Universidade do Minho
Escola de Engenharia

Cátia Cristina Pereira Oliveira

Benchmarking de técnicas de Business Analytics em Big Data

Julho de 2020

Benchmarking de técnicas de Business Analytics em Big Data

Cátia Oliveira

UMinho | 2020

Universidade do Minho
Escola de Engenharia

Cátia Cristina Pereira de Oliveira

Benchmarking de técnicas de *Bussiness Analytics* em *Big Data*

Dissertação de Mestrado
Mestrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do(s)
Professor Doutor Manuel Filipe Vieira Torres dos Santos
Professor Doutor Carlos Filipe da Silva Portela

Julho de 2020

AGRADECIMENTOS

O desenvolvimento e a conclusão desta dissertação só foi possível graças ao apoio e incentivo de várias pessoas.

Ao Professor Doutor e orientador Manuel Filipe Vieira dos Santos um obrigada pela dedicação e conselhos no desenvolvimento desta dissertação.

Ao Professor Doutor e coorientador Carlos Filipe da Silva Portela um obrigada pela paciência e pelo conhecimento partilhado.

Aos meus amigos que me acompanharam de perto durante esta etapa, um muito obrigada pela motivação e paciência nos momentos mais difíceis.

Ao meu namorado, obrigada pela paciência, nos bons e maus momentos, e companheirismo nesta etapa da minha vida.

Aos meus pais e irmãos, um muito obrigada por todo o investimento e apoio demonstrado durante todo o percurso académico.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Os desenvolvimentos tecnológicos e a crescente dependência das organizações e da sociedade no mundo da internet levaram ao crescimento e variedade de dados. Esse crescimento e variedade, tornaram-se num desafio para os manipuladores de dados, uma vez que o processamento de uma grande quantidade de dados pode ser um desafio, porque pode despender muito tempo. Assim, veio a criação do conceito *Big Data*. *Big Data* pode ser entendido como um grande conjunto de dados com várias estruturas, que a tecnologia tradicional não consegue lidar, tendo dificuldade de armazenamento e de processamento. Nesta dissertação, serão definidos dois conceitos.

Portanto, esta dissertação foca nos desafios que o *Big Data* coloca ao *Data Mining*,

Nesta dissertação foi analisado um estudo de seleção de ferramentas de *Data Mining*, onde foram utilizadas duas metodologias tendo em consideração vários critérios de avaliação. Posteriormente, com base nos resultados do estudo anterior, foram selecionadas as duas melhores ferramentas, *KNIME* e *RapidMiner*.

Nesta dissertação também são apresentadas algumas sugestões de boas práticas quando lidamos com dados.

Depois de selecionadas as ferramentas, foi analisado um estudo referente à performance das ferramentas *KNIME* e *RapidMiner* em ambiente *Big Data*.

No inicio deste documento é apresentado um enquadramento do projeto e qual o seu objetivo. De seguida, é apresentado a revisão de literatura onde são descritos os principais conceitos e tópicos relacionados com a dissertação. Posteriormente, são apresentadas as abordagens metodológicas utilizadas nesta dissertação, assim como de que forma foram utilizadas. De seguida, são apresentados os desafios de *Big Data Mining*. Seguidamente, é apresentado o estudo de seleção de ferramentas de *Data Mining*, assim como as experiências de comparação de performance das ferramentas selecionadas. Por fim, é apresentada a discussão dos resultados, onde também é apresentada uma análise SWOT, e a conclusão.

Palavras-Chave: Benchmarking; Técnicas; Business Analytics; *Big Data*;

ABSTRACT

Technological developments and the growing dependence of organizations and society in the world of the internet, led to the growth and variety of data. This growth and variety has become a challenge for data handlers, since processing a large amount of data can be challenging because it can take a great deal of time. Thus, came the creation of the *Big Data* concept. *Big Data* can be understood as a large set of data with various structures, which the traditional technology can not handle, having difficulty in storage and process them. In this dissertation, two concepts will be defined.

Therefore, this dissertation focuses on the challenges that *Big Data* puts the *Data Mining*.

In this dissertation was analyzed a study of selection of *Data Mining* tools, where they were used two methodologies taking in consideration various criteria of evaluation. Subsequently, based on the results of the previous study, were selected the two best tools, *KNIME* and *RapidMiner*. In this dissertation are also presented some suggestions of good practice when dealing with data. After the tools were selected, was analyzed a study on the performance of the tools *KNIME* and *RapidMiner* in the *Big Data* environment. At the beginning of this document is presented the context of this project and its purpose. Then, the literature review is presented describing the main concepts and topics related to the dissertation. Subsequently, the methodological approaches used in this dissertation are presented, as well as how they were used. Then, is presented the *Data Mining* tool selection study, as well as the performance comparison experiments of the selected tools. Finally, a discussion of the results is presented, which also presents a SWOT analysis, and the conclusion.

KEYWORDS: *Benchmarking; Techniques; Business Analytics; Big Data*

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

BD – *Big Data*

BA – *Business Analytics*

BI – *Business Intelligence*

PDME – *Pervasive Data Mining Engine*

DSS – *Decision Support Systems*

OLAP – *Online Analytical Processing*

ETL – *Extract, Transform, Load*

TI – *Tecnologias de Informação*

SI – *Sistemas de Informação*

KPI – *Key Performance Indicators*

DM – *Data Mining*

DW – *Data Warehouse*

1. INTRODUÇÃO

1.1Enquadramento

A rápida evolução tecnológica levou à criação de novas fontes de dados e, consequentemente a um aumento no volume dos mesmos. Este volume de dados continua a aumentar, havendo assim ainda mais potencial para extrair desses dados que tragam vantagens para o negócio. Com este aumento explosivo dos dados, a nível global, levou à criação do *Big Data (BD)*. Este termo é utilizado para descrever os dados, em relação ao seu volume, variedade, a velocidade a que são criados, entre outras características.

Big Data é uma combinação de tecnologias de gestão de dados que foram evoluindo ao longo do tempo. *Big Data* permite às organizações gerir, manipular e armazenar grandes quantidades de dados. A partir desses dados é possível determinar se existem padrões ocultos, através de técnicas analíticas, que podem indicar, previamente, uma mudança importante.

Nos dias de hoje as organizações dependem dos dados para terem sucesso e adquirir vantagem competitiva. Assim, é vital para as organizações darem sentido a esses dados, através da sua análise, de forma oportuna. De forma a tomar decisões atempadamente, é necessário recolher os dados e analisá-los em tempo real para gerar informação preditiva com base em modelos matemáticos e estatísticos.

Esta mudança na forma de tomar decisões, levou à evolução do tradicional conceito de *Business Intelligence* (BI) para *Business Analytics* (BA). O *Business Analytics* inclui uma variedade de métodos de análise de dados, como análise estatística. Envolve a exploração iterativa e metódica dos dados de uma organização, com o objetivo de ajudar as organizações na tomada de decisão.

1.2 Objetivos e Resultados

Esta dissertação tem como objetivo responder à seguinte questão de investigação:

Quais os desafios que o Big Data coloca às diferentes abordagens de Business Analytics?

O objetivo principal desta dissertação é a identificação dos desafios que o *Big Data* coloca às diferentes abordagens de Business Analytics. No entanto, apenas foi selecionado o *Data Mining* ao nível preditivo para este estudo.

Ao longo deste documento são identificados vários desafios que o *Big Data* coloca ao *Data Mining*, assim como análises de estudos realizados da performance de duas ferramentas, *KNIME* e *RapidMiner*, na área de *Big Data*. Também se pode encontrar sugestões de boas práticas que poderão ser aplicadas quando se lida com dados.

Como resultado, conclui-se que ambas as ferramentas demonstraram uma boa performance na análise de dados em ambiente *Big Data*, assim como ambas as ferramentas apresentam soluções muito semelhantes para a área de *Data Mining* e *Big Data*.

1.3 Estrutura do documento

O presente documento está estruturado da seguinte forma:

- **Introdução** – Neste capítulo é apresentado o enquadramento do projeto assim como, os objetivos e resultados esperados.
- **Revisão de Literatura** – Neste capítulo está presente o estado da arte dos temas abordados nesta dissertação. O primeiro tema abordado será o *Business Analytics,* de seguida o *Big Data,* e por fim, o *Big Data Analytics* e uma pequena abordagem acerca da plataforma de *Pervasive Data Mining Engine (PDME)*.
- **Abordagem Metodológica** – Neste capítulo é apresentada uma descrição acerca das metodologias a utilizar neste projeto de dissertação.
- **Desafios em Mining *Big Data***- Neste capítulo são apresentados os vários desafios que o *Big Data* coloca ao *Data Mining*, assim como possíveis soluções.
- **Avaliação de ferramentas de *Data Mining***- Neste capítulo é analisado um estudo de avaliação e seleção de várias ferramentas de *Data Mining*.

- **Ferramentas de *Data Mining* – *RapidMiner* e *KNIME* –** Neste capítulo as ferramentas *RapidMiner* e *KNIME* são analisadas.
- **Benchmarking –** Neste capítulo é apresentado o benchmarking realizado na área de *Big Data* e *Data Mining* referente às ferramentas *RapidMiner* e *KNIME*.
- **Discussão –** Neste capítulo é apresentada uma análise de todo o estudo realizado, assim como uma análise SWOT.
- **Conclusão e Trabalho Futuro –** Neste capítulo é apresentada a conclusão desta dissertação, assim como sugestões de trabalhos futuros dentro deste tema.
- **Referências**- Neste capítulo será apresentada uma lista de referências utilizadas no desenvolvimento desta dissertação.

2. REVISÃO DE LITERATURA

2.1 Estratégia de Pesquisa

Para o desenvolvimento desta dissertação, foram utilizados vários motores de pesquisa para a realização da revisão de literatura, onde continham vários artigos científicos publicados, livros, alguns capítulos de livros, entre outros. Assim sendo, os motores de pesquisa utilizados foram:

- *Google Scholar;*
- *Springer;*
- *Scopus;*
- *ScienceDirect;*
- *Google;*
- *RepositóriUM.*

Os termos e palavras utilizadas para a pesquisa foram as seguintes:

- *Business Analytics;*
- *Big Data;*
- *Big Data Analytics;*
- *Benchmarking;*
- *Case Study.*

No entanto, estes termos e palavras foram variando conforme os resultados das pesquisas. Os artigos foram selecionados, tendo em consideração a data dos artigos, citações e o conteúdo do mesmo. Em relação às datas, grande parte dos artigos utilizados têm datas entre 2002 e 2017, no entanto existem algumas exceções.

2.2Business Analytics

Muitas organizações utilizam a informação como um meio de atingir vantagem competitiva. No passado, a análise dos dados era realizada manualmente através de construção de equipas onde se juntavam pessoas da área de estatística, modeladores e analistas. No entanto, com o crescimento dos dados este tipo de abordagem tornou-se inviável (Provost & Fwacett, 2013). A vantagem competitiva é, nos dias de hoje, um aspeto importante para as organizações. Essa vantagem pode ser adquirida através da capacidade de realizar decisões precisas, eficazes e de forma atempada, com o intuito de responder às preferências dos clientes, por exemplo. Para isso, as organizações começaram a utilizar análises avançadas que permitia ter uma visão completa acerca das suas atividades, assim como dos seus clientes (Bose, 2009). Assim sendo, o *Business Analytics* (BA) começou a ser uma ferramenta importante para melhorar a eficiência, a competitividade e a rentabilidade de empresas (Oliveira, McCormack, & Trkman, 2012).

2.2.1 Definição

O *Business Analytics* inclui uma variedade de métodos de análise de dados (Shmueli, Bruce C, Yahav, Patel, & Lichtendahl Jr., 2017) e pode envolver ou não a utilização de software. Para melhorar as suas capacidades de tomada de decisão, os responsáveis pela tomada de decisão utilizam tecnologias como o OLAP (*Online Analytical Processing*) e *dashboards*. O *Business Analytics* inclui também, técnicas como a análise estatística, visualização de dados, modelos de previsão e sistemas de previsão. BA pode incluir sistemas como *Business Intelligence* (BI), *Big Data* (BD) e *Decision Support Systems* (DSS) (Watson et al., 2010).

O BI permite às organizações perceber o que aconteceu e o que está a acontecer numa organização, através de visualização dos dados e relatórios. Esta análise é realizada através da exibição de gráficos, tabelas e *dashboards* onde é possível explorar os dados (Shmueli et al., 2017).

Através da implementação do BA, os gestores conseguem melhorar a performance da organização, identificar oportunidades de negócio e realizar melhores decisões, pois o BA permite integrar dados de várias fontes e, a partir dessa integração, podem descobrir mais informação importante para o negócio (Bayrak, 2015).

De acordo com Watson et al., (2010), este define BA como "uma ampla categoria de aplicações, tecnologias, e processos para reunir, armazenar, aceder, e analisar dados para ajudar os utilizadores a tomar as melhores decisões." Esta definição de *Business Analytics* é a considerada para o desenvolvimento desta dissertação.

2.2.2 Modelo de Business Analytics

O modelo de BA tem como objetivo fornecer uma estrutura geral para perceber e criar BA em qualquer tipo de organização com sucesso. Assim sendo, o modelo de BA é utilizado como referência na criação de BA que permite perceber a interação das pessoas e a interação na criação de informação e posterior consumo (Laursen & Thorlund, 2010) .

O BA envolve a aquisição de conhecimento através da análise de dados e informação, aplicando esse conhecimento para desenvolver e implementar ações competitivas de criação de valor para a organização (Phillips-Wren, Iyer, Kulkarni, & Ariyachandra, 2015).

Como se pode verificar na Figura 1, a criação de BA envolve muitas competências, pessoas e processos. Na segunda camada do modelo, *business-driven environment*, é desenvolvida uma estratégia de informação com base na estratégia de negócios da organização. Na camada seguinte, *operational decision makers*, a seleção da informação e conhecimento é realizada com base na estratégia selecionada para a organização de forma que suporte a mesma. A camada a seguir, *analysts, controllers, and report developers*, são os responsáveis na criação da informação e conhecimento para ser utilizado pelos responsáveis pela tomada de decisão na área operacional, tendo como objetivo a inovação e otimização das atividades do dia a dia da organização. Na camada, *ETL developers and database specialists* no *tecnhnically oriented environment,* os dados são enriquecidos e combinados pelos especialistas em ETL (*Extract, Transform, Load*) para posteriormente serem acessíveis à organização. Na camada seguinte, *IT professionals*, os sistemas de gestão de dados da organização são desenvolvidos e executados pelos profissionais de TI (Tecnologias de Informação) (Laursen & Thorlund, 2010).

