

**ANÁLISE DA APLICAÇÃO DO *BIG DATA* NA ÁREA DE VENDAS EM UMA  
EMPRESA DO SEGMENTO DE BELEZA: PROPOSTAS PARA A AÇÃO**

Patrícia Ferreira Azevedo

Renan Medrado Pacheco

Projeto de Graduação apresentado ao Curso de Engenharia de Produção da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Renato Flório Cameira

Rio de Janeiro  
Fevereiro de 2021

ANÁLISE DA APLICAÇÃO DO *BIG DATA* NA ÁREA DE VENDAS EM UMA  
EMPRESA DO SEGMENTO DE BELEZA: PROPOSTAS PARA A AÇÃO

Patrícia Ferreira Azevedo

Renan Medrado Pacheco

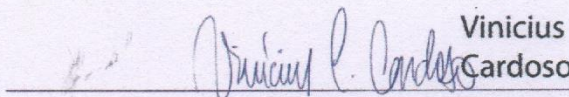
PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA DE PRODUÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO DE PRODUÇÃO.

Examinado por:



Renato  
Flório  
Cameira

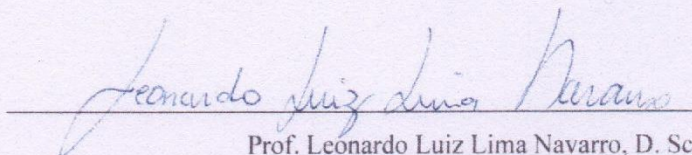
Prof. Renato Flório Cameira, D. Sc.



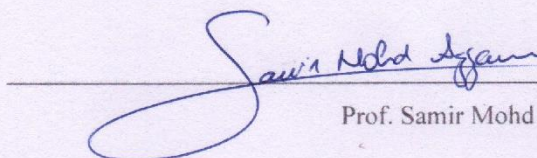
Vinicius  
Cardoso

Assinado de forma digital  
por Vinicius Cardoso  
Dados: 2021.03.08  
20:50:34 -03'00'

Prof. Vinicius Carvalho Cardoso, D. Sc.



Prof. Leonardo Luiz Lima Navarro, D. Sc.



Prof. Samir Mohd Azzam, M. Sc.

RIO DE JANEIRO, RJ – BRASIL

FEVEREIRO de 2021

Pacheco, Renan Medrado

Azevedo, Patrícia Ferreira

Análise da aplicação do *Big Data* na área de vendas em uma empresa do segmento de beleza: Propostas para a ação/  
Patrícia Ferreira Azevedo e Renan Medrado Pacheco - Rio de Janeiro: UFRJ/ Escola Politécnica, 2021.

IX, 84 p.: il.; 29,7 cm.

Orientador: Renato Flório Cameira

Projeto de Graduação - UFRJ/ POLI/ Curso de Engenharia de Produção, 2021.

Referências Bibliográficas: p.92-95.

1. *Big Data* 2. Análise de dados 3. Base de dados. I. Cameira, Renato Flório II. Universidade Federal do Rio de Janeiro, UFRJ, Curso de Engenharia de Produção. III. Análise da aplicação do *Big Data* na área de vendas em uma empresa do segmento de beleza: Propostas para a ação

## **Agradecimentos Patrícia Ferreira Azevedo**

Gostaria de primeiramente agradecer a minha mãe, Rosimere, pelo exemplo de dedicação e cuidado com a família, por ter me dado o suporte necessário para lidar com o dia a dia da universidade e me dar forças para não desistir. Também, a minha avó, Maria Cristina (*in memoriam*), por sempre ter acreditado em mim, na minha capacidade e ter me estimulado a correr atrás de todos os meus sonhos. Vocês foram essenciais para que eu chegasse até aqui.

Ao meu namorado Jean, um super agradecimento por ter sido meu grande parceiro nos momentos de alegrias e comemorações, mas também de apoio nos momentos difíceis. Sua companhia durante esses anos tornou todo esse processo mais leve e possível.

Também, aos meus colegas de graduação, não poderia deixar de agradecer por toda a parceria nos trabalhos, estudos em grupo, desabafos e cuidado. Em especial, obrigada Ana Sofia, Leandro, Victor, Carol, Juliana e Bernardo por todos os momentos compartilhados. Quero levar a amizade de vocês para muito além da universidade.

Ao Renan, meu parceiro neste trabalho, um agradecimento especial por toda a parceria não apenas na construção deste projeto, mas de todas as disciplinas que desenvolvemos juntos. Sua companhia foi imprescindível para que este trabalho se tornasse o que é.

Gostaria de agradecer aos meus colegas da Fluxo Consultoria que me ajudaram a iniciar a vida profissional e a explorar as engenharias além do bloco F. Obrigada especialmente a Thais, Luiz e Victor pela grande amizade.

Por fim, gostaria de agradecer a todos os professores que me deram a honra de aprender, me desenvolver e me tornar uma engenheira. Um agradecimento especial ao nosso orientador, Renato Cameira, pela sua dedicado ao magistério, por todos os aprendizados e lições durante toda a graduação.

## **Agradecimentos Renan Medrado Pacheco**

Primeiramente gostaria de agradecer aos meus pais, Maria Tereza e Pedro Paulo, por todo o empenho que tiveram para me formar como cidadão e por todos os esforços, que nunca foram poupados, para que eu tivesse uma educação que me permitiu chegar até este momento. Gostaria de agradecer ao meu irmão, Allan, por ter me acompanhado por todos esses anos sendo um suporte sempre que precisei.

Gostaria de agradecer também aos meus avós, Adélia e Sebastião, que mesmo sem estudo formal entendiam o poder da educação e sempre me incentivaram a focar nos estudos e a alcançar voos que a vida nunca os tinha proporcionado. Avós, acho que nunca falei isso, mas vocês são a história que sempre conto quando perguntam quem são as minhas inspirações, muito obrigado por todo o empenho.

Não tem ninguém que me ajudou mais durante a faculdade que meu namorado Gabriel. A ele devo muitos agradecimentos por todas as noites viradas que passou comigo para entregar os trabalhos no prazo certo, estudar para as provas ou até mesmo para me acalmar quando tudo estava nebuloso. Sem você tudo teria sido muito mais difícil.

Gostaria de agradecer a minha parceira neste trabalho, Patrícia, por todos os esforços que tivemos para tornar este sonho realidade. Ninguém mais do que a gente sabe o quanto foi difícil chegar neste momento.

Gostaria de agradecer ao meu amigo João Pedro por todos os momentos de carona em conjunto até o fundão e todos os momentos que sempre esteve disposto a ajudar nos trabalhos e nos estudos. Também gostaria de agradecer as minhas amigas Thayza e Ana Carolina por todos os trabalhos que fizemos em conjunto e que me fizeram aprender muito. Gostaria de agradecer também ao Fred, meu colega de trabalho, por todos os ensinamentos que me tornaram o engenheiro que sou hoje.

Por fim, gostaria de agradecer a todos os professores desta estimada instituição por todo empenho em gerar conhecimento para o crescimento deste país. Em especial gostaria de agradecer ao Renato Cameira, por ser um exemplo de professor e profissional, que me ensinou enormemente e que tenho o orgulho de poder tê-lo tido como orientador deste trabalho.

Resumo do Projeto de Graduação apresentado à Escola Politécnica/UFRJ como parte dos requisitos necessários para a obtenção do grau de Engenharia de Produção.

## **ANÁLISE DA APLICAÇÃO DO *BIG DATA* NA ÁREA DE VENDAS EM UMA EMPRESA DO SEGMENTO DE BELEZA: PROPOSTAS PARA A AÇÃO**

Patrícia Ferreira Azevedo

Renan Medrado Pacheco

Fevereiro/2021

Orientador: Renato Flório Cameira

Curso: Engenharia de Produção

O *Big Data* – termo criado para denominar grandes volumes de dados - está cada vez mais presente em pesquisas acadêmicas e sendo explorado pelas organizações. Neste contexto, tem-se como objetivo deste trabalho o entendimento do *Big Data* e suas características para posterior análise de como se aplica ao contexto empresarial, buscando entender os principais pontos de atenção em seu funcionamento e propor ações de correção para melhor extração de valor em sua análise. Para a parte de exploração teórica foi realizada uma pesquisa bibliométrica com os termos *Big Data* e *Data Analytics* e posterior aprofundamento em suas características, que comumente é explorado na academia através dos V's - volume, velocidade, variedade, veracidade, valor, variabilidade e verificação.

Como resultado deste trabalho foi possível compreender, limitando-se a uma empresa, que o uso abordado na academia ainda está distante da realidade, mesmo em uma grande organização multinacional. Chegou-se então a algumas propostas de ação para correção dos problemas e redução do hiato teoria-prática.

Palavras chave: *Big Data*, Análise de dados, Base de dados

Abstract of Undergraduate Project presented to POLI/UFRJ as a partial fulfillment of the requirements for the degree of Engineer.

**ANALYSIS OF BIG DATA'S APPLICATION IN THE SALES DEPARTMENT IN A  
BEAUTY SEGMENT COMPANY: PROPOSALS FOR ACTION**

Patrícia Ferreira Azevedo

Renan Medrado Pacheco

February/2021

Advisor: Renato Flórido Cameira

Course: Engenharia de Produção

Big Data - a term created to refer to large volume of data - is increasingly present in academic research and explored by organizations. In this context, the aim of this work is to understand Big Data and its characteristics for further analysis of how it applies to the business context, to understand the main points of attention in its operation and to propose corrective actions for better value extraction in its analysis. For the theoretical exploration part, a bibliometric research was carried out with the terms Big Data and Data Analytics and further deepening its characteristics, which is commonly explored in the academy through V's - volume, velocity, variety, veracity, value, variability, verification.

As a result of this work, it was possible to understand, limited to one company, that the use addressed in the academy is still far from reality, even in a large multinational organization. Then, some action proposals were arrived at to correct the problems and reduce the theory-practice gap.

Keywords: Big Data, Data analytics, Database

## LISTA DE FIGURAS

|  |    |
|--|----|
| Figura 1: Ilustração da metodologia utilizada. ....  | 15 |
| Figura 2: Quadro de entrevistas.....   | 17 |
| Figura 3: Resultado da pesquisa utilizando <i>Data Analytics</i> e <i>Big Data</i> .....   | 20 |
| Figura 4: Publicações sobre big data ano a ano.....  | 21 |
| Figura 5: Mínimo de 20 cocitações para elaboração do Mapa. ....  | 21 |
| Figura 6: Afunilamento da pesquisa bibliométrica.....  | 22 |
| Figura 7: Visualização do mapa de citação sobre big data.....  | 23 |
| Figura 8: Artigos mais citados da pesquisa utilizando <i>Data Analytics</i> e <i>Big Data</i> .....  | 24 |
| Figura 9: Quadro de artigos por assunto e cluster. ....  | 25 |
| Figura 10: 5 métricas de classificação do big data. ....   | 29 |
| Figura 11: 6 V's do <i>Big data</i> (valor, volume, velocidade, variedade, veracidade e variabilidade, que também se aplicam aos dados da saúde. ....                | 31 |
| Figura 12: Resumo da evolução do BI&A .....  | 33 |
| Figura 13: Possibilidades de aplicação do Big Data na cadeia de suprimentos. ....  | 34 |
| Figura 14: Dimensões da qualidade de dados. ....   | 35 |
| Figura 15: <i>Framework</i> de sistema de manufatura preditiva .....   | 37 |
| Figura 16: Uso da computação em nuvem no <i>big data</i> . ....  | 39 |
| Figura 17: Evolução da discussão por cluster. ....   | 40 |
| Figura 18: 3 V's da IBM.....   | 42 |
| Figura 19: 4 V's mais citados do Big Data.....   | 45 |
| Figura 20: Caminho percorrido pelo dado até disponibilização.....  | 49 |
| Figura 21: Exemplificação de sell in, sell out e sell through.....   | 52 |
| Figura 22: Quadro de período de disponibilização das métricas de <i>sell out</i> , <i>sell in</i> , estoque e positivação no banco de dados de <i>sell out</i> ..... | 54 |
| Figura 23: Origens de informação do banco de dados de sell out. ....   | 56 |
| Figura 24: Quadro resumo dos V's frente ao caso de estudo.....   | 62 |
| Figura 25: Estado atual dos envolvidos nos problemas do banco de dados.....  | 66 |
| Figura 26: Quadro de problemas com categorização. ....   | 71 |
| Figura 27: Quadro de premissas e características de contorno.....  | 74 |
| Figura 28: Quadro de análise de Fit entre condições de contorno.....   | 77 |
| Figura 29: Quadro de propostas de soluções frente as condições de contorno e análise de <i>fit</i> .80   |    |
| Figura 30: Quadro de plano de ação das soluções de curto prazo. ....   | 86 |



|  |    |
|--|----|
| Figura 31: Quadro de plano de ação das soluções de médio prazo. .... | 87 |
| Figura 32: Quadro de plano de ação das soluções de médio prazo. .... | 88 |

## SUMÁRIO

|  |    |
|--|----|
| <b>1. INTRODUÇÃO</b>   | 12 |
| <b>1.1. Motivação e Premissas</b>  | 12 |
| <b>1.2. Objetivos</b>  | 13 |
| 1.2.1. Objetivo Geral  | 13 |
| 1.2.2. Objetivos Específicos   | 13 |
| <b>1.3. Estrutura do Trabalho</b>  | 13 |
| <b>1.4. Metodologia</b>  | 14 |
| <b>1.5. Delimitações</b>   | 18 |
| <b>1.6. Limitações</b>   | 18 |
| <b>2. BIG DATA</b>   | 20 |
| <b>2.1. Pesquisa Bibliométrica</b>   | 20 |
| <b>2.2. Evolução</b>   | 24 |
| 2.2.1. <i>Cluster</i> dos métodos analíticos do <i>Big Data</i> (Vermelho) | 25 |
| 2.2.2. <i>Cluster</i> de logística e cadeia de suprimentos (Verde)         | 32 |
| 2.2.3. <i>Cluster</i> de <i>Internet of Things</i> (Azul)                  | 36 |
| 2.2.4. <i>Cluster Big Data</i> em nuvem (Amarelo)                          | 38 |
| 2.2.5. Evolução entre <i>clusters</i>                                      | 40 |
| <b>3. OS V'S DO BIG DATA</b>   | 42 |
| <b>3.1. Volume</b>   | 43 |
| <b>3.2. Variedade</b>  | 43 |
| <b>3.3. Velocidade</b>   | 44 |
| <b>3.4. Valor</b>  | 44 |
| <b>3.5. Veracidade</b>   | 46 |
| <b>3.6. Variabilidade</b>  | 46 |
|  | 10 |

|        |   |    |
|--------|---|----|
| 3.7.   | Verificação   | 46 |
| 4.     | ESTUDO DE CASO: EMPRESA RP COSMETICS  | 47 |
| 4.1.   | Caracterização da Empresa   | 47 |
| 4.2.   | Estrutura do <i>Big Data</i> da Companhia   | 47 |
| 4.3.   | Caracterização frente aos V's   | 53 |
| 4.3.1. | Volume  | 54 |
| 4.3.2. | Variedade   | 55 |
| 4.3.3. | Velocidade  | 58 |
| 4.3.4. | Valor   | 60 |
| 4.3.5. | Veracidade  | 60 |
| 4.3.6. | Síntese dos Vs  | 62 |
| 5.     | OBSERVAÇÃO DOS PROBLEMAS  | 63 |
| 5.1.   | Problemas de Big Data Comuns ao Mercado   | 63 |
| 5.2.   | Aprofundamento dos Problemas da <i>RP Cosmetics</i>                               | 66 |
| 5.3.   | Características de Contorno   | 71 |
| 6.     | ANÁLISE DE FIT  | 74 |
| 7.     | PROPOSTAS DE SOLUÇÃO  | 79 |
| 7.1.   | Problemas Organizacionais   | 81 |
| 7.2.   | Problemas de Pessoas  | 81 |
| 7.3.   | Problemas de Processos  | 82 |
| 7.4.   | Problemas de Tecnologia da Informação   | 83 |
| 8.     | PLANO DE AÇÃO   | 85 |
| 9.     | CONSIDERAÇÕES FINAIS  | 89 |
|        | APÊNDICE A: Guia de entrevistas da categorização do banco de dados frente aos V's | 92 |
|        | REFERÊNCIAS BIBLIOGRÁFICAS  | 93 |

## 1. INTRODUÇÃO

O primeiro capítulo do presente projeto tem por objetivo apresentar primeiramente as motivações que levaram a sua construção e as premissas que serão avaliadas durante o seu desenvolvimento para final verificação. Em seguida, são apresentados o objetivo geral e os objetivos específicos, a estrutura que será seguida e toda a metodologia aplicada. Ao fim, são destacadas as limitações e delimitações deste trabalho.

### 1.1. Motivação e Premissas

Os conceitos de *Big Data* e *Data Analytics* se tornaram uma constante nos assuntos pertinentes às empresas devido à grande importância que o acesso à informação tomou, se tornando um verdadeiro ativo para a tomada de decisão. O grande contraponto disso é a dificuldade que as empresas vêm encontrando de trabalhar seus dados e extrair todo o valor deles.

Conforme reforçado pelo diretor de análises da *Caesars Entertainment* em entrevista para a *McKinsey* (2016), as ferramentas de análises estão em constante desenvolvimento e por isso há muitas dúvidas dos resultados gerados e de como utilizá-las da forma correta para decisões acuradas. Isso é um agravante ainda maior para empresas que não surgiram no ambiente tecnológico e analítico, tendo que se reconstruir para se adaptar a essa realidade.

Ainda no que diz respeito ao *Big Data*, sua aplicação tem sido muito explorada nos últimos anos e algumas de suas características mais importantes, também: variedade, velocidade e volume. Ao longo do tempo, mais termos foram acrescentados à essa lista, e eles serão melhor explorados neste trabalho.

Com base no esclarecimento acima e no estudo de caso que será desenvolvido, algumas premissas sobre aplicação do *Big Data* e *Data Analytics* foram construídas.

A primeira premissa é a de que ter acesso às ferramentas corretas é suficiente para uma análise de dados acurada. Em complemento, a segunda premissa diz que ter acesso à um grande volume de dados garante uma orientação clara para a decisão.

A terceira premissa é de que para obter valor dos dados é necessária uma estrutura de banco de dados com diferentes fontes de informação e a quarta premissa diz respeito ao peso da veracidade do dado, sendo ele o fator principal a ser garantido em uma estrutura de análise de dados. Ainda, a quinta premissa elaborada é a de que a velocidade não é mais uma questão para a análise de dados em grandes empresas, considerando o avanço da tecnologia e importância do assunto no meio corporativo.

Como sexta premissa, que fundamenta o estudo de caso, temos que, com base na percepção dos problemas que serão abordados neste trabalho e de acordo com as premissas anteriores, o uso da estrutura teórica do *Big Data* resolveria todos os problemas encontrados na análise de dados da empresa estudada.

## **1.2. Objetivos**

### **1.2.1. Objetivo Geral**

O objetivo geral deste trabalho é entender os conceitos de *Big Data* e *Data Analytics* e suas aplicações perante o meio acadêmico, a partir disso explorar a categorização do *Big Data* através dos V's e como isso se desenvolve e se aplica em um banco de dados em um ambiente empresarial. Assim, busca-se entender a origem dos problemas enfrentados e encontrar soluções para mitigar estes problemas.

### **1.2.2. Objetivos Específicos**

Para alcançar o objetivo geral, há a definição dos objetivos específicos a seguir:

- Compreender como os conceitos de *Big Data* e *Data Analytics* são vistos no meio acadêmico ao longo do tempo a partir de livros e artigos científicos;
- Aprofundar na caracterização do *Big Data* através dos V's;
- Realizar o estudo de banco de dados em uma empresa visando entender a aplicação e dificuldades que envolvem os conceitos estudados;
- Entender quais são os principais problemas que envolvem as principais características do *Big Data*;
- Identificar as melhores oportunidades e propor soluções existentes atualmente que visam solucionar ou ao menos mitigar a percepção desses problemas.

## **1.3. Estrutura do Trabalho**

Após esta breve introdução, que busca explorar os interesses e bases principais que deram origem a este trabalho, será apresentado no segundo capítulo o estudo teórico sobre *Big Data*. Nele, será tratado com detalhes a pesquisa bibliométrica desenvolvida e as definições dos conceitos de *Big Data* e *Data Analytics* encontrados na literatura acadêmica a partir da estrutura de *clusters*.

No capítulo seguinte serão explorados os V's do *Big Data* com a visão dos autores dos artigos já explanados no capítulo anterior. Em seguida, o capítulo quatro apresenta o estudo de

caso, com a caracterização da empresa e detalhamento das principais características do banco de dados estudado frente aos V's.

O quinto capítulo aborda a análise dos problemas levantados através de uma primeira visão focada nos problemas de *Big Data* comuns ao mercado. Em seguida, explora-se os problemas da companhia estudada com mais detalhes, categorizando-os, para que então sejam levantadas as premissas e características de contorno para o estudo em questão.

O próximo capítulo aborda a metodologia de análise de *fit* que busca entender que características de contorno são ou não compatíveis para proposição de soluções, que são apresentadas no capítulo sete. Neste, as propostas de solução são descritas respeitando a categorização de problemas.

Por fim, tem-se um capítulo com a apresentação de um plano de ação das soluções propostas anteriormente, com visão de etapas e prazos, seguido do capítulo final de conclusão e das referências bibliográficas. Nela consta todas as referências utilizadas e mencionadas durante o desenvolvimento do presente trabalho.

#### **1.4. Metodologia**

Nesta seção serão abordados os métodos utilizados para a elaboração do presente trabalho e como ele está presente em cada seção posterior. Na figura 1, tem-se o esquema que demonstra o caminho metodológico seguido.

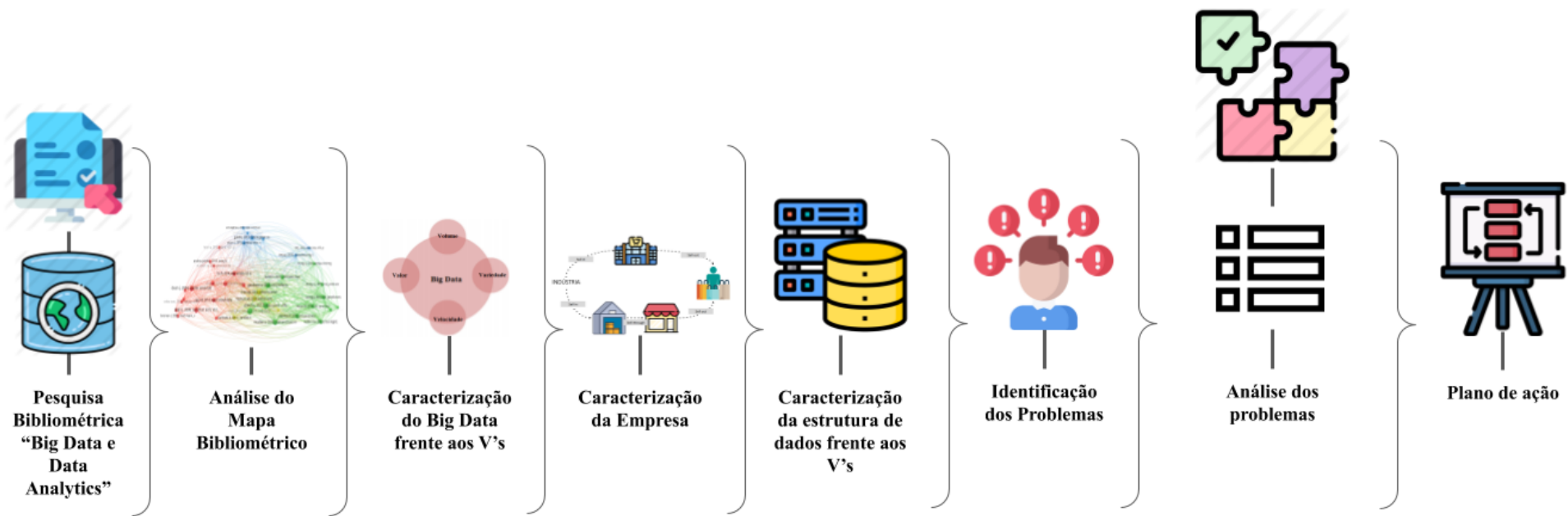


Figura 1: Ilustração da metodologia utilizada.

Fonte: Elaboração Própria.

A metodologia utilizada neste trabalho se baseia inicialmente em uma pesquisa teórica qualitativa a partir de artigos e livros levantados em uma pesquisa bibliométrica na base de publicações *Web of Science (WoS)*, um dos portais disponíveis na base de periódicos CAPES.

De acordo com SILVA *et al.* (2011, p.4), “o princípio da bibliometria constitui em analisar a atividade científica ou técnica pelos estudos quantitativos das publicações.”, já ROSTAING (1996, p.17) define bibliometria como “a aplicação dos métodos estatísticos ou matemáticos sobre o conjunto de referências bibliográficas”.

Uma das aplicações da bibliometria estão a seleção de livros e publicações periódicas, a identificação das características temáticas da literatura, a evolução da bibliografia e outros (OKUBO, 1997).

Com as referências levantadas, um mapeamento foi desenvolvido utilizando o software *VOS Viewer*. Esse software é uma ferramenta que constrói mapas nos quais é possível visualizar ligações entre publicações acerca de um tema utilizando como fonte uma base de dados de artigos científicos de portais e bibliotecas virtuais disponíveis na internet. Assim, a pesquisa teórica foi aplicada com base nos artigos obtidos no mapa de referências.

Também, para além do levantamento de referências por essa automação do *VOS Viewer*, foram lidos dinamicamente artigos que não estavam dentro das características escolhidas para este trabalho, conforme será mencionado na parte 2.1, mas que eram altamente citados nas referências dos principais artigos obtidos de modo a expandir a análise e não restringir o estudo apenas aos artigos co-citados identificados. Essa mecânica foi aplicada principalmente no aprofundamento dos V's.

Em seguida, o trabalho se aprofunda nas características do *Big Data*, que comumente na literatura sobre o tema é abordado em palavras que iniciam com a letra “V”. Para tal pegamos cada um dos artigos encontrados na pesquisa bibliométrica, e os artigos citados em suas referências, que falam sobre o tema e explicamos o que cada “V” significa a partir da visão dos respectivos autores.

Ademais, o trabalho foi acrescido de uma pesquisa prática desenvolvida através de um estudo de caso em uma grande empresa. O estudo de caso buscou fazer uma análise exploratória em uma companhia escolhida com o objetivo de entender qual o papel do *Big Data* para a garantia de minimização de problemas na análise de dados. O banco de dados da empresa foi explorado e entendido para que uma união de pesquisa teórica e prática pudesse ser concluída.



