens do SQLstream s-Server, como baixo custo, escalável para dados de alto volume e alta velocidade, a baixa latência e a alta complexidade são algumas das desvantagens [34].

3.2.4. Apache Kafka

O Apache Kafka é usado para gerenciar grandes quantidades de dados de *streaming* por meio de análise *in-memory* para tomada de decisão. A ferramenta possui quatro características: mensagens persistentes, estruturas de disco, processamento distribuído e alta taxa de transferência. O Apache Kafka fornece soluções analíticas *ad hoc*, combinando o processamento *off-line* e *on-line*. Apesar das muitas vantagens do Apache Kafka, como alta taxa de transferência, alta eficiência, estabilidade, escalável e tolerante a falhas, no entanto, a API de alto nível é uma das principais preocupações [34].

3.3. TÉCNICAS DE ANÁLISE *BIG DATA*

Com discutido no Capítulo 2, são necessárias técnicas de análise no contexto de *Big Data* para analisar eficientemente grandes quantidades de dados dentro de um período de tempo limitado. As Técnicas de Análise utilizadas para *Big Data* serão discutidos nas seções seguintes. O Quadro 3 compara e dá uma breve visão geral das técnicas discutidas, destacando suas aplicações e ferramentas relacionadas.

Quadro 3 - Comparação das Técnicas de Big Data

Comparação das Técnicas de Análise *Big Data*				
Técnicas de Análise *Big Data*	**Descrição**	**Uso em algumas aplicações multidisciplinares**	**Algoritmos / Técnicas**	**Ferramentas Disponíveis**
Data Mining	Para encontrar padrões consistentes e/ou relações sistemáticas entre variáveis	- Biomedicina - Saúde	- K-Mean - Fuzzy C-Mean - CLARA - CLARANS - BETULA	- Excel - Rapid-I - Rapidminer-R - KNIMINE - Weka/Pentaho
Análise de Redes Sociais	Para ver as relações sociais em termos de teoria de redes	- Antropologia - Mídia Social	- PCA - LTSA - LLE - Autoencoder	- Cytoscape - Gephi - Cuttlefish - MeerKat
Web Mining	Para descobrir padrões de uso em grandes repositórios da web	- *E-learning* (ensino eletrônico) - Bibliotecas Digitais - E-governo	- LOGML - Apriori	- KXEN - LIONsolver - Dataiku
Machine Learning	Permitir que computadores desenvolvam comportamentos baseados em dados empíricos	- Saúde - Serviço ao cliente	- Reconhecimento de padrões - Redes neurais artificiais	- Weka - Scikit-Learn - PyMc - Shogun
Métodos de otimização	Para resolver problemas quantitativos	- Ciência de Redes sociais - Biologia Computacional	- Paralelização - Recozimento simulado - Recozimento quântico - Otimização de enxame de partículas	- Matlab

Fonte: [34]

3.3.1. Data Mining

Data mining (ou Mineração de Dados), em ciência da computação, é o processo de descobrir padrões e relações úteis, em grandes volumes de dados. Essa análise combina ferramentas de estatística e inteligência artificial (como redes neurais e *machine learning*) com gerenciamento de banco de dados para analisar grandes conjuntos de dados. A Mineração de Dados é amplamente usada em negócios (seguros, bancos, varejo), pesquisa científica (astronomia, medicina) e segurança do governo (detecção de criminosos e terroristas) [53].

Muitas ferramentas para mineração e análise de *Big Data* estão disponíveis, incluindo *softwares* profissionais e amadores, *softwares* comerciais com alto custo e *softwares open-source*. O Quadro 4 compara os cinco *software* mais usados, por meio de uma pesquisa feita com 798 profissionais, que responderam à pergunta: "Qual software de análise de dados você usou nos últimos 12 meses para um projeto real?" [54].

Quadro 4 - Comparação das diferentes ferramentas para *Data Mining*.