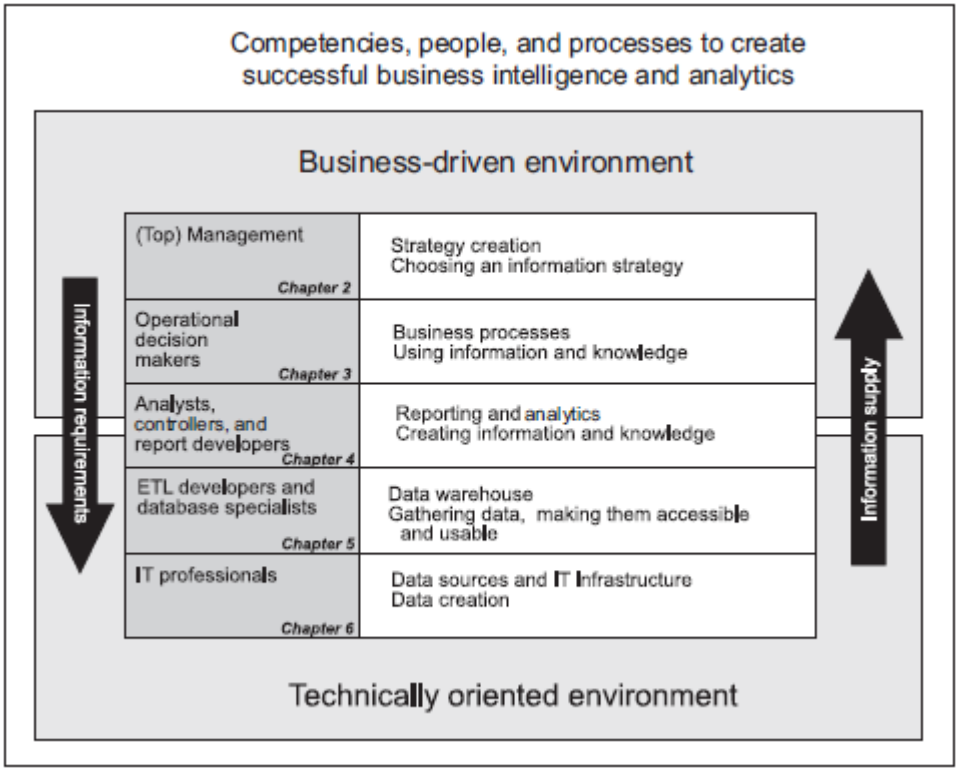

Figura 1- Modelo de Business Analytics .Retirado de (Laursen & Thorlund, 2010)

Posto isto, é importante que os gestores tenham uma visão global do mercado onde a organização está inserida e os seus desafios. A estratégia e os objetivos da organização devem ser bem definidos para conseguir fazer face a esses desafios (Laureano, Miguel da Silva Laureano, & Grencho, 2016).

2.2.3 Capacidades do Business Analytics

Como foi referido no ponto anterior, modelo de *Business Analytics*, as pessoas, os processos e as tecnologias estão envolvidas na aquisição, análise e transformação dos dados utilizada para o apoio na tomada de decisão.

De forma a perceber como as Tecnologias de Informação (TI) trazem vantagens e criam valor para a organização foi criado um modelo de capacidades do *Business Analytics*.

Este modelo foi baseado em estudos realizados anteriormente por vários autores, com objetivo de explicar porque é que a implementação do BA traz benefícios para as organizações. Estudos realizados acreditam que existe uma relação entre as capacidades do Sistemas de Informação (SI) e valor organizacional e vantagem competitiva (Cosic, Shanks, & Maynard, 2015).

Assim sendo, as capacidades do *Business Analytics*, pode ser definido como as interações entre tecnologias de informação, processos, pessoas e outros recursos, para a realização de uma tarefa (Cosic, Shanks, & Maynard, 2012).

Como se pode verificar na Figura 2, de acordo com Cosic et al.,(2015), após realizados os estudos, identificaram 16 capacidades do *Business Analytics* que foram agrupadas em quatro áreas: Governança (*Governance),* Cultura *(Culture),* Pessoas *(People) e* Tecnologia *(Tecnhology).* Governança pode ser entendido como a gestão de recursos de BA e a atribuição de direitos de decisão e responsabilidades, Cultura, são normas organizacionais e padrões comportamentais que se adquirem ao longo do tempo que levam a formas organizadas de recolha e análise de dados. De acordo com Cosic et al.,(2015), Pessoas são aqueles que utilizam o BA como parte do seu trabalho e, por fim, Tecnologia é o desenvolvimento e a utilização de software/hardware e de dados nas atividades de BA (Cosic et al., 2015).

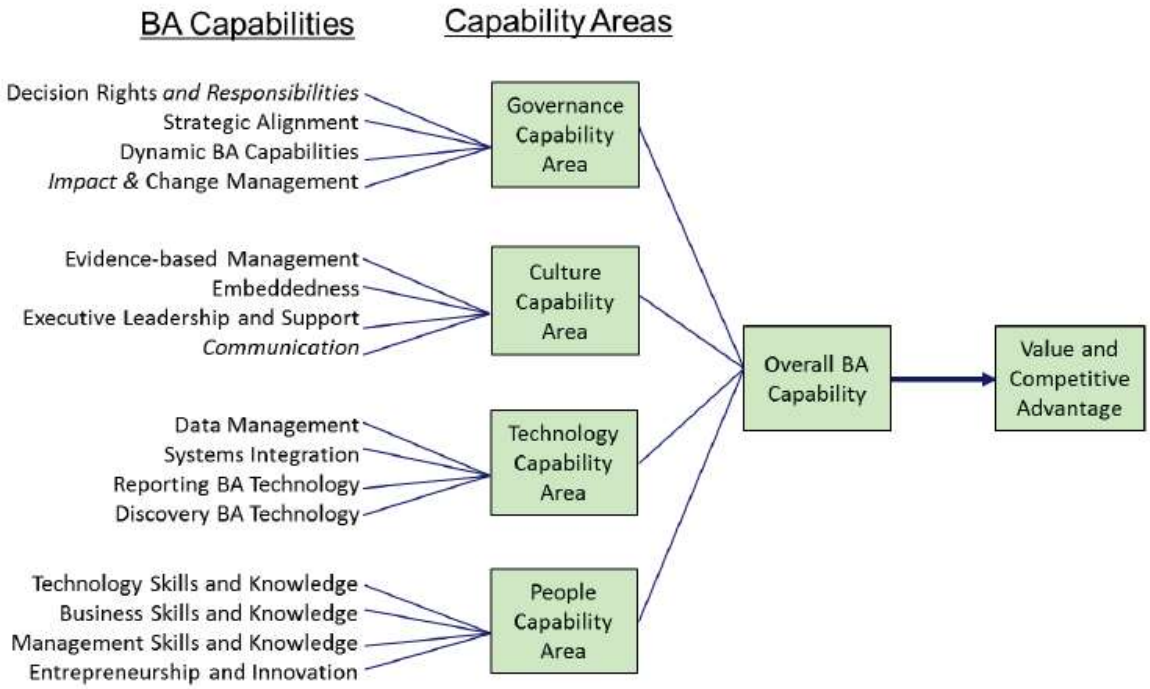

Figura 2-Modelo das capacidades do BA com base no estudo Delphi. Retirado de (Cosic et al., 2015)

De acordo com Cosic et al., (2012), quanto mais madura for a capacidade de BA, maior valor e vantagem competitiva sustentável é a alcançada pela organização.

Segundo Holsapple, Lee-Post, & Pakath, (2014), as capacidades do BA podem incluir:

- A utilização de técnicas que são quantitativas, qualitativas e mistas;
- A utilização de técnicas estatísticas;
- A utilização do raciocínio sistemático;
- Trabalhar com modelos que são: Descritivos/Explicativo, preditivos ou prospetivo;
- Trabalhar com evidências (Exemplo: documentos, sensores, mapas, etc.).

2.2.4 Tipos de Análise

Como já foi referido, o BA ajuda as organizações a tomar as melhores decisões fornecendo aos gestores, de forma mais intuitiva, uma visão do negócio. O BA é um processo que se inicia com um conjunto de dados relacionados com o negócio, sendo composto por três tipos de análises: análise descritiva, análise preditiva e análise prospetiva.

- Análise descritiva responde á questão "O que é que aconteceu?", através da consolidação de dados utilizando *Business Intelligence* e *Data Mining (DM)* para fornecer informações sobre o passado ou acontecimentos do presente e quais as suas possíveis causas (Appelbaum, Kogan, Vasarhelyi, & Yan, 2017; Bayrak, 2015). Esta informação pode ser representada através de *dashboards*, indicadores de desempenho (KPI's – *Key Performance Indicators*) ou outro tipo de visualização. O objetivo deste tipo de análise é obter uma visão geral dos dados, para perceber por exemplo, a frequência em que ocorre determinados eventos (Bayrak, 2015). Este tipo de análise ajuda aos gestores a perceber qual o comportamento do desempenho da organização ao longo do tempo ajudando assim, a realizar uma melhor gestão do negócio.
- Análise preditiva utiliza um conjunto de técnicas e modelos que se enquadram na categoria de *Data Mining*, para prever as tendências futuras com base na análise descritiva, respondendo á questão "O que poderá acontecer?" (Appelbaum et al.,2017). A análise preditiva utiliza dados acumulados ao longo do tempo para a realização de cálculos para prever os possíveis comportamentos ou eventos num futuro próximo, tendo por base esses dados do passado (Philpott, 2010).
- Análise prospetiva responde á questão "O que poderá ser feito?" com base nos resultados da análise descritiva e preditiva. De acordo com Appelbaum et al.,(2017), este acredita que a análise prospetiva vai para além da análise descritiva e preditiva,

pois esta oferece uma ou várias sugestões de soluções mostrando ainda o resultado de cada uma delas. Estas sugestões podem ser sob forma de resposta de "sim" ou "não" para problemas específicos ou, por exemplo, um plano completo de produção (Sharda, Asamoah, & Ponna, 2013). A análise prospetiva ajuda a alocar os recursos com base nas oportunidades que foram previstas para adquirir vantagem (Schniederjans, Schniederjans, & Starkey, 2014).

2.2.5 Desafios do Business Analytics

As organizações que implementam BA devem ter em consideração os desafios e obstáculos para beneficiar ao máximo o BA. Segundo Davenport (2006) as organizações para beneficiar ao máximo o BA e tirar o melhor dos dados que constantemente recolhem e armazenam, devem construir as culturas certas, contratar as pessoas certas e utilizar as tecnologias certas. A segurança e privacidade dos dados tornou-se uma preocupação para os consumidores. Pois o BA envolve o manuseamento de dados pessoais dos consumidores. Com situações de ameaça de roubo de identidade, o manuseamento dos dados dos consumidores torna-se limitada, condicionando a forma como as organizações podem criar listas de clientes, como os podem contactar e construir as suas mensagens. Por isso, o compartilhamento de dados por todas as áreas da organização é desafiante, pois os dados necessitam de ser protegidos e é necessário manter a privacidade e confidencialidade de alguns dados mais sensíveis (Bose, 2009).

Outro desafio identificado por Bose (2009), é a própria utilização das tecnologias. As pessoas precisam de ser treinadas para compreenderem e utilizarem essas tecnologias, caso contrário não conseguem retirar o seu potencial ao máximo. Para além disso, o autor refere também que o resultado final da análise dos dados é desafiante. O resultado precisa de ser simples, conciso, legível e utilizável.

2.3*Big Data*

A evolução da tecnologia nos últimos anos teve como uma das consequências o crescimento de dados em grande escala. Este crescimento fez também com que fosse necessário a criação de novas bases de dados e de tecnologias que fossem capazes de lidar com esta enorme quantidade de dados.

Nos dias de hoje as pessoas lidam com diferentes tipos de sistemas eletrónicos. Com o desenvolvimento de tecnologias como a internet sem fios, *smartphones, laptops e tablets* fez com que estar conectado fosse um fator preponderante no quotidiano da sociedade. Simples atividades como verificar a caixa de *e-mail* ou partilhar conteúdos nas redes sociais geram dados que de alguma forma diz qual a rotina do utilizador e o seu comportamento. Estes dados podem ser utilizados para estudar o comportamento dos utilizadores assim, as empresas podem adaptar os seus serviços e/ou produtos às reais necessidades do seu público(C. Lima & Calazans, 2013).

Devido a esta enorme quantidade de dados e por vivermos numa sociedade que faz uso crescente das tecnologias, começaram a surgir dificuldades em relação ao seu armazenamento e processamento surgindo assim o conceito *Big Data*.

2.3.1 Definição

Existem várias definições para o termo *Big Data*, ou seja, este pode ser entendido de diferentes formas.

Uma das definições mais comum define *Big Data* como sendo uma grande quantidade de dados complexos onde a tecnologia tradicional não tem capacidade para guardá-los, processá-los e visualizá-los para posteriores análises(Sagiroglu & Sinanc, 2013).

Também Chen, Mao & Liu (2014) declara que *Big Data*, de modo geral, é um conjunto de dados que não podem ser entendidos, adquiridos, geridos e processados pelas técnicas tradicionais de tecnologias de Informação e ferramentas de software/hardware.

Para Gupta & Chaudhri (2015), *Big Data* é definido como grande volume, grande velocidade e grande variedade que exigem formas inovadoras e económicas de processamento da informação de forma a melhorar a tomada de decisão.

No entanto, Hurtwitz, Nugent, Halper & Kaufman (2013), referem-se ao *Big Data* como não sendo uma única tecnologia, mas sim uma combinação de tecnologias antigas e novas. Com

esta combinação de tecnologias, segundo o autor, ajuda as organizações a obter conhecimento prático.

As definições de *Big Data* destes autores são muito semelhantes, no entanto há um consenso em relação à incapacidade das ferramentas tradicionais em lidar com este crescimento de dados em grande escala. Porém, Gupta & Chaudhari (2015), afirmam que a definição original de *Big Data* se foca em dados estruturados, mas que os investigadores se aperceberam que a grande parte da informação se encontra em formato não estruturado, sendo que a maioria em forma de texto e imagem.

Para além destas definições acima mencionadas, podemos definir *Big Data* com um conjunto de propriedades associados ao conceito assim como os seus desafios.

É importante referir que, para o desenvolvimento desta dissertação, a definição considerada para *Big Data* foi um grande volume de dados complexos, onde a tecnologia tradicional não tem capacidade para os processar, armazenar e visualizar para posteriores análises.

2.3.2 Características e Desafios

Os autores referidos acima caracterizam o *Big Data* como tendo três principais componentes, conhecidos como os três Vs do *Big Data*: variedade, velocidade e volume. No entanto, alguns autores foram mais longe e acrescentaram outras componentes como variabilidade, complexidade e valor (Katal, Wazid, & Goudar, 2013). Para além destes componentes há autores que consideram a veracidade como uma componente do *Big Data*.

Posto isto, na lista seguinte estão definidas as componentes que caracterizam o *Big Data*:

a) Volume

Grande parte das definições de *Big Data* foca-se no tamanho dos dados. Assim sendo, é obvio que o Volume é uma das principais caraterísticas do *Big Data*. Os dados são criados na escala dos *terabyte*, às vezes *pentabyte*, através de várias fontes e dispositivos (L. C. B. de Lima, 2014). As redes socias produzem, por dia, dados na ordem dos *terabytes* (Katal et al., 2013). Lidar com este grande volume de dados torna-se num importante desafio, pois este não pára de crescer e é difícil lidar com estes dados usando os sistemas tradicionais.

De acordo com Gupta & Chaudhari (2015), é o volume de dados que determina o valor e o potencial dos dados e se pode ser considerado como *Big Data* ou não. O autor afirma também que o nome '*Big Data'* por si só já contém o termo que está relacionado com o tamanho.

Uma das funções do *Big Data* é o processamento de grandes volumes de dados de baixa densidade, dados que não têm valor, como por exemplo cliques em páginas de Web, tráfego de rede, entre outros, transformando esses dados em dados de alta densidade, ou seja, dados com valor(Heller, Piziak, & Knudsen, 2016).

b) Velocidade

Os dados são criados a alta velocidade, consequentemente os dados têm de ser processados com maior rapidez. De acordo com Katal et al., (2013), esta característica não é restringida apenas à criação dos dados, mas também a velocidade a que os dados fluem. Por exemplo, os dados provenientes de sensores estão em constante movimento.