Foi utilizado o método MIASP, também conhecido como *QC Story*, método japonês de identificação, análise e solução de problemas para o estudo de caso. O método, de acordo com Campos (1992), é executado em 8 etapas - identificação do problema, observação, análise, plano de ação, ação, verificação, padronização e conclusão. Neste trabalho, por limitação na execução do plano de ação, iremos até este ponto.

Como parte da identificação dos problemas foram feitas entrevistas não estruturadas com a equipe de inteligência comercial da RP *Cosmetics*. Na figura 2 a seguir há o detalhamento de quantas entrevistas foram realizadas, com quais temas e que cargos estiveram envolvidos.

| Tema  | Entrevistas Realizadas | Quantidade de Participantes | Cargos  |
|---|------------------------|-----------------------------|---|
| Funcionamento do banco de dados                 | 2                      | 2                           | Analista de TI e analista de TI externo (suporte)   |
| Caracterização do banco de dados frente aos V's | 6                      | 6                           | 2 analistas de excelência comercial (responsáveis pelo banco de dados), 3 analistas de administração de vendas (usuários) e 1 coordenador (usuário) |
| Revalidação de problemas                        | 3                      | 6                           | 2 analistas de excelência comercial (responsáveis pelo banco de dados), 3 analistas de administração de vendas (usuários) e 1 coordenador (usuário) |

Figura 2: Quadro de entrevistas.

Fonte: Elaboração Própria.

Para a parte de observação foi feito um paralelo entre o banco de dados da empresa e os V's do *Big Data*. Também foi levantado a partir de *softwares* disponíveis na internet os principais problemas de uma estrutura de *Big Data* e, a partir disto, buscou-se entender os principais problemas na análise de dados que a equipe de inteligência comercial enfrenta. Com tal, foi possível realizar uma comparação entre isto e os casos explorados na pesquisa.

Na parte de análise dos problemas foi feita uma categorização dos problemas encontrados e, junto a equipe de inteligência comercial e suas expectativas futuras do *Big Data* da companhia, foram construídas as condições de contorno para as soluções.

A partir das condições de contorno, foi realizado uma análise de *fit*, onde buscou-se entender se as condições de contorno tinham conflito entre si e se sim, se esse conflito era ajustável ou não.

Na etapa seguinte foram propostas soluções para as situações indesejadas, de acordo com as condições de contorno levantadas com a equipe, buscando contornar os conflitos encontrados na seção anterior. Por último a partir das soluções e conflitos entre os mesmos foi

proposto um plano de ação considerando um período de curto, médio e longo prazo para resolução dos problemas.

### **1.5. Delimitações**

Como delimitação, o presente trabalho busca explorar a visão de *Big Data* e *Data Analytics* na literatura acadêmica através dos artigos mais relevantes no quesito de número de citações encontrados no *Web of Science*. A partir disso, o trabalho explora de forma mais consistente as visões dos V's do *Big Data* para cada um dos autores que o mencionam.

Além disso, é entendido como que o banco de dados da empresa estudada se aplica e se comporta dentro as características dos V's mais relevantes, não adentrando no estudo dos V's menos citados. Também, o estudo de caso foi feito analisando apenas um dos bancos de dados da companhia escolhida, não todos existentes.

Também, o trabalho não tem por objetivo explorar características técnicas avançadas de programação e sistemas, bem como o entendimento aprofundado do funcionamento técnico do banco de dados explorado ou o desenvolver técnico das propostas de ação apresentadas.

Por fim, tem-se como objetivo o levantamento e proposta de ações adequadas aos problemas identificados, não apresentando um projeto ou plano de ação completo de desenvolvimento e implementação das soluções.

### **1.6. Limitações**

Como limitação, por conta da pandemia de COVID-19 experienciada em 2020, algumas das entrevistas foram feitas a partir de plataformas digitais e envolvendo menos participantes do que se esperava inicialmente, o que pode caracterizar algum tipo de viés nos problemas encontrados e na proposta de solução dos mesmos. Também, devido a pandemia, o presente trabalho se tornou mais alongado temporalmente, o que ocasionou em um hiato entre a pesquisa da literatura realizada em 2019 e a defesa realizada no início de 2021.

Sendo necessária a confidencialidade da empresa, tem-se como limitação a exploração de dados numéricos, de modo que os números utilizados neste trabalho são os disponibilizados publicamente.

Além disso, considerando o levantamento de soluções e propostas de ação, não poderá ser acompanhado o desenvolvimento e implementação delas na companhia. Também, por haver a necessidade de envolver diferentes áreas e cargos de liderança, não será possível validar com

a companhia de modo geral se todas as propostas de solução deste trabalho são cabíveis de implementação.

Na próxima seção será abordada a pesquisa bibliográfica feita através de um mapeamento bibliométrico que serve como base teórica para o presente trabalho.

## 2. **BIG DATA**

Nesta seção é apresentado o mapeamento bibliométrico realizado a partir do *software VOS Viewer*, base teórica para o trabalho. O mapeamento e suas conexões permitem uma análise por grupos semelhantes (*cluster*) das publicações e uma posterior visão de como estes se comunicam.

### 2.1. **Pesquisa Bibliométrica**

No *Web of Science*, ao se pesquisar pela coleção *Science Citation Index* pelo termo *Big Data*, considerando os artigos de todos os anos disponíveis foi obtido 13.386 resultados de artigos e 221.704 citações ao pesquisar pelo termo *Data Analytics* na mesma base e também todos os anos disponíveis se obteve 2.269 resultados de artigos que possuíam 89.994 citações.

Ambos os termos fazem parte do objetivo do estudo, porém a pesquisa individual em cada um dos dois termos retornou resultados demais. Para restringir a quantidade de resultados foi definido pesquisar ambos os termos em conjunto, *Big Data* e *Data Analytics* (Figura 3).

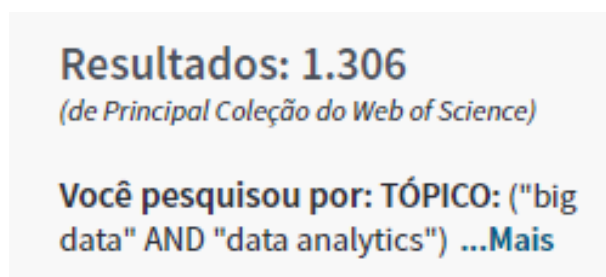


Figura 3: Resultado da pesquisa utilizando *Data Analytics* e *Big Data*.

Fonte: Elaboração Própria.

Nesta pesquisa, onde só teríamos como resultado artigos que citassem ambos, limitada na coleção *Science Citation Index Expanded* e considerando artigos de todos os anos disponíveis tivemos 1.306 publicações, como visto na Figura 3, o que levou a 57.547 citações.

Ao verificar a evolução de publicações acerca dos temas pesquisados, é notável o grande aumento ano a ano. O ano de 2019 foi omitido da análise e na figura 4 por não estar completo, porém na pesquisa foram identificadas 168 publicações para ele. As bases foram extraídas em 18 de maio de 2019.

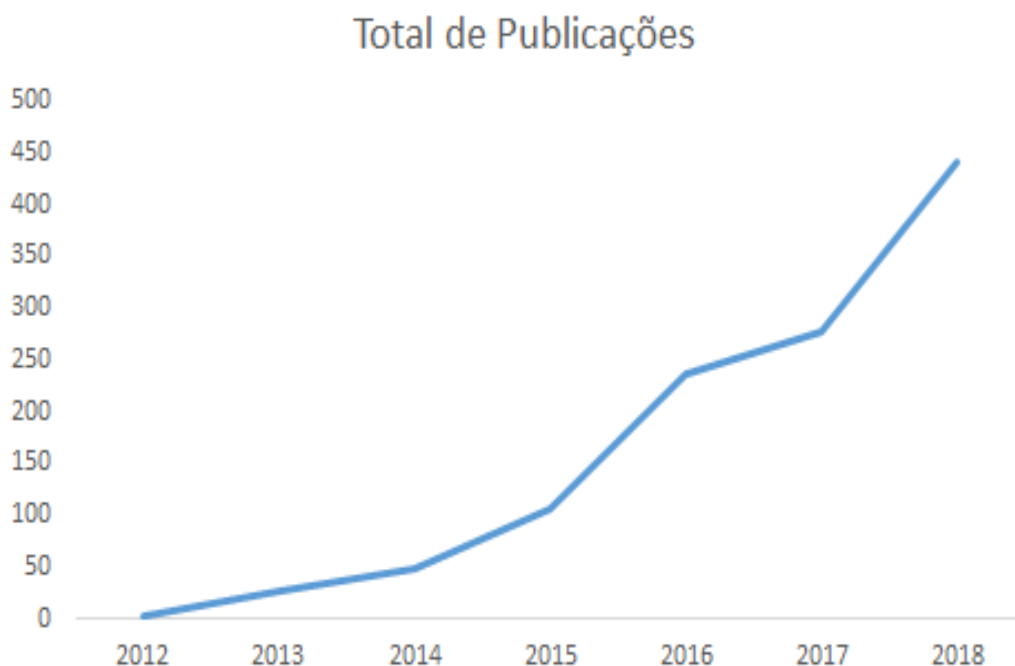


Figura 4: Publicações sobre big data ano a ano.

Fonte: Elaboração própria.

Para a elaboração do mapa da pesquisa bibliométrica no *VOS Viewer*, foi feita uma limitação de no mínimo 20 cocitações em artigo (figura 5), chegando assim em 40 publicações selecionadas. O número de 20 cocitações foi escolhido como forma de restringir o mapa ao máximo de 50 publicações, buscando ter resultados consistentes e em bom número. Todo o afunilamento da pesquisa pode ser visto na figura 6.

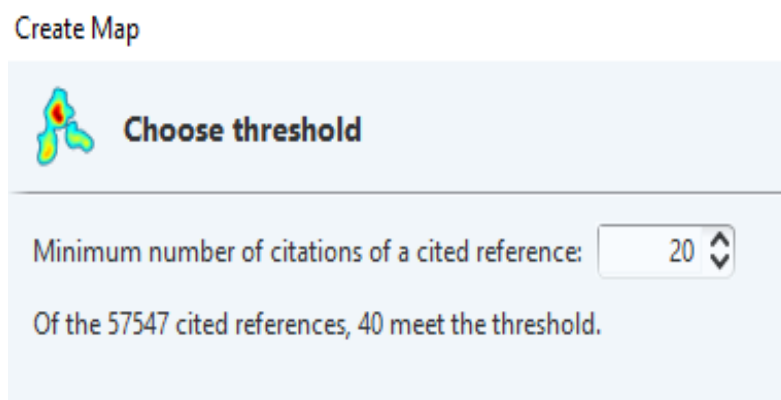


Figura 5: Mínimo de 20 cocitações para elaboração do Mapa.

Fonte: Elaboração Própria.

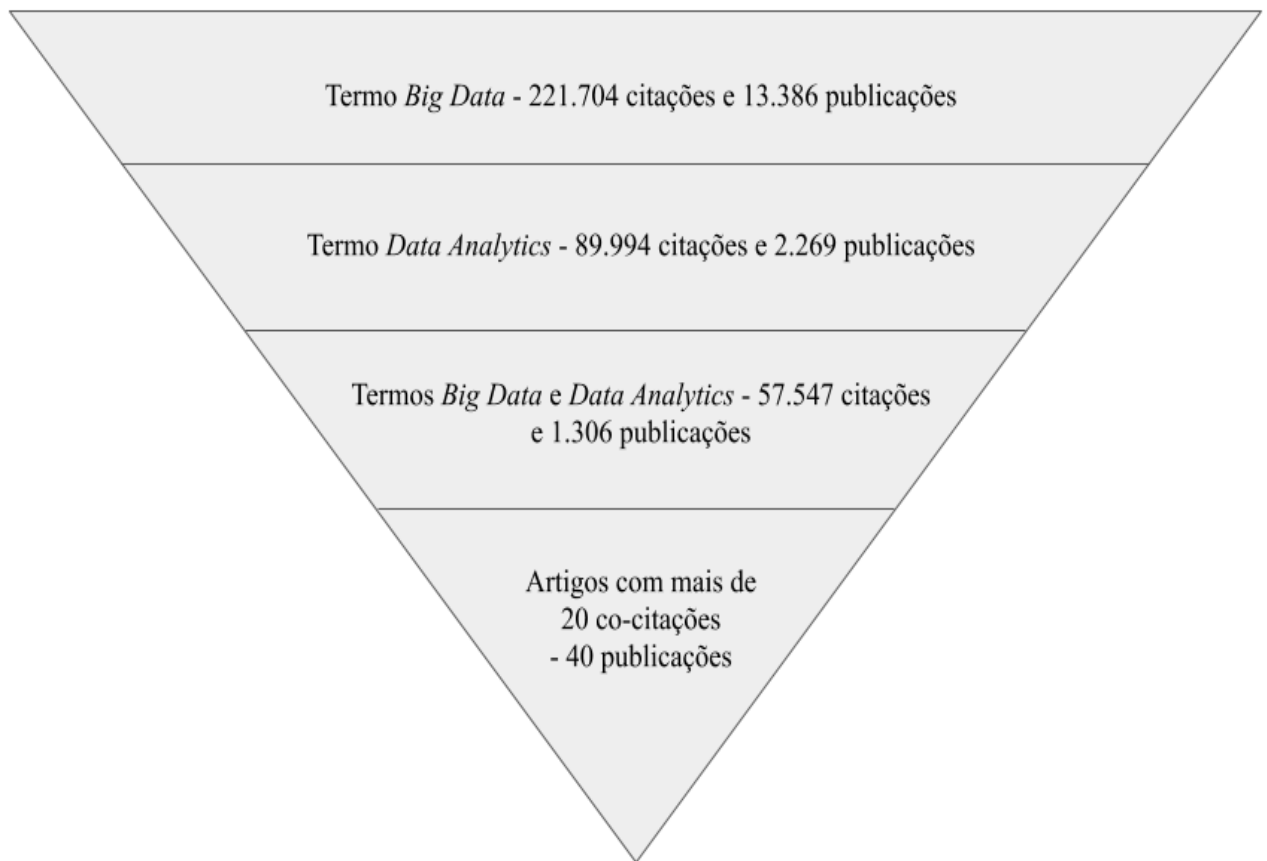


Figura 6: Afunilamento da pesquisa bibliométrica.

Fonte: Elaboração Própria.

No mapa elaborado (figura 7) cada cor sinaliza um *cluster* e cada um deles é determinado como um conjunto de artigos que são co-citados juntos. Cada esfera representa um artigo, as linhas demonstram a ligação entre eles através das cocitações e o seu tamanho é determinado pelo número de vezes em que foi co-citado nas publicações. Na figura 8 temos quais os artigos mais citados e quantas vezes cada um foi citado.

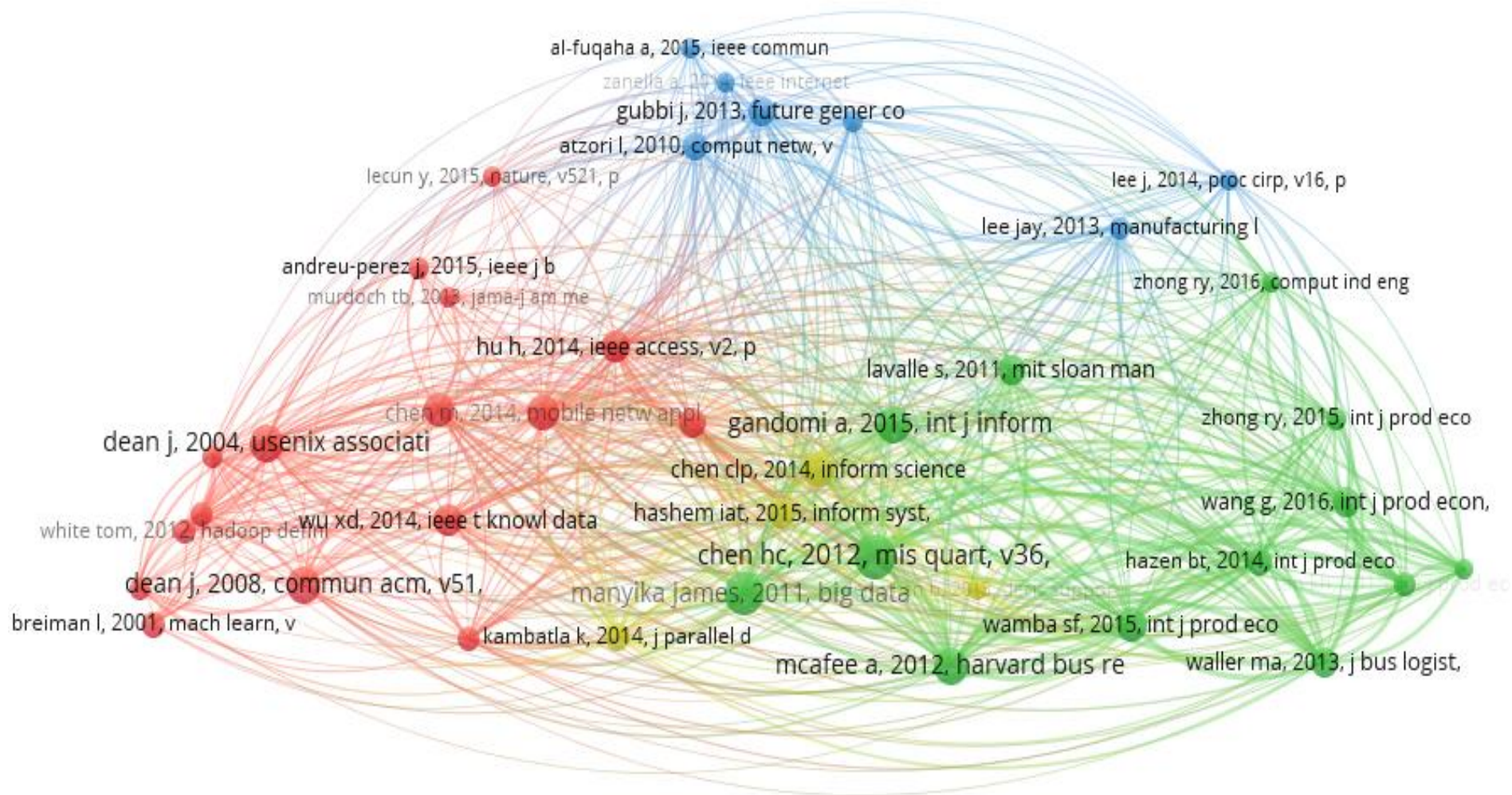



Figura 7: Visualização do mapa de citação sobre big data.

Fonte: Elaboração própria.

 **Verify selected cited references**

| Selected                            | Cited reference  | Citations | Total link strength |
|-------------------------------------|--|-----------|---------------------|
| <input checked="" type="checkbox"/> | chen hc, 2012, mis quart, v36, p1165                     | 86        | 273                 |
| <input checked="" type="checkbox"/> | manyika james, 2011, big data next fronti                | 83        | 265                 |
| <input checked="" type="checkbox"/> | gandomi a, 2015, int j inform manage, v35, p137, ...     | 59        | 261                 |
| <input checked="" type="checkbox"/> | mcafee a, 2012, harvard bus rev, v90, p60                | 60        | 251                 |
| <input checked="" type="checkbox"/> | wamba sf, 2015, int j prod econ, v165, p234, doi 1...    | 49        | 243                 |
| <input checked="" type="checkbox"/> | chen clp, 2014, inform sciences, v275, p314, doi 1...    | 57        | 237                 |
| <input checked="" type="checkbox"/> | hazen bt, 2014, int j prod econ, v154, p72, doi 10.1...  | 37        | 218                 |
| <input checked="" type="checkbox"/> | waller ma, 2013, j bus logist, v34, p77, doi 10.1111...  | 42        | 213                 |
| <input checked="" type="checkbox"/> | wang g, 2016, int j prod econ, v176, p98, doi 10.10...   | 40        | 211                 |
| <input checked="" type="checkbox"/> | chen m, 2014, mobile netw appl, v19, p171, doi 10...     | 55        | 202                 |
| <input checked="" type="checkbox"/> | hu h, 2014, ieee access, v2, p652, doi 10.1109/acce...   | 46        | 187                 |
| <input checked="" type="checkbox"/> | tan kh, 2015, int j prod econ, v165, p223, doi 10.10...  | 28        | 168                 |
| <input checked="" type="checkbox"/> | kambatla k, 2014, j parallel distr com, v74, p2561, ...  | 37        | 147                 |
| <input checked="" type="checkbox"/> | hashem iat, 2015, inform syst, v47, p98, doi 10.101...   | 38        | 146                 |
| <input checked="" type="checkbox"/> | lavage s, 2011, mit sloan manage rev, v52, p21           | 42        | 138                 |
| <input checked="" type="checkbox"/> | raghupathi w, 2014, health inf sci syst, v2, doi 10.1... | 49        | 130                 |
| <input checked="" type="checkbox"/> | dean j, 2004, usenix association proceedings of th...    | 61        | 126                 |
| <input checked="" type="checkbox"/> | wu xd, 2014, ieee t knowl data en, v26, p97, doi 10...   | 47        | 122                 |

< Back    Next >    Finish    Cancel

Figura 8: Artigos mais citados da pesquisa utilizando *Data Analytics* e *Big Data*.

Fonte: Elaboração Própria.

## 2.2. Evolução

Na figura 9 abaixo temos um quadro com a divisão por assunto de cada *cluster* do mapa bibliométrico gerado (figura 7), dentro de cada assunto temos os artigos que são enquadrados. Entre parênteses, tanto no *cluster* quanto nos assuntos temos a quantidade de artigos presentes naquela divisão.



| Cluster          | Assuntos   | Artigos   |
|------------------|--|---|
| Vermelho<br>(16) | Métodos analíticos para <i>Big Data</i> (8)                      | Breiman (2001); Dean, Ghemawat (2004); Dean, Ghemawat (2004); Thusoo <i>et al.</i> (2009); Chansler <i>et al.</i> (2010); Zaharia <i>et al.</i> (2010); White (2012); Wu (2014) |
|                  | <i>Big Data e Data Analytics</i> (4)                             | Zikopoulos (2011); Chen <i>et al.</i> (2014); Hu (2014); Mayer-Schönberger, Cukier (2014)   |
|                  | <i>Big Data</i> aplicado a saúde (3)                             | Murdoch, Detsky (2013); Raghupathi, Raghupathi (2014); Andreu-Perez <i>et al.</i> (2015)  |
|                  | <i>Big Data e Machine Learning</i> (1)                           | LeCun <i>et al.</i> (2015)  |
| Verde (11)       | <i>Big Data e Data Analytics</i> (4)                             | Manyika <i>et al.</i> (2011); Lavallo <i>et al.</i> (2011); McAfee, Brynjolfsson (2012); Chen <i>et al.</i> (2012)  |
|                  | Métodos analíticos para <i>Big Data</i> não estruturado (1)      | Gandomi, Haider (2015)  |
|                  | <i>Big Data</i> aplicado a operações de emergência (1)           | Wamba <i>et al.</i> (2015)  |
|                  | <i>Big Data</i> aplicado a Logística e Cadeia de Suprimentos (5) | Waller, Stanley (2013); Hazen <i>et al.</i> (2014); Zhong <i>et al.</i> (2015); Zhong <i>et al.</i> (2016); Wang <i>et al.</i> (2016)   |
| Azul (6)         | <i>Internet of Things</i> (5)                                    | Atzori <i>et al.</i> (2010); Lee <i>et al.</i> (2013); Gubbi <i>et al.</i> (2013); Zanella <i>et al.</i> (2014); Al-Fuqaha <i>et al.</i> (2015)                                 |
|                  | <i>Big Data</i> aplicado a indústria 4.0 (1)                     | Lee <i>et al.</i> (2014)  |
| Amarelo<br>(4)   | <i>Big data</i> em nuvem (2)                                     | Demirkan, Delen (2013); Hashem <i>et al.</i> (2015)   |
|                  | <i>Big data</i> e aplicações (1)                                 | Chen, Zhang (2014)  |
|                  | Tendências do <i>Big Data</i> (1)                                | Kambatla <i>et al.</i> (2014)   |

Figura 9: Quadro de artigos por assunto e cluster.

Fonte: Elaboração Própria.

### 2.2.1. Cluster dos métodos analíticos do *Big Data* (Vermelho)

O cluster vermelho contém 14 artigos e 2 livros e apresenta as publicações mais antigas encontradas na pesquisa. Os temas centrais das publicações giram em torno do *big data* e *data analytics* em uma visão mais ampla do assunto, aplicações analíticas para o *big data*, focadas em diferentes técnicas e tecnologias, além de 3 artigos que tratam da aplicação de *big data* no setor de saúde e 1 artigo que discorre sobre *machine learning*.

Separando o *cluster* em 4 temas centrais, o maior dentre eles traz técnicas e tecnologias aplicáveis ao *big data* e se inicia com o artigo “*Random Forest*” de Breiman (2001) que possui características mais técnicas por abordar o funcionamento do algoritmo da floresta aleatória em aprendizagem supervisionada. O algoritmo cria uma “floresta” aleatória que nada mais é do que uma combinação de árvores de decisão que podem ser utilizadas para classificação e regressão. O artigo aborda o seu uso e apresenta o resultado após a execução de testes, analisa resultados e explora os efeitos de ruídos em *outputs*, além de dados que possuem muitos *inputs* fracos, com baixo embasamento analítico, como por exemplo os diagnósticos médicos. Leo Breiman (2001), durante os testes, analisa os resultados empíricos da correlação no erro de generalização e, por fim, desenvolve outra aplicação voltada para regressão.

“*MapReduce: Simplified Data Processing on Large Clusters*” de Jeffrey Dean e Sanjay Ghemawat (2004) é um artigo que aparece duas vezes no *cluster* vermelho por ter sido republicado em 2008 em uma versão mais simplificada, porém mantendo o mesmo conteúdo principal. O *MapReduce* “é um modelo de programação e uma implementação associada para processamento e geração de grandes conjuntos de dados” (DEAN, GHEMAWAT, 2004, p.137). O artigo aborda a visão do modelo ao apresentar exemplos e mostrar como ocorre a sua implementação através das etapas de execução, estrutura de *master data*, tolerância a falhas, localização, granularidade das tarefas e tarefas de *backup*. É falado também sobre como é feito o refinamento, performance e experiência do modelo desenvolvido. Segundo Jeffrey Dean e Sanjay Ghemawat, o modelo de *MapReduce* é escalonável e sua facilidade de uso foi notada, levando a vários programas implementados e trabalhos executados no Google.