Comparação das diferentes ferramentas para *Data Mining*		
Ferramenta de *Data Mining*	Descrição	Percentual de Uso
Excel	Ele fornece recursos poderosos de processamento de dados e análise estatística	29,80%
Rapid-I RapidMiner	É usado para mineração de dados, *machine learning* e análise preditiva	26,70%
R	É usado para mineração de dados / análise e visualização	30,70%
KNIME	É usado para integração de dados, processamento de dados, análise de dados e mineração de dados	21,80%
Weka/Pentaho	Ele fornece funções, como processamento de dados, seleção de recursos, classificação, regressão, clustering, regra de associação e visualização	14,80%

Fonte: Adaptado de [54] [16] [34].

3.3.1.1. Excel

O Excel, um componente central do Microsoft Office, fornece poderosos recursos de processamento de dados e análise estatística. Quando o Excel é instalado, alguns *add-ins* avançados, com funções poderosas para análise de dados, são integrados inicialmente, mas eles só podem ser usados se os usuários os habilitarem. O Excel é o único *software* comercial entre os cinco principais [16].

Os *add-ins* são programas que adicionam recursos e comandos opcionais aos recursos tradicionais do Microsoft Excel. O Excel criou *add-ins* para várias finalidades: análise de dados, apresentação, investimento, negócios, pessoal, utilitários e ferramentas de produtividade e organização. Na análise de dados, alguns dos suplementos mais populares incluem o *Analysis Toolpak*, *Solver* (gratuitos) e *MegaStat* (pago) [55].

3.3.1.2. Rapid-I RapidMiner

O Rapid-I RapidMiner é um *software open-source* usado para mineração de dados, *machine learning* e análise preditiva. Os programas de mineração de dados e *machine learning* fornecidos pelo RapidMiner incluem extração, transformação e carga (ETL), pré-processamento de dados e visualização, modelagem, avaliação e implantação. O fluxo de mineração de dados é descrito em XML e exibido por meio de uma interface gráfica do usuário (*Graphic User Interface* - GUI). O RapidMiner é escrito em Java. Integra o método de aprendizado e de avaliação do Weka, e trabalha com R [16].

3.3.1.3. R

R é um software estatístico gratuito e *open-source*. Foi criado em 1993, por Robert Gentleman e Ross Ihaka. Em 2014, contava com mais de 2 milhões de usuários [55]. Durante a execução de tarefas intensivas, códigos programados com C, C ++ e Fortran podem ser chamados no ambiente R. Além disso, usuários experientes podem chamar diretamente objetos R em C. Na verdade, R é uma realização da linguagem S, que é uma linguagem interpretada desenvolvida pela AT&T Bell Labs e usada para exploração de dados, análise estatística e gráficos de desenho. Comparado ao S, o R é mais popular, já que é *open-source*. Devido à popularidade do R, os fabricantes de bancos de dados, como Teradata e Oracle, lançaram produtos que suportam R [16].

RAJESWARI et al [51] define algumas vantagens e desvantagens do uso do R:

Vantagens:

- As habilidades gráficas de R são extraordinárias, dando um design completamente programável que supera a maioria dos outros pacotes gráficos;
- R não tem restrições de licença. É possível executá-lo em qualquer lugar e a qualquer momento, e até vendê-lo sob as condições da licença;

Desvantagens:

- É necessário aprender muito bem a ferramenta antes de usar. Caso contrário, não poderá ser usado de forma eficaz.
- Todos os pacotes usados em R nem sempre dão resultado perfeito [51].

3.3.1.4. KNIME

É uma plataforma de integração, processamento, análise de dados e mineração de dados, fácil de usar, inteligente e aberta. Ele permite que os usuários criem fluxos ou canais de dados de maneira visual. O KNIME foi escrito em Java e fornece mais funções como *plug-ins*. Os usuários podem inserir módulos de processamento para arquivos, imagens e séries temporais e integrá-los em vários projetos *open-sources*, como por exemplo, R e Weka. O KNIME controla a integração de dados, *cleaning*, conversão, filtragem, estatísticas, mineração e, finalmente, visualização de dados. Todo o processo de desenvolvimento é conduzido sob um ambiente visualizado. O KNIME é projetado como uma estrutura expansível e baseada em módulos [16].

3.3.1.5. Weka/Pentaho

Técnica mencionada na seção 3.1.4.