Segundo Maier (2013), o grande desafio da velocidade é o processamento, mas de acordo com Portela, Lima & Santos (2016), esse desafio pode ser resolvido com as capacidades de processamento, como o paralelismo e investimentos em hardware.

c) Variedade

Como já foi mencionado acima, os dados são criados a alta velocidade. Mas, para além disso, estes dados têm origem de uma grande variedade de fontes e, geralmente existe em três tipos: dados estruturados, semi-estruturados e não estruturados. O tipo de dados estruturados são aqueles onde a informação está organizada, como se cada coluna e linha fossem uma etiqueta. Desta forma, é mais fácil analisar os dados. O contrário acontece nos dados não estruturados, por exemplo texto, imagens, entre outros, isto porque, estes tipos dados são aleatórios e difíceis de analisar (Garg, Singla, & Jangra, 2016). Os dados semi-estruturados é uma mistura entre dados estruturados e não estruturados, não têm necessariamente um esquema fixo, mas pode se auto descrever (Hurwitz, Nugent, Halper, & Kaufman, 2013).

De acordo com Lima (2014), esta variedade dos dados é desafiante no que diz respeito ao processamento, armazenamento e gestão dos dados.

d) Variabilidade

Esta característica refere-se à inconsistência dos dados, dificultando a sua gestão.

O carregamento dos dados torna-se difíceis de gerir, especialmente com o aumento da utilização das redes sociais que normalmente geram picos no carregamento dos dados (Katal et al., 2013).

e) Veracidade

A qualidade dos dados pode variar. Para aqueles que analisam os dados a precisão da sua análise depende da veracidade da origem dos dados (Gupta & Chaudhari, 2015). Isto porque, os dados têm origem de várias fontes e essas fontes podem conter dados repetidos, dados não verificados ou dados que não servem para nenhum propósito, entre outros. Para os dados serem úteis estes devem ser confiáveis e limpos, assim sendo é necessário despender tempo em tornar os dados limpos e confiáveis para o uso (Garg et al., 2016).

f) Valor

É necessário saber realizar as questões certas aos dados. O utilizador pode executar certas consultas aos dados e deduzir resultados importantes. Estes resultados, segundo o autor, ajudam as pessoas a encontrar tendências nos negócios, podendo assim, mudar as suas estratégias no negócio (Katal et al., 2013).

No entanto, encontrar valor nos dados requer analistas inteligentes e perspicazes, pois o grande desafio do *Big Data* é o utilizador que está a aprender a fazer as questões certas, através da realização de suposições e a previsão de comportamentos (Heller et al., 2016).

g) Complexidade

A gestão dos dados pode ser um processo complexo devido ao grande volume de dados.

Segundo Katal et al.,(2013), é necessário conectar e correlacionar relacionamentos, hierarquias e as múltiplas ligações dos dados

O rápido crescimento de dados traz consigo desafios em relação ao armazenamento, processamento, gestão e a sua análise. Os sistemas tradicionais de gestão e análise são baseados em bases de dados relacionais, no entanto este tipo de bases de dados só se aplica a dados estruturados, ou a dados não estruturados, ou, ainda a dados semi-estruturados (Chen, Mao, & Liu, 2014).

Como já foi dito, os sistemas tradicionais não conseguem lidar com grandes volumes e heterogeneidade dos dados. Segundo o autor, existem dados com vários níveis de heterogeneidade como estrutura, tipo, semântica organização, granularidade e acessibilidade. Assim sendo, a representação dos dados torna-se num desafio, pois o seu objetivo é tornar os dados mais significativos para análise computacional, assim como para

interpretação do utilizador. Mas se essa representação for realizada de forma incorreta pode enviesar os dados originais e levar a uma interpretação errada (Chen et al., 2014).

Outro desafio considerado é a gestão dos dados, pois estes dados são utilizados para tomada de decisão. Posto isto, é necessário que os dados sejam disponibilizados atempadamente e de forma completa (Katal et al., 2013).

Processar uma grande quantidade de dados também pode ser desafiante pois pode levar a uma grande quantidade de tempo. No entanto, para evitar isso, aquando a obtenção e armazenamento dos dados a criação de índices reduz o tempo de processamento (Katal et al., 2013). Outra forma de reduzir o tempo de processamento, e de modo a minimizar custos de rede, passa pelo processamento no local de armazenamento (Gupta & Chaudhari, 2015).

As ferramentas e técnicas essenciais para lidar com *Big Data* inclui gestão de base de dados, como *Data Warehousing (DW)*, *Data Mining*, *dashboards* e as tecnologias associadas (Davis, 2014).

Para além destes desafios, o *Big Data* conduz a outros processos. Esses processos são a agregação de tecnologias e análises que são utilizadas para definir o valor dos dados (Ohlhorst, 2012).

As melhores tecnologias e conceitos definidos como categorias de análise são as seguintes (Ohlhorst, 2012):

- ***Business Intelligence tradicional* (BI):** consiste em um conjunto de categorias de aplicações e tecnologias para adquirir, armazenar, analisar e providenciar acesso aos dados. Em BI são realizadas análises em profundidade de dados detalhados do negócio, adquiridos através de bases de dados, dados de aplicações entre outros. Em certas circunstâncias, *Business Intelligence* oferece visões históricas, atuais e preditivas de operações de negócio.
- ***Data Mining***: os dados são analisados de diferentes perspetivas para obter dados considerados úteis. As técnicas de *Data Mining* são geralmente utilizadas na modelação e descoberta de conhecimento para previsão.
- **Aplicações Estatísticas:** utilizam algoritmos baseados em princípios estatísticos e geralmente, foca-se em dados relacionados com pesquisas, censos entre outros conjuntos de dados estatísticos. As aplicações estatísticas, tem como ideia principal o

estudo de amostras que podem ser utilizadas para estudar um conjunto de dados com a finalidade de estimar, testar e realizar análise preditiva.

- **Análise Preditiva**: está relacionado com aplicações estatísticas onde são examinados um conjunto de dados com o objetivo de realizar previsões, baseadas em tendências e informação recolhida de bases de dados. Esta análise tem como objetivo identificar riscos e oportunidades para os processos de negócio, mercados e indústrias.
- **Modelação de dados:** aplicação de cenários *"what-if"* através de algoritmos, que podem ser aplicados a vários conjuntos de dados.

Estas categorias de análise são apenas alguns exemplos de porquê o *Big Data* ter valor intrínseco nas organizações e como ajudam as organizações a obter vantagem competitiva.

2.3.3 Tecnologias *Big Data*

Nesta secção serão apresentadas algumas das várias tecnologias *Big Data*.

2.3.3.1 Apache *Hadoop*

Esta *framework* permite o processamento distribuído de grandes dados em *clusters* [1], ou seja, o processamento é distribuído por vários nós de forma a que o processo de maior consumo seja executado no nó mais disponível ou subdividido por vários. Com isto, o sistema continua a trabalhar no caso de um nó falhar (Thillaieswari, Phil, & Ed, 2017). O facto de o Apache *Hadoop* trabalhar em *cluster* significa que todo o sistema tem uma configuração idêntica, ou seja, trabalha em um ambiente homogéneo (J. Singh & Singla, 2015).
O Apache *Hadoop* inclui o HDFS (*Hadoop Distributed File System*) e o *Map Reduce:*

- **HDFS:** O *Hadoop* Distributed File System é um sistema de ficheiros distribuído para armazenar grandes dados em máquinas distribuídas de forma confiável, uma vez que deteta falhas e as recupera automaticamente (Thillaieswari et al., 2017). A arquitetura deste sistema de ficheiros é baseada na arquitetura "master slave". Esta arquitetura consiste em três nós sendo estes, Name Node, Data Node e Secondary Name Node. O Name Node, que representa o master e vários Data Node que representam os slaves,

[1] Fonte: http://*Hadoop*.apache.org/
Data de acesso: 27/08/2018

é o responsável pela forma como os arquivos são divididos em blocos e armazenados pelos vários Data Node. Esses blocos são replicados para proporcionar confiabilidade e disponibilidade. Nos Data Nodes também são executadas tarefas de leitura e escrita controladas pelo Name Node. O Secondary Name Node é o responsável pela leitura periódica do sistema de arquivos, fazendo o registo das alterações com o intuito de ajudar o Name Node nas atualizações, mas não é considerado como um backup (Ghazi & Gangodkar, 2015; Prasad & Rajesh, 2017).

- **MapReduce:** Esta plataforma permite o processamento de grandes dados em paralelo nos clusters. Os dados são divididos em grandes blocos independentes, a cada bloco é atribuído uma "key-value" que são posteriormente processados pela tarefa "Map". O "Map" utiliza como entrada um par de "key-value" e gera uma lista de pares de "key-value" intermediários. De seguida, esses "key-value" intermediários são agrupados com aqueles que têm a mesma chave intermediária. Por fim, a tarefa "Reduce" une as "key-value" que têm a mesma chave e produz o resultado final (Inoubli, Aridhi, Mezni, Maddouri, & Mephu Nguifo, 2017).

2.3.3.2 *Apache Spark*

O *Apache Spark* é um framework de rápido processamento de *Big Data*. Esta plataforma é um sub projeto *Hadoop*, sendo este uma alternativa ao MapReduce. O *Hadoop* suporta ambas plataformas (Jonnalagadda, Srikanth, Thumati, Nallamala, & Dist, 2016).

A velocidade de processamento é um fator importante quando se fala em grandes dados, pois determina a forma como esses dados são explorados, de forma interativa ou esperar minutos ou horas. Esta plataforma oferece processamento em memória e disco, sendo esta mais eficiente do que o MapReduce. Segundo Jonnalagadda et al.,(2016), o Spark ajuda a processar uma aplicação no *Hadoop*, uma vez que é 100x mais rápido no processamento em memória e 10x mais rápido em processamento em disco (Belouch, El Hadaj, & Idlianmiad, 2018; Jonnalagadda et al., 2016).

Para além de ser uma alternativa ao MapReduce, o Spark também oferece um conjunto de ferramentas como o Spark Core onde são construídas outras funcionalidades e extensões, Spark Streaming, Spark SQL para manipular dados usando linguagem SQL, Spark MLLib que

oferece uma biblioteca de algoritmos de machine learning e GraphX para computação gráfica (Inoubli et al., 2017).

2.3.3.3 Apache Storm

Esta framework permite o processamento de grandes dados em tempo real. O Apache Storm é tolerante a faltas, permite a análise em tempo real, machine learning, computação continua, ETL, entre outros (Inoubli, Aridhi, Mezni, Maddouri, & Mephu Nguifo, 2018).
Segundo Morais (2015), o cluster do Storm é ligeiramente semelhante ao do *Hadoop*, uma vez que no *Hadoop* são executados tarefas de MapReduce e no Storm são executadas "topologias". A diferença entre estas execuções é o facto de que uma execução de uma tarefa de MapReduce eventualmente termina, enquanto que uma topologia nunca acaba, a plataforma continua a processar os dados conforme vão chegando, a não ser que seja dada a ordem para terminar.

2.3.3.4 Apache Flink

Esta framework permite o processamento de dados em tempo real e *batch.* O Apache Flink processa os dados a uma grande velocidade. Devido à sua arquitetura pipeline os dados em tempo real são processados com maior rapidez e com menor latência do que as arquiteturas "micro-batch" (Jena, 2017). Esta framework permite ainda que sejam processados dados em tempo real provenientes de outras ferramentas como Flume e Kafka (Inoubli et al., 2017).
O Apache Flink além de ser considerado um processador de dados completo e eficiente, possui uma API especializada para o processamento de dados estáticos, oferecendo funções como join, grouping e bibliotecas para análise de gráficos e machine learning (Carbone et al., 2015).

2.3.3.5 Pig

Pig consiste em uma linguagem de alto nível para expressar programas de análise de dados, tem como uma das características principais o processamento em paralelo que permite lidar com grandes dados. Através do compilador Pig, são produzidas sequências de programas MapReduce. A linguagem Pig consiste em uma linguagem textual denominada Pig Latin (The Apache Software Foundation, 2018).

2.3.3.6 HBase

O HBase é uma base de dados orientada a colunas, tolerante a faltas que é executado sobre o HDFS. Permite gravar e ler dados no sistema de ficheiros HDFS em cenário de tempo real(J. Singh & Singla, 2015).

Os dados no HBase são organizados em tabelas, essas tabelas são compostas por vários arquivos e blocos HDFS, onde cada um é replicado pelo *Hadoop* (Vora, 2011).

Em semelhança com o HDFS e MapReduce, o HBase também apresenta uma arquitetuta baseada em "master slave". O HMaster (master) é o responsável por atribuir regiões a HRegionServers (slave) e por recuperar falhas de HRegionServer. O HBase não suporta linguagem SQL(Vora, 2011).

2.3.3.7 *Hive*

Hive é um software de data warehouse, estando este assente sobre *Hadoop*, utilizado para consultar, gerir e analisar grandes dados. A linguagem utilizada é o *Hive*QL, sendo esta semelhante à linguagem SQL (Structured Query Language) utilizada na consulta de petabytes de dados. É utilizado para a analise de dados no HDFS e oferece suporte para MapReduce (Narasimhan & Bhuvaneshwari, 2014).

Os dados no *Hive* são organizados em tabelas, partições e buckets. Cada tabela tem associada um diretório do HDFS. Cada tabela pode ter uma ou mais partições que determinam a distribuição dos dados dentro dos subdiretórios do diretório da tabela. Os dados em cada partição, podem ser divididos em intervalos com base no hash de uma coluna na tabela. Cada bucket é armazenado como um arquivo no diretório da partição(Thusoo et al., 2009).

2.4*Big Data* Analytics

Como já foi referido, os desafios do *Big Data* passa pela sua análise, armazenamento e processamento e, nos dias de hoje as organizações armazenam e recolhem mais dados, pois é uma parte essencial para obter vantagem competitiva. Assim sendo, a capacidade de analisar grande volume de dados traz vantagens para as organizações.

As organizações necessitam de informações relevantes, para isso precisam de processos eficientes para transformar grande volume de dados em informação relevante para o negócio. Este processo pode ser dividido em dois grupos: gestão de dados e análise. A gestão de dados passa pelo armazenamento e aquisição de dados, através de processos e tecnologias, para posteriores análises. A análise de dados envolve técnicas para adquirir valor a partir de dados importantes. Na Tabela 1 estão descritas algumas técnicas, que abrangem outras técnicas, para a análise de dados estruturados e não estruturados (Gandomi & Haider, 2015):

Tabela 1- Técnicas de Big Data Analytics. Adaptado de (Gandomi & Haider, 2015)

Técnicas	**Descrição**
Text Analytics **(Análise de texto)**	A análise de texto refere-se a técnicas que extraem informações de dados textuais, como por exemplo e-mails, respostas a pesquisas, *feeds* de redes sociais, etc. As análises realizadas envolvem análises estatísticas, *machine learning* e linguagem computacional.
Audio Analytics **(Análise de Áudio)**	A análise de áudio analisa e extrai a informação de dados de áudio não estruturados. Quando aplicado à linguagem humana, é referido também como análise da fala. Os centros de atendimento ao cliente e os cuidados de saúde são as áreas de aplicação primária de análise de áudio.
Video Analytics **(Análise de Vídeo)**	A análise de vídeo envolve técnicas para monitorar, analisar e extrair informações significativas de fluxos de vídeo. A análise de vídeo pode ser utilizada como uma forma de detetar problemas em zonas restritas, identificar objetos removidos, entre outros.

Técnicas	Descrição
Social Media Analytics **(Análise de Redes Sociais)**	A análise de redes sociais refere-se à análise de dados estruturados e não estruturados. As redes socias abrangem várias plataformas *online* que permite aos utilizadores criar e trocar conteúdos.
Predictive Analytics **(Análise Preditiva)**	A análise preditiva envolve várias técnicas que prevê resultados futuros com base em dados históricos e atuais. Pode ser utilizado, por exemplo, para prever os próximos movimentos dos clientes com base naquilo que eles compram e quando compram.