No quarto artigo, “*Hive - A Warehousing Solution Over a Map-Reduce Framework*”, publicado em 2010 pelos autores Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff e Raghotham Murthy possui conexão com o artigo citado anteriormente ao tratar do *Hive*, uma solução *warehouse* de código aberto em estruturas de *MapReduce* aplicado na plataforma de software livre *Hadoop*, utilizada para armazenar e processar grandes volumes de dados. O *Hive* busca resolver o problema do *Hadoop*, já que “o modelo de programação de redução de mapa é muito baixo e requer que os desenvolvedores escrevam programas personalizados que são difíceis de manter e reutilizar” (THUSOO *et al.*, 2010, p.1). Essa solução busca auxiliar a partir do momento em que entende que o setor de BI agrega um tamanho muito grande de dados e necessita de solução de armazenamento mais ágeis e com bom custo x benefício.

“*The Hadoop Distributed File System*”, publicado em 2010 por Robert Chansler, Hairong Kuang, Sanjay Radia e Konstantin Shvachko também fala sobre o software *Hadoop*, porém focado no *Hadoop Distributed File System (HDFS)*. O HDFS é o sistema de armazenamento do *Hadoop* e “foi projetado para armazenar conjuntos de dados muito grandes de forma confiável e para transmitir esses conjuntos de dados em alta largura de banda para aplicativos do usuário” (CHANSLER *et al.*, 2010, p.1). O artigo, além de apresentar o HDFS, aborda a experiência de utilizá-lo para gerenciar dados no *Yahoo!* que totalizavam 25 *pentabytes*.

Já no artigo “*Spark: Cluster Computing with Working Sets*” publicado por Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker e Ion Stoica em 2010 é apresentada uma alternativa de *software* ao *Hadoop*, o sistema e modelo de programação *Spark*. Durante o seu desenvolvimento, o objetivo era criar algo que superasse o *Hadoop* entregando um fluxo de dados que poderiam ser reutilizados em outras operações paralelas, já que o *Hadoop* trabalha em um modelo de dados acíclico. O sistema possui as mesmas características prometidas pelo *MapReduce* e entrega valor para aplicações que trabalham com “muitos algoritmos de aprendizado de máquina iterativos, bem como ferramentas interativas de análise de dados” (ZAHARIA *et al.*, 2010, p.1)

A terceira edição do livro “*Hadoop: The definitive guide*” publicado em 2012 pelo Tom White é um guia que busca apresentar o *Apache Hadoop*, “uma estrutura de *software* de código aberto que simplifica drasticamente a escrita de aplicativos distribuídos com grande quantidade de dados” (WHITE, 2012, p.601), além de apresentar material do modelo de programação *MapReduce*. São apresentados estudos de casos desenvolvidos com o *Hadoop* para aplicações específicas, demonstrando sua usabilidade para conjunto de dados de variados tamanhos e para configuração e execução de clusters de dados.

Para encerrar os artigos que abordam técnicas e tecnologias de aplicação de *big data*, temos o “*Data mining with big data*”, artigo publicado em 2014 pelos autores Xindong Wu, Xingquan Zhu, Gong-Qing Wu e Wei Ding. Nele é apresentado o teorema HACE, um modelo para processamento de *big data* que utiliza como base o *data mining*. “Esse modelo orientado por dados envolve a agregação de fontes de informação orientada por demanda, mineração e análise, modelagem de interesse do usuário e considerações de segurança e privacidade” (WU *et al.*, 2014, p.1). O teorema HACE se refere às fontes homogêneas e autônomas que geram dados de grandes volumes para utilização no *big data* que irá explorar as relações complexas e evolutivas dentre eles. O artigo fala também sobre três camadas de processamento no *big*

*data* que consideram a computação e acesso dos dados, a privacidade dos dados e o domínio do conhecimento e algoritmos de mineração e *big data*.

Em seguida, a pesquisa resulta quatro artigos que abordam de modo geral o que é o *big data* e *data analytics*. O primeiro deles, “*Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*” dos autores Paul Zikopoulos, Chris Eaton, Tom Deutsch, Dirk Deroos e George Lapis publicado em 2011, fala sobre a aplicação de *big data* na empresa de tecnologia IBM. Nele é apresentado como a empresa está utilizando o *big data* junto com suas tecnologias, são abordadas as três características do *big data* (volume, variedade e velocidade), que serão mais abordados na seção 3 deste trabalho, além da apresentação de casos da indústria. Ele também traz uma introdução sobre o *Hadoop* e como a IBM está estruturando-o para sua utilização.

O artigo seguinte, “*Big data: a survey*” de 2014 foi publicado por Min Chen, Shiwen Mao e Yunhao Liu e é uma revisão do estado da arte e evolução do *big data*. Nele, são tratadas diversas características e definições do assunto, revisadas tecnologias relacionadas como internet das coisas e novamente o *Hadoop*, além de tratar das “quatro fases da cadeia de valor de *big data*, ou seja, geração de dados, aquisição de dados, armazenamento de dados e análise de dados” (CHEN *et al.*, 2014, p.1). Os autores, ao fim, discutem as possíveis direções do tema e quais problemas poderão surgir.

Em 2014 foi publicado o volume 2 do artigo “*Toward Scalable Systems for Big Data Analytics: A Technology Tutorial*” pelos autores Han Hu, Yonggang Wen, (Senior Member, IEEE), Tat-Seng Chua e Xuelong Li, (Fellow, IEEE) que apresentam uma pesquisa bibliográfica de *big data* e um tutorial de sistema para plataformas de análise de *big data*. Assim como no artigo anterior, ele quebra o *big data* nos quatro módulos (ou quatro etapas) de modo a compreender a cadeia de valor. Nele também são apresentadas algumas tecnologias de aplicação e há um espaço de enfoque na entrega de valor do *Hadoop* para lidar com os desafios do *big data*. Ao tratar das características, já existe a consideração de 5 delas, visto na figura 10, (volume, variedade, velocidade, escalabilidade e limitação relacional) e não mais 3 como no artigo “*Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*”. O artigo também aborda todo o processo de transmissão, pré-processamento, armazenagem, gerenciamento do *framework* e análise dos dados e como cada processo desse se divide.

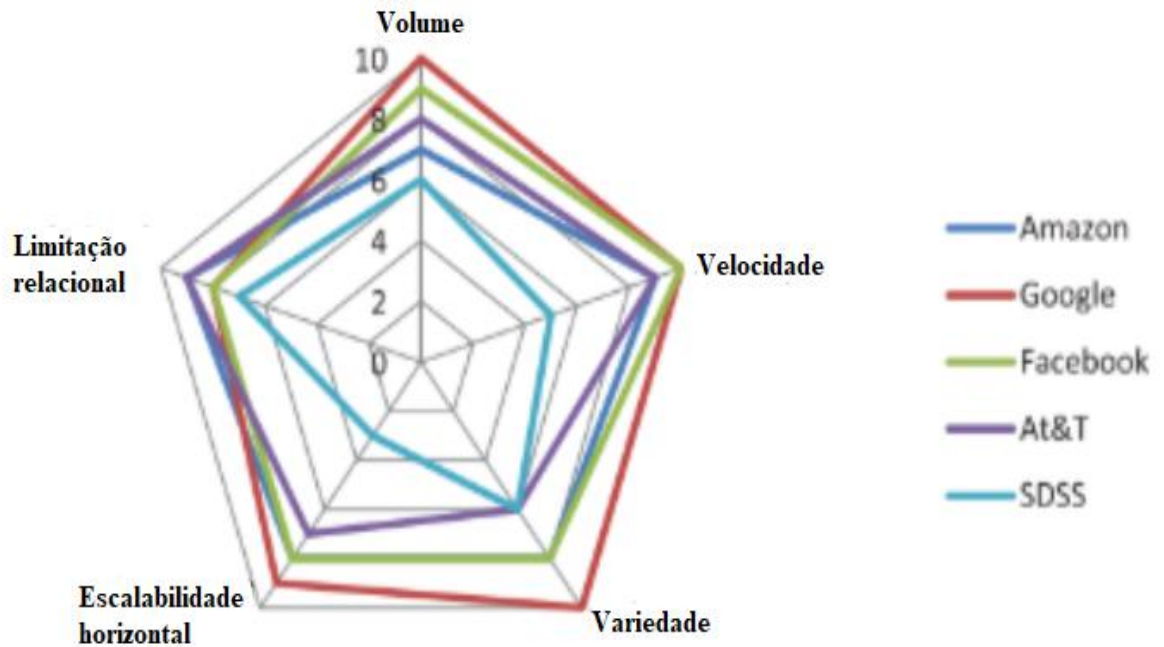


Figura 10: 5 métricas de classificação do big data.

Fonte: Hu *et al* (2014, p.661)

O segundo livro do *cluster*, “*Big Data: A Revolution That Will Transform How We Live, Work and Think*”, foi publicado em 2014 pelos autores Viktor Mayer-Schonberger e Kenneth Cukier e busca trazer a definição de *big data* e a visão dos autores de como ele será capaz de mudar a vida das pessoas e como será possível se proteger dos perigos causados por ele. O livro possui uma característica crítica quanto ao interesse de se obter mais e mais dados sem a garantia de que sua qualidade está intacta. Ele também traz uma visão de como o *big data* pode auxiliar as empresas e principalmente os serviços de saúde, porém não deixando de ressaltar o perigo de um grande volume de informação nas mãos da sociedade.

Como terceiro tema central, há o *big data* aplicado à saúde que foi representado no *cluster* por três artigos. O primeiro deles, “*The Inevitable Application of Big Data to Health Care*”, foi publicado pelos autores Travis B. Murdoch e Allan S.Detsky em 2013. Nele é tratado como o *big data* pode auxiliar na melhora da qualidade e eficiência dos serviços de saúde e para isso ele elenca quatro maneiras de isso ocorrer. A primeira trata sobre a expansão de novos conhecimentos através de bases de evidências observacionais para questões clínica e a segunda aborda o fato de que o *big data* auxilia na disseminação de conhecimento, já que mesmo com a digitalização dos estudos, hoje há um número muito grande de pesquisas, em diversas línguas, que cria empecilhos para o desenvolvimento de tratamentos adequados.

“Essa abordagem está sendo usada na colaboração entre o supercomputador *Watson* da IBM e o *Memorial Sloan-Kettering Cancer Center* para ajudar a diagnosticar e propor opções de tratamento para pacientes com câncer.” (MURDOCH, DETSKY, 2013, p.1352). A terceira maneira fala sobre a oportunidade de utilizar recursos analíticos para integrar biologia de sistemas com dados de registros eletrônicos de informações dos prontuários e a quarta discursa sobre fornecendo informações aos pacientes, permitindo um papel mais ativo do paciente com sua saúde. Para os autores, a privacidade de informações será uma questão para essa aplicação.

No artigo “*Big data analytics in health care: promise and potential*” dos autores Wullianallur Raghupathi e Viju Raghupathi publicado em 2014, além da discussão sobre os impactos da utilização do *big data* no setor da saúde, é discutida a importância dos 4 V’s. Para esse artigo, foram considerados como as principais características do *big data* os já citados volume, velocidade e variedade, acrescentando a eles a veracidade, tão importante quando o assunto é dado relacionado à saúde. O artigo também aborda as diferentes ferramentas e códigos disponíveis para aplicação e apresenta resultados de aplicações reais de *big data* em hospitais.

O último artigo que trata de *big data* e saúde é o “*Big Data for Health*” dos autores Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong and Guang-Zhong Yang publicado em 2015. O artigo apresenta uma visão sobre o desenvolvimento do *big data* aplicado na saúde, destacando suas características e quais serão os benefícios provenientes de conexões com ampla variedade de fontes de dados, sendo eles estruturados ou não, de indivíduos no longo prazo. Acredita-se que o *big data* vai criar possibilidades no gerenciamento de doenças, desde o diagnóstico até o tratamento personalizado. Por outro lado, ele também aborda os desafios relacionados à privacidade e segurança. Nesse artigo são consideradas 6 V’s característicos do *big data*, sendo eles volume, velocidade, veracidade, validade, variedade e variabilidade, visto na figura 11 abaixo.

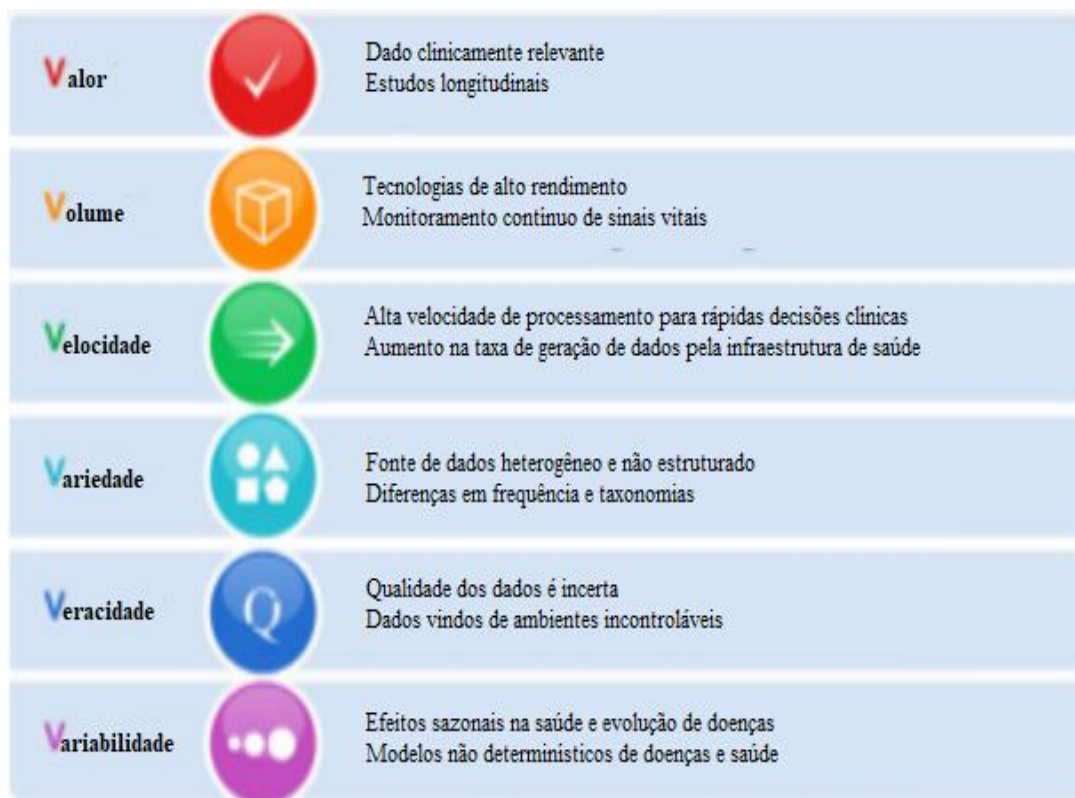


Figura 11: 6 V's do *Big data* (valor, volume, velocidade, variedade, veracidade e variabilidade, que também se aplicam aos dados da saúde.

Fonte: Andreu-Perez *et al.* (2015 p.2).

O último artigo do *cluster*, que tem como tema central o *machine learning*, é o “*Deep Learning*”, publicado em 2015 por Yan LeCun, Yoshua Bengio e Geoffrey Hinton. O artigo busca passar uma visão da definição do *machine learning* e como ocorre o seu funcionamento. São apresentadas as diversas formas de aprendizado de máquina, mas o foco é voltado principalmente para a aprendizagem supervisionada, que é a mais comum. O modelo do *deep learning* capacita modelos com várias camadas de processamento a representar dados com diferentes níveis de abstração, ou seja, “o aprendizado profundo descobre uma estrutura complexa em grandes conjuntos de dados usando o algoritmo de retropropagação para indicar como uma máquina deve alterar seus parâmetros internos que são usados para calcular a representação em cada camada da representação na camada anterior” (LECUN *et al.*, 2015, p.1)

### 2.2.2. Cluster de logística e cadeia de suprimentos (Verde)

Dentro do *cluster* verde tem-se 11 artigos, onde 4 discorrem focados em *Big Data* e *Data Analytics*, não se especificando em nenhuma aplicação ou técnica.

No artigo de MANYIKA *et al.* (2011), “*Big Data: The next frontier for innovation, competition and productivity*”, se demonstra como se dá a criação de valor através do *big data* em diversos setores de bens de consumo, serviços e políticas públicas. MANYIKA *et al.* (2011) também defende que um concorrente que não conseguir desenvolver bem as capacidades necessárias atreladas a nova tecnologia será esquecido pelo mercado, onde o *big data* se tornará padrão da indústria.

LAVALLE *et al.* (2011), no artigo “*Big Data, Analytics and the Path From Insights to Value*”, discorre sobre como a análise de dados favorece a melhor tomada de decisão, sendo refletido no mercado, visto que as empresas com topo de eficiência no mercado são mais orientadas a dados, e como isto se dá em cada setor.

Em “*Big Data: The management revolution*”, segundo MCAFEE e BRYNJOLFSSON (2012, p.6.) “usar *Big Data* leva para previsões melhores e melhores previsões trazem melhores decisões” e as empresas que souberem dominar esse conhecimento com dados terão destaque frente aos seus concorrentes. Porém a visão humana do negócio não pode ser esquecida após a adoção dos dados, e os dados não devem ser direcionados por escolhas previamente tomadas e sim justificar uma escolha após a análise.

Já CHEN *et al.* (2012), no artigo “*Business Intelligence and Analytics: From Big Data to Big Impact*”, apresentam a aplicação do *Big Data* e *Analytics* conjuntamente e como eles evoluíram do BI&A 1.0, passando pelo BI&A 2.0 até o BI&A 3.0, como visto na figura 12. O termo *Business Intelligence* começou a ser utilizado no ambiente de Tecnologia da Informação evoluindo para o *Business Analytics* como um mecanismo analítico de TI até chegar ao *Big Data*, solução mais em foco atualmente devido à complexidade e volume dos dados capturados.



| BI&A 1.0  | BI&A 2.0   | BI&A 3.0  |
|---|--|---|
| Dados estruturados e armazenados em bancos de dados relacionais, representados através de indicadores e dashboards. Análises feitas com base em métodos estatísticos e técnicas de data mining. | Dados não estruturados, principalmente textos, e coletados via internet. Através do uso e interação de usuários na internet, é possível obter informações sobre o interesse dos clientes e assim explorar oportunidades. | Dados obtidos através de dispositivos móveis sensores. Visto como próximo passo para o mercado de BI. |

Figura 12: Resumo da evolução do BI&A

Fonte: Adaptado de Chen, Chiang e Storey (2012, p.1169)

Posteriormente, como exceção ao *cluster*, temos um artigo discorrendo sobre métodos analíticos aplicados a *Big Data*. GANDOMI, HAIDER (2015), no artigo “*Beyond the hype: Big data concepts, methods, and analytics*”. Este artigo destaca a necessidade de desenvolver métodos analíticos apropriados e eficientes para alavancar volumes massivos de dados heterogêneos em formatos de texto, áudio e vídeo não estruturados. Além disso adentra em alguns métodos para analisar esses formatos de dados sempre tangenciando o uso para análise preditiva.

WAMBA *et al.* (2015) realiza um estudo sistemático da literatura sobre *Big Data* e posteriormente contextualiza aplicando a teoria em um estudo de caso em operações de emergência. Neste o *Big Data* cria valor a partir do compartilhamento de dados em tempo real, geolocalização de todos os voluntários sendo posicionados de acordo com a demanda, integração de vários bancos de dados para análise de risco de ações preventivas, todos os recursos disponíveis em tempo real para a tomada de decisão.

Em WALLER, STANLEY (2013), no artigo “*Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management*”, se inicia a discussão sobre a aplicação de *Big Data* na cadeia de suprimentos. Nesse é defendido a aplicação não agnóstica de domínio, onde as pessoas responsáveis pela análise devem também saber sobre o tópico de estudo, evitando cair em falsos positivos. Também se explora as possibilidades do *Big Data* na cadeia de suprimentos, como demonstra a figura 13 abaixo:

| Do utilizador  | Previsão  | Gestão de Inventário   | Gestão de Transporte  | Recursos humanos   |
|----------------|---|--|---|--|
| Transportadora | Hora da entrega, fatoração do tempo, características do motorista, hora do dia e data | Disponibilidade de capacidade em tempo real  | Roteamento ideal, levando em consideração o tempo, o congestionamento de tráfego e as características do motorista        | Redução na rotatividade de driver, atribuição de driver, usando análise de dados de sentimento       |
| Fabricante     | Resposta antecipada ao sentimento extremamente negativo ou positivo do cliente        | Redução no encolhimento, resposta eficiente do consumidor, resposta rápida e inventário gerenciado do fornecedor | Melhoria na notificação do tempo de entrega e disponibilidade, dados de vigilância para melhorar o gerenciamento do pátio | Monitoramento mais eficaz da produtividade, sensores médicos para segurança do trabalho nas fábricas |
| Varejista      | Dados de sentimento do cliente e uso de dispositivos móveis nas lojas                 | Melhoria na precisão do sistema de inventário perpétuo   | Vinculando o congestionamento de tráfego local e o clima para armazenar o tráfego   | Redução de mão de obra devido à redução do estoque extraviado  |

Figura 13: Possibilidades de aplicação do Big Data na cadeia de suprimentos.

Fonte: Adaptado de Waller e Stanley (2013, p.82)

O segundo artigo sobre cadeia de suprimentos, “*Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications*”, HAZEN *et al.* (2014) discorre sobre a necessidade e monitoramento da qualidade de dados em *Big Data*, apresentando e discutindo métodos que auxiliem essa necessidade. Em seguida, na figura 14 abaixo, demonstra a aplicação desses métodos de controle de qualidade de dados a produtos de dados em um ambiente de cadeia de suprimentos.

| Dimensão de qualidade de dados | Descrição                                   | Exemplo de cadeia de fornecimento  |
|--------------------------------|---|--|
| <b>Precisão</b>                | Os dados estão livres de erros?             | O endereço de envio do cliente em um sistema de gerenciamento de relacionamento com o cliente corresponde ao endereço no pedido do cliente mais recente        |
| <b>Pontualidade</b>            | Os dados estão atualizados?                 | O sistema de gerenciamento de estoques reflete os níveis de estoque em tempo real em cada local de varejo  |
| <b>Consistência</b>            | Os dados são apresentados no mesmo formato? | Todas as datas de entrega solicitadas são inseridas em um formato DD/MM/AA   |
| <b>Completeness</b>            | Os dados necessários estão faltando?        | O endereço de envio do cliente inclui todos os pontos de dados necessários para concluir uma remessa (ou seja, nome, endereço, cidade, estado e código postal) |

Figura 14: Dimensões da qualidade de dados.

Fonte: Adaptado de HAZEN *et al.* (2014, p.74)

ZHONG *et al.* (2015) em “*A big data approach for logistics trajectory discovery from RFID-enabled production data*”, propõe o *Big Data* aplicado ao chão de fábrica habilitado a partir de *Radio Frequency Identification (RFID)* ou Identificação por radiofrequência, em português. A partir de um fabricante de peças automotivas o *RFID* foi utilizado para visualizar, gerenciar e rastrear cada material em tempo real. A posição de cada item em cada instante, em tempo - quase - real, “revelam um rico conhecimento para tomadas de decisão avançadas como *MRP* e *APS*. Além disso, as principais descobertas e observações são convertidas em implicações gerenciais, pelas quais os usuários são capazes de tomar decisões precisas e eficientes sob diferentes situações.” (ZHONG *et al.*, 2015, p.269).

ZHONG *et al.* (2016) faz uma análise sobre os principais caminhos que as discussões de *Big Data* aplicado a cadeia de suprimentos de manufatura e serviços está tomando no mundo. Faz essa análise com um recorte por continentes e pelas 10 principais empresas de cadeia de suprimentos do mundo.

WANG *et al.* (2016) classifica a literatura de *Big Data* aplicado a logística e cadeia de suprimentos com base na natureza da análise (descritivo, preditivo, prescritivo) e o foco operacional e estratégico da gestão logística e da cadeia de suprimentos. Desta forma discute as capacidades da análise de *Big Data* na cadeia de suprimentos, propondo uma estrutura de maturidade da tecnologia baseada na revisão da literatura. Conclui assim que o uso de *Big Data* aplicado a gestão logística e da cadeia de suprimentos ajuda as organizações a medir o desempenho de várias áreas e fornece a elas a capacidade de estabelecer uma referência para

determinar operações de valor agregado, se aprofundando em problemas de baixo desempenho dando insumos para a descobertas das causas raízes.

Ou seja, podemos perceber que a discussão sobre aplicação do *Big Data* na cadeia de suprimentos se inicia na discussão das possibilidades de usos e ganhos, passando pela garantia da qualidade dos dados que serão utilizados, chegando a estudo de caso de aplicações específicas como habilitadoras do *Big Data*, neste caso o *RFID*, e depois, de forma mais macro, como que os grandes *players* estão interagindo com esta tecnologia, tendo pôr fim a análise da adoção dessa tecnologia em relação ao que a literatura propõe.

Sendo assim o *cluster* verde claramente serve de insumo para o estudo da aplicação do *Big Data* na cadeia de suprimentos e logística, explorando seu uso do macro, desde artigos que apenas discorrem sobre *Big Data* de forma geral, até às micro aplicações, evolução e análise da tecnologia neste ambiente.

### 2.2.3. Cluster de *Internet of Things* (Azul)

O *cluster* azul é mais coeso, onde de início fica claro que o *Big Data* é utilizado como uma tecnologia habilitada a partir da internet das coisas, sendo representados por 5 dos 6 artigos do *cluster*. O artigo que destoa discorre sobre o uso do *big data* na indústria 4.0, termo que se refere a quarta revolução industrial.

ATZORI *et al.* (2010), no artigo “*The Internet of Things: A survey*”, compara as diferentes visões da literatura acerca do objeto estudado, quais são as principais tecnologias capacitadoras da *Internet of things (IoT)* ou Internet das coisas, em português, e descreve as principais aplicações inferindo sobre o uso futuro da tecnologia, mas mostrando o que já foi realizado para habilitar a tecnologia.

LEE *et al.* (2013) discorre sobre manufatura preditiva e sistemas ciberfísicos habilitados a partir da internet das coisas. Neste contexto os dados se tornam mais acessíveis e onipresentes contribuindo para o *Big Data*. Propõe a melhoria de sistemas de informações de manufatura a partir da análise das funções 5C (*Connection, Cloud, Content, Community e Customization*), conexão a partir de sensores e redes, nuvem com dados sob demanda a qualquer momento, conteúdo com correlação e significado destes dados, comunidade com a parte de compartilhamento e social e customização com personalizar e criar valor. Todo este ambiente pode ser visualizado na figura 15 abaixo.