3.3.2. Análise de Rede Social

A técnica de análise de redes sociais (*Social Network Analysis* - SNA) é utilizada para ver as relações sociais no ambiente das Redes Sociais. [34]. A coleta de dados pode ser de modo qualitativo ou quantitativo. A coleta de dados de modo qualitativo pode ser feita através de entrevistas, questionários ou mesmo da observação do pesquisador de um determinado ambiente ou grupo. As coletas quantitativas de dados geralmente focam em bases de dados preexistentes. Nesses casos, é bastante comum que se utilizem ferramentas construídas pelos pesquisadores [56]. Essas ferramentas de coleta automática de dados são denominadas *crawlers*. RECUERO [56] em seu trabalho, destaca os seguintes *crawlers* para coleta de dados:

- **yTK:** *crawler* para Twitter, está um pouco desatualizado, é bem complicado de instalar e precisa de um servidor e uma conexão bastante estável para não sair do ar. Ele coleta tweets do momento em que se iniciou a busca em diante, e tende a coletar uma quase totalidade dos tweets no período em que estiver rodando.
- **NodeXL**: Não precisa servidor. Tem uma interface bastante amigável e funciona como um *layer* para o Excel. Apresenta *crawler* para Twitter, Facebook, Youtube e Flickr.
- **NetVizz**: Funciona via Facebook e para o Facebook. Coleta dados de grupos e de páginas de busca e permite que sejam exportados. Seu uso é relativamente simples, mas a coleta de dados pode ser bastante demorada.
- **Gephi**: O Gephi é um aplicativo popular para a visualização e manipulação dos dados. Exige que se tenha uma autorização para o acesso à API.

Uma vez coletados os dados de mídia social, inicia-se então o processo de análise. “A análise dá-se através da aplicação das métricas de rede para os dados que foram coletados.” A visualização dos dados pode ser utilizada de modo paralelo, buscando compreender essas estruturas. RECUERO [56] também discute as ferramentas de análise e visualização de dados para Redes Socias:

- **NodeXL**: possui várias métricas, mas a visualização dos dados é um pouco limitada, com poucos algoritmos. É útil para trabalhar com grafos pequenos (até 20 mil nós). Acima disso, começa a ficar muito lento. Funciona apenas em Windows.

- **Gephi**: é uma das ferramentas mais utilizadas por ser gratuita, aberta e incluir diferentes *plugins*, feitos por usuários. Entretanto, possui uma interface bastante difícil e uma curva de aprendizagem maior. Entre suas partes mais fortes está a visualização de dados e a inclusão frequente de novas métricas. É preciso aumentar a memória dedicada a ele manualmente para conseguir trabalhar com dados maiores.
- **NetDraw**: é uma ferramenta relativamente limitada, mas gratuita. Funciona no Windows. É uma das ferramentas mais simples de ser compreendida por um usuário novato.
- **Pajek**: é uma das ferramentas mais conhecidas para análise de redes sociais. Também tem uma interface um pouco difícil, mas há várias dezenas de tutoriais e livros de ajuda para quem quer começar. Os grafos possuem uma quantidade bem útil de métricas e formas de visualização.
  [56].

3.3.3. *Web Mining*

Web mining (Mineração da Web) é uma técnica empregada para descobrir padrões em grandes repositórios da web. Pode ser dividido em dois tipos:

- Mineração de conteúdo da Web: ajuda a extrair informações úteis do conteúdo da web (áudio, vídeo, texto e imagens). A Mineração da Web envolve o desenvolvimento de sistemas de Inteligência Artificial sofisticados que podem atuar de forma autônoma ou semi-autônoma, em nome de um usuário específico, para descobrir e organizar informações baseadas na web.
- Mineração da estrutura da Web: é utilizada para analisar nós e a estrutura de conexão de um site, por meio da teoria dos grafos [34].

3.3.4. *Machine Learning*

Segundo ZHOU et al, as técnicas de *Machine Learning* (Aprendizado de Máquina - ML) geram enormes impactos sociais em uma ampla gama de aplicações, como área computacional, processamento de fala, compreensão de linguagem natural, neurociência, saúde e Internet das Coisas [17].