Segundo McKinsey & Company (2011), existem inúmeras técnicas para a análise de dados baseadas, por exemplo, em estatística e ciências da computação. Na Tabela 2, serão descritas apenas algumas técnicas mais especificas, que podem ser aplicadas a grandes e variados conjuntos de dados.

Tabela 2- Técnicas de análise de Big Data. Adaptado de(McKinsey & Company, 2011)

Técnica	Descrição
Association rule learning	É um conjunto de técnicas utilizadas para descobrir relacionamentos. Consistem numa variedade de algoritmos para gerar e testar possíveis regras. Por exemplo, um vendedor pode determinar quais os produtos que normalmente são comprados juntos e utilizar essa informação para comercialização (por exemplo: os compradores que compram fraldas também costumam comprar cerveja).
Machine Learning	Está relacionado com o desenvolvimento de algoritmos que permitem, aos computadores, evoluir comportamentos com base em dados empíricos. Em *Machine Learning* o objetivo é aprender automaticamente a reconhecer padrões complexos e tomar decisões inteligentes. Um exemplo de *Machine Learning* é o processamento de linguagem natural.

Técnica	Descrição
***Neural Networks* (Redes Neurais)**	São modelos computacionais inspirados no funcionamento de redes neurais biológicas, como conexões dentro do cérebro. São utilizados para o reconhecimento e otimização de padrões.
Network analysis	É utilizada para analisar qualquer tipo de rede, por exemplo redes sociais. Na análise de redes socias são mapeados os relacionamentos entre indivíduos.
Visualização	São técnicas para criar imagens, diagramas, gráficos ou animações para comunicar, perceber e melhorar os resultados de análise de grandes dados.

Portanto, o *Big Data Analytics* é a aplicação de técnicas de análise avançada que operam em grandes conjuntos de dados (Russom, 2011) . No entanto, *Analytics* é um termo abrangente para aplicações de análise de dados (Watson, 2014).

Arunachalam, Kumar & Kawalek (2017), afirmam existir um padrão de evolução em relação às terminologias e ao desenvolvimento de capacidades adequadas para a tomada de decisões apoiadas nos dados.

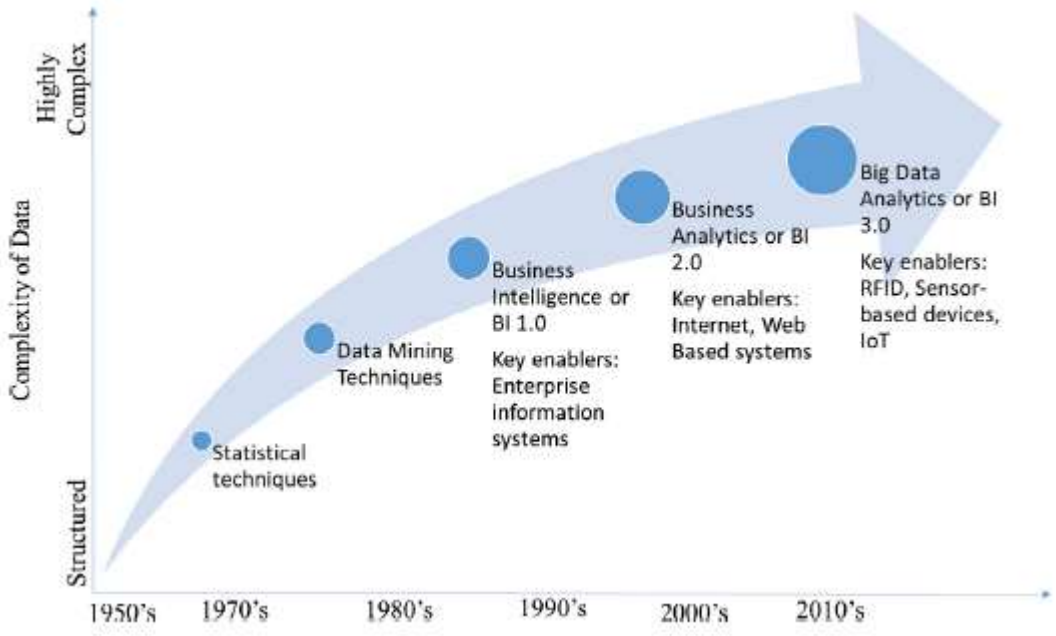

Figura 3- Evolução do Big Data Analytics. Retirado de (Arunachalam, Kumar, & Kawalek, 2017).

Entre as soluções tradicionais de BI e as tecnologias de *Big Data*, a principal diferença identificada por Arunachalam et al,.(2017), é a escalabilidade e a capacidade de armazenar uma variedade de tipos de dados em tempo real. Uma vez que, a maioria dos sistemas tradicionais de BI não são adequados e só podem armazenar e analisar dados estruturados, agregados em intervalos de tempo específicos.

Como se pode ver na Figura 3, o *Big Data Analytics* não é novo, foi evoluindo ao longo do tempo de modo a responder às novas necessidades de processamento de informação das organizações, pois de 1950 a 2010 a complexidade dos dados aumentou gradualmente (Arunachalam et al., 2017).

2.5Plataforma Pervasive *Data Mining* Engine (PDME)

Relativamente à análise de grandes dados, tem vindo a ser desenvolvida uma plataforma on-line que permite realização de análises estatísticas.

Esta plataforma foi desenvolvida com o intuito de tornar mais fácil a utilização de motores de *Data Mining*. Este novo conceito reúne as características gerais dos motores de *Data Mining* com as características da computação generalizada (Peixoto, Portela, & Santos, 2016) .

A construção desta solução foi realizada como auxilio da ferramenta R, base de dados contruída em *MySQL* e das linguagens de programação *HTML, JavaScript, PHP e JQuery* (Ribeiro, 2017).

O objetivo principal desta plataforma é fornecer funcionalidades de *Data Mining* e os seus resultados automaticamente e em tempo real, para qualquer pessoa, em qualquer momento e lugar. Neste estudo foram utilizados dados provenientes do Centro Hospitalar do Porto (CHP) (Peixoto et al., 2016).

Esta plataforma permite realizar análises estatísticas aos dados de forma automática, mostrando aos utilizadores, de forma agradável, simples e de fácil compreensão, os resultados. Realiza ainda diferentes análises conforme o tipo de dados, qualitativos ou quantitativos (Ribeiro, 2017).

2.6 Aplicações do *Big Data* Analytics

Vários setores de indústria já utilizam os "grandes dados" com o intuito de analisar as evoluções de vendas, assim como melhorar ou adaptar os seus produtos/serviços às novas necessidades dos clientes, sejam elas necessidades atuais ou futuras

2.6.1 *Big Data* Analytics na saúde

Na área da saúde é fundamental registar informações acerca dos pacientes, como anotações médicas, relatórios laboratoriais, histórico clínico, relatórios de raio X, regime de dieta, assim como lista de médicos e enfermeiros de um determinado hospital (Archenaa & Mary Anita, 2015).

Por isso, a área da saúde gera, desde sempre, uma grande quantidade de dados provenientes do atendimento ao paciente, manutenção de registos, entre outros (Raghupathi & Raghupathi, 2014). De acordo com os autores, esta grande quantidade de dados promete melhorar a qualidade de prestação de cuidados de saúde à população assim como, apoiar a tomada de decisão nas funções médicas.

De acordo com Raghupathi e Raghupathi (2014), com base nesta grande quantidade de dados, gerados na área da saúde, é possível prever resultados através do estudo de tendências, diagnósticos, tratamentos, entre outros, e assim obter melhores resultados no que diz

respeito ao cuidado do paciente. Com isto, é possível detetar doenças em fases iniciais, onde ainda podem ser facilmente tratadas.

Segundo Archenaa & Mary Annita (2015), os "grandes dados" são necessários na área da saúde para melhorar a qualidade do tratamento do paciente. Isso, segundo os mesmos autores, é possível através dos seguintes aspetos:

- Fornecer serviços centrados no paciente – Através de um tratamento baseado em evidências, detetando doenças nas fases inicias com base nos dados clínicos disponíveis, conduz a um alívio mais rápido ao paciente fornecendo medicamentos eficientes e nas doses necessárias evitando efeitos colaterais.
- Deteção de doenças que se propagam com maior rapidez – Através da análise dos registos sociais dos pacientes, que sofrem de uma determinada doença numa geo-localização especifica, é possível identificar doenças e tomar medidas preventivas necessárias.
- Avaliar a qualidade dos hospitais – Através de uma avaliação periódica é possível verificar se os hospitais estão de acordo com as normas estabelecidas, possibilitando o governo a tomar as medidas necessárias contra a desqualificação dos hospitais.
- Melhorar os métodos de tratamento – Através da análise de dados de pacientes que já sofreram determinados sintomas, ajuda ao médico tomar uma melhor decisão oferecendo assim, um tratamento mais eficaz ao paciente.

Posto isto, a atual segurança e facilidade na partilha de informação auxilia a realização de estudos e ainda, a criação de novas abordagens na prestação de cuidados através da análise de dados históricos e atuais(Groves, Knott, Kayyali, & Van Kuiken, 2013).

2.6.2 *Big Data* Analytics no Governo

Ao contrário do que acontece em grande parte das organizações, no governo as decisões demoram muito mais tempo a serem tomadas pois, normalmente necessitam de consentimento mutuo de um grande conjunto de atores (Kim, Trimi, & Chung, 2014).

De acordo com Archenaa & Anita (2015), os "grandes dados" ajudam o governo a melhorar determinados serviços, como:

- Qualidade na educação – Através dos dados acerca da população, o governo tem conhecimento das crianças que já têm idade para frequentar a escola e assim, consegue avaliar as suas necessidades educacionais.
- Reduzir o desemprego – De forma a minimizar a taxa de desemprego, são realizados estudos preditivos, como o número de estudantes que saem da universidade em cada ano, com o intuito de apresentar a taxa de alfabetização e oferecer cursos especiais para diminuir a taxa de desemprego e de alfabetização.
- Proporcionar as necessidades básicas com maior rapidez – Através da análise diária dos grandes dados as pessoas com necessidades serão detetadas com maior rapidez, com isto o problema será resolvido imediatamente.

Estes são alguns exemplos de como o "*Big Data*" pode ajudar o governo a proporcionar uma melhor qualidade de vida aos cidadãos. Através da sua análise diária os problemas serão detetados mais rapidamente e, consequentemente, serão resolvidos com prontidão.

3. ABORDAGEM METODOLÓGICA

3.1Descrição das abordagens metodológicas

3.1.1 Case Study

A metodologia *Case Study* envolve um estudo intensivo e detalhado de uma entidade bem definida. Esta metodologia é também designada para divulgar os pontos de vista dos participantes, utilizando várias fontes de dados (Coutinho & Chaves, 2002; Tellis, 1997).

De acordo com Noor (2008), o objetivo de um caso de estudo não é estudar uma organização inteira, por exemplo, mas sim se concentrar em uma questão específica, característica ou unidades de análise.

De acordo com Runeson & Host (2009), existem três metodologias, que estão relacionadas com o caso de estudo. Na Tabela 3 estão descritas cada uma delas.

Tabela 3- Definição de metodologias Survey, Experiment e Action Research. Adaptado de (Runeson & Host, 2009)

	Definição
Survey	Fornece uma descrição quantitativa ou numérica de tendências, atitudes ou opiniões de uma população através do estudo de uma amostra de uma população (Creswell, 2014).
Experiência *(Experiment)*	Experiência ou experiência controlada caracteriza-se pela medição dos efeitos de uma variável em outra variável. Testa o impacto de uma intervenção sobre um resultado (Creswell, 2014).
Investigação Ação *(Action Research)*	Investigação Ação tem como objetivo influenciar ou mudar algo, está focada no processo de mudança.

Para Yin citado por Coutinho & Chaves (2002) o caso de estudo pode ter como objetivo explorar, descrever ou explicar.

A metodologia caso de estudo, de acordo com Yin citado por Tellis (1997), apresenta quatro áreas de aplicação, sendo elas:

- Explicar ligações casuais complexas nas intervenções da vida real;
- Descrever o contexto da vida real em que ocorreu a intervenção;

- Descrever a própria intervenção;
- Explorar as situações nas quais, a intervenção avaliada, não possui um conjunto claro de resultados.

3.1.2 Benchmarking

Benchmarking é um processo para medir e comparar, continuamente, processos de uma organização em relação aos melhores, tendo como objetivo melhorar o seu desempenho através de informações úteis (Madeira, 1999). De uma forma mais direta, o *Benchmarking* trata-se de um processo de avaliação e melhoria do desempenho (Maire & Buyukozkan, 1997). De acordo com Elmuti e Kathawala(1997), *Benchmarking* é mais do que um meio para reunir informação sobre o desempenho de uma organização em relação a outras. Para além disso, o *Benchmarking* pode ser utilizado como uma forma de identificar novos modos de melhorar os processos e identificar novas ideias.

De acordo com Sarkis (2001), a literatura sobre metodologias de *Benchmarking* apoiam a abordagem de melhoria contínua composta por quatro fases:

- **Fase 1 – Planear:** Consiste no planeamento do projeto onde é definido o âmbito do projeto, a abordagem de recolha de dados e requisitos e a definição dos critérios (APQC, 2017).

  Esta fase envolve a identificação da estratégia do negócio ou processo a ser comparado, assim como a sua compreensão. É importante começar por identificar quais os pontos fortes e quais os problemas (Maire & Buyukozkan, 1997);
- **Fase 2 – Recolher:** Nesta fase é realizada a recolha dos dados com base na abordagem estabelecida no planeamento (APQC, 2017).

  Esta etapa envolve a recolha de informação sobre a performance e práticas das melhores organizações. A recolha de dados do site é também importante para perceber, de forma mais aprofundada os processos da organização para as melhorias a realizar (B. Singh, Grover, & Singh, 2013);
- **Fase 3 – Analisar:** Esta fase consiste na análise e validação da informação recolhida, para identificar os níveis de desempenho, indicadores e modelos de desempenho (APQC, 2017).

Esta fase permite identificar falhas de desempenho e as possíveis causas, assim como as melhores práticas (Sarkis, 2001; B. Singh et al., 2013) .

- **Fase 4 – Adaptar:** Esta última fase consiste em desenvolver um plano de ação para a mudança (APQC, 2017).

  Esta última fase envolve esforços de mudança da organização com o objetivo de melhorar o seu desempenho de acordo com as conclusões do estudo (Madeira, 1999; Sarkis, 2001).

  Estas fases de *Benchmarking* incorporam um processo como se pode ver na Figura 4.