Figura 15: Framework de sistema de manufatura preditiva

Fonte: Adaptado de LEE *et al.* (2013 p.40)

GUBBI *et al.* (2013), no artigo “*Internet of Things (IoT): A vision, architectural elements, and future directions*” entende a internet das coisas como sendo a proliferação de dispositivos interconectados em uma rede de comunicação ativa, onde sensores e atuadores se misturam com o ambiente. Assim apresenta as tendências atuais da pesquisa em *Internet of Things* centrada na nuvem, realizando um estudo de caso de análise de dados na plataforma de nuvem Aneka / Azure.

No artigo “*Internet of Things for Smart Cities*” de ZANELLA *et al.* (2014) explora o conceito de *Smart Cities*, sendo apoiada pelo uso de *Internet of Things*, que visa explorar as mais avançadas tecnologias de comunicação para apoiar serviços de valor agregado aos cidadãos. Ou seja, o artigo busca discutir a estrutura geral para um projeto de *Internet of Things* urbana, inclusive analisando o caso da *Padova Smart City*, uma implementação de *Internet of Things* em uma ilha na cidade de Padova, Itália.

No artigo “*Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications*”, de AL-FUQAHA *et al.* (2015), de acordo com os autores a premissa básica de *Internet of Things* é ter sensores inteligentes interagindo, sem envolvimento humano, para entregar uma nova classe de aplicações. Assim sendo, a primeira fase é composta pela internet,

mobilidade e comunicação máquina a máquina, porém os autores preveem que *Internet of Things* interligue diversas tecnologias que vão permitir novas aplicações, apoiando a tomada de decisão. Além disso o artigo se propõe a trazer uma discussão mais técnica e específica, expondo os padrões e protocolos tecnológicos atuais de *Internet of Things*.

Por fim, LEE *et al.* (2014), no artigo “*Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment*” se propõe a entender como as atuais indústrias de manufatura evoluem para a indústria 4.0 habilitado a partir do *Big Data*. Dessa forma as máquinas se conectam em uma comunidade colaborativa, inclusive com automanutenção, reduzindo quebras e interrupções de produção, com a informação percorrendo a linha de produção, facilitando a tomada de decisão, reduzindo custo de trabalho e eventualmente de energia elétrica.

Portanto se torna claro que o *cluster* azul desenvolve a tecnologia da *Internet of Things*, utilizando por fundo o *Big Data* como consequência dessa tecnologia. Além disso também se nota que a *Internet of Things* e *Big Data* são habilitadores da quarta revolução industrial ou manufatura inteligente.

A discussão no *cluster* evolui do menos específico, onde discute-se *Internet of Things* como tecnologia de forma ampla, até o mais específico como seu uso na indústria 4.0, como fomentador de cidades inteligentes e seus padrões e protocolos.

#### 2.2.4. *Cluster Big Data* em nuvem (Amarelo)

Dentre o *cluster* amarelo, existem 4 artigos que foram divididos em três temas centrais: *big data* em nuvem, *big data* e aplicações e tendências do *big data*. O artigo “*Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud*” dos autores Haluk Demirkan e Dursun Delen publicado em 2013 fala sobre os sistemas de suporte à decisão (*DSS*, em inglês) orientados à serviço aplicados em nuvem. São apresentados os requisitos para se aplicar o *DSS* e como sua utilização pode contribuir para a aplicação em produtos e serviços de engenharia na nuvem. “Quando definimos dados, informações e análises como serviços, vemos que os mecanismos de medição tradicionais, que são principalmente orientados por tempo e custo, não funcionam bem. As organizações precisam considerar o valor do nível de serviço e a qualidade, além do custo e da duração dos serviços prestados. O *DSS* na nuvem permite economias de escala, escopo e velocidade” (DEMIRKAN, DELEN, 2013, p.412). Segundo os autores, os dados são ativos primários para uma empresa nessa nova realidade.

“The rise of ‘big data’ on cloud computing: Review and open research issues”, artigo publicado em 2015 por Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani e Samee Ullah Khan aborda a necessidade de infraestrutura exigida pelo *big data* para o processamento e análise do grande volume de dados e como a computação em nuvem é capaz de auxiliar nesse quesito, como visto na figura 16. Além disso, o artigo estuda a tecnologia do *Hadoop*, como artigos anteriores citados, e investiga os diferentes desafios atrelados ao *big data*, como os já mencionados, privacidade e qualidade de dados, por exemplo. Também são apresentados estudos de caso de aplicação de *big data* em computação em nuvem ao final.

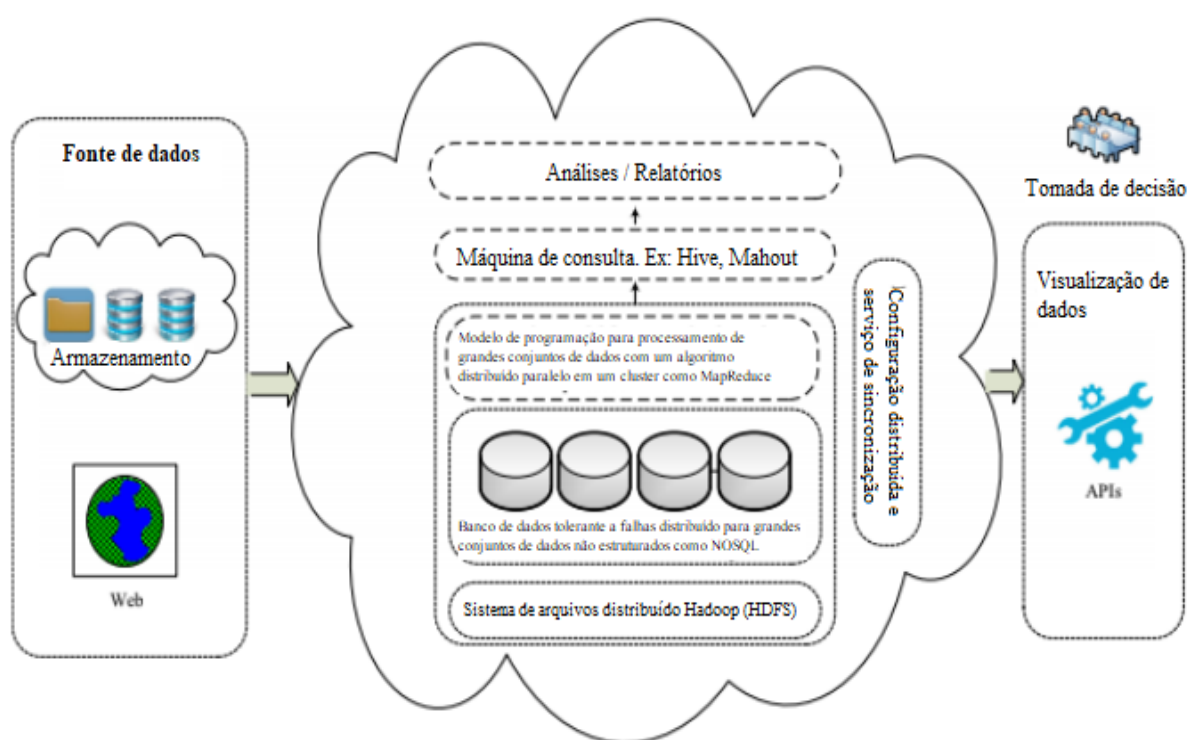


Figura 16: Uso da computação em nuvem no *big data*.

Fonte: HASHEM *et al.* (2015, p.103)

O artigo que é focado em *big data* e aplicações no *cluster* amarelo é o “Data intensive applications, challenges, techniques and technologies: A survey on Big Data” de 2014 publicado pelos autores C.L. Philip Chen e Chun-Yang Zhang. Assim como os artigos que tem esse mesmo tema central, ele aborda as características inerentes ao *big data*, porém com uma abordagem voltada a descoberta científica intensiva em dados (*DISD*), que é o estudo dos problemas inerentes ao *big data*. “Por um lado, o *Big Data* é extremamente valioso para produzir produtividade nas empresas e avanços evolutivos nas disciplinas científicas, o que

nos dá muitas oportunidades de fazer grandes progressos em muitos campos. (...) Por outro lado, o *Big Data* também surge com muitos desafios, como dificuldades na captura, armazenamento, análise e visualização de dados” (CHEN, ZHANG, 2014, p.314). O artigo também traz discussões sobre técnicas e tecnologias aplicadas ao *big data*, como *machine learning* e *Hadoop*, além de tratar sobre sua relação com computação em nuvem.

“*Trends in big data analytics*”, dos autores Karthik Kambatla, Giorgos Kollias, Vipin Kumar e Ananth Grama publicado em 2014 é o último artigo do *cluster* amarelo e traz uma visão sobre as tendências do *big data*. São abordadas tendências em escala e de aplicativos de análise de *big data*, além das tendências atuais e futuras em *hardware* e das técnicas de *software* atualmente utilizadas e possibilidades para o futuro.

#### 2.2.5. Evolução entre *clusters*

Com a análise de cada *cluster*, os anos de publicação dos artigos e as ligações do mapa bibliométrico (Figura 7), é possível perceber um relacionamento e evolução na discussão de *Big Data* e *Data Analytics*, essa relação foi simplificada na figura 17.

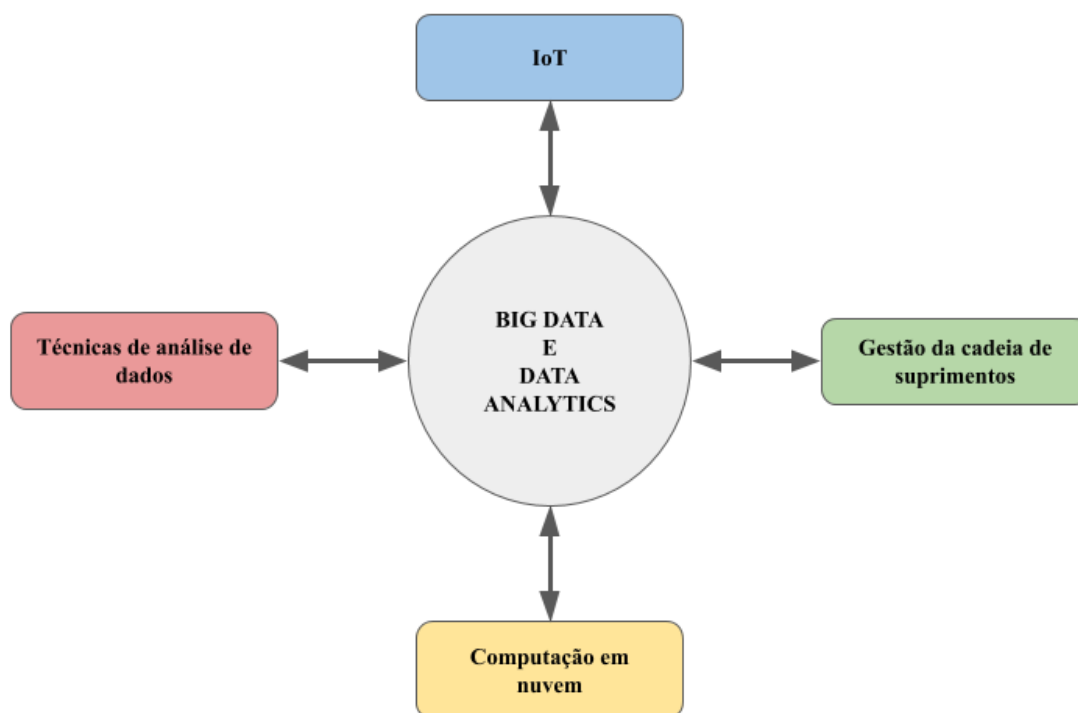


Figura 17: Evolução da discussão por cluster.

Fonte: Elaboração própria.



O *cluster* vermelho, como dito, aborda o *Big Data* por dentro, explorando os métodos de uso e as técnicas de análises, o que serve de alicerce para todas as aplicações de *Big Data*, inclusive as na área de saúde, que estão presentes no *cluster*.

Na outra ponta do mapa tem-se o *cluster* verde, que também possui artigos básicos sobre *Big Data* e *Data Analytics*, que servem de base interna para as posteriores discussões sobre a aplicação em logística e cadeia de suprimentos - uma das grandes áreas de aplicação do *Big Data* e *Analytics* - como por exemplo o uso do *RFID*, debatido por ZHONG *et al.* (2015).

Ambos os *clusters*, vermelho e verde, servem de base para as discussões dos *clusters* azul e amarelo. Esse primeiro habilita o *Big Data* a partir da *Internet of Things*, ou seja, o grande volume de dados gerados por equipamentos interconectados, que utilizam destes próprios dados para otimizar vários parâmetros do ambiente ao redor, se autoajustando, como no caso das propostas de *Smart Cities* explorado no artigo de ZANELLA *et al.* (2014) e culminando na Indústria 4.0 (quarta revolução industrial) explorado por LEE *et al.* (2014).

No *cluster* amarelo tem-se as tendências desta tecnologia. Uma delas é o uso da computação na nuvem para reduzir a infraestrutura necessária de processamento, análise e armazenamento dos dados. Além disso o cluster trata de outras tecnologias emergentes como *Machine Learning*, e KAMBATLA *et al.* (2014) trata exatamente das tendências atuais e futuras em *hardware* e *software* a partir de *Big Data*.

A partir da análise desta seção, abordaremos na próxima os artigos que citam a caracterização do *Big Data* a partir dos V's.

### 3. OS V'S DO BIG DATA

Dentro da pesquisa bibliométrica abordada no tópico anterior, temos uma característica comum a alguns artigos: Eles caracterizam o *Big Data* utilizando alguns conceitos que se iniciam com a letra V. Neste tópico será abordado a definição a partir de diferentes autores presentes na pesquisa ou citados em artigos da pesquisa sobre esses conceitos e quais são suas convergências e divergências.

A discussão se inicia, pelos artigos mais antigos em 3 V's: variedade, velocidade e volume, vide figura 18, e depois outros são incluídos na discussão. A autoria da definição de *Big Data* a partir dos V's é dada a IBM pelos artigos mais antigos, ou seja, o conceito se inicia na empresa para depois chegar à academia científica. Atualmente, em ZIKOPOULOS (2011), a IBM já inclui veracidade na lista de características do *Big Data*.

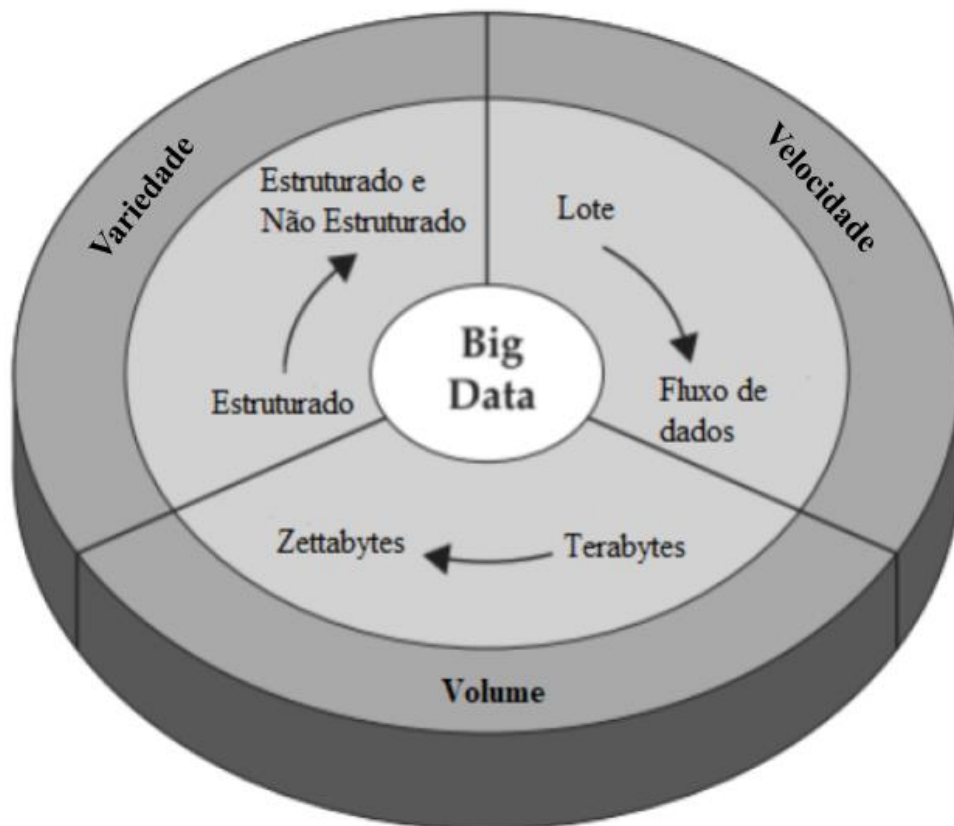


Figura 18: 3 V's da IBM.

Fonte: Traduzido de Zikopoulos 2011 p.33

### 3.1. Volume

Volume é a principal característica do *Big Data*, tanto que é presente até no nome, onde *big* se refere ao grande volume de dados. RUSSOM (2011) define o volume dos dados em relação ao seu espaço ocupado na memória, sendo da ordem de *terabytes* até *petabytes*. Porém, ZIKOPOULOS (2011) chegou a uma questão interessante quanto a isso: definir o volume do *Big Data* em relação ao volume ocupado na memória é certeza de que essas definições ficarão desatualizadas em muito pouco tempo. GANDOMI, HAIDER (2015) converge para a mesma linha de raciocínio, afirmando que a definição do volume é relativa e varia de acordo com fatores como o tempo e o tipo de dado, o que, para eles torna impraticável a definição de um limite específico para grandes volumes de dados.

Ou seja, a definição de volume é pouco precisa, o certo é que para ser considerado *Big Data* deve possuir uma quantidade de espaço ocupado de uma ordem de grandeza superior ao comumente usado.

### 3.2. Variedade

Variedade refere-se tanto as fontes de origem quanto aos tipos de dados. RUSSOM (2011) exemplifica a variedade considerando as novas fontes da *web*, como fluxo de cliques e mídias sociais e quando se refere aos tipos de dados categoriza em 3 tipos: estruturados, semiestruturados e não estruturados - além disso considera difícil de categorizar os provenientes de áudio, vídeo e outros dispositivos.

Dados estruturados são os mais tradicionais e já utilizados normalmente para análise. O *Big Data* se expande além das análises tradicionais, possibilitando análises de dados não estruturados, como texto e linguagem humana e semiestruturados (*XML* e *RSS Feed*).

ZIKOPOULOS (2011) diz que 80% dos dados do mundo são desestruturados ou não estruturados, e isso que traz um grande valor ao *big data*, visto que há muita informação que não seria extraída somente com as análises rotineiras de dados estruturados.

RAGHUPATHI (2014) discorre sobre a questão dessa variedade de dados na área da saúde. De acordo com o artigo o atendimento médico gera muitos dados não estruturados como registros médicos, anotações manuscritas de enfermeiros e médicos, além de várias imagens como tomografia computadorizada e ressonância magnética. Ou seja, com tecnologia comum pouco desses dados conseguem ser capturados, armazenados, organizados e analisados.

ZHONG *et al.* (2016) ao caracterizar em relação a *Supply Chain*, foca como origem de dados diversos tipos de sensores, desde os locais de fabricação, passando pelas rodovias, lojas

e casas. “A integração desses dados em uma formatação padrão requer uma linguagem de maquiagem mais complexa” (ZHONG *et al.*, 2016, p.573).

### **3.3. Velocidade**

A questão de velocidade está se referindo a taxa de geração dos dados ou de entrega dos dados. RUSSOM (2011) exemplifica a partir dos dados da web, que são gerados em “tempo real” e devem ser analisados e servir de base para ações, no limite, também em “tempo real”. “A vanguarda do *big data* é o *streaming* de dados” (RUSSOM, 2011, p.7).

ZIKOPOULOS (2011) traz uma nova característica para velocidade, de acordo com o artigo também se refere a taxa de armazenamento. “Sugerimos que você aplique esta definição aos dados em movimento: a velocidade com que os dados estão fluindo”. (ZIKOPOULOS, 2011, p.8)

É um consenso aos artigos de que a velocidade é muito importante para se criar valor a partir do *Big Data*. ZIKOPOULOS (2011) diz a criação de uma vantagem competitiva pode estar na identificação de uma oportunidade um pouco antes de seus concorrentes. MCAFEE, BRYNJOLFSSON (2012) exemplifica esta questão em relação aos analistas de *Wall Street* na bolsa de valores.

### **3.4. Valor**

Aqui entra o quarto V mais citado, o valor, como ilustrado na figura 19.

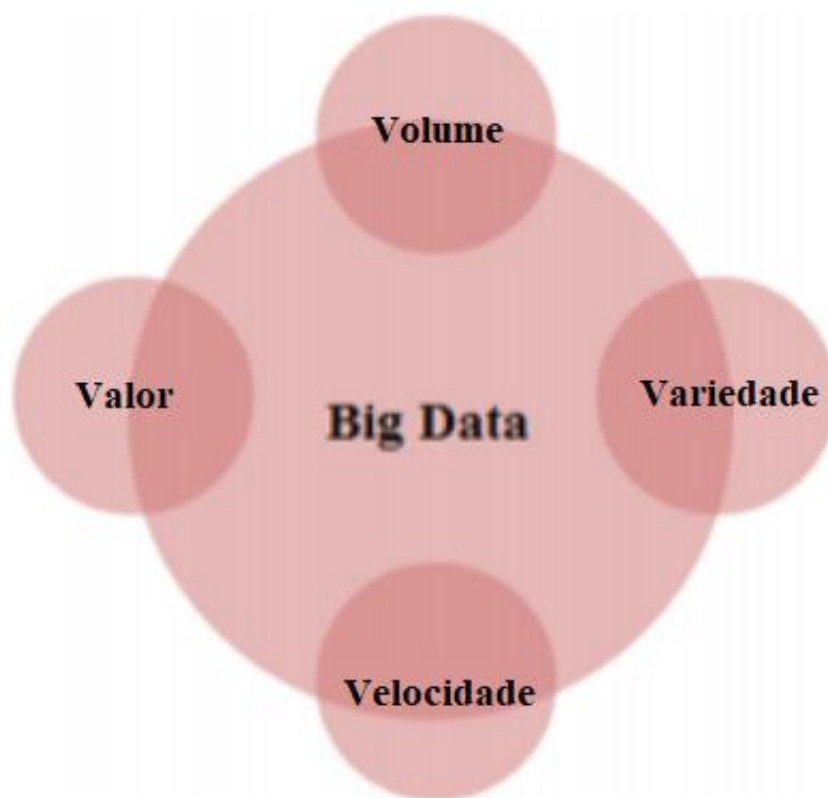


Figura 19: 4 V's mais citados do Big Data.

Fonte: Traduzida de HASHEM *et al.* (2015)

É um consenso entre os autores que usam valor para definir *Big Data* que este é presente em grande quantidade, porém em baixa densidade, ou seja, são dados que em baixo volume trariam muita pouca informação para a tomada de decisão.

De acordo com CHEN *et al.* (2014), o incremento do valor como definição ao *Big Data* serve para destacar o significado e a necessidade dessa tecnologia, afinal o *big data* é usado para se extrair valores ocultos, que só se consegue tendo acesso a um grande volume de dados. Já WAMBA *et al.* (2015) considera valor como demonstração da importância de se extrair benefícios econômicos do *Big Data*.

ZHONG *et al.* (2016) vê o valor por outro ponto, de acordo com ele, num primeiro momento, é difícil extrair valor do *big data* por conta de sua magnitude e complexidade. Num segundo momento é desafiador analisar qual é o real impacto das ideias, relatórios, estatísticas e por fim decisões devido à grandes influências nas perspectivas micro e macro em uma organização.

### **3.5. Veracidade**

A IBM utiliza atualmente a veracidade como sendo o quarto V, junto com volume, velocidade e variedade. De acordo com GANDOMI, HAIDER (2015) esse V representa a falta de confiabilidade ou imprecisão de algumas fontes de dados. O artigo também exemplifica em relação a análise de sentimentos de clientes em rede social, que apesar de conter informações valiosas, são de natureza incerta por envolver julgamento humano.

WAMBA *et al.* (2015, p.235) traz uma definição um pouco diferente, ao citar Lavallo (2009) que diz que “alguns analistas estimam que 1 em cada 3 líderes de negócios não confiam nas informações que usam para tomar decisões”, ele diz que a veracidade é incluída na definição para destacar a importância de dados de qualidade e o nível de confiança em várias fontes de dados.

ANDREY-PEREZ *et al.* (2015) exemplifica com relação ao setor de saúde, em relação a exames e medições realizadas em ambientes menos controlados que se tornam menos confiáveis em comparação a dados clínicos coletados por profissionais treinados.

Ou seja, a veracidade se torna presente para a definição de *big data* para que incertezas e imprecisões tanto das fontes quanto inerentes ao dado não atrapalhem a análise.

### **3.6. Variabilidade**

Somente dois autores da pesquisa citam variabilidade. ANDREY-PEREZ *et al.* (2015) define o termo como a consistência dos dados ao longo do tempo, já GANDOMI, HAIDER (2015) define em relação a velocidade, como variação nas taxas de fluxo de dados.

### **3.7. Verificação**

Somente um autor da pesquisa utiliza de verificação na definição de *big data*. ZHONG *et al.* (2016) ao exemplificar o *big data* para uso em cadeia de suprimentos, afirma que há uma grande variedade de dados inválidos, como por exemplo, ruídos, atributos imprecisos, etc. Neste cenário a verificação se dá como ferramenta para reduzir problemas de qualidade e conformidades dos dados.

Na próxima seção será abordado a caracterização da empresa na qual se dará o estudo de caso e como funciona sua estrutura de base de dados para o setor comercial e sua caracterização frente aos V's do *Big Data* que foram abordados nesta seção.

## **4. ESTUDO DE CASO: EMPRESA RP COSMETICS**

Nesta seção será abordado a caracterização da empresa RP *COSMETICS* na qual se dará o estudo de caso. Depois da caracterização da empresa será descrito sua estrutura de dados para o setor comercial e como essa estrutura se caracteriza frente aos V's do *Big Data*.

### **4.1. Caracterização da Empresa**

O estudo de caso será desenvolvido em uma empresa multinacional que atua no segmento da beleza reconhecida mundialmente. No Brasil, país que detém o quarto maior mercado mundial de beleza, a empresa atua há mais de 50 anos oferecendo diferentes categorias de produtos, como cuidados com os cabelos e com a pele, maquiagens e fragrâncias.