> "O advento da era Big Data estimulou amplos interesses no ML. (...) Por um lado, o *Big Data* fornece informações ricas e sem precedentes para que os algoritmos de ML extraiam padrões subjacentes e criem modelos preditivos. Por outro lado, os algoritmos ML tradicionais enfrentam desafios críticos, como a escalabilidade, para realmente liberar o valor oculto do *Big Data*. (...) o ML precisa crescer e avançar para transformar grandes dados em inteligência acionável [17]."

As técnicas de ML permitem que os usuários façam previsões a partir de grandes conjuntos de dados. O ML desenvolve por meio de técnicas de aprendizagem eficientes (algoritmos), dados ricos e grandes ambientes de computação. Assim, o ML tem um grande potencial e é uma parte essencial da análise de big data [17].

3.3.5. Métodos de Otimização

"Os métodos de otimização são utilizados para resolver problemas quantificáveis. Estes métodos são usados em campos multidisciplinares [34]". A fim de abordar os problemas de otimização global, são utilizadas diferentes estratégias, altamente eficientes porque exibem paralelismo. Possuem alta complexidade e consomem tempo. Como ferramenta de otimização, temos o uso do Matlab.

As áreas da engenharia e TI estão usando o MATLAB para desenvolver sistemas avançados de BDA, desde manutenção preditiva e telemática até sistemas avançados de assistência ao motorista e análise de sensores. A MathWorks, acredita que o MATLAB é escolhido porque oferece recursos essenciais, não encontrados em sistemas de BI ou em linguagens *open-source* [57]:

- Dados do mundo físico: o MATLAB possui suporte nativo para sensores, imagem, vídeo, telemetria, binário e outros formatos em tempo real.
- *Machine Learning*, redes neurais, estatísticas e etc: o MATLAB oferece um conjunto completo de estatísticas e funcionalidades, além de métodos avançados como otimização não linear, identificação de sistema e milhares de algoritmos pré-construídos para processamento de imagem e vídeo, modelagem financeira e *design* de sistema de controle
- Processamento de alta velocidade de grandes conjuntos de dados: processamento paralelo em *clusters* e nuvem.
- Implantação on-line e em tempo real: o MATLAB integra-se a sistemas corporativos, clusters e nuvens, e pode ser direcionado para hardware embarcado em tempo real [57].

3.4. CONSIDERAÇÕES FINAIS

Neste Capítulo foi apresentado as ferramentas mais usuais para processamento *Batch*, as ferramentas mais usuais para processamento *Stream* e as ferramentas de análise para Big Data. Foi discutido os cenários que os métodos e ferramentas podem atuar, destacando seus pontos forte e seus pontos fracos. O próximo Capítulo encerra este Trabalho de Conclusão de Curso 2, fazendo uma conclusão do que foi visto nessa pesquisa realizada.

CAPÍTULO 4
CONSIDERAÇÕES FINAIS

Big Data é um método de análise de dados que se diferencia dos métodos tradicionais devido à complexidade dos dados ou dos processos. Vimos durante este trabalho, que o tamanho da soma de todos os dados do mundo vem num crescente exponencial, com expectativa de chegar a 175ZB em 2025. A Internet da Coisas, as mídias sociais, o avanço da computação, tudo isso impulsiona a geração de dados e, consequentemente, a necessidade de tratamento dessa informação. A computação em nuvem fornece a base tecnológica para que o *Big Data* possa ser desenvolvido, uma vez que os *datacenters* tradicionais não são mais suficientes para o número de dados produzidos no mundo. A análise *Big Data* vem ganhando destaque em diversas áreas, como na economia, na política, na saúde, na agricultura, no social e na engenharia.

Para analisar esses "grandes dados", existe um grande número de ferramentas disponíveis. Dentre elas, temos em o Hadoop, o Excel, o R e o Matlab. Métodos como *data mining, machine learning* e de otimização são intimamente ligados e dependentes do conceito *Big Data*. Antes de iniciar o processo de análise, é importante entender seus dados e seus objetivos. A escolha da técnica e ferramenta é uma etapa primordial para o sucesso da análise. Entender sua técnica e ter domínio da ferramenta é fundamental no processo do *Big Data*.

Como proposta de trabalhos futuros, temos uma revisão aprofundada das técnicas de análise descritas. Desenvolver, por exemplo, pesquisas atualizadas de uso dos *software* (similar ao que foi feito no Quadro 4), ou ainda sobre número de publicações sobre o tema, como feito na Figura 5. Outra proposta de trabalho futuro é o desenvolvimento de um estudo de caso ou aprofundar em uma categoria/aplicação, listando abordagens, resultados obtidos e ferramentas utilizadas.