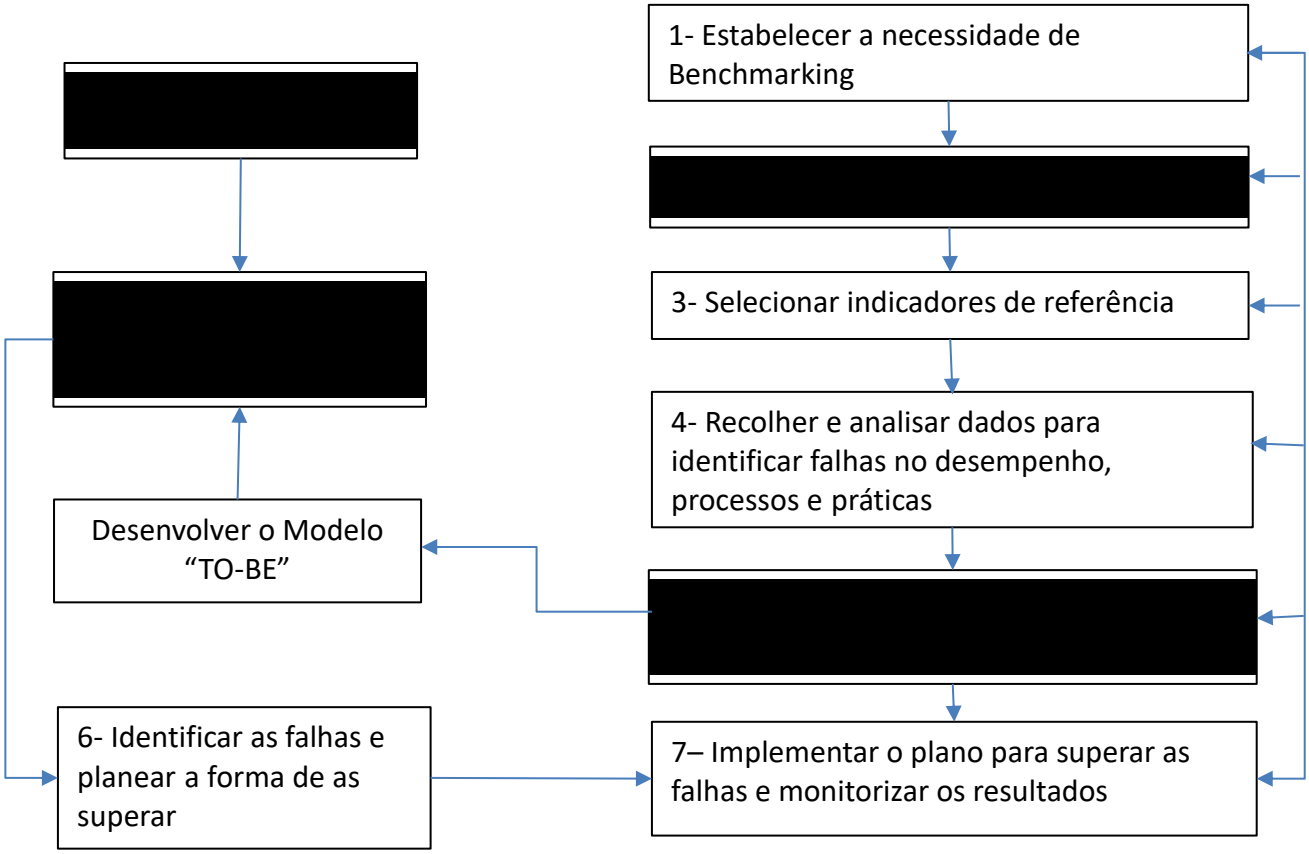

Figura 4 - Processo de Benchmarking. Adaptado de (Singh, Grover, & Singh, 2013)

3.2 Aplicação da abordagem metodológica

3.2.1 Metodologia Case Study

Inicialmente, depois de definidas as fontes de dados a utilizar, foi realizada a recolha de informação referente a Business Analytics, *Big Data* e as suas aplicações. Devido às vastas abordagens de Business Analytics foi definido qual a área de Business Analytics a abordar, assim sendo foi selecionada a área de *Data Mining* em contexto *Big Data*. De seguida, foram analisados estudos realizados referentes às diferentes ferramentas de *Data Mining* e quais dessas ferramentas permitem a análise de *Big Data*.

3.2.2 Metodologia Benchmarking

Nesta metodologia, o objetivo consistiu em encontrar experiências comparativas de desempenho na análise de *Big Data*, de ferramentas de *Data Mining* que permitissem a análise de dados em ambiente *Big Data*.

Como descrito no ponto 3.1.2 foram seguidas 4 fases. Na fase 1 (Planear), foram definidas algumas palavras-chave para a pesquisa de experiências comparativas:

- *Benchmarking;*
- *Performance comparison;*
- *Comparison;*
- *Comparative study;*
- *Comparative analysis*.

As principais fontes utilizadas para a pesquisa, foram as seguintes:

- *Google Scholar;*
- *Science Direct;*
- *Google*.

Nesta pesquisa apenas foram selecionados artigos que não excedessem 5 anos da data de publicação, de modo a haver um maior rigor.

Na fase 2 (Recolher), foram recolhidas algumas experiências com base nos resultados apresentados utilizando as palavras-chave definidas na fase 1. Foram pesquisadas vários estudos comparativos e benchmarkings realizados às ferramentas previamente selecionadas para o benchmarking.

De seguida, na fase 3 (Analisar) as experiências recolhidas na fase anterior foram analisadas e selecionadas aquelas que mais se enquadravam com o objetivo desta dissertação e as ferramentas selecionadas.

Por fim, na fase 4 (Adaptar), foram apresentadas um conjunto de sugestões de boas práticas assim como, foram identificados os prós e contras desta abordagem em contexto organizacional.

4. DESAFIOS EM MINING *BIG DATA*

Analisar grandes conjuntos de dados pode ser desafiante, isto porque muitas das ferramentas não são capazes de lidar com tais dados. Com as redes socias como o Facebook, Twitter, Instagram, entre outras, são gerados grandes quantidades de dados dia após dia a grande velocidade. Há dados que precisam de ser processados em tempo real e que podem ser utilizados para a prever o comportamento de um mercado de ações, por exemplo. Para a análise de dados em tempo real são necessárias ferramentas que suportam esse tipo de dados (Hashmi & Ahmad, 2016). Alguns dados necessitam de ser processados logo após o evento, caso contrário perdem o seu valor e já não terão utilidade. MapReduce permite o processamento de um conjunto fixo de dados mas não é apropriado para processamento de *streaming* em tempo real. Os sistemas de processamento de *streaming* têm uma grande capacidade de processar vários eventos (Ounacer, Talhaoui, Ardchir, Daif, & Azouazi, 2017). Assim sendo, o processamento de streaming é importante porque permite o processamento dos dados aquando da sua produção. MOA (Massive Online Analysis) e SAMOA (Scalable Advanced Massive Online Analysis) são exemplos de ferramentas utilizadas para processamento de s*treaming.*

Assim sendo, as tradicionais metodologias ou ferramentas de *Data Mining* não possibilitam o armazenamento, gestão e análise de *Big Data*. No entanto surgiu o *Big Data Mining*, que é a capacidade de extrair informação útil de um grande conjunto de dados ou streams de dados que, devido ao seu volume, variedade e velocidade a que são criados, não era possível antes (Jaseena & David, 2014).

No entanto, lidar com *Big Data* por si só já é desafiante, a sua mineração continua a ser um desafio. Os autores identificaram como desafios de *Big Data Mining* (Jaseena & David, 2014):

- Heterogeneidade;
- Escalabilidade;
- *Timeliness*;
- Complexidade;
- Privacidade.

4.1Heterogeneidade

Como já foi referido, *Big Data* é definido pela sua variedade, podendo ser estruturado, semi-estruturado e não estruturado. No entanto, a presença de diferentes regras ou padrões nos dados torna-se num desafio para a sua análise. A conversão de dados não estruturados para dados estruturados de acordo com Jaseena & David (2014), é um desafio para *Big Data Mining* Dados incompletos também pode ser considerado um desafio, devido às incertezas que estes podem criar. Em *Data Mining* existe uma forma de lidar com dados incompletos, essa forma baseia-se em ignorar os dados em falta ou inserção de dados. Na inserção de dados, as falhas são preenchidas com o objetivo de melhorar o modelo criado com os dados originais (Jaseena & David, 2014).

4.2 Escalabilidade e Complexidade

Como já foi dito, as ferramentas tradicionais não são adequadas para lidar com este crescimento dos dados e para processamento em tempo real, lidar com o volume dos dados por si só já é um desafio. Assim sendo, a organização, análise de dados, recuperação e modelação dos dados são também um desafio devido à escalabilidade e complexidade dos dados a serem analisados (Jaseena & David, 2014).

Este desafio da escalabilidade pode ser ultrapassado através do processamento em paralelo. Em *Big Data* é utilizado o modelo MapReduce criado pela Google, que, numa forma muito breve, permite o processamento de várias entradas de dados de forma paralela(Che, Safran, & Peng, 2013).

4.3 Timeliness and Privacidade

Devido ao volume e à complexidade dos dados, a análise destes pode consumir muito tempo. No entanto, existem situações onde os resultados da análise dos dados são necessários no momento. Devido à importância dos dados na tomada de decisão, a análise dos dados deve ser realizada num certo período de tempo, caso contrário os resultados já não terão o mesmo

valor ou podem até ser considerados inúteis (Jaseena & David, 2014)(Atanassov & Al-Barznji, 2017).

A privacidade é também um desafio porque nas ferramentas utilizadas para análise, armazenamento, gestão, etc., são analisados dados de várias fontes. Posto isto, existe o risco de exposição de dados pessoais, tornando-os vulneráveis. Assim sendo, os analistas devem utilizar os dados com muito cuidado (Jothi, Amudha, & J, 2018).

Cada vez mais dados pessoais, como o número de cartão de crédito, número de segurança social, entre outros, estão em risco de ser expostos, isto porque nos dias de hoje tudo é feito pela internet, desde compras online ao acesso às contas bancárias. Com isto, cada vez mais é necessário seguir todas as políticas de segurança de modo a minimizar o risco de exposição de dados (Che et al., 2013).

4.4Lidar com os dados

A análise de um grande volume de dados pode trazer alguns desafios. No entanto, alguns deles podem ser ultrapassados com a aquisição de novas tecnologias, hardware, software, entre outros. Porém, é necessário olhar para todos os dados e saber tirar o máximo deles, com base em um objetivo previamente definido. A seguir são apresentadas algumas sugestões de boas práticas quando lidámos com *Big Data Mining*:

- **Colocar as questões certas:** numa organização circulam vários dados de diversas origens, assim sendo é necessário saber aquilo que a organização necessita. Posto isto, é necessário saber colocar certas questões, dependendo do propósito e requisitos da organização, e também aquilo que ser quer como resultado. Posteriormente, os dados têm de ser trabalhados de acordo com o que se definiu anteriormente, isto é, os dados é que devem guiar o processo e não as tecnologias. No entanto, as pessoas também são importantes em uma organização e precisam de estar preparadas para possíveis mudanças que possam ocorrer na organização.
- **Saber que dados utilizar:** para além de saber o que uma organização necessita, é necessário saber que dados utilizar. Devido à imensidão dos dados, nem todos podem ser analisados e explorados, assim como pode existir dados irrelevantes para o objetivo geral da organização ou mesmo para o objetivo do momento. Por vezes pode

ser possível alcançar um objetivo com pequenos grupos de dados, não sendo necessário analisar todos os dados. Com isto, a organização deve ser capaz de identificar que dados são os mais relevantes e que podem responder às questões previamente colocadas, com isto o objetivo da organização poderá ser alcançado mais facilmente.

- **Seleção de ferramentas:** as ferramentas necessitam de ser adaptadas às necessidades da organização. Adquirir novas ferramentas pode ser dispendioso, no entanto pode trazer várias vantagens. Com o tempo as necessidades mudam, sendo necessário adaptar as ferramentas às novas necessidades, analisando para quais funções as ferramentas atuais são utilizadas, se são fáceis de utilizar e intuitivas, e se possui todos os requisitos desejados para uma análise de dados com qualidade.

5. AVALIAÇÃO DE FERRAMENTAS DE *DATA MINING*

Os dados têm cada vez mais uma grande importância e impacto nas organizações, e a capacidade destas em retirar valor dos mesmos é crucial no seu crescimento.

Assim sendo, Ventura et al.,(2017) realizou um estudo comparativo entre 19 ferramentas open-source de *Data Mining* e descoberta de conhecimento (Knowledge Discovery). Este estudo tem como objetivo avaliar as 19 ferramentas e apresentar todos os requisitos que uma ferramenta deve satisfazer. A avaliação foi realizada com base em duas metodologias, a primeira metodologia é baseada em pontuações dadas por especialistas da área de forma a avaliar as ferramentas em vários critérios. Na segunda metodologia é feita uma análise objetiva onde em cada ferramenta é verificado se satisfaz determinado requisito.

As duas melhores ferramentas de *Data Mining* consideradas neste estudo, serão posteriormente analisadas em relação à área de *Big Data*.

Ao longo deste ponto será descrito com mais detalhe este estudo.

5.1Ferramentas avaliadas

Como já foi referido, neste estudo realizado por Ventura et al., (2017) foram avaliadas as seguintes 19 ferramentas:

ADaM – Algorithm Development and Mining;

ADAMS – Advanced *Data Mining* and Machine learning Systems;

AlphaMiner;

Cramer Modelling Segmentation and Rules;

Databionic ESOM;

DataMelt;

ELKI – Environment for developing KDD-applications supported by Index-structures;

GDataMine - The gnome data mine tools;

KELL – Knowledge Extraction based on Evolutionary Learning;

KNIME – Konstanz Information Miner;

MiningMart;

ML-Flex;

Orange;

RapidMiner;

Rattle – the R Analytical Tool To Learn Easily;

SPMF – Sequential Pattern Mining Framework;

Tanagra;

VW - Vowpal Wabbit;

WEKA – Waikato Environment for Knowledge Analysis.

Esta lista de ferramentas foi retirada do KDnuggets, sendo este um website líder em análise de negócios, *Big Data*, mineração de dados, ciência de dados e aprendizado de máquina.
Cada uma destas ferramentas foi avaliada segundo as duas metodologias anteriormente referidas, com base em vários critérios.

5.2 Metodologias e critérios de avaliação

Como já foi referido, na avaliação destas 19 ferramentas de *Data Mining* foram utilizadas duas metodologias. Na primeira metodologia, denominada como procedimento de pontuação, especialistas da área avaliaram as ferramentas dando-lhes pontuações.

Nesta metodologia, as ferramentas foram avaliadas segundo quatro grandes categorias:

- Desempenho;
- Funcionalidade;
- Usabilidade;
- Suporte para atividades suplementares.

Em cada uma destas quatro categorias, foram considerados vários critérios como a sua capacidade para diferentes tamanhos de dados, compatibilidade com várias plataformas, suporte para validação de modelos, entre muitos outros.
Na segunda metodologia utilizada para avaliar as ferramentas, denominada como procedimentos de caracterização, as ferramentas foram avaliadas segundo quatro categorias:

- Requisitos do Sistema;

- Tipos de Abordagens;
- Atividades Complementares;
- Características da Interface do Utilizador.

Da mesma forma como na metodologia anterior, foram considerados vários critérios como que ferramenta trabalha em várias plataformas, tem capacidade para regressão, associações, deteção de anomalias, capacidade para transformação dos dados, análise ROC, entre muitos outros.

5.3Resultados do estudo

Na aplicação da primeira metodologia, as pontuações foram dadas por especialistas que tiveram como referência a ferramenta WEKA para avaliação das restantes. Foram atribuídas pontuações numa escala de 1 a 5, sendo posteriormente dado pesos apropriados a cada um dos critérios. Os resultados deste estudo, realizado por Ventura et al.,(2017) estão apresentados na Figura 5.