Para não ter sua razão social exposta diretamente por motivos de sigilo empresarial a empresa será caracterizada ao longo deste trabalho como RP *Cosmetics*. Segundo reportes oficiais, em 2018 a empresa teve um faturamento de quase 27 bilhões de euros contando com 86 mil funcionários no mundo.

Seus produtos são distribuídos em um portfólio acessível para diferentes classes sociais em todo o país. Suas unidades administrativas e de produção estão alocadas nas grandes capitais, contando com centros de distribuição em cidades estratégicas.

Relativo ao negócio, a SEBRAE divulgou que dentre as grandes tendências para 2020, temos o mercado de higiene e cosméticos, citando a preocupação com a beleza e autoestima. O cenário de crise e pandemia reforça essa preocupação, demonstrando-se favorável para algumas categorias. Por outro lado, maquiagem e fragrância são mais impactadas negativamente.

### **4.2. Estrutura do *Big Data* da Companhia**

Devido ao tamanho da empresa e o seu mercado bastante disputado, existem grandes investimentos voltados para a captura, gestão e análise de dados, que servem de base para a tomada de decisão. Atualmente a RP *Cosmetics* possui 6 grandes bancos de dados dentre os seus principais.

Para obtenção das informações referentes à estrutura e funcionamento do banco de dados estudado, foram feitas entrevistas com o analista de TI responsável pela ferramenta, de modo que o máximo de informações e explicações fossem obtidas. Durante as entrevistas, o funcionamento técnico da ferramenta era exemplificado pelo analista que demonstrava as

etapas explanadas. Grande parte delas foram acompanhadas do suporte de TI externo que auxiliava o analista nas atividades de interface da área comercial com TI.

A interface utilizada para a visualização e manipulação dos dados é a *OLAP*, que permite análises *ad hoc*, ou seja, consultas do que se deseja sem ter a visão do que há por trás conectando os dados. Essa tecnologia se conecta e trabalha juntamente com o *Data Warehouse (DWH)* com o objetivo de dar visibilidade com rapidez aos dados que são armazenados nele.

Todos os bancos de dados são acessados pelos usuários através de um *template* no *Microsoft Excel*, com a apresentação de dimensões e métricas dispostas em uma tabela dinâmica. Sua estrutura é feita em *SQL Server* e o integrador utilizado é o *Power Center*.

Sua estrutura conta com 3 áreas de armazenamento nas quais os dados transitam até serem disponibilizados para os usuários: *Staging (STG)*, *Data Warehouse (DWH)* e *Datamart (DMT)*. O que permite esse trânsito de dados pelo banco de dados é o processo *ETL*.

O *ETL* (Extrair, Transformar e Carregar - *Load*, em inglês) é um dos processos utilizados no *data warehouse*, que por sua vez é um dos métodos existentes para a integração através do armazenamento de dados de diferentes fontes que não são disponibilizados em tempo real. Sua principal função é auxiliar na leitura dos dados que são capturados no seu formato bruto. O primeiro passo desse processo é extrair os dados, logo, ao se conectar com as origens de informação, os dados são coletados, analisados e processados. Em seguida, os dados são transformados em um padrão pré-definido que atenda o seu uso futuro através de regras estabelecidas. Por fim, esses dados são carregados no seu banco de dados de armazenamento.

O *STG* é a área de armazenamento intermediário que faz parte do *ETL* e onde todos os dados carregados ficam alocados em tabelas temporárias e lidos em seguida pelas rotinas do cubo, sejam elas automáticas ou forçadas por algum detentor dessa função. Nesse momento, a estrutura do arquivo carregado se mantém a mesma em termos de organização de colunas e formatação.

Quando o dado migra para o *DWH*, as tabelas são alocadas em uma única tabela principal, que possível formato específico necessário para que a carga seja feita corretamente no banco de dados. Nesta tabela, só são armazenados os dados de relevância estratégica para a companhia. Eles se organizam como em contêineres, onde cada um deles armazena os dados por partições de um mês. O armazenamento nesse estágio é permanente, ou seja, os dados só podem ser excluídos ou manipulados no banco de dados caso o analista de TI interfira manualmente.



Após a integração dos dados no *DWH*, eles estão prontos para serem utilizados no *DMT*. Este possui estrutura similar ao *DWH* e é onde são atualizados os cálculos e a visualização dos dados do dia anterior no cubo. O *workflow* desse último passo é executado todos os dias durante a madrugada. Assim, além de carregar os dados de *sell out*<sup>1</sup> e estoque, por exemplo, o *DMT* também é responsável por atualizar os cálculos de positivação<sup>2</sup> dos pontos de venda com base nas informações de *sell out* recebidas no dia anterior. Como esses cálculos são gravados em uma tabela do banco de dados (*DMT*), a utilização do cubo pelos usuários se torna mais otimizada. Em seguida, como último passo, é feita a carga desses dados no cubo *OLAP* para que então possam ser utilizados no banco de dados.

Existe uma outra maneira de fazer cálculos no banco de dados que é através da linguagem *MDX*, que desempenha os cálculos durante a visualização, a partir das métricas e dimensões selecionadas pelo usuário no *template* da tabela dinâmica. Por se tratar de um cálculo executado no momento da seleção dos filtros, seu desempenho é comprometido, tornando a manipulação mais lenta. Atualmente a definição sobre quais informações serão geradas no *DMT* e quais serão executados em *MDX* é determinada a partir da frequência de uso e importância para os usuários, já que a aplicação do cálculo no *DMT* exige mais horas dedicadas de uma equipe de TI para desenvolvimento.

A figura 20 é uma representação de como o dado transita entre as áreas de armazenamento e como o ETL atua nesse processo.



Figura 20: Caminho percorrido pelo dado até disponibilização.

Fonte: Elaboração própria.

<sup>1</sup> *Sell out* são as vendas feitas diretamente ao consumidor final do produto.

<sup>2</sup> A positivação de um ponto de venda (PDV) é um retorno binário que sinaliza se aquele PDV realizou algum *sell out* no dia informado de determinado produto.

Neste estudo de caso, o foco principal será dado ao cubo de *sell out*, com certas explicações sobre o cubo que detém as informações de *sell in*<sup>3</sup>, já que existe uma conectividade e troca de dados entre eles.

O cubo de *sell out* é principalmente utilizado pelas áreas de Comercial, *Trade Marketing* e Logística. Os dados disponibilizados nele são utilizados para geração de diferentes relatórios da divisão e imprescindíveis para tomada de decisão.

A governança dos seus dados é de responsabilidade da área de inteligência comercial, que possui interface direta com a área de TI, para a sua manutenção e melhoria contínua. A área desenvolve diferentes projetos junto ao analista de TI, em prol da constante tentativa de tornar os dados disponibilizados no banco de dados mais corretos e confiáveis.

O banco de dados de *sell out* foi desenvolvido com o objetivo de ser uma ferramenta integradora de dados. Os dados são originários de 7 diferentes origens de informação: banco de dados de *sell in*, SAP, fichas manuais e 4 diferentes provedores de informações. Todas essas origens se conversam em algum nível dentre as 3 dimensões utilizadas no cubo: produto, loja e data.

#### 4.2.1. Origens de Informação

O banco de dados de *sell in* possui informações que são geradas na própria empresa, sem origem de informação externa. Ele tem conexão direta com o *ERP SAP*, software de gestão empresarial, e o *Salesforce, CRM* aplicado a vendas, que juntos provêm todas as informações que dizem respeito aos produtos, cadastro de clientes e *sell in* feito para eles.

Assim, o banco de dados de *sell in* é um importante provedor de informações para o banco de dados de *sell out*, pois ao torná-los conectados e cruzar as informações de *sell in* e *sell out* em um mesmo cubo, é possível traçar estratégias a partir do entendimento de estoque e desempenho de cada cliente à nível mês/produto. Também é responsável por passar todas as informações de segmentação de clientes entre canais, circuitos e regionais.

O SAP também possui conexão direta com o banco de dados de *sell out* com o objetivo de manter o cadastro de produtos sempre atualizado. Assim, sempre que é criado um novo item no SAP ou algum item é descontinuado, dentre outras possíveis alterações à nível de cadastro, essa informação é atualizada no cubo em uma rotina que ocorre semanalmente.

---

<sup>3</sup> *Sell in* são as vendas consideradas diretas, ou seja, vendas feitas da RP Cosmetics para outra empresa (cliente).

Com relação ao dado principal do cubo que é o *sell out*, existem 2 formas de carregá-los: Fichas manuais e cargas automática de provedores conectados.

As fichas são uma forma manual e limitada de se ter acesso as vendas daqueles clientes que ainda não possuem uma conexão direta com provedor de dados ou não possuem interesse em compartilhar esse tipo de informação com um provedor. Assim, esses dados de *sell out*, em unidades de produtos e as vezes em valor (R\$), que podem vir ou não acompanhados da informação de estoque final do mês, são compartilhados pelo cliente via e-mail através de planilhas em *Excel*.

Essas planilhas então são formatadas para atender o formato necessário de leitura do *STG* e algumas conferências são realizadas antes de esses dados serem carregados no banco de dados: Conferir se o código de barras são de produtos *RP Cosmetics*, conferir se as lojas informadas já possuem cadastro no banco de dados, e se não, criá-los, além de verificar se o valor de *sell out* informado está de acordo com o nível de venda histórico do cliente e se contempla todas as marcas as quais ele faz *sell in*. Após esse processo, essas planilhas são salvas em formato *.TXT* e então carregadas no *STG* através de uma rotina criada pela TI.

Quando essa planilha é lida, o resultado da análise retorna por e-mail sinalizando se houve algum código de barras não identificado. Caso isso aconteça, o arquivo precisa ser revisto, senão, basta rodar uma segunda rotina para que os dados sejam transportados do *STG* para o *DWH*, e então sigam o *workflow* padrão para estarem disponíveis no banco de dados de *sell out* no dia seguinte.

Para os casos em que os clientes possuem conexão com provedores de dados, a obtenção da informação é mais simples, já que não necessita de interferência humana constante. Esses provedores possuem rotinas automáticas de envio de dados diários para o banco de dados de *sell out* de modo a manter o dado sempre atualizado em D-2 (dois dias antes do dia vigente). Comparado com as fichas, essa forma permite um acompanhamento mais claro do desempenho do cliente, já que do primeiro modo o *sell out* só é lido no mês seguinte.

A companhia atua principalmente com 2 provedores de dados, que internamente são entendidos como 3, que são: Provedor X que fornece dados do canal direto em uma plataforma e dados do canal indireto em outra plataforma, provedor Y que fornece dados do canal direto e provedor Z que fornece dados de grande parte dos clientes do setor de farmácias. Recentemente o provedor Y se fundiu ao provedor X tornando-se uma única empresa e seu processo de fusão está em andamento, portanto ainda são tratados como diferentes dentro da companhia.

Para melhor entendimento das diferenças entre os provedores e a necessidade de haver conexão com mais de um, vê-se necessário tratar sobre o canal direto e indireto primeiramente. O canal direto é conhecido como os casos de vendas de grandes empresas comerciantes, como varejos, por exemplo. São aqueles clientes que fazem o *sell in* diretamente com a RP *Cosmetics* e o *sell out* diretamente com o consumidor final do produto.

Já o canal indireto é aquele em que pertencem os atacadistas e distribuidores. São os clientes que fazem *sell in* com a RP *Cosmetics* e realizam venda para uma outra empresa comerciante, geralmente de menor porte, que é conhecido como *sell through*. Nesses casos, o *sell out* é feito pelo comerciante que fez a compra no atacadista ou distribuidor para o consumidor final, que é um dado mais difícil de se ter acesso. Logo, quando se fala em dado do canal indireto, está-se falando de *sell through* e não *sell out*.

Na figura 21 há uma representação gráfica desses diferentes termos para exemplificação e maior clareza.

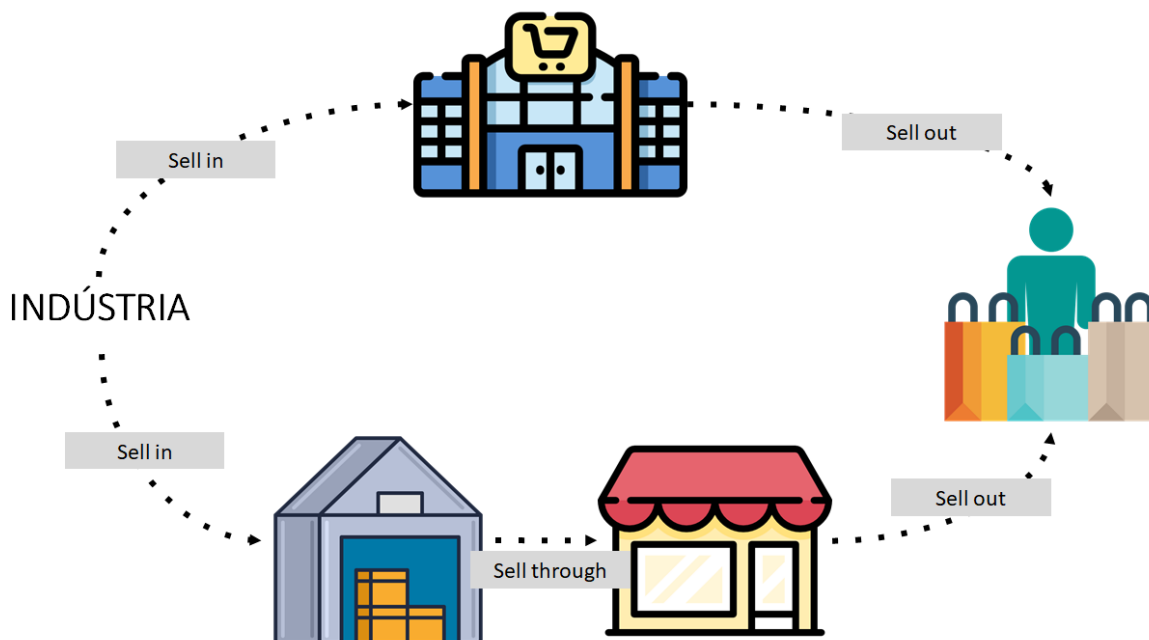


Figura 21: Exemplificação de sell in, sell out e sell through.

Fonte: Elaboração própria.

O provedor X é o único provedor que a divisão possui com informações do canal indireto, porém ele também fornece dados do canal direto, assim como o provedor Y. Esses 2 provedores fornecem informações de diferentes clientes diretos, porém no passado já se teve

dado de mesmo cliente nos dois provedores à título de comparação e entendimento da veracidade do dado.

Um dos principais motivos de se ter conexão com mais de um provedor para um mesmo canal é mitigar a dependência dos dados de apenas uma empresa, o que garante melhores condições de negociação e garantias na qualidade do dado, além de permitir a troca de provedores caso existam muitos problemas.

O provedor Z é importante por agregar dados de quase todos os clientes farmácia, o que é estratégico para algumas marcas da divisão. Além disso, eles possuem dados de clientes que outros provedores ainda não conseguiram conectar ou não possuem interesse, por serem de pequeno porte, permitindo aumentar o percentual de leitura de vendas.

Todos eles trazem dados de venda de *sell out* diários, em unidades e valor, e estoque em unidades. Eles possuem conexão com os clientes para conseguir a captura dos dados sempre de D-1 (um dia antes do dia vigente) para que no envio automático o banco de dados de *sell out* esteja sempre alimentado com dados de D-2 (dois dias antes do dia vigente). As vendas e estoque são informadas ao menor nível de granularidade de item e cliente, que são código de barras e CNPJ respectivamente.

Esses dados são informados em 4 diferentes arquivos: um arquivo de vendas, um arquivo de estoque, um arquivo de lojas e um arquivo de produtos. No primeiro é informado todo o *sell out* em unidades por data e código de barras (EAN), que referencia o produto, e nele também consta um código que referencia a loja daquela venda, podendo ser ou não o CNPJ. No arquivo de estoque a informação vem de forma similar e no arquivo de lojas constam todas as informações para preenchimento dessa dimensão, como endereço completo e CNPJ, por exemplo. Da mesma forma, o arquivo de produto fornece as demais descrições relacionadas à cada EAN, como nome, marca e categoria.

#### **4.3. Caracterização frente aos V's**

Com base no que foi detalhado anteriormente, o banco de dados da RP *Cosmetics* terá suas características avaliadas considerando os 5 V's do Big Data: Volume, Variedade, Velocidade, Valor e Veracidade. Para tal, foram feitas entrevistas não estruturadas com pessoas que são da área de inteligência comercial (2 pessoas) assim como alguns usuários que utilizam a informação para relatórios e análises (4 pessoas).

Estas foram conduzidas com uma breve explicação do objetivo do presente trabalho e do que compõe a definição de cada V frente ao *Big Data*. Em seguida, as entrevistas foram

executadas seguindo alguns tópicos que serviram de guia para obtenção dos pontos de vistas de cada um sobre as características e principais problemas observados. Uma exemplificação do guia dessas entrevistas encontra-se no apêndice A. Com isso, foram construídos os pontos de vista a seguir detalhados para construção da visão de problemas.

#### 4.3.1. Volume

Ao pensar em volume de dados, o banco de dados de *sell out* possui essa característica muito marcante entre todos os seus usuários devido às diferentes informações que são acessíveis através dele.

Como já mencionado, o cubo funciona com 3 diferentes dimensões que são loja, produto e data. Sendo assim, ao analisar o universo de dados, atualmente o banco de *sell out* possui dados de aproximadamente 5.000 EANs da companhia que podem ter números reportados através de unidades de *sell out*, valor (R\$) de *sell out*, unidades de *sell in*, unidades de estoque e positivação de lojas (que o venderam), todos eles diariamente.

Com base no histórico existente, conforme figura 22 abaixo, há registro de cerca de 200 clientes, os quais pode-se ler *sell out* unidade e valor desde 2013 e as demais métricas já mencionadas desde 2018, com granularidade diária. Além disso, essas métricas podem ser lidas no nível loja, tornando o dado ainda mais granular.

| Métrica                           | Histórico |
|-----------------------------------|-----------|
| <i>Sell out</i> (unidade e valor) | 2013      |
| <i>Sell in</i> unidades           | 2018      |
| Estoque                           | 2018      |
| Positivação                       | 2018      |

Figura 22: Quadro de período de disponibilização das métricas de *sell out*, *sell in*, estoque e positivação no banco de dados de *sell out*.

Fonte: Elaboração própria.

A título de dimensionamento, foi utilizada uma amostra com os dados de *sell out* unidades de janeiro de 2020. Houve reporte de 140 clientes que representam aproximadamente 110.000 lojas, com *sell out* de 1.834 EANs. Considerando que todos eles fizeram venda de pelo menos 70% dos EANs e que nenhum dos clientes faz venda no fim de semana (o que não é a realidade), com 23 dias úteis de vendas, temos um volume médio de 3.248.014.000 valores gerados mensalmente apenas para a métrica de *sell out* unidades.

O volume de dados é uma questão de grande influência quando são necessários ajustes na estrutura do cubo, já que quanto maior o volume maior o esforço para sua gestão. Uma alteração de cálculo em uma métrica de uma tabela específica exige que todo o histórico seja recalculado e isso afeta todo o desempenho da ferramenta.

Além disso, muitas visualizações de dados são criadas no cubo em um determinado momento e depois de um tempo elas deixam de ser úteis, porém não desconsideradas. Assim, conforme relatado pelos entrevistados, hoje é grande o número de pastas, dimensões e métricas existentes no banco de dados de *sell out* que não são utilizadas, o que torna a manipulação da ferramenta muitas vezes confusa, já que a existência de tantas opções de campos gera dúvidas em novos usuários.

Considerando esse ponto juntamente com o já mencionado problema de atualização de todas as métricas após ajustes, tem-se a visão de que existem campo sem uso que são atualizados, gerando volume desnecessário no sistema.

#### 4.3.2. Variedade

A variedade do dado pode ser considerada tanto com relação à sua fonte quanto ao tipo do dado, conforme mencionado em capítulo anterior. O tipo do dado lido pelo banco de dados de *sell out* através de sua rotina automática são todos estruturados, já que existe um padrão de formatação para sua leitura e integração. Esse padrão é definido pela TI responsável e é acordado com todos os provedores de dados antes da conexão.

De acordo com o mercado, o formato de leitura desse tipo de dado (vendas e estoque) possui um padrão já comum entre os provedores, porém como pode ocorrer formas diferentes de interpretação e organização, toda vez que um novo cliente vai se conectar à um provedor, assim como toda vez que vamos iniciar a contratação de um novo fornecedor de dados, há uma etapa inicial de definição de *layout*, para que quando finalizada a integração, não ocorram problemas relacionados a isso. Uma definição de *layout* seria ter na 1º coluna o reporte de data, na 2º, o EAN, e na 3º, o volume de *sell out* unidades, por exemplo.

Quando ocorre a necessidade de ler um dado que ainda não é encontrado no cubo, assim como dado de uma outra fonte, além da etapa de padronização de *layout*, há também a necessidade de padronizar o formato do dado, para que a integração ocorra fluidamente quando iniciada. Um exemplo simples de formatação de dado é a data, que pode ser reportada como 01/01/2020 ou 01-jan-20 ou 1º de janeiro de 2020, dentre muitas outras, o que dificulta a leitura.

Para a integração atual no banco de dados, a data precisa ser reportada em um formato padrão específico para que ela seja lida corretamente durante a integração.

Qualquer alteração de formato de dado, ou mudança de coluna no reporte de um dado, afeta a integração e impede a carga dos dados no cubo. Isso demonstra uma limitação da ferramenta atual que depende dessas definições asseguradas para que a leitura do dado ocorra e que leva a diversos questionamentos sobre a demora de novas conexões.

Com relação às fontes de dados, o cubo é alimentado diariamente por 6 fontes, além das fichas manuais que são carregadas mensalmente, conforme exemplificado na figura 23. Para o dado de *sell out*, as fontes são as fichas manuais e os provedores externos de dados.

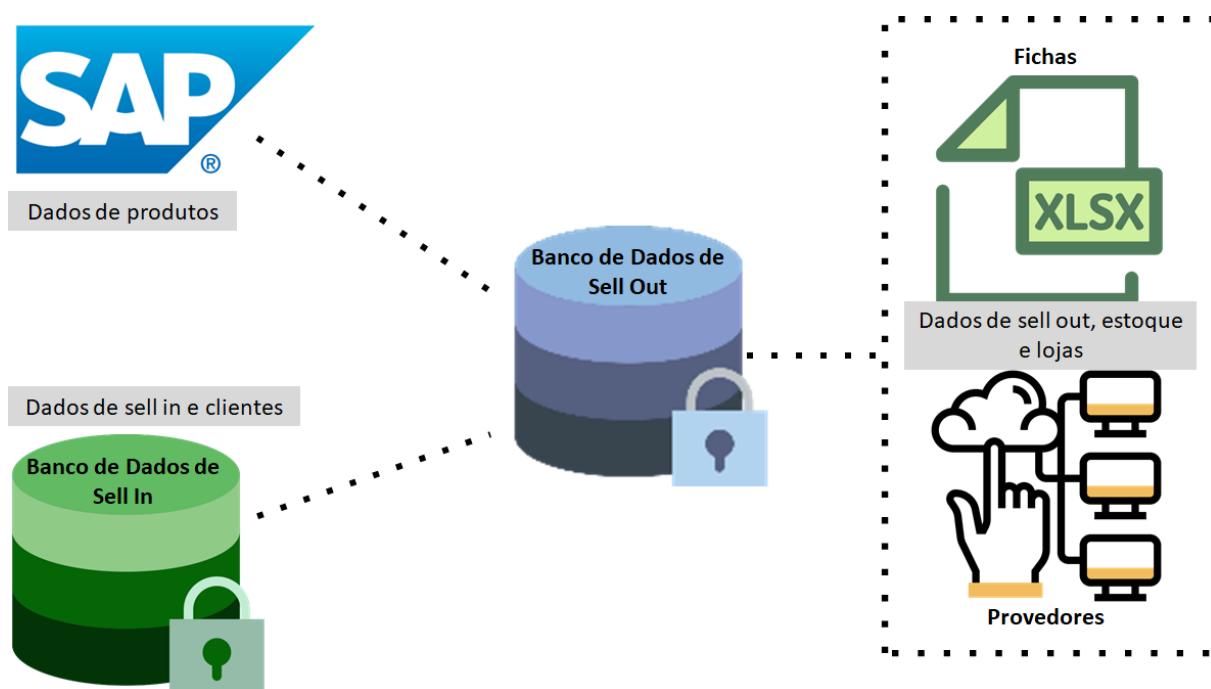


Figura 23: Origens de informação do banco de dados de sell out.

Fonte: Elaboração própria.

As fichas manuais são as mais passíveis de controle interno com relação à seguridade do número, porém também de erros humanos durante a carga. Todo mês, o *sell out* de cerca de 20 clientes são enviados via e-mail com leitura por EAN, e alguns deles por loja, com o compilado de venda no mês. Em alguns casos, o cliente também compartilha o estoque por EAN. Esses arquivos são tratados e formatados no layout exato definido para carga de fichas no cubo.

Caso ocorra algum erro de coluna ou qualquer erro relacionado à posição e/ou formatação incorreta, a carga é afetada, sendo carregada com erro ou não carregada. Atualmente



existe uma única ferramenta de verificação prévia que é a de EAN. Quando o arquivo da ficha é disponibilizado na pasta especificada e a rotina é rodada pelo responsável, há uma verificação de EAN que sinaliza por e-mail se algum dos códigos não foi identificado.

Sendo assim, qualquer outra informação que seja carregada erroneamente só será identificada no dia seguinte, quando for verificado que o dado não está sendo exibido ou está sendo exibido diferente do esperado. Por exemplo, conforme vivência relatada por um entrevistado, ao carregar a ficha de um cliente que contenha uma loja nova, ainda não cadastrada no sistema, a informação será carregada, porém o *link* entre o cliente e a loja se perderão na visualização. Problema similar ocorre caso o arquivo seja carregado com a data em formato diferente do padronizado, o que faz com que as fichas não sejam disponibilizadas para visualização. Não haver outros tipos de verificações de erros comuns gera atraso na correção, já que o erro sendo identificado no dia seguinte à carga, só será corrigido no dia posterior, atrasando a disponibilização do dado.

Para as cargas feitas através de provedores, a ocorrência de erros é menos frequente, já que há menos interferência humana. De todo modo, são notados problemas de integração gerados por qualquer alteração nos arquivos de carga fora do que foi especificado. Por exemplo, se um arquivo é disponibilizado com nomenclatura errada, ou com tamanho muito superior ao comum, a rotina automática do cubo não consegue processá-lo, de modo que naquele dia o dado não é disponibilizado. Esse tipo de erro ocorre com uma frequência maior do que a ideal e gera transtorno para todas as áreas, já que a não integração de um provedor em um dia de fechamento afeta toda a divisão.