REFERÊNCIAS

[1] HASHEM, I. A. T. et al. **The rise of "big data" on cloud computing: Review and open research issues**. Information Systems 47: 98-115, jul. 2014.

[2] EKBIA, H. et al. **Big Data, Bigger Dilemmas: A Critical Review**. Journal of the Association for Information Science & Technology, 2014.

[3] REINSEL, D., GANTZ, J., RYDNING, E. **Data Age 2025: The Evolution of Data to Life-Critical**. Seagate, 2017.

[4] DAVENPORT T. H., BARTH, P., BEAN, R. **How 'Big Data' Is Different**. MIT Sloan. 30 jul. 2012. Disponível em: < https://sloanreview.mit.edu/article/how-big-data-is-different/>. Acesso em: 05 mai. 2018

[5] LETOUZÉ, E. **Big Data for Development: Challenges & Opportunities**. UN Globo Pulse, mai. 2012. Disponível em: <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf>. Acesso em: 06 de mai 2018.

[6] LUVIZAN, S. S., MEIRELLES, F. S., DINIZ, E. H. **BIG DATA: EVOLUÇÃO DAS PUBLICAÇÕES E OPORTUNIDADES DE PESQUISA**. São Paulo, mai. 2014.

[7] REINSEL, D., GANTZ, J., RYDNING, E**. The Digitization of the World: From Edge to Core**. Seagate, 2018. Disponível em: < https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf >. Acesso em 26 de novembro de 2018.

[8] GANTZ, J., REINSEL, D. **THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Grow th in Far East**. IDC, dez. 2012. Disponível em: < https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf>. Acesso em 08 mai. 2018.

[9] MAÇADA, A. C. G., BRINKHUES, R. A., JUNIOR, J. C. F. **Big data e as capacidades de gestão da informação**. Com Ciência: Revista Eletrônica de Jornalismo Científico. Rio Grande do Sul, 09 jul. 2015. Disponível em: <http://www.comciencia.br/comciencia/handler.php?section=8&edicao=115&id=1388&tipo=1>. Acesso em: 08 mai. 2018.

[10] IDC. **The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things,** abr. 2014. Disponível em: <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>. Acesso em: 20 nov 2018.

[11] JIN, X. et al, **Significance and Challenges of Big Data Research**. Big Data Research 2: 59–64, 2015.

[12] FRAMINGHAM, M. **Revenues for Big Data and Business Analytics Solutions Forecast to Reach $260 Billion in 2022, Led by the Banking and Manufacturing Industries, According to IDC.** 15 ago. 2015. Disponivel em: <https://www.idc.com/getdoc.jsp?containerId=prUS44215218, 15 ago 2015>. Acesso em: 05 out. 2018.

[13] EREVELLES, S., FUKAWA, N., SWAYNE, L. **Big Data consumer analytics and the transformation of marketing**. Journal of Business Research 69: 897–904, 2016.

[14] DEMCHENKO, Y. **Addressing Big Data Issues in the Scientific Data Infrastructure**. University of Amsterdam. Maastricht, 2013.

[15] IBM. **What is big data?** 2017. Disponível em: <https://developer.ibm.com/dwblog/2017/what-is-big-data-insight/>. Acesso em 05 out. 2018.

[16] CHEN, M., MAO, S., LIU,Y. **Big data: a survey**. Nova Iorque, 2014.

[17] ZHOUA, L. et al. **Machine learning on big data: Opportunities and challenges**. Neurocomputing 237: 350–361, 2017.

[18] RATHOREA, M. M. U., AHMAD, A., PAUL, A. **Urban planning and building smart cities based on the Internet of Things using Big Data analytics**. Computer Networks, mar. 2016.

[19] O'LEARY, D. E., **'Big Data', The 'Internet Of Things' And The 'Internet of Signs'.** Intell. Sys. Acc. Fin. Mgmt. 20, 53–65, 2013.