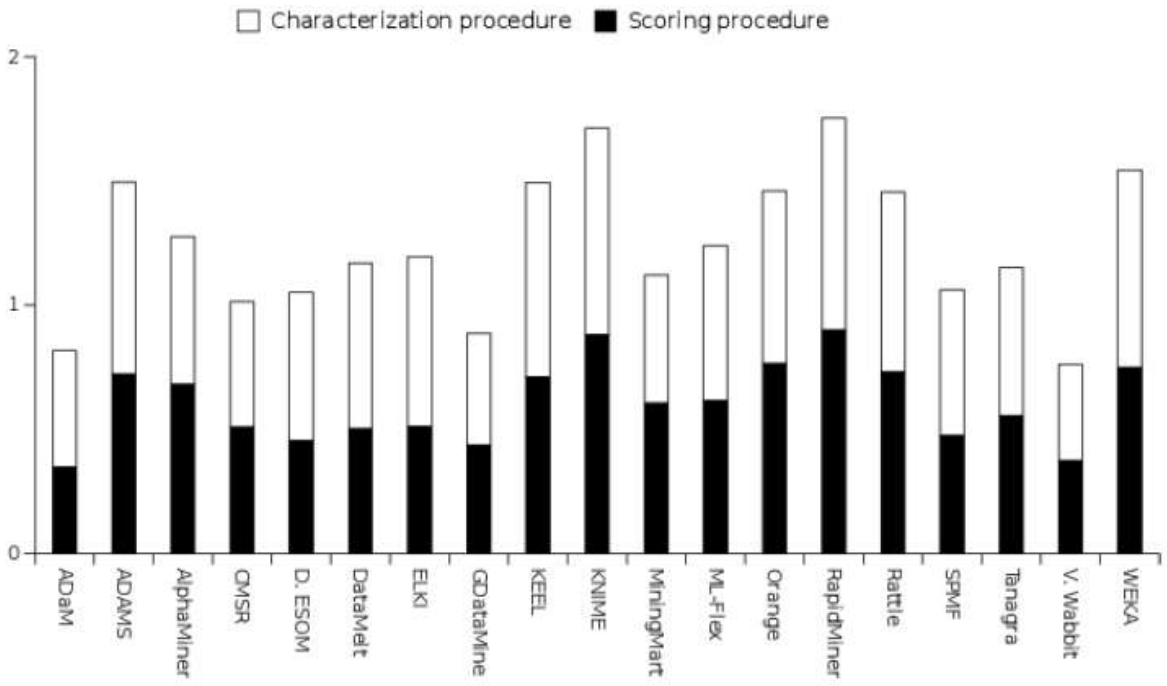

Figura 5- Resultado da aplicação das metodologias - Retirado de Ventura et al.,(2017).

Como se pode ver pelo gráfico da figura 5, as duas ferramentas que tiveram melhores resultados na aplicação das duas metodologias foram o *KNIME* e *RapidMiner*. Quer isto dizer

que estas ferramentas tiveram a pontuação mais alta, na aplicação da primeira metodologia, e são as mais completas, na aplicação da segunda metodologia.

Assim sendo, estas duas ferramentas foram analisadas em relação à área de *Big Data*, tanto em relação à sua performance na análise de dados *Big Data*, assim como à sua oferta em relação às opções para análise de *Big Data*.

6. FERRAMENTAS DE *DATA MINING – RAPIDMINER* E *KNIME*

Neste ponto pode-se encontrar uma descrição geral dos vários produtos que as ferramentas *KNIME* e *RapidMiner* oferecem, assim como das soluções *Big Data* de cada uma das ferramentas.

6.1*RapidMiner*

RapidMiner é uma plataforma de análise de dados que oferece ao utilizador uma libraria de 1500 algoritmos de *machine learning* e funções para construir modelos preditivos mais precisos[2].

Esta ferramenta foi desenvolvida em 2001 por Ralf Klinkenberg, Ingo Mierswa e Simon Fisher(Bisht, Negi, Mishra, & Chauhan, 2018). O *RapidMiner* é composto por quatro produtos: *RapidMiner* Studio, *RapidMiner* Server, *RapidMiner* Radoop e *RapidMiner* Cloud. Na Tabela 4 estão descritos, de forma sucinta, cada um destes produtos, assim como algumas das suas capacidades.

Tabela 4- Descrição da ferramenta RapidMiner

Produto	***Descrição***
RapidMiner Studio	Este produto oferece um ambiente gráfico (GUI - graphical user interface), ou seja, permite interagir com a ferramenta através de elementos gráficos. Este tipo de interface ajuda na produtividade dos utilizadores pois, torna o processo de criação de modelos muito mais simples. Cada análise é um processo, cada transformação ou cada ação na análise é um operador. Deste modo, a compreensão do processo torna-se mais fácil. Permite aceder, carregar e analisar qualquer tipo de dados até mesmo, transformar dados não estruturados em dados estruturados. Permite preparar os dados através de operadores como join, merge, append, union, entre outros. Permite a criação de vários modelos através de algoritmos

[2] Fonte: https://*RapidMiner*.com/products/studio/
Data de acesso: 12/08/2018

	como Naive Bayes, Regressão, entre outros, assim como avaliar a performance dos mesmos[3].
Produto	***Descrição***
RapidMiner Server	Esta plataforma permite aos utilizadores partilhar conhecimento assim como, as melhores práticas por toda a organização. É ainda possível aumentar a velocidade do processo de modelação devido ao hardware de alto desempenho[4].
RapidMiner Radoop	O Radoop é uma das extensões do *RapidMiner* Studio, sendo este voltado para análises *Big Data*, e permite a execução de operadores complexos no *Hadoop*. Suporta Cloudera, Hortonworks, MapR, Amazon EMR, Apache, Microsoft's Azure HDInsight. Oferece vários operadores para a realização de ETL (*Extract*, *Transform, Load*), modelos como o *Naive Bayes*, *Decision Tree* entre outros[5].
RapidMiner Cloud	Esta plataforma permite ao utilizador armazenar e executar os modelos na *cloud*[6].

Esta plataforma utiliza uma metodologia designada *"drag-and-drop"*, ou seja, o utilizador pode selecionar os vários operadores que a ferramenta oferece, como já foi mencionado em cima, e combiná-los de diferentes formas, tornando mais intuitivo o processo de criação (Aggarwal, 2015) .

[3] Fonte: https://*RapidMiner*.com/products/studio-2/feature-list/#application
[4] Fonte: https://*RapidMiner*.com/products/server/
[5] Fonte: https://*RapidMiner*.com/products/radoop/feature-list/
[6] Fonte: https://*RapidMiner*.com/products/cloud/
Data de acesso: 12/08/2018

6.2*KNIME*

Esta plataforma, designada *Konstanz Information Miner* (*KNIME*), é uma ferramenta *open-source* de análise de dados, que auxilia os utilizadores a descobrir o potencial dos dados assim como realizar previsões (Bisht et al., 2018).

Esta ferramenta foi desenvolvida em 2004 na Universidade de Konstanz, por uma equipa de desenvolvedores de software do Silicon Valley especializados em aplicações farmacêuticas. A primeira versão do *KNIME* saiu em 2006 e foi sendo utilizada por várias empresas farmacêuticas[7].

O *KNIME* oferece quatro produtos: *KNIME* Analytics Plataform, *KNIME* Server, *KNIME* Extensions e *KNIME* Integrations. Na Tabela 5 estão descritos cada um destes produtos, de forma sucinta, assim como algumas das suas capacidades.

Tabela 5- Descrição da ferramenta KNIME

Produto	***Descrição***
KNIME Analytics Plataform	Esta plataforma oferece ao utilizador uma interface gráfica (GUI- Graphical User Interface) do estilo *drag-and-drop.* Oferece ainda mais de 2000 módulos para criação de *workflows.* Possibilita combinação de vários formatos de ficheiros, desde texto simples a dados não estruturados ou séries temporais e, ainda, possibilita a conexão a várias bases de dados e *data warehouses* como Apache *Hive*, entre muitas outras. Permite a realização do processo de limpeza dos dados, através de conversão de tipo de dados, manipulação de valores ausentes e, ainda, a deteção *outliers* e anomalias. É possível a criação de modelos de *machine learning* como classificação, regressão, entre outros.[8]

[7] Fonte: https://www.*KNIME*.com/*KNIME*-open-source-story
Data de acesso: 22/08/2018
[8] Fonte: https://www.*KNIME*.com/*KNIME*-software/*KNIME*-analytics-platform

Produto	*Descrição*
KNIME Server	Esta plataforma permite aos utilizadores partilhar dados, *workflows,* práticas recomendadas por toda a equipa, assim como possibilita hospedar o *KNIME* Server na *cloud.*[9]
KNIME Extensions	*KNIME Extensions* oferece funcionalidades adicionais à ferramenta *KNIME Analytics Plataform.* Algumas dessas funcionalidades são: integração de dados da Amazon Athena e Redshift, *Hive*, Impala, etc., e ainda, infraestruturas *Big Data* como a *Apache Spark*. Permite trabalhar com ficheiros de vários formatos.[10]
KNIME Integrations	*KNIME Integrations* oferece acesso a grandes projetos *open-source* como o *Apache Spark* para o processamento de *Big Data*. Possibilita o acesso, importação e exportação de dados no *Hive, Impala ou HDFS (Hadoop Distributed File System)*, através do *KNIME Analytics Plataform*, entre muitas outras funcionalidades.[11]

Para além destas funcionalidades, esta plataforma oferece exemplos já construídos e a explicação de cada operador, o que ajuda o iniciante no mundo do *Data Mining.*

[9] Fonte: https://www.*KNIME*.com/*KNIME*-software/*KNIME*-server
[10] Fonte: https://www.*KNIME*.com/*KNIME*-software/*KNIME*-extensions
[11] Fonte: https://www.*KNIME*.com/*KNIME*-software/*KNIME*-integrations

7. BENCHMARKING

7.1 *RapidMiner* vs *KNIME*

Na Tabela 6, são apresentadas algumas das várias características, algumas delas já mencionadas nos pontos 6.1 e 6.2, que estas ferramentas oferecem para a análise de *Big Data*.

Tabela 6-Características KNIME e RapidMiner

Características	***KNIME***	***RapidMiner***
Interface	- Interface gráfica intuitiva com modo "*drag and drop*"; - Suporte para *scripting* em R e *Phyton;* - Oferece mais de 2000 módulos para criação de *workflows.*	- User-friendly com operadores "*drag and drop*"; - Mais de 1500 operadores para transformação e análise; -Suporte para ambiente *scripting* como R.
Acesso a dados	-Acesso a vários tipos de dados como CSV, XML, JSON entre outros; -Acesso a várias bases de dados como Microsoft SQL, Oracle entre outros; -Acesso a dados de várias origens como *Azure, Twitter* entre outros;	-Acesso, carregamento e analise de qualquer tipo de dados; -Acesso ao armazenamento em nuvem como a Dropbox; -Suporte a todas conexões de bases de dados JBDC.
Transformação de dados	-Módulos como *aggregate, sort, filter e join;* -Deteção de *outliers* e anomalias; -Limpeza dos dados através de normalização, conversão de dados entre outros;	-Operadores como *join, merge and append;* - Filtração de *outliers;* -Vários operadores de transformações de dados como normalização, conversão, rotação de datasets, entre outros;

	- Módulos de estatística como medias, quartis, desvio padrão entre outros.	- Eliminar atributos sem utilidade e com valores em falta;
Características	***KNIME***	***RapidMiner***
Machine Learning	- Permite através de algoritmos avançados, como algoritmos *deep learning*, métodos *tree-based,* a criação de módulos de *machine learning* para classificação, regressão, redução de dimensão, *clustering,* redes-neuronais entre outros; - A otimização de módulos pode ser realizada através da otimização *hyperparameter, boosting, building complex enables,* entre outros.	-Permite criar módulos de *machine learning* para *clustering*, regressão, redes-neuronais, *support vector machines*, árvores de decisão, entre outros. - Oferece modelos como ada *boost, Bayesian boosting, bagging,* classificação por regressão, entre outros.
Validação	-A validação do modelo pode ser realizada através de *accuracy, ROC curve, AUC, R2* e *cross validation* para a estabilidade do modelo.	-Oferece critérios de validação como *accuracy*, *AUC, ROC*, classificação de erro, *recall* entre outros; -Permite a comparação de modelos através do *t-test and anova.*
Big Data	**Extensões *Big Data*** -Acesso a Apache *Hadoop*/HDFS via *Hive* ou Impala, através *KNIME* Platform, sem necessidade de codificação	**Extensão Radoop**: -Integração de SparkR e scripts PySpark; -Suporta *Hive* em Spark e *Hive*-on-Tez;

	Permite: -Mover dados entre *KNIME* platform e *Hive*/Impala; -Escrever queries SQL em *Hive*/Impala; -Processamento de queries SQL diretamente no *Hive* e Impala; -Extensão para *Apache Spark*; - Permite escrita e leitura, importação e exportação, acesso a dados no HDFS, *Hive* e *Impala*; - Análises Preditivas e *scoring* na *Apache Spark* utilizando modelos PMML desenvolvidos no *KNIME* platform; - Oferece novos *KNIME* nodes para I/O, manipulação, *machine learning*, pontuação estatisticas, entre outros, para permitir a criação e execução de aplicações do *Apache Spark* no *KNIME* plataform.	-Acesso ao cluster *Hadoop*; - Oferece operadores para ETL como *join, aggregate, replace* entre ouros; **Permite:** -Leitura, armazenamento, anexo de e para *Hive*; -Leitura de ficheiros CSV do HDFS, *Amazon S3*, *Azure Blob* e do sistema de ficheiros local; -Armazenamento e combinação de resultados no *Hive* ou Impala; -Gestão de tabelas *Hive* ou Impala; -Tranformação de dados no *Hive*QL ou *Pig*.

Como se pode verificar estas ferramentas apresentam muitas semelhanças nas soluções *Big Data* que oferecem, como a sua interface, operadores para transformação dos dados, algoritmos de *machine learning*, assim como as bases de dados *Big Data*.

7.2*RapidMiner* Radoop vs *RapidMiner* Analytics

Gaspar et al., (2015)realizaram uma experiência para comparar a performance entre o *Radoop*, uma extensão do *RapidMiner*, e o *RapidAnalytics* (RA), um servidor *open-source* que suporta todo o processo de ETL utilizado como referencia. Esta experiência consistiu na execução de várias tarefas de transformações de dados no *Hadoop* cluster (*Radoop*) e no *RapidAnalytics* (in memory).

Nesta experiência, foi avaliado o desempenho do *Radoop* e o seu comportamento na mudança do volume dos dados e o número de nodes de processamento. Foram utilizados 4 a 16 nodes e dados de 128MB a 8GB no *Radoop* e 128MB a 1GB no *RapidAnalytics*.

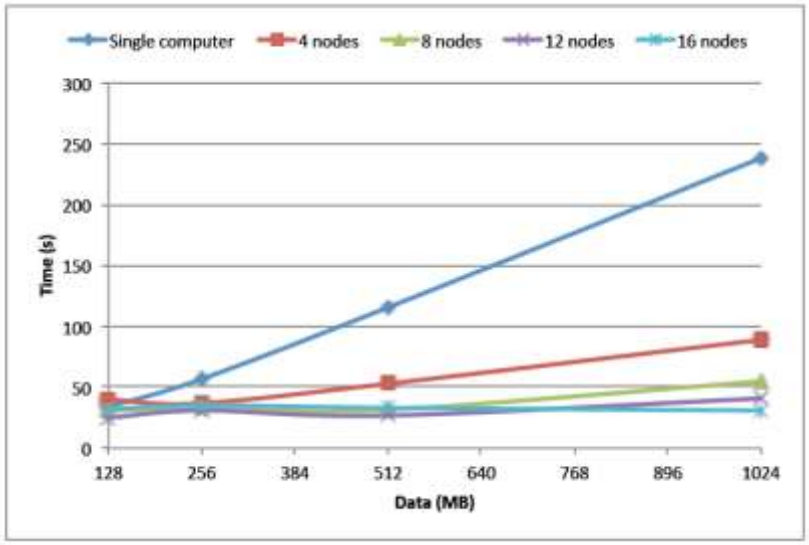

Figura 6- Resultado do tempo de processamento de pequenos dados (RA). Retirado de Gaspar et al., (2015).