O cubo atualmente atualiza todos os dias os dados de até 40 dias para trás, independentemente de o provedor enviar mais do que essa quantidade de dias. Assim, quando a rotina de carga se inicia, primeiro são apagados os dados do cubo desses 40 dias, para então carregar os valores disponibilizados. O ponto mais grave relacionado à essa característica da integração é quando o arquivo disponibilizado pelo provedor não pode ser lido ou está em branco, fazendo com que os dias apagados retornem nenhum dado de vendas. Assim, por pelo menos um dia, alguns clientes ficam sem informação disponível para os usuários.

O banco de dados de *sell out* possui diferentes ferramentas de validação interna para garantir que os dados dos provedores sejam exibidos de melhor forma para o usuário final e o mais verossímil possível com o que foi disponibilizado. Sendo assim, todo provedor disponibiliza diariamente um arquivo de *sell out*, um arquivo de estoque, um arquivo de lojas e um arquivo de produtos. Durante a integração, o cubo confirma se todos os produtos enviados

são da companhia para garantir a integração apenas de venda da RP *Cosmetics*, além de verificar toda a lista de lojas, para que aquelas que ainda não possuem cadastro no cubo sejam adicionadas.

No canal indireto é muito comum um cliente reportar venda para uma loja no arquivo de *sell out*, mas no arquivo de lojas, não passar as informações referentes a ela, como CNPJ e endereço, por exemplo. Com o objetivo de permitir que a leitura desse *sell out* seja possível no banco de dados, o cubo processa essa loja como “inexistente” e disponibiliza a sua venda para ser considerada nas análises, mesmo que não seja possível identificar a loja. Isso possui o lado positivo de se garantir que a venda total do cliente seja visível, porém atrapalha nas análises de apuração de pontos de vendas, já que são lojas não identificáveis.

O SAP e o banco de dados de *sell in* também são fontes de informações para o banco de dados de *sell out*, porém essas integrações são extremamente sólidas, por se tratar de 3 ferramentas internamente gerenciadas pela TI, o que permite baixa taxa de erros.

O SAP fornece toda a base de produtos ativos e inativos para o banco de dados e o banco de dados de *sell in*, além de fornecer os dados de *sell in* unidades por cliente, também passa informações de segmentações internas de cada cliente, como a região, por exemplo. Como ambos são fornecedores de dados para o cubo de *sell out*, eles são atualizados primeiro. Assim, quando ocorre atraso ou erro na atualização do banco de dados de *sell in*, a atualização do banco de dados de *sell out* também é comprometida.

#### 4.3.3. Velocidade

A velocidade, de acordo com a explicação teórica anterior deve ser analisada como aquela em que os dados estão fluindo. Desse modo, serão analisadas tanto a velocidade no que diz respeito a obtenção dos dados pelo cubo de origens externas quanto a velocidade com que os dados são visíveis no cubo para os usuários.

Conforme já mencionado, o cubo atualiza toda madrugada para que seus dados estejam disponíveis pela manhã, em torno das 8h, com dados de D-2. Alguns dias, quando ocorrem eventuais erros essa atualização atrasa podendo ser necessário aguardar até a tarde para sua finalização.

Isso ocorre porque não há pré-programação de uma rotina suficiente a ser seguida caso certos erros ocorram, sendo muito comum a interrupção do processo até que uma interferência humana possa ser feita para corrigir o ocorrido. Com isso, pode acontecer de um dia de trabalho iniciar sem os dados de vendas atualizados ou métricas recalculadas, o que gera atraso para

todos os usuários. Os motivos mais comuns são arquivos disponibilizados em formatos diferentes ou com *layout* desconfigurado ou também algum erro no servidor durante leitura e carga.

Durante o recebimento e leitura de dados, há diferentes mecanismos utilizados no cubo que permitem que atualizações de métricas sejam travadas durante a execução de algumas rotinas de leituras de dados. Ainda assim, como isso exige muito do sistema, é preferível sempre rodar toda e qualquer atualização em horários não comerciais. Quando é necessário forçar qualquer leitura fora dessa regra, de acordo com o relatado por todos os entrevistados, os impactos são perceptíveis, apresentando um desempenho comprometido ou muitas vezes impedido.

Além disso, considerando que os dados recebidos são sempre de D-2, há sempre uma defasagem de 2 dias de informação no mínimo. Considerando também a atualização de 40 dias para trás todos os dias, os dados têm um período de atualização de até 40 dias permitidos, o que faz com que a velocidade de recebimento dos dados fechados para análise e tomada de decisão seja lenta, ou seja, até 40 dias para frente ele pode ser atualizado.

No que diz respeito ao uso, como os dados são acessados pelos usuários via *Excel*, a variar pelo nível de granularidade que se deseja visualizar as consultas, o tempo de disponibilização da informação é bastante variável, podendo até se tornar inconveniente. Além disso, quanto maior o número de acessos ao cubo, maior o seu nível de lentidão, principalmente quando se considera que os períodos de geração de relatório são similares para grande parte das áreas (início e fim do mês).

Muito relacionado ao tópico de volume, entende-se que ambas características são capazes de se impactar mutuamente. Como já mencionado, caso necessária a visualização de métricas que possuem cálculo em *MDX* a velocidade de processamento dos dados é ainda mais comprometida. Por isso que a definição de uma métrica em *MDX* ou *DMT*, como já mencionado, é definida com base na sua importância e uso para a divisão.

Além disso, quando são feitas grandes alterações como criação de uma nova dimensão, alteração de uma propriedade ou mudança na estrutura de uma métrica é definida uma data no mês de menor impacto, no geral entre os dias 15 e 20, para que o cubo seja retirado do ar e todas as atualizações sejam feitas de modo a minimizar as chances de erros, tanto durante o processamento das mudanças quanto com relação aos dados acessados pelos usuários durante esse processo. Essa data é comunicada a todos com dias de antecedência visando impactar ao mínimo o trabalho das áreas usuárias.

#### 4.3.4. Valor

O valor é definido de diferentes formas a variar do autor, porém nesse estudo de caso ele será visto como o impacto que os dados presentes no banco de dados podem ser capazes de gerar valor para a divisão e conseqüentemente para a companhia.

Ao analisar os momentos levantados por ZHONG *et al.* (2006), entende-se que a empresa se encontra nos dois momentos, a variar do setor. Como já tratado, o volume de dados do comercial é gigante e atualmente muitos são os relatórios reportados com base nesses dados.

Com eles, diversas estratégias são desenvolvidas e acompanhadas, demonstrando o entendimento de valor que se tem por esses dados. Disso vem a importância e exigência da garantia de veracidade que será tratada a seguir.

Por outro lado, questiona-se se há o total aproveitamento da informação que pode ser obtida não apenas com os dados já disponíveis, como daquelas que poderiam ser extraídas caso eles fossem cruzados com demais fontes que hoje não são disponibilizados em um mesmo sistema.

A partir disso, hoje se vive uma realidade que abrange a necessidade de muitas bases de relacionamento que conectam a informação de um sistema com outro (planilhas de de-para), tornando todas as análises muito dispendiosas. Isso ocorre porque ainda hoje não existem chaves em comum que conectem todas as informações, como por exemplo, a garantia de CNPJ para as bases de lojas que conecte o *sell out* e estoque com a base de implementação de material de *merchandising*.

Isso também se deriva do fato de que ainda se trabalha com um grande número de bancos de dados diferentes e são poucos que possuem conexões entre si, como explicado que existe entre o banco de dados de *sell out* e o banco de dados de *sell in*. Desse modo, entende-se que muitos setores são detentores de diferentes dados e que se houvesse uma maior integração entre eles, mais valor poderia ser gerado.

#### 4.3.5. Veracidade

Com base no que já foi dito e considerando a importância dos dados do banco de dados de *sell out* para o negócio, a veracidade é um dos pontos críticos e chave nesse processo.

Atualmente as fontes de informação são muitas e aquelas mais difíceis de garantia de veracidade são as de origem externa. Assim, para garantia desses dados, alguns mecanismos de defesa são utilizados.

Existe todo um processo de *data quality* desenvolvido por analistas com o uso de ferramentas de *Power BI* com o objetivo de garantir que as informações obtidas e divulgadas façam sentido no todo. Assim, alguns exemplos de análise são os comparativos de *sell in* e *sell out* no nível EAN e quantidade para entender se as informações se conversam, ou verificação de dados no comparativo banco de dados de *sell out* e portal dos provedores para entender se estão reportando o mesmo número, dentre muitos outros.

Além disso, um dos mecanismos do sistema para garantia dos dados, como já mencionado, é a atualização diária de dados dos provedores externos para D-40. Isso se explica porque o reporte de um cliente pode ser modificado em período retroativo e são admissíveis no processo atual alterações de até 40 dias para trás no banco de dados de *sell out* de forma automática e diária. Esse período foi definido pela própria área de inteligência comercial alinhada com a estruturação de TI.

Nesse processo de atualização de dados retroativos é importante ressaltar que cada provedor atua de uma forma, tendo aqueles que enviam dado retroativo de 90 dias, porém apenas de *sell out*, não atualizando variação de estoque, e aqueles que enviam de 40 dias retroativos atualizando todas as informações. Dessa forma, a atualização de D-40 é uma trava criada no sistema para respeitar o limite aceitável comum de variação nos reportes.

Quando alguma inconsistência de dados de um cliente é divulgada ou identificada com um período maior do que esse torna-se necessário iniciar o reprocesso dos dados, que permite a atualização de qualquer período desejado, porém com interferência do analista de TI responsável. Assim, quando isso ocorre, uma solicitação é direcionada a ele para que, durante o período de processamento do cubo, esses dados sejam também corrigidos.

O impacto de uma alteração desse nível é enorme quando se trata de um cliente que é analisado estrategicamente e reportado em relatório, já que ocorre a alteração de todo o histórico. Logo, esse processo é visto de forma cirúrgica e evita-se ao máximo que ocorra, por isso as ferramentas de *data quality* seguem sendo implementadas e aprimoradas.

Isso também exige um contato muito próximo dos provedores e clientes que acaba por gerar uma rede, com o objetivo de garantir um dado cada vez mais verídico e que consequentemente seja capaz de gerar informações acuradas para o negócio, não só da empresa estudada, como também desses *stakeholders*.

Assim, entende-se que há diferentes aplicações do que foi definido anteriormente para cada V no banco de dados da RP *Cosmetics*. Muitas questões foram levantadas e identificadas, permitindo que sejam construídas análises de problemas a seguir.

#### 4.3.6. Síntese dos Vs

A partir do que foi explorado, e com o intuito de delinear as demais partes e focar nos principais ofensores identificados, uma síntese deles na visão de cada V foi elaborada. Essa síntese encontra-se na figura 24.

| <b>V</b>   | <b>Resumo dos Ofensores</b>  |
|------------|--|
| Volume     | Cubo com grande quantidade de dados devido diversas métricas e propriedades que não são mais utilizadas e atrapalham o uso.    |
| Variedade  | Muitas fontes de dados alimentam o cubo com dados de formato padrão.   |
| Velocidade | Cubo possui desempenho comprometido pelo software.   |
|            | Cubo tem defasagem de D-2 na disponibilização dos dados e possibilidade de alteração de até 40 dias.                           |
| Valor      | Dados de grande importância são mantidos, porém apresenta-se dificuldade de relacionamento entre eles e outras bases de dados. |
| Veracidade | Cubo se conecta com diferentes fontes de dados que atualizam de formas diferentes.   |

Figura 24: Quadro resumo dos V's frente ao caso de estudo.

Fonte: Elaboração Própria.

Nos próximos capítulos, esses ofensores serão desenvolvidos e equiparados com o que é encontrado atualmente no mercado em outras empresas que também se deparam com eles. Com isso, será construída uma visão mais sólida de como eles podem ser tratados ou evitados na análise dos problemas.

## 5. OBSERVAÇÃO DOS PROBLEMAS

Esta parte visa apresentar, frente as características do Big Data já apresentadas, quais são os problemas mais comuns enfrentados por grandes empresas do mercado de tecnologia. Em seguida, a partir desses problemas, são aprofundados aqueles já identificados na RP *Cosmetics* e então são apresentadas as características de contorno necessárias para a proposta de soluções.

### 5.1. Problemas de Big Data Comuns ao Mercado

Primeiramente, para a construção de um quadro de problemas sólido da RP *Cosmetics*, foi feita uma breve exploração para entendimento de como esses problemas são comumente vistos no mercado por outras companhias.

Os três *ERPs* pesquisados foram os da *Oracle*, da *SAP* e da *Totvs*, que foram escolhidos por serem os três mais populares *ERPs* do Brasil (MEIRELLES, 2020, p.14). Nas empresas de tecnologia foram procuradas as ferramentas *Google Cloud* e *Amazon Web Services* por estarem entre as maiores empresas de tecnologia do mundo e que possuem plataformas na nuvem para uso de *Big Data* (TRUCKER, 2020).

De acordo com *Oracle* (2019), existem 3 grandes problemas no uso da tecnologia do *Big Data*. O primeiro é inerente ao conceito do *Big Data*, o grande volume. É difícil encontrar formas de lidar e armazenar com eficiência dados muito grandes e com crescimento acelerado.

O segundo problema citado é a organização dos dados. Manter os dados limpos e organizados apesar de toda sua complexidade para conseguir análises realmente significativas demandam muito trabalho. “Cientistas de dados gastam de 50 a 80 por cento de seu tempo curando e preparando os dados” (ORACLE, 2019).

Como terceiro problema tem-se a evolução muito rápida da tecnologia, portanto manter-se atualizado e utilizando as melhores técnicas e ferramentas é um desafio contínuo para as organizações.

Já a *SAP*, na forma de conteúdo patrocinado pela empresa e de autoria da consultoria *ENTERPRISE MANAGEMENT ASSOCIATES (EMA)*, descreve oito problemas da análise de dados no contexto do *Big Data*.

- Caos da informação

Este item tem relação com a complexidade na vinculação de informações internas e externas a empresa, a partir do uso de várias origens de armazenamento, de

múltiplas plataformas, com contratos e *SLAs (Service Level Agreements)* de acesso diferentes.

- Inchaço da plataforma de dados

Este ponto conversa um pouco com o anterior. Há uma complexidade em se ter várias plataformas de gerenciamento de dados. Cada plataforma nova aumenta o custo e complexidade de forma exponencial. Neste tópico SANTAFERRARO (2020, p.4) dá alguns resultados de pesquisas feitas, onde “85% dos usuários finais têm de 2 a 8 tipos diferentes de plataformas de gerenciamento de dados” e “99% usariam plataformas de integração destes dados”.

- Entrega de *insights*

Este tópico se inicia com um questionamento em SANTAFERRARO (2020, p.5): “O *insight* produz uma ação que leva a criação de valor?”. O processo de análise de dados no contexto do *Big Data* demanda muitos recursos de engenharia e analistas para acontecer, além de tempo. Portanto as perguntas podem não ter respostas em tempo hábil para a tomada de decisão. Este ponto fica mais claro em empresas digitais onde não é possível esperar insights do analista para que se decida agir. As mudanças de processos devem ser mais dinâmicas.

- Privacidade e segurança de dados

A velocidade e volume que o *Big Data* requer dificulta a proteção dos dados. “Com enormes quantidades de dados as organizações precisam garantir estejam protegidos contra perda, roubo ou acesso inadequado” (SANTAFERRARO, 2020, p.6).

- Gestão de dados

Os dados passaram de uma tática da organização para fazer parte da estratégia de grande parte das organizações. “No entanto muitos ainda apresentam problemas básicos na gestão dos dados como consistência, certificação, uso ativo de dados mestre e avaliação de dados” (SANTAFERRARO, 2020, p.7). Estes problemas básicos de gestão de dados são grande empecilho para a transição de tático para estratégico.



- Inteligência artificial

A inteligência artificial entra como necessidade do aumento da eficiência no uso do *Big Data*. É necessário se analisar cada vez mais dados em menos tempo, além de necessidade de ações automáticas em tempo real em alguns casos.

- Automação de operações de dados

As organizações ainda possuem muitas operações manuais de dados, porém estas operações demandam muito tempo e recursos humanos para execução, o que reduz a capacidade de rápida aplicação dos dados. A tendência é a automação desses processos, a partir do uso de inteligência artificial e aprendizado de máquina, focado na “hipervelocidade e automação de processos de dados complexos e multidirecionais” (SANTAFERRARO, 2020, p.11).

Em relação a *Amazon*, existe o produto *Amazon Web Services (AWS)* que fornece algumas plataformas de serviços de gerenciamento de dados. No *White Paper* “Opções de análise de *big data* na *AWS*” publicado em janeiro de 2016 é possível extrair alguns problemas comuns no uso das plataformas do mercado de acordo com a *Amazon*.

O primeiro a ser citado é sobre a demanda na disponibilidade dos dados. Alguns dados são necessários em tempo real para uso rápido, outros podem ser agrupados em lotes e executados de forma diária. De acordo com a *Amazon Web Services* (2016, p.38) ao utilizar as mesmas ferramentas para ambos os conjuntos de dados citados anteriormente terá um serviço mais caro além de mais demora, talvez até inviabilizando os dados em tempo real. Portanto deve-se utilizar a ferramenta certa de acordo com a velocidade necessária de cada conjunto de dados.

Entretanto, outro problema comum é fragmentar demais os dados em vários sistemas ao invés de utilizar um grande. De acordo com a *AWS* (2016, p. 39) essa abordagem traz maior custo e dificuldade de gerenciamento. Neste último ponto converge também com a *SAP* quando esta cita o inchaço na plataforma de dados como um problema comum.

Não foram encontrados na *Totvs* e no *Google Cloud Platform* material sobre o tema explorado.

## 5.2. Aprofundamento dos Problemas da RP *Cosmetics*

Com base no que foi pesquisado e apresentado anteriormente, e considerando todos os relatos feitos pelos 6 respondentes das entrevistas, esta sessão aborda no detalhe os problemas enfrentados pelos usuários do banco de dados decorrentes das características apresentadas na parte 4.

Conforme o que já foi desenvolvido, podemos quebrar os problemas percebidos em 3 envolvidos: a área de inteligência comercial e TI como responsável pelo bom funcionamento do banco de dados, as áreas usuárias e as áreas (e usuários) estratégicos dos dados. Na figura 25 encontra-se um resumo do estado atual de cada um desses envolvidos, como eles se relacionam, o que demonstra a grande interligação e dependência entre eles para que o banco de dados seja bem sucedido como um todo. Cada estado é desenvolvido e detalhado a seguir.

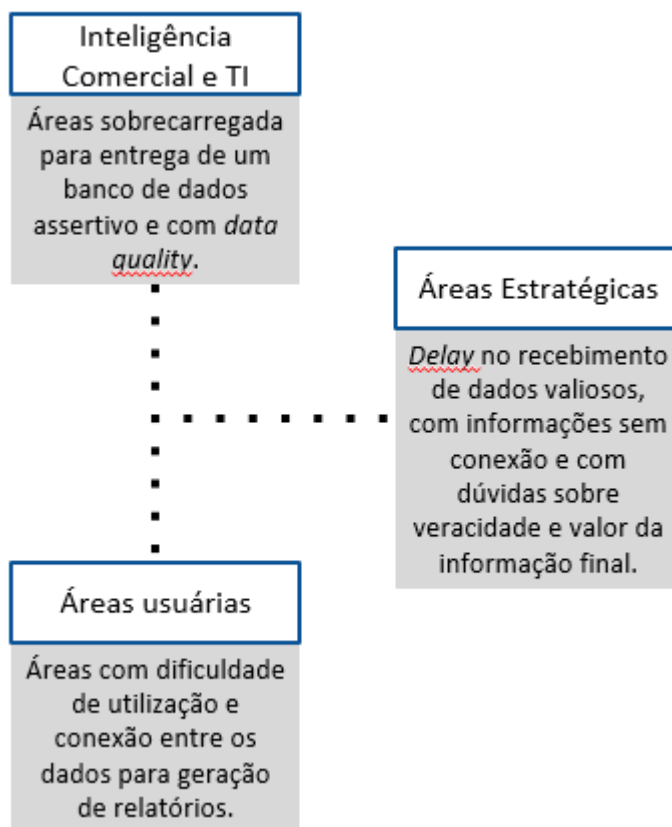


Figura 25: Estado atual dos envolvidos nos problemas do banco de dados.

Fonte: Elaboração Própria.

Assim como identificado pela *Oracle*, o banco de dados estudado neste trabalho também enfrenta problemas derivados da dificuldade em se manter dados organizados considerando volume e velocidade. Uma das maiores dificuldades é a manipulação dos dados por qualquer

usuário. Existe uma grande quantidade de métricas, propriedades e campos que não são utilizados e/ou atualizados e que mesmo assim não foram eliminados do sistema.

Por isso, há a necessidade de diferentes treinamentos, o que ainda assim não é capaz de tornar as pessoas seguras para criarem análises sozinhas. Isso se dá porque a função e uso do banco de dados de *sell out* não estão 100% definidos e desenhados, logo há muitas exceções à regra e também há constantes mudanças no funcionamento de métricas e filtros.

A consequência disso é que a equipe responsável pelo banco de dados (área de inteligência comercial) se torna sobrecarregada por receber muita demanda para dar explicações ou montar *templates* personalizados para a necessidade da área. Assim, o usuário só precisa atualizar toda vez que necessitar de um relatório recente, não tendo o *know-how* para montar novas visões ou fazer qualquer alteração sozinho. Nesses momentos, recorrem novamente à equipe para ajudá-los nas alterações.

Devido às constantes mudanças, não há a manutenção de um guia atualizado sobre o uso do sistema. As alterações frequentes fizeram com que a atualização regular se tornasse onerosa e fosse deixada de lado, tornando a consequência mencionada ainda mais persistente. Ainda, com o aumento da necessidade de informação cada vez mais intensa, a tendência é que a demanda pelo uso do banco de dados seja reforçada, agravando cada vez mais a dependência dos demais usuários à equipe de planejamento.

A manipulação dos dados também é uma questão no quesito velocidade de retorno, já que a elaboração de certos relatórios é comprometida e incômoda devido à lentidão do sistema em retornar os dados solicitados. Pelo fato de os dados serem visualizados através de uma tabela dinâmica em *Microsoft Excel*, encontra-se a dificuldade toda vez que o nível de granularidade gera um volume que ultrapassa o número máximo de linhas permitidas pelo *software*.

Na elaboração de relatórios complexos, as áreas responsáveis precisam recorrer a outros programas que auxiliem na extração de dados do banco de dados. Por muito tempo, o *Access* foi utilizado para essa função, sendo atualmente substituído pelo *Power BI*. Com a elaboração de uma *query*, a extração se torna automática e conseqüentemente menos dispendiosa para o usuário que pode dedicar seu tempo ao desenvolvimento de análises. Isso demonstra a questão de que, além de existir um sistema de difícil uso, tem-se a necessidade de agregar conhecimento em outras ferramentas para que seu trabalho possa ser desenvolvido.

No mais, pelo fato de o *Excel* não ser um software feito para esse uso, principalmente considerando o aumento exponencial do volume de dados, ele se torna um dos pontos de dor do banco de dados.

Para a velocidade de recebimento do dado, temos o oposto do que é citado pela *Amazon* como necessário, que é a disponibilidade instantânea do dado para a tomada de decisão. Considerando o longo tempo para recebimento e atualização do dado de *sell out* e estoque, o desenvolvimento de *insight* e conseqüentemente a tomada de decisão ficam prejudicados, fazendo com que as ações sejam tomadas com certo *delay* de tempo.

Considerando a informação existente no banco de dados de *sell out* hoje e o que poderia existir caso o banco de dados fosse capaz de ler dados semi ou não estruturados, temos a visibilidade de um grande *gap* de informações. Atualmente, por exemplo, a companhia tem acesso aos dados de um fornecedor que, dentre diversas informações, compartilha dados através de imagens. Esse não é um fornecedor que hoje possui um *layout* combinado e estruturado no formato que o banco de dados é capaz de ler e por isso suas informações são manipuladas manualmente e de forma separada das demais que existem no banco de dados. Assim, nota-se uma subutilização de informações que seriam mais ricas se combinadas com outras de forma mais prática.

Um outro exemplo são as pesquisas feitas pelo time de expositores que trabalha diariamente nos pontos de venda e que possuem dentre suas diferentes atividades a necessidade de responder essas pesquisas que falam sobre como está a execução das marcas nas lojas visitadas. Ao final dessas pesquisas, o colaborador insere uma imagem que demonstra o que foi respondido. Essas imagens são uma forma de verificarmos se as respostas estão ou não corretas, porém por não haver hoje um sistema que seja capaz de fazer a leitura das imagens, faz-se uso apenas das respostas obtidas na pesquisa.

Além disso, é válido destacar que todas os dados obtidos através dessas pesquisas são inseridos em um banco de dados diferente, criado apenas para essa função. Esses dados são de suma importância para entendimento de performance das marcas no mercado e servem de indicadores que junto com as informações de *sell out* e estoque, por exemplo, são capazes de direcionar melhor a tomada de decisão. Entretanto, por serem dados alocados em bancos de dados diferentes, eles possuem estruturas e fontes diferentes e conseqüentemente não conversam entre si.

As 3 dimensões necessárias no banco de dados de *sell out* (produto, loja e data) são diferentes para esse outro banco de dados. Nesse, o EAN não é exigido como chave dos produtos, o CNPJ não é exigido como chave das lojas, o que faz com que não exista uma chave capaz de conectar os dados disponíveis em ambos os bancos de dados sem validação prévia.

Assim, é reforçado o problema também identificado pela SAP e a *Amazon*, onde uma companhia é munida de muitos dados que garantem informação, mas que por não se comunicarem entre si, em diferentes plataformas e com diferentes chaves, a geração de conhecimento fica comprometida devido esses *gaps* de conectividade.

Isso impacta diretamente no problema identificado para a análise de valor dos dados. Por existirem diferentes fontes de informações, com diferentes dados disponíveis, porém sem conexão entre eles, muito valor não é gerado. Essa constatação vem do fato de que, caso houvesse essa conexão, seria possível gerar análises mais completas, com indicadores que fossem capazes de explicar uns aos outros, gerando decisões corretas e de forma mais rápida. Esse é exatamente o problema citado pela SAP ao confrontar o tempo de tomada de decisão versus os *insights* gerados a partir de dados.

A maior dificuldade em trabalhar isso hoje é o fato de que o volume histórico existente ser muito grande e novos dados chegarem todos os dias em também grande volume. Assim, a correção de informações se torna muito trabalhosa, além de trazer os questionamentos sobre o impacto que uma alteração pode trazer no histórico e na comunicação entre o dado corrigido e o dado incorreto do histórico.