[20] NIYATO, D. **Market Model and Optimal Pricing Scheme of Big Data and Internet of Things (IoT)**. 2016. Disponível em: < https://ieeexplore.ieee.org/document/7510922>. Acesso em 05 out. 2018.

[21] MARJANI, M**.** et al. **Big IoT Data Analytics: Architecture,Opportunities, and Open Opportunities, and Open**, mar. 2017. Disponível em: < https://www.researchgate.net/publication/316240052_Big_IoT_Data_Analytics_Architecture_Opportunities_and_Open_Research_Challenges> Acesso em: 20 nov. 2018.

[22] AGRAWA, D., DAS, S., ABBADI, A. E. **Big Data and Cloud Computing: Current State and Future Opportunities**. University of California, Santa Barbara, 2011.

[23] PANDEY, S., NEPAL, S. **Cloud Computing and Scientific Applications — Big Data, Scalable Analytics, and Beyond**. Future Generation Computer Systems Volume 29: 1774-1776, set. 2013.

[24] DAVENPORT, T. H. **Competing on Analytics**. Harvard Business Review, jan. 2006. Disponível em: < https://hbr.org/2006/01/competing-on-analytics>. Acesso em: 08 jun. 2018.

[25] CHEN, H., CHIANG, R. H. L., STOREY, V. C. **Business Intelligence And Analytics: From Big Data To Big Impact**. MIS Quarterly Vol. 36 No. 4: 165-1188, dez. 2012.

[26] OLIVEIRA, D. T., PEREIRA, O. J. **Um estudo do Business Intelligence no ambiente empresarial**.

[27] DEBORTOLI, S., MULLER, O., BROCKE, J. V., **Comparing Business Intelligence and Big Data Skills**, 15 ago. 2014.

[28] BRANKOVIC, L., ESTIVILL-CASTRO, V. **Privacy Issues In Knowledge Discovery And Data Mining**. Jan, 1999.

[29] XU, L. et al. **Information Security in Big Data: Privacy and Data Mining**. IEEE, 09 out. 2014

[30] EMC/IDC, **Brazil - The Digital Universe of Opportunities**. 2014. Disponível em: <https://brazil.emc.com/collateral/analyst-reports/idc-digital-universe-2014-brazil.pdf.>. Acesso em: 01 dez. 2018.

[31] PERERA, C. **Privacy of Big Data in the Internet of Things Era.** 2015.

[32] ZICARI, R. V. **Big Data: Challenges and Opportunities**. ODBMS, 2012.

[33] FILHO, A. D. P. C. **Uso de big data em saúde no Brasil: perspectivas para um futuro próximo**. 2015. Disponível em: <https://www.scielosp.org/article/ress/2015.v24n2/325-332/pt/>. Acesso em: 08 mai. 2018.

[34] YAQOOB et al. **Big data: From beginning to future**. 2016. Disponível em: <https://www.researchgate.net/publication/305736330_Big_Data_From_Beginning_to_Future/download>. Acesso em: 07 dez. 2018

[35] PAI, V. **Big Data New Challenges, Tools And Techniques.** Department of Information Technology, Srinivas Institute of Management Studies, Mangalore, Karnataka, 2016.

[36] KALLA, S. **What is statistics?** Disponível em: <https://explorable.com/what-is-statistics.>. Acesso em 01 dez. 2018.

[37] DAVIDSON, J. **What is Statistics?** Disponível em: <https://www.sscc.edu/home/jdavidso/mathadvising/AboutStatistics.html.>. Acesso em: 01 dez. 2018

[38] IIT Madras - Mechanical Engineering Department, **Optimization Methods.** Disponível em: <https://mech.iitm.ac.in/nspch52.pdf.>. Acesso em 07 dez. 2018.

[39] MURPHY, K. P. **Machine Learning: A Probabilistic Perspective.** Massachusetts Institute of Technology, 2012.

[40] NIELSEN, M. A. **Neural Networks and Deep Learning**. Determination Press, 2015.

[41] IIT KANPUR – NPTEL. **Digital Signal Processing - Introdução**. Disponível em: <https://nptel.ac.in/courses/Webcourse-contents/IIT-ANPUR/Digi_Sign_Pro/pdf/ch1.pdf>. Acesso em 28 nov. 2018.