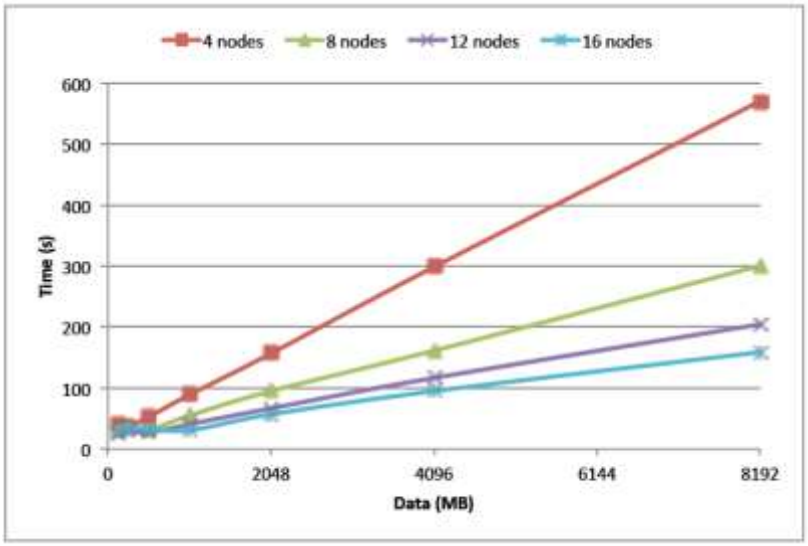

Figura 7- Resultado do tempo de processamento de grandes dados- Radoop. Retirado de Gaspar et al., (2015).

Como se pode verificar pelos gráficos das Figuras 6 e 7, ambas RA e *Radoop* escalam linearmente com o aumento do tamanho dos dados, no entanto o *Radoop* termina o processo muito mais rápido, mesmo com 4 nodes de processamento. Porém, nos dados mais pequenos, como 128MB, o tempo de processamento é o mesmo que no *Hadoop* cluster porque cada bloco de 64MB (bloco padrão) é processado por apenas um node, ou seja, um bloco com 128MB é processado por dois nodes.

Na segunda experiência, referente a transformações de dados, depois de carregados os dados foram selecionados vários atributos, filtrados alguns exemplos e agregados de acordo a um atributo, posteriormente, as novas colunas foram renomeadas e os resultados foram salvos.

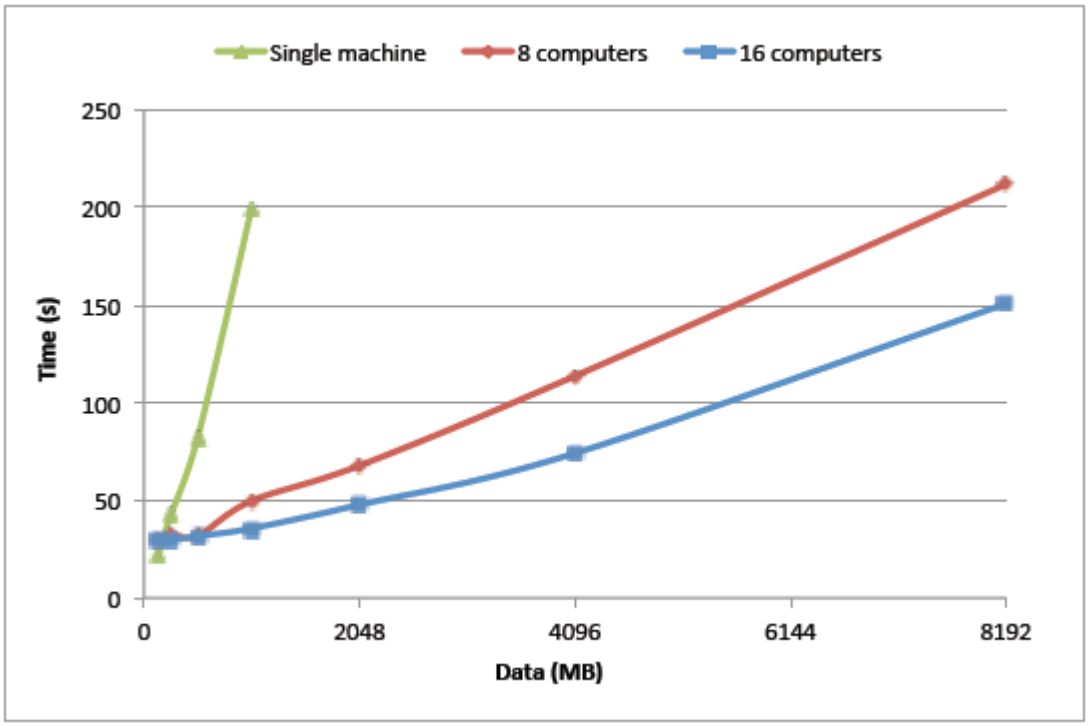

Figura 8- Resultado do tempo de processamento de transformações complexas dos dados. Retirado de Gaspar et al.,(2015).

Como se pode ver pelo gráfico da Figura 8*, Radoop* tem um melhor tempo de processamento do que o RA, mesmo em dados mais pequenos. Quer isto dizer que há pouca sobrecarga na distribuição de tarefas.

7.3*KNIME* – Apache *Hive* based on MapReduce vs Apache *Hive* based in Tez

Koetter (2015), realizou uma experiência com o intuito de comparar o tempo de processamento de dois diferentes mecanismos de execução *Hive, Hive* baseado no *MapReduce* e *Tez*, através da repetição da execução de uma *query SQL* em cada mecanismo. *Apache Tez*[12] é uma *framework* para a criação de aplicações de alto desempenho em *batch* e processamento de dados interativo. *MapReduce* é uma *framework* para a escrita de aplicações que processam uma grande quantidade de dados estruturados e não estruturados armazenados no HDFS.

Nesta experiência foi utilizado o *Hortonworks Sandbox* versão 2.2, sendo esta uma plataforma com todos os recursos integrada com uma máquina virtual, e a "*SQL Generation*" para a criação de operações de base de dados e criação de instruções SQL complexas.

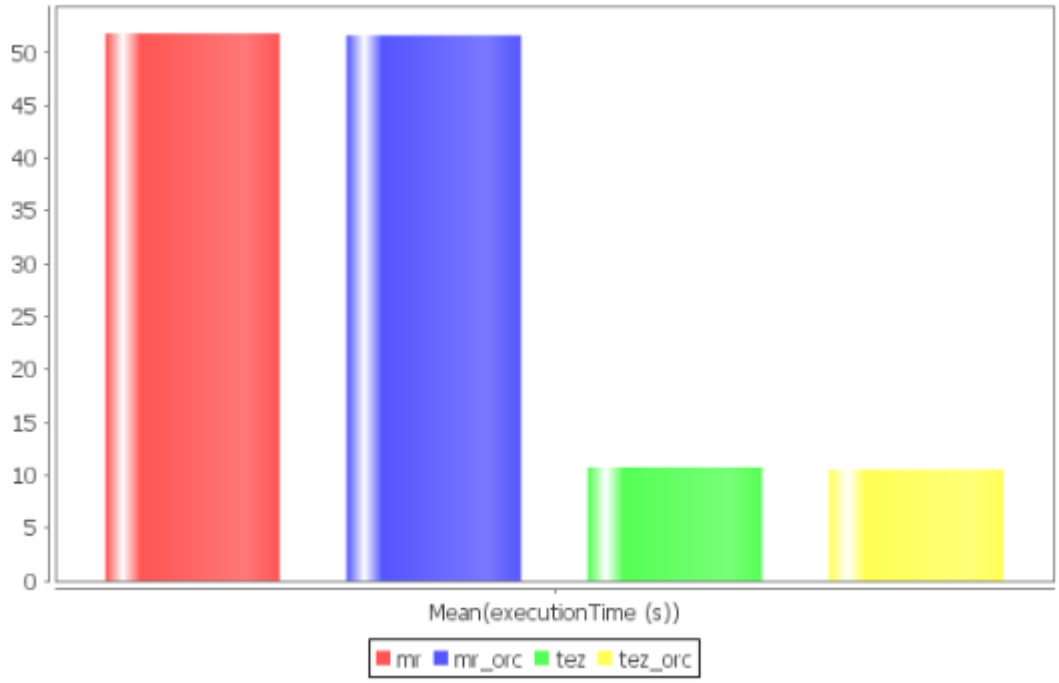

Figura 9-Tempo médio de execução dos mecanismos de execução e formatos de arquivos. Retirado de Koetter (2015).

Como se pode ver pelo gráfico da Figura 9, a *Apache Tez* apresenta um melhor desempenho do que o *MapReduce*, de referir que nesta experiência foi utilizado um pequeno data set. Foi também comparado o desempenho do formato padrão com o formato *Optimized Row Columnar*, onde cada *query* foi executada 15 vezes.

[12] https://hortonworks.com/apache/tez/

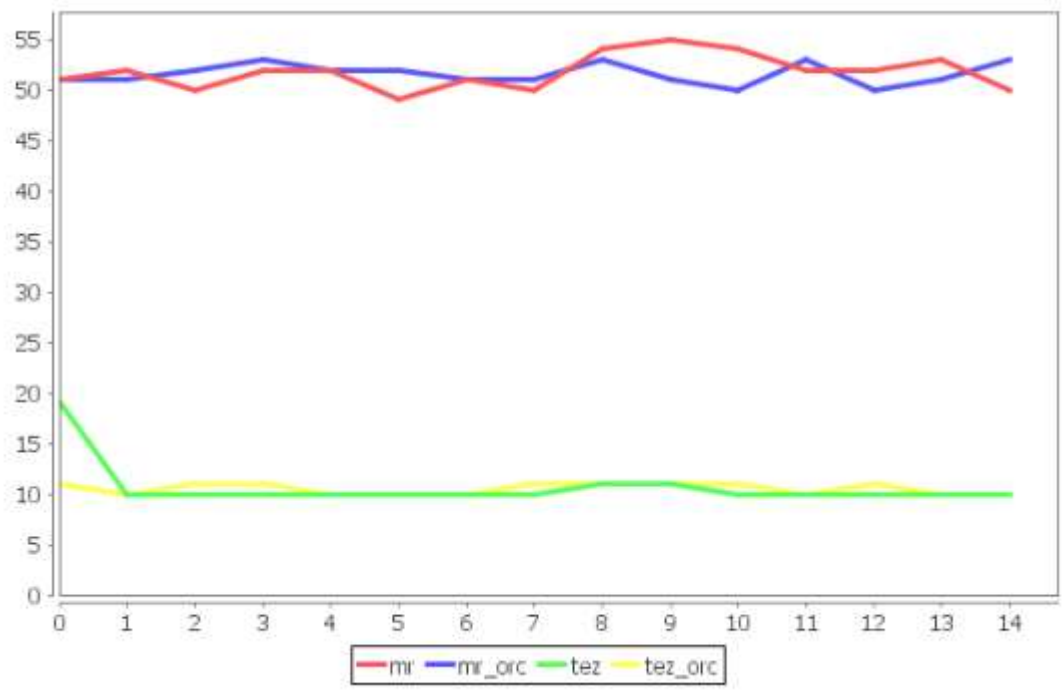

Figura 10- Tempo de execução das 15 iterações. Retirado de Koetter (2015).

Como mostra o gráfico da figura 10, o tempo de execução no *MapReduce* e do *Tez* com o formato *orc* mantêm-se relativamente estáveis durante as 15 iterações. O tempo de execução em cada iteração na *Apache Tez* varia muito pouco, enquanto que a *Apache MapReduce* mostra algumas irregularidades, mas não acentuadas.

Com base nesta experiência, é possível concluir que a *Apache Tez* tem um melhor desempenho comparativamente com a *Apache MapReduce*.

8. DISCUSSÃO

Apesar de estes estudos apresentarem bons resultados, é necessário analisar outros parâmetros quando lidamos com dados, tecnologias e organizações. Assim sendo, é necessário ter em consideração as oportunidades que o *Big Data Mining* pode oferecer, quando aplicado em ambiente organizacional, assim como as suas fraquezas. Com isto, com o intuito de contribuir para esta análise, foi realizada uma análise SWOT (*Strengths, Weaknesses, Opportunities, Threats*), representada na Tabela 7, que mostra o impacto que o *Big Data Mining* pode ter em um ambiente organizacional.

Tabela 7- Análise SWOT

	Forças	**Fraquezas**
Fatores Internos	• Aumento de dados em tempo real; • Aumento da necessidade de realizar as melhores decisões com as melhores técnicas; • Permite melhorar o desempenho das organizações; • Permite melhorar o serviço oferecido aos consumidores através da análise de *feedback*; • Existência de várias ferramentas *open-source* e dispositivos de hardware poderosos; • *Big Data* continua a crescer assim como a necessidade da sua análise.	• Infraestruturas não são adequadas às necessidades da organização ou existe um fraco suporte financeiro na aquisição de tecnologias *Big Data Mining*; • Profissionais com pouco conhecimento nas várias áreas de *Big Data* e *Data Mining* e a sua possível resistência à mudança; • Seleção de tecnologias ou técnicas que não estão de acordo com as necessidades da organização.

	Oportunidades	Ameaças
Fatores Externos	• O *Big Data Mining* permite a análise de dados de várias origens, permitindo assim a extração de mais informações úteis para a organização; • Permite a análise do comportamento do mercado e permite realizar previsões de possíveis mudanças e tendências do Mercado; • Através da análise de dados em tempo-real é possível tomar decisões atempadamente; • Permite prever novas necessidades dos clientes através da análise de redes sociais; • Permite oferecer experiências personalizadas a cada cliente.	• Impossibilidade de acesso a dados externos devido a politicas de segurança; • Vários algoritmos complexos; • Realização de previsões incorretas devido a dados incompletos ou incorretos podendo levar a decisões erradas; • Restrições de privacidade devido aos dados pessoais dos clientes levam a preocupações na sua partilha e possível violação de privacidade.

Este estudo apresenta algumas características que estas ferramentas, *KNIME* e *RapidMiner*, podem oferecer assim como influencias positivas e negativas quando lidamos com *Big Data Mining*.

É importante garantir que a organização tenha as tecnologias certas para fazer face às suas necessidades assim como, conhecimento sobre as oportunidades e ameaças que o *Big Data Mining* pode trazer para a organização. É também importante referir que a evolução das tecnologias sempre irá trazer mais desafios para ultrapassar.

Para além dos critérios utilizados neste estudo, na seleção da melhor ferramenta, é necessário ter em atenção vários outros critérios quando lidamos com tecnologias, dados e organizações, pois o melhor desempenho pode não ser sinónimo de melhores resultados.

9. CONCLUSÃO E TRABALHO FUTURO

No desenvolvimento deste documento de projeto de dissertação foi possível identificar os principais desafios que o *Big Data* coloca às tradicionais ferramentas de análise de dados. *Big Data* tem vindo a ser aplicado em várias áreas como saúde, finanças, economia, entre outras. Os dados sempre irão ter um papel importante numa organização assim como as tecnologias e técnicas utilizadas para análise, transformando-os em informação com valor.

As organizações, nos dias de hoje, necessitam de tirar o melhor proveito dos dados, de forma a tomar as melhores decisões no tempo certo. O surgimento do *Big Data*, trouxe outro desafio referente à sua análise. Pois os dados vêm de diferentes fontes, de diferentes formas, velocidade e tamanhos.