Além disso, existe a questão de quantidade de recursos disponíveis para fazer isso funcionar, como também tratado no último tópico da SAP. Atualmente existe um responsável no TI que dá todo o suporte para a área de inteligência comercial com o banco de dados. Esse mesmo analista responde pelos outros bancos de dados da companhia, sendo assim, a concorrência de prioridades é muito acirrada, criando mais uma barreira de dificuldade. Apesar de haver uma estrutura de gerentes que cuida do banco de dados na América e demais países, as intervenções e manutenções diárias ficam a cargo desse único analista.

Com isso, existem todas as restrições que tornam a inserção e uso dos dados muito fechada. Assim, se o novo fornecedor não respeita o *layout* necessário para carga do banco de dados, esse dado não será disponibilizado no sistema. Do mesmo modo, se o dado de outro fornecedor não atende à todas as dimensões de produto, loja e data, esse dado não será disponibilizado no sistema. Todas essas restrições criam uma limitação de análise, dependência de tabelas de de-para para correspondência e acabam por gerar até mesmo a geração de informações diferentes, já que cada usuário terá sua forma e suas fontes para fazer uma análise.

Por fim, a veracidade dos dados é uma das questões mais presentes e importantes para um banco de dados relatada pelos entrevistados. Por existirem diferentes fontes de dados que

não são geradas pelos sistemas internos da companhia, a dificuldade de garantir a veracidade é muito grande.

Muito são os trabalhos desenvolvidos com os provedores para a garantia do dado. Validações são feitas mensalmente, os dados são verificados diariamente e o contato com esses provedores é constante para que seja identificado qualquer problema sistêmico, seja na companhia, no provedor ou no cliente, que atrapalhe a divulgação de informações.

No que diz respeito à veracidade que está nas mãos da empresa, as conferências diárias servem para entender se algum processamento do banco de dados apresentou erro e impediu que dados não fossem carregados ou fossem carregados incorretamente. O próprio sistema tem alguns mecanismos que sinalizam quando uma carga não foi realizada corretamente.

Além disso, as métricas calculadas no *MDT* precisam ser constantemente verificadas, porque pode ocorrer alguma alteração no dado ou na forma que ele é reportado e que não foi mapeada na construção da métrica, fazendo com que ela não desempenhe conforme o esperado. Um simples exemplo disso é o caso de um cliente não ter venda em determinado dia, e no passado o provedor enviar isso com um zero no campo e depois passar a enviar isso com o campo em branco. Caso a métrica não esteja considerando esse cenário, erros nos cálculos podem ocorrer e afetar todo o resultado. Por isso, elas também são validadas constantemente.

Esse último ponto se alinha muito com a questão da evolução tecnológica identificada pela *Oracle* e também mencionada pela *SAP* quando cita que a não automação dos processos com uma dificuldade para o atingimento do objetivo “mais dados em menos tempo”. Processos automáticos são necessários para que o tempo de resposta seja reduzido e a geração de dados não seja comprometida.

Assim, em resumo, foi gerada a figura 26 com um quadro de problemas que reúne todos os pontos descritos bem como a sua categoria principal.

| <b>Categorização</b>     | <b>Problema</b>  |
|--------------------------|--|
| Organizacional           | Sobrecarga na montagem de templates para áreas usuárias do banco de dados.                                   |
|                          | Desgaste diário para validação dos dados de origens externas.  |
| Pessoas                  | Necessidade de treinamentos constantes sobre o uso da ferramenta.  |
|                          | Pouco recurso de TI focado no trabalho do banco de dados.  |
| Processos                | Guia de utilização desatualizado.  |
|                          | Delay entre a informação e tomada de decisão devido longo tempo de disponibilização e atualização dos dados. |
| Tecnologia da Informação | Lentidão para utilização do banco de dados em <i>Excel</i> para análises mais detalhadas.                    |
|                          | Utilização de <i>softwares</i> auxiliares para viabilizar o uso do banco de dados.                           |
|                          | Subutilização de dados devido estrutura de leitura do banco de dados.  |
|                          | Ausência de chaves de dados comuns para relacionamento de bases de banco de dados diferentes.                |
|                          | Dependência de arquivos de de-para.  |
|                          | Alta dependência de ajustes manuais por parte da TI.   |
|                          | Dificuldade para conexão de novas fontes de dados no cubo.   |

Figura 26: Quadro de problemas com categorização.

Fonte: Elaboração própria.

Para determinar as categorias utilizadas, pensou-se nos principais departamentos organizacionais que possuem interface com o caso estudado, que foram TI, processos e pessoas. Além destes serem os que englobam os principais pontos de dores relatados, foi considerada uma quinta categoria que completa os relatos que dizem respeito aos problemas de origem organizacional. O objetivo deste agrupamento foi viabilizar o desenvolvimento de soluções que atendam ao máximo todo o grupo de problemas de uma categoria de forma conjunta e integrada.

Assim, a partir desse quadro é possível trabalhar as diferentes categorias de modo a visualizar as melhores soluções capazes de atingir as causas principais e conseqüentemente eliminar esses problemas. A seguir, para auxiliar na identificação de soluções, as características de contorno da RP *Cosmetics* serão apresentadas e analisadas.

### **5.3. Características de Contorno**

Com base nos problemas já apresentados, um último passo antes de analisar as soluções ideais é a definição das características de contorno da RP *Cosmetics* que servirão de guia para

o desenho de soluções. Para construção delas, utilizou-se o material obtido nas entrevistas anteriores como suporte para a definição dos principais pontos de dor e conseqüentemente quais seriam as características de contorno ideais.

Para isso, primeiramente foram definidas as premissas que precisam ser consideradas para a determinação das características de contorno. Todas elas foram levantadas com a área de inteligência comercial que tem experiência com o banco de dados de *sell out* e todos os processos em que ele está envolvido, desde funcionamento, usabilidade e fornecedores. Estas são:

- Qualquer mudança no banco de dados proposta pela área de inteligência comercial deve ser passada para a TI avaliar a viabilidade e possibilidade de seguir adiante;
- Por se tratar de uma empresa multinacional que presa pela sua segurança e conectividade tecnológica, qualquer grande mudança em estrutura de banco de dados deve ser aprovada e difundida pelo tomador de decisão global da área de TI;
- A contratação de novos fornecedores e/ou alterações de sistemas demandarão longo período de estudo e de análise do custo vs. ganho financeiro e operacional que as mudanças trarão;
- A contratação de fornecedores passa por um longo processo de análise de compras e jurídico para além de entendimento do serviço oferecido e valor cobrado, se a empresa está em linha com as políticas sociais e ambientais às quais a RP *Cosmetics* segue;
- Atuais fornecedores de dados possuem contratos com prazos fixados e qualquer cancelamento deve ser previamente mapeado e comunicado à área de compras e jurídico para que não existe nenhuma pendência com o fornecedor;
- Os dados do banco de dados de *sell out* são utilizados por toda a companhia e qualquer alteração em estrutura de dados deve ser previamente comunicada e muito bem planejada para que não ocorram longos períodos sem a disponibilização do dado, bem como longos períodos de instabilidade;
- A manutenção dos dados históricos, bem como mantê-los utilizáveis é de grande importância, sendo inviável perdê-los;
- A manipulação dos dados deve ser acessível a todos da empresa, com ferramentas de fácil uso, inclusive em aparelhos móveis através de aplicativos,



já que esse é o uso mais comum dos cargos de campo (comerciais) que são grande parte dos usuários dos dados;

- A RP *Cosmetics* é uma empresa que preza pelos treinamentos e esse tipo de atividade é estimulada, porém é necessário garantir a qualidade e domínio do conteúdo, sendo possível qualquer colaborador oferecer o treinamento de algo do seu domínio para toda a companhia;
- A definição de escopo de funções e a manutenção dele a fim de futuras avaliações e evolução do colaborador no cargo é importante para a companhia que enxerga o crescimento do funcionário na empresa como algo interessante para a retenção de talentos para futuros cargos estratégicos.

A partir das premissas necessárias apresentadas e considerando todos os problemas já levantados e categorizados, chegou-se as seguintes características de contorno.

- Organização da rotina da área de inteligência comercial;
- Treinamentos frequentes da ferramenta;
- Facilidade de acesso ao manual de uso da ferramenta;
- Garantia de atendimento às demandas técnicas do banco de dados;
- Garantia de dados confiáveis e atualizados;
- Facilidade no acesso aos dados de *sell out*;
- Facilidade no relacionamento de dados de diferentes bases;
- Flexibilidade na leitura de novos dados e fornecedores;
- Flexibilidade na variação mapeada dos dados.

Na figura 27 é apresentada a relação entre as premissas e quais características de contorno devem seguir para cada uma delas.

| Premissas   | Características de Contorno                                     |
|---|---|
| Qualquer mudança no banco de dados proposta pela área de inteligência comercial deve ser passada para a TI avaliar a viabilidade e possibilidade de seguir adiante.   | Garantia de atendimento às demandas técnicas do banco de dados. |
| Por se tratar de uma empresa multinacional que presa pela sua segurança e conectividade tecnológica, qualquer grande mudança em estrutura de banco de dados deve ser aprovada e difundida pelo tomador de decisão global da área de TI.   |   |
| A contratação de novos fornecedores e/ou alterações de sistemas demandarão longo período de estudo e de análise do custo vs. ganho financeiro e operacional que as mudanças trarão.   | Garantia de dados confiáveis e atualizados.                     |
| A contratação de fornecedores passa por um longo processo de análise de compras e jurídico para além de entendimento do serviço oferecido e valor cobrado, se a empresa está em linha com as políticas sociais e ambientais às quais a RP Cosmetics segue.                                  | Facilidade no acesso aos dados de sell out.                     |
| Atuais fornecedores de dados possuem contratos com prazos fixados e qualquer cancelamento deve ser previamente mapeado e comunicado à área de compras e jurídico para que não existe nenhuma pendência com o fornecedor.  | Facilidade no relacionamento de dados de diferentes bases.      |
| Os dados do banco de dados de sell out são utilizados por toda a companhia e qualquer alteração em estrutura de dados deve ser previamente comunicada e muito bem planejada para que não ocorram longos períodos sem a disponibilização do dado, bem como longos períodos de instabilidade. | Flexibilidade na leitura de novos dados e fornecedores.         |
| A manutenção dos dados históricos, bem como mantê-los utilizáveis é de grande importância, sendo inviável perdê-los.  | Flexibilidade na variação mapeada dos dados.                    |
| A manipulação dos dados deve ser acessível a todos da empresa, com ferramentas de fácil uso, inclusive em aparelhos móveis através de aplicativos, já que esse é o uso mais comum dos cargos de campo (comerciais) que são grande parte dos usuários dos dados.                             | Facilidade de acesso ao manual de uso da ferramenta.            |
| A RP Cosmetics é uma empresa que preza pelos treinamentos e esse tipo de atividade é estimulada, porém é necessário garantir a qualidade e domínio do conteúdo, sendo possível qualquer colaborador oferecer o treinamento de algo do seu domínio para toda a companhia.                    | Treinamentos frequentes da ferramenta.                          |
| A definição de escopo de funções e a manutenção dele a fim de futuras avaliações e evolução do colaborador no cargo é importante para a companhia que enxerga o crescimento do funcionário na empresa como algo interessante para a retenção de talentos para futuros cargos estratégicos.  | Organização da rotina da área de inteligência comercial.        |

Figura 27: Quadro de premissas e características de contorno.

Fonte: Elaboração própria.

Com essas características de contorno que determinam onde se deseja chegar com as propostas de soluções é possível construir uma análise de *fit* entre elas visando o entendimento dos conflitos e empecilhos que podem existir durante a proposição de soluções.

## 6. ANÁLISE DE FIT

O objetivo de construir uma análise de *fit* das condições de contorno, ou seja, entender quais condições de contorno são controversas ou independentes em seus alvos, é possível delinear melhor o caminho para a proposta de soluções. Segundo SMITH, REECE (1998), ao combinar a análise de *fit* dos elementos operacionais com a estratégia, nota-se uma importância maior do que apenas desenhar a estratégia.

Assim, as 9 características de contorno foram avaliadas entre elas, formando uma matriz de resultado, conforme figura 28.

A característica de contorno “facilidade de relacionamento dados de diferentes bases”, além de ser uma das mais importantes para os usuários do banco de dados em questão, foi considerada aquela que não apresenta conflito de proposta de solução com nenhuma outra. Sendo assim, ela possui *fit* com todas as 8 demais características de contorno.

A característica de contorno de “treinamentos frequentes da ferramenta” possui *fit* ajustável com a “organização da rotina da área de inteligência comercial”, pois será necessário acrescentar mais uma atividade frequente a rotina da área que no momento já passa por dificuldades, porém há formas de contornar a sobrecarga remanejando funções. Ademais, ela possui *fit* com as demais 7 características de contorno, não havendo impacto entre elas.

A “facilidade de acesso ao manual de uso da ferramenta” apresenta *fit* ajustável com a “organização da rotina da área de inteligência comercial” pelo mesmo motivo da característica de contorno anterior. Com relação às demais, ela possui *fit* com todas.

A característica de contorno de “garantia de dados confiáveis e atualizados” possui *fit* ajustável com a “organização da rotina da área de inteligência comercial”, por ser uma atividade que já demanda demais no dia-a-dia, porém ambas podem ser ajustadas com mudanças de processos que garantam o funcionamento da validação dos dados. Também possui *fit* ajustável com a “garantia de atendimento às demandas técnicas do banco de dados” por ser uma atividade que também demanda tempo da equipe de TI a cada eventual divergência a ser reprocessada, sobrecarregando o analista e demandando maior organização das demandas ou redistribuição de tarefas na área.

Ela também possui *fit* ajustável com a “flexibilidade na leitura de novos dados e fornecedores” por demandar novos processos de validação que garantam a seguridade do dado dentro dessa flexibilização e com a “flexibilidade na variação mapeada dos dados”, pelo mesmo motivo – é preciso assegurar a garantia do dado. Há *fit* com as demais características de contorno.

A “facilidade no acesso aos dados de *sell out*” possui *fit* ajustável apenas com a “garantia de atendimento às demandas técnicas do banco de dados”, já que a melhoria de software/aplicação do banco de dados demandaria um trabalho extra da equipe de TI resultando em uma sobrecarga não comum, porém temporária. Com uma divisão de funções, contratação de terceiros ou outra solução factível, seria possível realizar ambas, entretanto. Com as demais, essa característica de contorno possui *fit*, não apresentando conflitos.

A característica de contorno da “garantia de atendimento às demandas técnicas do banco de dados”, além dos *fits* ajustáveis já mencionados, não possui *fit* com a “flexibilidade na leitura de novos dados e fornecedores” e a “flexibilidade na variação mapeada dos dados”. Isso se dá porque, considerando o contexto atual da situação de uso que o banco de dados se encontra, e a sobrecarga da equipe de TI e da equipe de inteligência comercial, a grande mudança na estrutura do banco de dados para tornar possíveis as flexibilizações mencionadas afetam diretamente a rotina e o uso dos dados da forma atual. Essas 2 características de contorno demandam soluções mais robustas, burocráticas, de longo prazo e alto investimento, aspectos que não vão de encontro com a realidade buscada no curto prazo.

A “flexibilidade na leitura de novos dados e fornecedores” e “flexibilidade na variação mapeada dos dados”, além dos já mencionados *fits* ajustáveis e não possui *fit*, apresentam *fit* com as demais 6 características de contorno.

| <b>Análise de fit</b>  | <b>Organização da rotina do time de inteligência comercial</b> | <b>Treinamentos frequentes da ferramenta</b> | <b>Facilidade de acesso ao manual de uso da ferramenta</b> | <b>Garantia de atendimento às demandas técnicas do banco de dados</b> | <b>Garantia de dados confiáveis e atualizados</b> | <b>Facilidade no acesso aos dados de <i>sell out</i></b> | <b>Facilidade no relacionamento de dados de diferentes bases</b> | <b>Flexibilidade na leitura de novos dados e fornecedores</b> | <b>Flexibilidade na variação mapeada dos dados</b> |
|--|--|--|--|---|---|--|--|---|--|
| Organização da rotina do time de inteligência comercial        |  |  |  |   |   |  |  |   |  |
| Treinamentos frequentes da ferramenta                          | <i>Fit Ajustável</i>   |  |  |   |   |  |  |   |  |
| Facilidade de acesso ao manual de uso da ferramenta            | <i>Fit Ajustável</i>   | Possui <i>Fit</i>                            |  |   |   |  |  |   |  |
| Garantia de atendimento às demandas técnicas do banco de dados | Possui <i>Fit</i>  | Possui <i>Fit</i>                            | Possui <i>Fit</i>  |   |   |  |  |   |  |
| Garantia de dados confiáveis e atualizados                     | <i>Fit Ajustável</i>   | Possui <i>Fit</i>                            | Possui <i>Fit</i>  | <i>Fit Ajustável</i>  |   |  |  |   |  |
| Facilidade no acesso aos dados de <i>sell out</i>              | Possui <i>Fit</i>  | Possui <i>Fit</i>                            | Possui <i>Fit</i>  | <i>Fit Ajustável</i>  | Possui <i>Fit</i>                                 |  |  |   |  |
| Facilidade no relacionamento de dados de diferentes bases      | Possui <i>Fit</i>  | Possui <i>Fit</i>                            | Possui <i>Fit</i>  | Possui <i>Fit</i>   | Possui <i>Fit</i>                                 | Possui <i>Fit</i>  |  |   |  |
| Flexibilidade na leitura de novos dados e fornecedores         | Possui <i>Fit</i>  | Possui <i>Fit</i>                            | Possui <i>Fit</i>  | Não possui <i>Fit</i>   | <i>Fit Ajustável</i>                              | Possui <i>Fit</i>  | Possui <i>Fit</i>  |   |  |
| Flexibilidade na variação mapeada dos dados                    | Possui <i>Fit</i>  | Possui <i>Fit</i>                            | Possui <i>Fit</i>  | Não possui <i>Fit</i>   | <i>Fit Ajustável</i>                              | Possui <i>Fit</i>  | Possui <i>Fit</i>  | Possui <i>Fit</i>   |  |

Figura 28: Quadro de análise de Fit entre condições de contorno.

Fonte: Elaboração Própria.

Com base nessa análise de *fit* é possível desenhar respostas mais adequadas para cada característica de contorno desenhada, visando também contornar os conflitos identificados. A seguir, será feito o estudo de proposição de soluções.

## 7. PROPOSTAS DE SOLUÇÃO

Nesta seção serão apresentadas propostas de solução para os problemas abordados na parte anterior, de acordo com suas respectivas características de contorno e *fits*. Serão abordadas soluções separadas de acordo com sua categoria. Todos os problemas relatados no presente trabalho foram revalidados com os entrevistados conforme figura 2.

Como já abordado, sendo uma limitação do presente trabalho, não foi possível a validação das propostas de solução visto que envolveria diversas áreas e cargos de liderança da companhia nas quais não foi obtido acesso.

Segue a seguir a figura 29 com o resumo de todas as propostas de soluções levantadas. Chegou-se nestas através de *feedbacks* durante as entrevistas e conversas sobre o tema com os usuários do banco de dados.

| Situação Indesejada  | Categorização            | Características de Contorno                                    | Análise fit - Características de contorno ajustáveis   | Análise fit - Características de contorno divergentes   | Propostas de Solução   |
|--|--------------------------|--|--|---|--|
| Sobrecarga na montagem de templates para áreas usuárias do banco de dados.                                   | Organizacional           | Organização da rotina do time de inteligência comercial        | -Treinamentos frequentes da ferramenta<br>-Facilidade de acesso ao manual de uso da ferramenta<br>-Garantia de dados confiáveis e atualizados<br>-Facilidade no acesso aos dados de sell out   |   | Uso de ferramenta mais específica para análise de dados no volume proposto.  |
| Desgaste diário para validação dos dados de origens externas.  |                          | Organização da rotina do time de inteligência comercial        | -Treinamentos frequentes da ferramenta<br>-Facilidade de acesso ao manual de uso da ferramenta<br>-Garantia de dados confiáveis e atualizados<br>-Facilidade no acesso aos dados de sell out   |   | Etapa de validação antes de disponibilização dos dados   |
| Necessidade de treinamentos constantes sobre o uso da ferramenta.  | Pessoas                  | Treinamentos frequentes da ferramenta                          | -Organização da rotina do time de inteligência comercial   |   | Manual de uso atualizado da ferramenta de análise e documento com a explicação sobre todas as bases, métricas e filtros.                   |
| Pouco recurso de TI focado no trabalho do banco de dados.  |                          | Garantia de atendimento às demandas técnicas do banco de dados | -Garantia de dados confiáveis e atualizados<br>-Facilidade no acesso aos dados de sell out   | -Flexibilidade na leitura de novos dados e fornecedores<br>-Flexibilidade na variação mapeada dos dados | Divisão das responsabilidades sobre os bancos de dados para outros colaboradores.  |
| Guia de utilização desatualizado.  | Processos                | Facilidade de acesso ao manual de uso da ferramenta            | -Organização da rotina do time de inteligência comercial   |   | Manual de uso atualizado da ferramenta de análise e documento com a explicação sobre todas as bases, métricas e filtros.                   |
| Delay entre a informação e tomada de decisão devido longo tempo de disponibilização e atualização dos dados. |                          | Garantia de dados confiáveis e atualizados                     | -Organização da rotina do time de inteligência comercial<br>-Garantia de atendimento às demandas técnicas do banco de dados<br>-Flexibilidade na leitura de novos dados e fornecedores<br>-Flexibilidade na variação mapeada dos dados |   | Se aprofundar na análise para entender onde se encontra o gargalo temporal do processo.  |
| Lentidão para utilização do banco de dados em Excel para análises mais detalhadas.                           | Tecnologia da Informação | Facilidade no acesso aos dados de sell out                     | -Organização da rotina do time de inteligência comercial<br>-Treinamentos frequentes da ferramenta<br>-Garantia de atendimento às demandas técnicas do banco de dados<br>-Garantia de dados confiáveis e atualizados                   |   | Uso de ferramenta mais específica para análise de dados no volume proposto.  |
| Utilização de softwares auxiliares para viabilizar o uso do banco de dados.                                  |                          | Facilidade no acesso aos dados de sell out                     | -Organização da rotina do time de inteligência comercial<br>-Treinamentos frequentes da ferramenta<br>-Garantia de atendimento às demandas técnicas do banco de dados<br>-Garantia de dados confiáveis e atualizados                   |   | Uso de ferramenta mais específica para análise de dados no volume proposto.  |
| Sub-utilização de dados devido estrutura de leitura do banco de dados.                                       |                          | Facilidade no acesso aos dados de sell out                     | -Organização da rotina do time de inteligência comercial<br>-Treinamentos frequentes da ferramenta<br>-Garantia de atendimento às demandas técnicas do banco de dados<br>-Garantia de dados confiáveis e atualizados                   |   | Diversificar a variedade dos dados da plataforma. Sensoriamento ou Inteligência Artificial (Software).                                     |
| Ausência de chaves de dados comuns para relacionamento de bases de banco de dados diferentes.                |                          | Facilidade no relacionamento de dados de diferentes bases      |  |   | Curto Prazo: Entender as perdas para corrigir o histórico e estruturar as bases.<br>Longo Prazo: Investimento em <i>Machine Learning</i> . |
| Dependência de arquivos de de-para.  |                          | Facilidade no relacionamento de dados de diferentes bases      |  |   | Curto Prazo: Entender as perdas para corrigir o histórico e estruturar as bases.<br>Longo Prazo: Investimento em <i>Machine Learning</i> . |
| Alta dependência de ajustes manuais por parte da TI.   |                          | Flexibilidade na variação mapeada dos dados                    | -Garantia de dados confiáveis e atualizados  | -Garantia de atendimento às demandas técnicas do banco de dados   | Etapa de validação antes de disponibilização dos dados   |
| Dificuldade para conexão de novas fontes de dados no cubo.   |                          | Flexibilidade na leitura de novos dados e fornecedores         | -Garantia de dados confiáveis e atualizados  | -Garantia de atendimento às demandas técnicas do banco de dados   | Estruturação de processo, gerando um manual, para conexão de novas fontes de dados   |

Figura 29: Quadro de propostas de soluções frente as condições de contorno e análise de fit.

Fonte: Elaboração Própria.



### **7.1. Problemas Organizacionais**

São dois os problemas organizacionais: a sobrecarga da equipe de inteligência comercial na assistência às demais áreas usuárias e o desgaste diário para validação de dados de origem externa.

O primeiro problema se dá pela falta de conhecimento da equipe que utiliza a ferramenta de criar novas visualizações de dados. Como possível solução para este problema está a utilização de outra ferramenta mais específica para análise de dados que o Excel, que permita para os usuários finais, de forma mais simplificada, o melhor manejo das métricas e filtros para a criação de diferentes *templates*.

No segundo caso fica claro uma falta de padrão na qualidade dos dados. É necessário ter uma etapa de validação destes antes de subi-los na base *sell out*. Nesta etapa deve ser possível eliminar erros para evitar a redução da qualidade da base.

Ambos problemas organizacionais possuem em sua característica de contorno um *fit* ajustável com "Treinamentos frequentes da ferramenta", "Facilidade de acesso ao manual de uso da ferramenta", "Garantia de dados confiáveis e atualizados" e "Facilidade no acesso aos dados de *sell out*". A questão do *fit* ajustável é por essas características de contorno gerarem demandas para a equipe de suporte já sobrecarregada. Neste sentido as soluções propostas contornam esse conflito, visto que ambas facilitam o atingimento das outras condições de contorno.

### **7.2. Problemas de Pessoas**

Problemas de pessoas também são dois. Tem-se a necessidade de treinamento constante no uso da plataforma e dos significados das métricas e filtros e somente um colaborador como responsável pelo banco de dados na área de TI.

Parte do primeiro problema também pode ser resolvido com o uso de uma nova ferramenta que seja mais específica para análise de dados, onde seria mais fácil para a RP *Cosmetics* manter um treinamento padrão de uso. Além disso é importante que haja um documento onde se registre o significado das métricas e dos filtros e que seja atualizado a cada inserção ou atualização, de forma a se reduzir a dificuldade de interpretação dos usuários.

Esta situação indesejada tem *fit* ajustável com "Organização da rotina da área de inteligência comercial", porém ao se ter um manual estruturado e uma ferramenta mais focada na finalidade da operação é reduzido o tempo consumido da área de inteligência comercial.