[42] SHIFF, L. **Real Time vs Batch Processing vs Stream Processing: What's The Difference?** Abril, 2018. Disponível em: <https://www.bmc.com/blogs/batch-processing-stream-processing-real-time/>. Acesso em: 07 dez. 2018.

[43] Streamlio. **Understanding Batch, Microbatch, and Streaming**. Disponível em: <https://streaml.io/resources/tutorials/concepts/understanding-batch-microbatch-streaming>. Acesso em 07 dez. 2018

[44] WHITE, T. **Hadoop: The Definitive Guide**. 4ª Edição. O'Reilly Media, abr. 2015.

[45] SHVACHKO, K., KUANG, H., RADIA, S. **The Hadoop Distributed File System**. Incline Village, NV, USA, 28 jun. 2010.

[46] DEAN, J., GHEMAWAT, S. **MapReduce: simplified data processing on large clusters**. Google, Inc, 2004.

[47] MUNSHI, A. A., YASSER, A., MOHAMED, R. I. **Big data framework for analytics in smart grids**, 2017. Disponível em <http://isiarticles.com/bundles/Article/pre/pdf/142004.pdf>. Acesso em 28 nov. 2018.

[48] EL-SHAFEIY, E. A., EL-DESOUKY, .A. I. **A Big Data Framework for Mining Sensor Data Using Hadoop**. 2017. Disponível em: <https://sic.ici.ro/wp-content/uploads/2017/10/SIC_2017-3-Art12.pdf>. Acesso em 03 nov. 2018.

[49] KUMAR, P., CHANDRASEKAR, E. S. **E-Commerce Trends and Future Analytics Tools**. 2016. Disponível em: <https://www.researchgate.net/publication/308038039_E-Commerce_Trends_and_Future_Analytics_Tools/download>. Acesso em: 03 nov. 2018.

[50] STRANG, K. D., SUN, Z. **Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics**. 2017. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/08874417.2016.1181497>. Acesso em: 03 nov. 2018.

[51] RAJESWARI, C., BASU, D., MAURYA, N. **Comparative Study of Big data Analytics Tools: R and Tableau**. 2017. Disponível em: <http://iopscience.iop.org/article/10.1088/1757-899X/263/4/042052/pdf>. Acesso em: 04 nov. 2018.

[52] BARTOLINI, I., PATELLA, E. M. **Comparing Performances of Big Data Stream Processing Platforms with RAM'S**. 2017. Disponível em: <http://ceur-ws.org/Vol-2037/paper_21.pdf>. Acesso em: 04 nov. 2018.

[53] Encyclopædia Britannica, Inc. **Data Mining**. 2018 Disponível em: <https://academic.eb.com/?target=%2Flevels%2Fcollegiate%2Farticle%2F437561>. Aceso em 15 out. 2018.

[54] KDNUGGETS. **What analytics data mining, big data software you used in the past 12 months for a real project?** 2012.

[55] OZGUR, C., KLECKNER, M., LI, E. Y. **Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities.** 2015.

[56] RECUERO, R. **Introdução à análise de redes sociais online**. EDUFBA, 2017.

[57] MathWorks. **Data Analytics**. 2018. Disponível em: < https://www.mathworks.com/solutions/data-analytics.html >. Acesso em 07 dez. 2018.

[58] EVANS, D. **The Internet of Things [INFOGRAPHIC]**. CISCO, 15 jul. 2011. Disponível em: < https://blogs.cisco.com/diversity/the-internet-of-things-infographic>. Acesso em: 07 jun. 2018.

[59] IBM. **What is a mainframe? It's a style of computing.** Disponível em: <https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zmainframe/zconc_whatismainframe.htm>. Acesso em: 28 nov. 2018.

[60] Intel. **Moore's Law and Intel Innovation.** Disponível em: <https://www.intel.com.br/content/www/br/pt/history/museum-gordon-moore-law.html>. Acesso em 28 nov. 2018.

[61] Fórum. **Entenda o que é Compliance e descubra os principais benefícios para as empresas**. Disponível em: <http://www.editoraforum.com.br/noticias/entenda-o-que-e-

compliance-e-descubra-os-principais-beneficios-para-as-empresas/>. Acesso em: 01 dez. 2018.