Com o volume de dados a aumentar cada vez mais, existe uma preocupação em garantir que os sistemas de informação consigam lidar com este crescimento e que tenham capacidade para tratar todos esses dados.

Existe uma grande variedade de tecnologias para a área de *Data Mining* e *Big Data*, no entanto com o desenvolvimento das tecnologias estas começam a integrar componentes para lidar com a área de *Big Data*.

É cada vez mais importante as organizações terem profissionais com experiência na área de *Big Data* e *Data Mining*, com o intuito de extrair dos dados informação que ajude na tomada de decisão, obtendo assim vantagem competitiva assim como o aperfeiçoamento dos serviços prestados aos clientes. Tudo isto é possível através da análise de dados, feedback do cliente e redes sociais. Para além disso, ainda é possível prever necessidades futuras da sociedade.

No entanto, o *Big Data Mining* pode trazer alguns desafios apresentados neste documento que é necessário ter em consideração, mas que podem ser facilmente ultrapassados com soluções já existentes.

A análise apresentada nesta dissertação, destaca a importância de colocar as questões certas e saber que dados utilizar para analisar. Para além disso, com o objetivo de melhorar o processo de tomada de decisão, é importante ter em consideração os fatores negativos e positivos em relação ao *Big Data Mining*, assim como é importante ter em consideração as necessidades da organização na aquisição de soluções tecnológicas.

O objetivo principal desta dissertação é a análise do desempenho das ferramentas de *Big Data Mining*, no sentido de perceber o seu comportamento quando lidam com *Big Data*, assim

como que opções oferecem, tanto para a área de *Data Mining* como para a área de *Big Data*, com o objetivo de perceber se suportam eventuais necessidades da organização.

Portanto, é possível concluir que ambas as ferramentas apresentaram um bom desempenho no processo de análise de *Big Data*, assim como oferecem uma grande variedade de operadores para análise de dados. Apresentam, também, formas similares no acesso a dados *Big Data*, assim como no seu tratamento.

No entanto, é necessário realizar mais *benchmarking* para perceber melhor que outros benefícios estas ferramentas podem trazer, no mesmo ambiente com as mesmas condições. Seria também interessante analisar outras técnicas de análise de dados com o objetivo de perceber as vantagens e desvantagens tecnológicas assim como organizacionais.

9.1Tabela de Riscos

Na Tabela 8 estão identificados alguns dos riscos identificados inicialmente e que ocorreram durante o desenvolvimento da dissertação, assim como o seu impacto e as ações de mitigação seguidas.

Tabela 8-Tabela de Riscos

Descrição	Ação de Mitigação	Grau (1-5)	Verificado
Incumprimento do plano de trabalhos	O plano de trabalhos foi ajustado.	3	Sim
Alteração dos objetivos e resultados esperados	O plano de trabalho foi ajustado e foi feito um reajustamento dos objetivos.	5	Sim
Ausência de informação relativa ao tema da dissertação	Prolongamento do tempo estabelecido no plano de trabalhos para a realização de mais pesquisas.	3	Sim

10. REFERÊNCIAS

Aggarwal, S. (2015). *Data Mining* Tools : A Comparative and Analytical Study, *2*(3), 5–9.

Appelbaum, D., Kogan, A., Vasarhelyi, M., & Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. *International Journal of Accounting Information Systems*. https://doi.org/10.1016/j.accinf.2017.03.003

Arunachalam, D., Kumar, N., & Kawalek, J. P. (2017). Understanding *Big Data* analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review*, 1–21. https://doi.org/10.1016/j.tre.2017.04.001

Atanassov, A., & Al-Barznji, K. (2017). A SURVEY OF *BIG DATA MINING* : CHALLENGES AND TECHNIQUES A SURVEY OF *BIG DATA MINING* : CHALLENGES AND TECHNIQUES, (December).

Bayrak, T. (2015). A Review of Business Analytics: A Business Enabler or Another Passing Fad. *Procedia - Social and Behavioral Sciences*, *195*, 230–239. https://doi.org/10.1016/j.sbspro.2015.06.354

Belouch, M., El Hadaj, S., & Idlianmiad, M. (2018). Performance evaluation of intrusion detection based on machine learning using *Apache Spark*. *Procedia Computer Science*, *127*, 1–6. https://doi.org/10.1016/j.procs.2018.01.091

Bisht, P., Negi, N., Mishra, P., & Chauhan, P. (2018). A Comparative Study on Various *Data Mining* Tools for Intrusion Detection, *9*(5), 1–8.

Bose, R. (2009). Advanced Analytics : opportunities and challenges. *Industrial Management & Data Systems*, *109*(2), 155–172.

Carbone, P., Ewen, S., Haridi, S., Katsifodimos, A., Markl, V., & Tzoumas, K. (2015). Apache Flink: Unified Stream and Batch Processing in a Single Engine. *Data Engineering*, *36*, 28–38. https://doi.org/10.1109/IC2EW.2016.56

Che, D., Safran, M., & Peng, Z. (2013). From *Big Data* to *Big Data Mining* : Challenges , Issues , and Opportunities, 1–15.

Chen, M., Mao, S., & Liu, Y. (2014). *Big Data*: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Cosic, R., Shanks, G., & Maynard, S. (2012). Towards a business analytics capability maturity model. *ACIS 2012 : Location, Location, Location : Proceedings of the 23rd Australasian*

Conference on Information Systems 2012, 1–11.

Cosic, R., Shanks, G., & Maynard, S. (2015). A business analytics capability framework. *Australasian Journal of Information Systems Cosic, Shanks & Maynard*, *19*, 5–19.

Coutinho, C., & Chaves, J. (2002). O estudo de caso na investigação em Tecnologia Educativa em Portugal. *Revista Portuguesa de Educação*, *15*(1), 221–243.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed method*. *Research design Qualitative quantitative and mixed methods approaches*. https://doi.org/10.1007/s13398-014-0173-7.2

Davenport, T. H. (2006). Competing on analytics. *Harvard Business Review*, *22*, 5–20. https://doi.org/Article

Davis, C. K. (2014). Beyond data and analysis. *Communications of the ACM*, *57*(6), 39–41. https://doi.org/10.1145/2602326

Gandomi, A., & Haider, M. (2015). Beyond the hype: *Big Data* concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garg, N., Singla, S., & Jangra, S. (2016). Challenges and Techniques for Testing of *Big Data*. *Procedia Computer Science*, *85*, 940–948. https://doi.org/10.1016/j.procs.2016.05.285

Gaspar, C., Henk, T., Makrai, G., & Prekopesák, Z. (2015). Radoop : Analyzing *Big Data* with *RapidMiner* and *Hadoop*, (March 2015).

Ghazi, M. R., & Gangodkar, D. (2015). *Hadoop*, mapreduce and HDFS: A developers perspective. *Procedia Computer Science*, *48*(C), 45–50. https://doi.org/10.1016/j.procs.2015.04.108

Groves, P., Knott, D., Kayyali, B., & Van Kuiken, S. (2013). The ' *Big Data* ' revolution in healthcare, (January).

Gupta, S., & Chaudhari, M. S. (2015). *Big Data* Issues and Challenges. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(2), 62–67. https://doi.org/10.1109/HICSS.2013.645

Hashmi, A. S., & Ahmad, T. (2016). *Big Data Mining* : Tools & Algorithms, *4*(1), 36–40.

Heller, P., Piziak, D., & Knudsen, J. (2016). An Enterprise Architect's Guide to *Big Data*, (March). Retrieved from http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics.

Decision Support Systems, *64*, 130–141. https://doi.org/10.1016/j.dss.2014.05.013

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data for Dummies*.

Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M., & Mephu Nguifo, E. (2017). An experimental survey on *Big Data* frameworks. *Future Generation Computer Systems*, *86*(April 2017), 546–564. https://doi.org/10.1016/j.future.2018.04.032

Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M., & Mephu Nguifo, E. (2018). An experimental survey on *Big Data* frameworks. *Future Generation Computer Systems*, *86*, 546–564. https://doi.org/10.1016/j.future.2018.04.032

Jaseena, K. U., & David, J. M. (2014). Issues, Challenges, and Solutions: *Big Data Mining*, 131–140.

Jena, R. K. (2017). *Big Data* Computing Framework : A Compact Review. *Ijedr*, *5*(2), 1781–1789. Retrieved from https://www.ijedr.org/papers/IJEDR1702280.pdf

Jonnalagadda, V. S., Srikanth, P., Thumati, K., Nallamala, S. H., & Dist, K. (2016). A Review Study of *Apache Spark* in *Big Data* Processing. *International Journal of Computer Science Trends and Technology(IJCST)*, *4*(3), 93–98.

Jothi, B., Amudha, S., & J, J. (2018). Research Challenges in Mining of *Big Data*: A survey, *118*(20), 241–247.

Katal, A., Wazid, M., & Goudar, R. H. (2013). *Big Data*: Issues, challenges, tools and Good practices. In *2013 6th International Conference on Contemporary Computing, IC3 2013*. https://doi.org/10.1109/IC3.2013.6612229

Kim, B. G., Trimi, S., & Chung, J. (2014). Big-Data Applications in the Government Sector, *57*.

Koetter, T. (2015). *Hive* execution engine comparison with the *KNIME* Analytics Platform, 3–5.

Laureano, R. M. S., Miguel da Silva Laureano, L., & Grencho, A. R. R. R. (2016). Framework to implement business analytics: Phases and critical success factors. *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*. https://doi.org/10.1109/CISTI.2016.7521422

Laursen, G. H. N., & Thorlund, J. (2010). The Business Analytics Model. In *Business Analytics for Managers* (pp. 1–12).

Lima, C., & Calazans, J. (2013). Performances Interacionais e Mediações Sociotécnicas PEGADAS DIGITAIS: " *BIG DATA* " E INFORMAÇÃO ESTRATÉGICA SOBRE O CONSUMIDOR 1.

Lima, L. C. B. de. (2014). *Big Data* for data analysis in financial industry. Retrieved from http://repositorium.sdum.uminho.pt/handle/1822/34919

Madeira, P. J. (1999). Benchmarking: A arte de copiar. *Jornal Do Técnico de Contas e Da Empresa*.

Maier, M. (2013). Towards a *Big Data* Reference Architecture, (October), 1–144.

Maire, J.-L., & Buyukozkan, G. (1997). Methods and tools for first five steps of benchmarking process. *Innovation in Technology Management. The Key to Global Leadership. PICMET '97*, (November), 798. https://doi.org/10.1109/PICMET.1997.653643

McKinsey & Company. (2011). *Big Data*: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, (June), 156. https://doi.org/10.1080/01443610903114527

Morais, T. da S. (2015). Survey on Frameworks for Distributed Computing: *Hadoop*, Spark and Storm, (January), 95–105.

Narasimhan, R., & Bhuvaneshwari, T. (2014). *Big Data* – A Brief Study. *International Journal of Scientific & Engineering Research*, *5*(9), 350–353.

Oliveira, M. P. V. De, McCormack, K., & Trkman, P. (2012). Business analytics in supply chains - The contingent effect of business process maturity. *Expert Systems with Applications*, *39*(5), 5488–5498. https://doi.org/10.1016/j.eswa.2011.11.073

Ounacer, S., Talhaoui, M. A., Ardchir, S., Daif, A., & Azouazi, M. (2017). A New Architecture for Real Time Data Stream Processing, *8*(11), 44–51.

Peixoto, R., Portela, F., & Santos, M. F. (2016). Towards a pervasive *Data Mining* engine - Architecture overview. *Advances in Intelligent Systems and Computing*, *445*, 557–566. https://doi.org/10.1007/978-3-319-31307-8_58

Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business analytics in the context of *Big Data*: A roadmap for research. *Communications of the Association for Information Systems*, *37*.

Philpott, S. (2010). Advanced Analytics : Unlocking the Power of Insight. *Intelligence*, (April), 1–15.

Prasad, P. S., & Rajesh, K. (2017). Mining Analysis on Customers Data using *Big Data* tools. *International Journal of Computer Science & Engineering Technology*, *7*(6), 56–58.

Provost, F., & Fwacett, T. (2013). Data Science for Business - What you need to know about *Data Mining* and data analytic thinking.

Raghupathi, W., & Raghupathi, V. (2014). *Big Data* analytics in healthcare : promise and potential, 1–10.

Ribeiro, V. H. da S. (2017). Análise Estatística em R utilizando o Pervasive *Data Mining* Engine.

Russom, P. (2011). *BIG DATA* ANALYTICS TDWI besT pracTIces reporT Introduction to *Big Data* Analytics, 40. Retrieved from https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

Sagiroglu, S., & Sinanc, D. (2013). *Big Data*: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. https://doi.org/10.1109/CTS.2013.6567202

Sarkis, J. (2001). Article information : *Journal of Service Management*, *26*(2), 182–205. https://doi.org/10.1108/MBE-09-2016-0047

Schniederjans, M. J., Schniederjans, D. G., & Starkey, C. M. (2014). *Business Analytics: principles, concepts, and applications*.

Sharda, R., Asamoah, D., & Ponna, N. (2013). Business Analytics: Research and Teaching Perspectives. *Proceedings of the ITI 2013 35th International Conference on INFORMATION TECHNOLOGY INTERFACES*, 3–8. https://doi.org/10.2498/iti.2013.0589

Shmueli, G., Bruce C, P., Yahav, I., Patel, N. R., & Lichtendahl Jr., K. C. (2017). *Data Mining* for Business Analytics: Concepts, Techniques, and Applications in R.

Singh, B., Grover, S., & Singh, V. (2013). An Overview of Benchmarking Process: The Continuous Improvement Tool. *II) YMCAUST International Journal of Research*, *1*(July), 80–83.

Singh, J., & Singla, V. (2015). *Big Data* : Tools and Technologies in *Big Data*, *112*(15), 6–10.

Tellis, W. M. (1997). Application of a Case Study Methodology Application of a Case Study Methodology, *3*(3), 1–19.

The Apache Software Foundation. (2018). Welcome to Apache Pig ! *Apache Software Foundation*, 2–3.

Thillaieswari, B. M. S., Phil, M., & Ed, B. (2017). Comparative Study on Tools and Techniques of *Big Data* Analysis, *66*(61), 61–66.

Thusoo, A., Sarma, J. Sen, Jain, N., Shao, Z., Chakka, P., Anthony, S., … Murthy, R. (2009). *Hive - A Warehousing Solution Over a Map-Reduce Framework*. *Sort*, *2*, 1626–1629. https://doi.org/10.1109/ICDE.2010.5447738

Ventura, S., Altahi, A. H., Luna, J. M., & Vallejo, M. . (2017). Evaluation and Comparison of

Open Source Software Suites for *Data Mining* and Knowledge Discovery, (September). https://doi.org/10.1002/widm.1204

Vora, M. N. (2011). *Hadoop*-HBase for large-scale data. *Proceedings of 2011 International Conference on Computer Science and Network Technology, ICCSNT 2011*, *1*, 601–605. https://doi.org/10.1109/ICCSNT.2011.6182030

Watson, H. J. (2014). Tutorial : *Big Data* Analytics : Concepts , Technologies , and Applications Tutorial : *Big Data* Analytics : Concepts , Technologies , and Applications, *34*(April), 1247–1268. Retrieved from http://aisel.aisnet.org/cais/vol34/iss1/65/

Watson, H. J., Hiltbrand, T., Thomas, R., Halter, O., Passariello, S., Ramesh, R., … Williams, N. (2010). BI Training Solutions : As Close as Your Conference Room. *Business Intelligence Journal*, *15*(2), 1–57.