No segundo problema, fica claro a dependência da equipe em uma pessoa de suporte para o banco de dados, que atualmente fica sobrecarregada. Neste contexto é importante levantar todas as demandas desse analista para entender quais os fatores que mais o tomam tempo. No caso de serem tarefas que realmente necessitem de sua atenção e que não sejam reduzidas com as outras soluções propostas nesse trabalho, ainda assim é importante dividir as responsabilidades do setor de forma a garantir mais flexibilidade na demanda e evitar completa dependência de uma só pessoa.

Esta situação indesejada tem como condição de contorno a garantia de atendimento as demandas técnicas do banco de dados, que tem *fit* ajustável com "Garantia de dados confiáveis e atualizados" e "Facilidade no acesso aos dados de *sell out*". Essas condições são contornadas na medida em que se irá analisar o trabalho de forma a eliminar tarefas que sejam menos técnicas e possam ser realizadas por outra(s) pessoas além de buscar novos profissionais capacitados para redução da dependência, facilitando o foco da equipe especializada em demandas que realmente precisam da atenção deles, liberando tempo para questões mais estratégicas para a companhia.

### **7.3. Problemas de Processos**

Na parte de processos tem-se um guia de utilização desatualizado e um *delay* entre o surgimento da ação a ser medida/metrificada e a apresentação do dado para a tomada de decisão. No primeiro ponto é simples a solução, é necessário, como dito na subseção anterior, ter um treinamento constante sobre o uso da plataforma além de matérias para consulta que estejam atualizados para a equipe. Aqui temos a condição de contorno facilidade de acesso ao manual de uso da ferramenta atendida e seu *fit* ajustável contornado já que reduzirá a demanda rotineira da equipe de inteligência comercial.

Para o tempo de demora entre a tomada do dado e disponibilização do dado tem-se que se aprofundar para entender onde se dá o gargalo temporal deste processo. No modelo ideal de Big Data os dados são coletados por sensores ou há conexão com o ERP do cliente de forma a ter os dados em tempo real em que a venda ou outra ação é realizada – esta integração deve permitir que a base de análise da RP *Cosmetics* também esteja atualizada de forma a se ter alertas a equipe, ações automáticas a partir do dado e dados em tempo real para uma ação “inteligente” via analistas mais rápida. Desta forma é atendida a condição de garantia de dados confiáveis e atualizados e contornando os ajustes, já que a conexão via sensores e/ou *ERP* do cliente já estará programada para um padrão de dados, garantindo as flexibilidades citadas.

#### 7.4. Problemas de Tecnologia da Informação

Esta é a parte mais extensa, onde tem-se sete problemas para se propor soluções, porém alguns tem intercessão. Como primeiro temos a lentidão para utilização do banco de dados em Excel para análises mais detalhadas e como segundo a utilização de *softwares* auxiliares para viabilizar o uso do banco de dados.

Ambos se dão por uso não específico de um *software*, como também especificado na parte 7.1 e 7.2. Portanto é necessário a contratação de um *software* mais específico ao tipo de análise que se quer da equipe – os diversos problemas apresentados levam para a consequência essa busca por outras ferramentas pela equipe já que não é atendida.

Como terceiro problema tem-se a subutilização dos dados devido a estrutura de leitura do banco de dados. Aqui é possível visualizar o “V” de variedade do Big Data que de acordo com ZIKOPOULOS (2011) 80% são semiestruturados ou desestruturados. Com isso, ao somente ler dados estruturados há uma perda grande no valor a ser retirado a partir das decisões com base nesses dados. É importante nesse caso diversificar as variedades de dados que entram para as bases. Neste ponto há diversas formas de fazer isso – tanto por sensores quanto por novos *softwares* que te permitam isso de forma nativa.

Os três problemas acima têm como condição de contorno a facilidade no acesso aos dados de *sell out*, que foi atendido pelas soluções propostas. Os *fit* ajustáveis foram contornados já que as soluções reduziram demanda de treinamento pós implementação, liberando tempo dos integrantes da área.

Como quarto ponto, se tem outro problema que limita a extração de valor – a falta de correlação entre as bases de dados. Aqui está um ponto de falta de planejamento nas bases de dados, onde o crescimento dos dados é tão grande que a equipe local não dá conta de resolver. Como primeiro passo para a solução deste problema está em se debruçar sobre a estrutura atual e entender onde há perdas de conexão para tentar corrigir se possível o histórico, mas se não pelo menos deste ponto para frente. Como solução a mais longo prazo, já explorada na seção 5.1, está em investir em inteligência artificial para entender a base e melhorar a organização dos dados de forma a evitar problemas desse tipo.

O quinto ponto conversa muito com o anterior. O uso de planilhas de de-para acontece pela falta de comunicação entre as fontes de dados internas. Portanto é necessário a mesma solução no curto e longo prazo. Aqui tem-se a resolução de acordo com a condição de contorno de facilidade no relacionamento de dados de diferentes bases.

Como sexto problema relatado temos a alta dependência de ajustes manuais pela TI. Neste ponto é necessário que as bases de dados tenham um padrão de forma a manter o entendimento das análises pela a equipe, porém deve ser o mais flexível possível para os fornecedores inserirem dados, como forma de não perder esses dados e nem ter entradas erradas. Para isso é preciso entender as possíveis formatações de entrada de dados, e assim como no problema de qualidade de dados, com solução na seção 7.2, ter uma etapa de validação dos dados, onde deve ser possível interpretar diferentes formatações e colocá-las no padrão antes de subir para as bases.

Como último problema temos a dificuldade de conexão de novas fontes de dados no cubo. Para isso é preciso ter um processo estruturado de inserção de novas bases, gerando um manual para orientar os clientes e colaboradores como é feito.

As duas situações indesejadas têm como *fit* ajustável a garantia de dados confiáveis e atualizados, porém ao realizar as soluções propostas acima a base de dados vai ser capaz de reconhecer e interpretar novos formatos de dados, reduzindo assim os erros na inserção dos dados, contornando assim o conflito.

Com base nas propostas realizadas nesta seção será abordada na próxima um plano de ação definindo a priorização, prazos e etapas de cada proposta.

## 8. PLANO DE AÇÃO

A partir de todas as soluções levantadas na parte anterior, considerando todas as características de contorno e a análise de *fit*, desenhou-se propostas de ação como sugestão para início da implementação. Para isso, pensou-se em dividi-lo em 3 partes: curto prazo, médio prazo e longo prazo com altos investimentos.

Para o curto prazo, pensou-se em atividades que unem menor complexidade, menor custo e de alta urgência para solução dos problemas. Considerou-se o prazo de até 6 meses úteis para conclusão e implementação nesses casos. Na figura 30 a seguir, o quadro com as propostas de solução de curto prazo, com o prazo e as etapas sugeridas, além do impacto esperado ao final do desenvolvimento.

Entende-se que o desenvolvimento do manual e do documento resumo das informações do banco de dados é um projeto de grande importância e de fácil resolução, demandando um tempo dedicado inicial e uma manutenção periódica posterior de menor demanda de tempo. Já a análise da comunicação entre as bases demanda mais tempo e atenção aos detalhes, bem como a contratação de novos colaboradores, que demanda todo um processo interno de RH, além das questões financeiras envolvidas.

| Propostas de Solução   | Prazo   | Custos Atrrelados  | Etapas   | Impacto Esperado  |
|--|---------|--------------------|--|---|
| Manual de uso atualizado da ferramenta de análise e documento com a explicação sobre todas as bases, métricas e filtros. | 3 meses | -                  | <ol style="list-style-type: none"> <li>1. Mapeamento de todas as métricas, filtros do banco de dados.</li> <li>2. Levantamento com fornecedores das informações de cada base de dados e internamente das bases internas.</li> <li>3. Criação de documento com resumo das informações.</li> <li>4. Criação de um manual de uso com as principais funcionalidades do banco de dados.</li> </ol>    | Redução da sobrecarga do time de inteligência comercial, maior facilidade de uso do banco de dados por qualquer usuário e melhor clareza do que compõe o banco de dados.          |
| Entender as perdas para corrigir o histórico e estruturar as bases.  | 6 meses | -                  | <ol style="list-style-type: none"> <li>1. Mapeamento de todas as bases da companhia que precisam se comunicar.</li> <li>2. Entendimento da origem das bases e como atualizar todas para ter uma chave em comum.</li> <li>3. Mapeamento de todas as perdas e riscos de alterações no histórico das bases.</li> <li>4. Desenho do plano de reestruturação das bases para implementação.</li> </ol> | Agilidade na comunicação entre bases de dados, reduzindo a necessidade de tabelas auxiliares e a alta demanda de conferências de dados para garantia de relacionamentos corretos. |
| Divisão das responsabilidades sobre os bancos de dados para outros colaboradores.  | 6 meses | Novas contratações | <ol style="list-style-type: none"> <li>1. Mapeamento de todas as atividades de TI desenvolvidas para o banco de dados.</li> <li>2. Segmentação das atividades em diferentes funções/áreas de conhecimento.</li> <li>3. Levantamento e seleção de novo(s) colaborador(es) para integração da equipe.</li> <li>4. Treinamento dos colaboradores na ferramenta.</li> </ol>                          | Maior rapidez e foco no atendimento as demandas de ajuste da área e na proposição de melhorias para a ferramenta reduzindo consideravelmente a sobrecarga da equipe de TI.        |

Figura 30: Quadro de plano de ação das soluções de curto prazo.

Fonte: Elaboração Própria.

O plano de médio prazo consiste nas propostas que podem ser desenvolvidas em 1 ano, conforme figura 31. O uso de ferramenta mais adequada para a análise é de grande valia para a companhia e apresenta uma oportunidade enorme de crescimento e aumento da produtividade, entretanto talvez seja necessária a compra de licenças de *softwares* que demandariam investimentos.

Além disso, entende-se que ainda no médio prazo é possível começar a explorar as razões que levam ao longo tempo entre a venda de um item e a sua visualização pelos usuários do banco de dados. Essa solução é importante para desenvolvimento de etapas e projetos futuros de melhoria.

| Propostas de Solução  | Prazo | Custos Atrrelados                                       | Etapas   | Impacto Esperado  |
|---|-------|---|--|---|
| Uso de ferramenta mais específica para análise de dados no volume proposto.             | 1 ano | Possível licença de <i>software</i> ainda não utilizado | <ol style="list-style-type: none"> <li>1. Levantamento das maiores dificuldades e necessidades na análise dos dados disponíveis no banco de dados estudado.</li> <li>2. Estudo de mercado das ferramentas disponíveis para as necessidades identificadas.</li> <li>3. Prospecção com atuais usuários da ferramenta para <i>benchmarking</i> e alinhamento de expectativas.</li> <li>4. Contratação da ferramenta ou adaptação de ferramenta já utilizada.</li> </ol>   | <p>Maior produtividade na rotina do time de inteligência comercial e também dos usuários, garantindo a utilização dos dados para qualquer tipo de análise e permitindo a manipulação conforme necessária para geração de relatórios sem a necessidade de uso de diferentes softwares.</p> |
| Se aprofundar na análise para entender onde se encontra o gargalo temporal do processo. | 1 ano | -   | <ol style="list-style-type: none"> <li>1. Agendamento de reuniões com fornecedores para entendimento a fundo de todos os processos que ocorrem desde a conexão com o cliente até o envio para o banco de dados da companhia.</li> <li>2. Agendamento de reuniões com alguns clientes de diferentes portes para entendimento de como funcionam os sistemas internos.</li> <li>3. Prospecção com outras companhias de setores diferentes e consultorias para entendimento de como as demais empresas do mercado lidam com isso.</li> </ol> | <p>Profundidade no entendimento dos entraves que existem para o grande intervalo de tempo entre a venda e a visualização do dado para que seja possível então pensar alternativas para melhorar ou ao menos amenizar essa questão.</p>  |

Figura 31: Quadro de plano de ação das soluções de médio prazo.

Fonte: Elaboração Própria.

No longo prazo estão as propostas que terão prazo acima de 1 ano, conforme a figura 32. A primeira proposta de longo prazo seria a estruturação do processo de conexão de novas fontes de dados, gerando um manual para a equipe. Isto tem como impacto uma conexão mais facilitada e rápida. Entende-se que a própria equipe interna, após as contratações propostas no curto prazo, é capaz de desenvolver um novo módulo de *input* das novas fontes.

Tem-se como segunda proposta diversificar a variedade de dados na plataforma, com prazo de dois anos. Esta proposta tem como custo o investimento em softwares e sensores para gerar como resultado um ganho de valor nas análises a partir destes dados de fontes com maior variedade e de diferentes estruturas (não apenas planilhas).

Como terceira proposta está a criação da etapa de validação dos dados antes da disponibilização. Nesta o prazo também é de dois anos e gera como custo o investimento em desenvolvimento e/ou customização de *software*. Neste ponto é de suma importância que seja um processo completamente automático, para não aumentar de forma significativa o tempo até a disponibilização dos dados. Com esse ponto implementado espera-se uma maior confiabilidade (veracidade de acordo com os V's), e conseqüentemente, a redução da demanda da equipe para validação dos dados e ajustes manuais.

Como quarta e última proposta tem-se o investimento em *machine learning* para a melhoria e automatização da organização da estrutura de dados, minimizando perdas de conexão entre as bases. Para este ponto tem-se como custo a contratação de profissionais especialistas nessa tecnologia ou a contratação de uma consultoria para desenvolvimento junto da equipe interna já existente.

| Propostas de Solução   | Prazo    | Custos Atrelados   | Etapas  | Impacto Esperado  |
|--|----------|--|---|---|
| Estruturação de processo, gerando um manual, para conexão de novas fontes de dados                     | 1,5 anos | -  | <ol style="list-style-type: none"> <li>1. Identificação dos gargalos atuais no processo</li> <li>2. Parametrização do módulo de <i>input</i> de fontes de dados</li> <li>3. Desenvolvimento do módulo</li> <li>4. Redesenho do processo</li> <li>5. Criação de manual para conexão de novas fontes</li> </ol>   | Conexão de novas fontes de forma mais rápida e fácil.   |
| Diversificar a variedade dos dados da plataforma. Sensoriamento ou Inteligência Artificial (Software). | 2 anos   | Investimento em software e sensores                        | <ol style="list-style-type: none"> <li>1. Identificação de indicadores de acordo com a estratégia da empresa</li> <li>2. Levantamento de onde nos processos esses indicadores serão medidos</li> <li>3. Levantamento e compra de equipamentos e softwares necessários</li> <li>4. Implementação dos equipamentos e softwares</li> <li>5. Conexão dos dados gerados nas fontes de dados</li> </ol> | Ganho de valor nas análises a partir de dados importantes de forma estratégica para a empresa                           |
| Etapa de validação antes de disponibilização dos dados   | 2 anos   | Investimento em desenvolvimento / customização de software | <ol style="list-style-type: none"> <li>1. Mapeamento de erros e consequências no <i>input</i> de dados</li> <li>2. Parametrização do tratamento dos erros</li> <li>3. Desenvolvimento ou customização de etapa de validação</li> <li>4. Processo contínuo das três etapas anteriores para novos erros</li> </ol>  | Maior confiabilidade / veracidade dos dados - consequente redução da demanda da equipe para validação e ajustes manuais |
| Investimento em Machine Learning   | 3 anos   | Investimento em equipe interna ou consultoria              | <ol style="list-style-type: none"> <li>1. Contratação de especialistas ou de consultoria especializada em <i>machine learning</i> para tratamento de dados</li> </ol>   | Organização da estrutura de dados minimizando perdas de conexão entre bases   |

Figura 32: Quadro de plano de ação das soluções de médio prazo.

Fonte: Elaboração Própria.

Após o plano de ação proposto, os próximos passos seriam a estruturação desse projeto de mudanças para planejamento da implementação. Como abordado anteriormente, as próximas etapas não serão passíveis de desenvolvimento e acompanhamento neste trabalho.

Na próxima seção apresentam-se as considerações finais deste estudo, com todas as conclusões atingidas, a verificação das premissas definidas na introdução e as recomendações de próximos passos.



## 9. CONSIDERAÇÕES FINAIS

Este trabalho teve por objetivo compreender os termos *Big Data* e *Data Analytics*, tanto no meio acadêmico, quanto na aplicação empresarial real. Com isto, pode-se chegar as principais características do *Big Data* que são conhecidas como os V's do *Big Data*. Elas serviram de base para embasar o trabalho a partir de então.

Com o aprofundamento de cada um dos termos, pode-se identificar o que cada um deles englobava e quais seriam seus principais desafios. Para embasar ainda mais o trabalho, visando a aplicação e estudo na prática, fez-se um estudo de caso em uma empresa multinacional para entendimento de como um grande banco de dados se comportava perante essas características.

É importante ressaltar que o caso foi escolhido considerando-se que há um grande volume de dados na empresa, mesmo que o banco de dados atualmente utilizado não possua todas as estruturas de um Big Data. Por se tratar de um estudo, ele foi escolhido como um exercício de uso dos conceitos estudados.

O nome da empresa foi descaracterizado por motivos de confidencialidade, porém foram feitas entrevistas não estruturadas com colaboradores da área de inteligência comercial e também usuários do banco de dados para garantia de melhor direcionamento no trabalho.

A partir de então foi possível verificar o que da literatura era ou não aplicado no dia a dia da companhia e quais eram os impactos advindos da má aplicação dos conceitos do *Big Data*. Também, para melhor embasamento, pesquisou-se online quais eram os principais problemas e oportunidades no uso de banco de dados em diferentes companhias referência de tecnologia. Concluiu-se que muitos problemas são comuns entre várias delas.

Os problemas da companhia estudada foram então categorizados e suas características de contorno foram identificadas através de entrevistas não estruturadas com funcionários da RP *Cosmetics*. Com isso, propostas de soluções foram levantadas e um plano de ação para implementação foi desenhado.

Ao verificar as premissas levantadas no início deste trabalho, não somos capazes de chegar à uma conclusão com relação a primeira, pois identificamos que a RP *Cosmetics* não conta com ferramentas corretas para a análise de dados. Sendo assim, não foi possível verificar se apenas isso é suficiente para uma análise acurada. Apesar disso, pelo que foi lido e encontrado em diferentes materiais expostos neste trabalho, não parece ser suficiente.

Podemos concluir que a segunda premissa não é verdadeira, pois apesar de possuir um grande volume de dados, isso não garante uma orientação clara para a tomada de decisão na RP

*Cosmetics*. Na verdade, há dificuldade neste quesito devido a necessidade de garantir a veracidade dos dados.

A terceira premissa era de que para obter valor dos dados é necessária uma estrutura de banco de dados com diferentes fontes de informação e concluímos que ela é verdadeira. Verificamos que há uma presença grande de diferentes dados que, conectados, poderiam agregar ainda mais valor à compreensão e tomada de decisão de vendas. A quarta premissa também foi entendida como verdadeira porque realmente a veracidade do dado é fator chave para a RP *Cosmetics*.

A penúltima premissa foi identificada como falsa, pois a velocidade é sim um problema, não apenas na RP *Cosmetics*, mas também para muitas empresas que trabalham com um grande volume de dados. A escolha da tecnologia mais adequada para a manipulação e obtenção de dados para o *Big Data* é um dos principais fatores no estudo de caso realizado.

Encerrando as premissas, a última também é falsa, pois garantir a aplicação correta da estrutura do *Big Data* na empresa ainda não resolveria todos os problemas identificados, já que há origens externas à tecnologia que afetam o todo. Sendo assim, identificou-se que há outras questões a serem solucionadas, juntamente com a correta estrutura do *Big Data*, para que o banco de dados opere na eficiência esperada.

Assim, entendeu-se que os conceitos de *Big Data*, apesar de amplamente conhecidos, ainda estão em desenvolvimento no que diz respeito à aplicação real. Questões organizacionais, processos e inclusive de ciência da computação básicas impedem o correto e ideal funcionamento do *Big Data*, o que pode ser verificado na RP *Cosmetics*.

Dada a relevância do tema estudado, trabalhos futuros podem explorar mais a fundo as maiores causas que originam os problemas mais recorrentes na aplicação do *Big Data*, bem como os principais passos a serem aplicados em empresas de grande porte que já operam com um grande banco de dados, mas de forma incorreta. Também, pode-se explorar a maior aplicação de *machine learning* e sensoriamento nesta aplicação específica de dados de vendas, visando entender como implementá-las nos diferentes setores e diferentes portes de comércio. Por fim, a partir da pesquisa bibliométrica realizada, pode-se construir uma estrutura de quadro conceitual de V's que orientaria o diagnóstico de *Big Data* de empresas perante as características de volume, variedade, velocidade, valor e veracidade.

Em conclusão, este trabalho compreende que sob a ótica do *Big Data* e *Data Analytics*, foi possível entender sua grande importância para o saudável desenvolvimento de uma empresa que sabe utilizar seus dados da forma correta. Sua significância acadêmica se demonstra a partir

do desenvolvimento da pesquisa bibliométrica e do estudo de caso envolvendo a análise de dados como estratégia de venda, além da aplicação de ferramentas de análise e caracterização de problemas para proposição de soluções e plano de ação.

## Guia de Entrevistas da Caracterização do Banco de Dados Frente aos V's

- Breve apresentação do objetivo do projeto de graduação
- Apresentação da definição dos V's que serão abordados
- Para cada V:
  - Como visualiza cada V no banco de dados estudado a partir da definição apresentada
  - Quais as características de cada V se destaca como mais marcante no banco de dados estudado
  - Principais dificuldades encontradas durante o uso
  - Principais oportunidades que visualiza como melhoria para o banco de dados
  - Quais pontos considera mais relevantes para uma melhoria do banco de dados

## REFERÊNCIAS BIBLIOGRÁFICAS

- AL-FUQAHA, Ala; GUIZANI, Mohsen; MOHAMMADI, Mehdi; *et al.* Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. **IEEE Communications Surveys Tutorials**, v. 17, n. 4, p. 2347–2376, 2015.
- AMAZON WEB SERVICES. Opções de análise de big data na AWS. Disponível em: <[https://d1.awsstatic.com/whitepapers/pt\\_BR/Big\\_Data\\_Analytics\\_Options\\_on\\_AWS.pdf](https://d1.awsstatic.com/whitepapers/pt_BR/Big_Data_Analytics_Options_on_AWS.pdf)>. Acesso em: 1 jul. 2020.
- ANDREU-PEREZ, Javier; POON, Carmen C. Y.; MERRIFIELD, Robert D.; *et al.* Big Data for Health. **IEEE Journal of Biomedical and Health Informatics**, v. 19, n. 4, p. 1193–1208, 2015.
- ATZORI, Luigi; IERA, Antonio; MORABITO, Giacomo. The internet of things: A survey. **Computer networks**, v. 54, n. 15, p. 2787–2805, 2010.
- BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.
- CAMPOS, Vicente Falconi. **TQC: Controle da Qualidade Total (no estilo japonês)**. 3. ed. Belo Horizonte, MG: Fundação Christiano Ottoni, Escola de Engenharia da UFMG: Bloch Editores S.A., 1992.
- CHEN, H.; CHIANG, R.; STOREY, V. Business Intelligence and Analytics: From Big Data to Big Impact. **MIS Q.**, 2012.
- CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big Data: A Survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014.
- DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: simplified data processing on large clusters. **Communications of the ACM**, v. 51, n. 1, p. 107–113, 2008.
- DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: simplified data processing on large clusters. **Communications of the ACM**, v. 51, n. 1, p. 107–113, 2008.
- DEMIRKAN, Haluk; DELEN, Dursun. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. **Decision Support Systems**, v. 55, n. 1, p. 412–421, 2013.
- FOSSO WAMBA, Samuel; AKTER, Shahriar; EDWARDS, Andrew; *et al.* How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, v. 165, p. 234–246, 2015.
- GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137–144, 2015.
- GUBBI, Jayavardhana; BUYYA, Rajkumar; MARUSIC, Slaven; *et al.* Internet of Things (IoT): A vision, architectural elements, and future directions. **Future Generation Computer Systems**, v. 29, n. 7, p. 1645–1660, 2013. (Including Special sections: Cyber-enabled

Distributed Computing for Ubiquitous Cloud and Network Services & Cloud Computing and Scientific Applications — Big Data, Scalable Analytics, and Beyond).

HASHEM, Ibrahim Abaker Targio; YAQOOB, Ibrar; ANUAR, Nor Badrul; *et al.* The rise of “big data” on cloud computing: Review and open research issues. **Information Systems**, v. 47, p. 98–115, 2015.

HAZEN, Benjamin T.; BOONE, Christopher A.; EZELL, Jeremy D.; *et al.* Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. **International Journal of Production Economics**, v. 154, p. 72–80, 2014.

HU, Han; WEN, Yonggang; CHUA, Tat-Seng; *et al.* Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. **IEEE Access**, v. 2, p. 652–687, 2014.

KAMBATLA, Karthik; KOLLIAS, Giorgos; KUMAR, Vipin; *et al.* Trends in big data analytics. **Journal of Parallel and Distributed Computing**, v. 74, n. 7, p. 2561–2573, 2014. (Special Issue on Perspectives on Parallel and Distributed Processing).

JUNIOR, Celso Carlino Maria Fornari. Aplicação da Ferramenta da Qualidade (Diagrama de Ishikawa) e do PDCA no Desenvolvimento de Pesquisa para a reutilização dos Resíduos Sólidos de Coco Verde. **INGEPRO-Inovação, Gestão e Produção**, v. 2, n. 9, p. 104–112, 2010.

LAVALLE, Steve; LESSER, Eric; SHOCKLEY, Rebecca; *et al.* How the smartest organizations are embedding analytics to transform information into insight and then action. Findings and recommendations from the first annual New Intelligent Enterprise Global Executive study. p. 13, .

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LEE, Jay; KAO, Hung-An; YANG, Shanhu. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. **Procedia CIRP**, v. 16, p. 3–8, 2014.

LEE, Jay; LAPIRA, Edzel; BAGHERI, Behrad; *et al.* Recent advances and trends in predictive manufacturing systems in big data environment. **Manufacturing Letters**, v. 1, n. 1, p. 38–41, 2013.

MANYIKA, James; CHUI, Michael; BROWN, Brad. **Big data: The next frontier for innovation, competition, and productivity** | McKinsey. Disponível em: <<https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation#>>. Acesso em: 6 jul. 2019.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big Data: A Revolution that Will Transform how We Live, Work, and Think**. [s.l.]: Houghton Mifflin Harcourt, 2013.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data: The Management Revolution. p. 9, .

MEIRELLES, Fernando S. Uso da TI - Tecnologia de Informação nas Empresas. **FGV**, n. 31, p. 162, 2020.

MEIRELES, Manuel. **Ferramentas Administrativas Para Identificar Observar E Analisar Problemas**. [s.l.]: Arte & Ciência, 2001.

MURDOCH, Travis B.; DETSKY, Allan S. The inevitable application of big data to health care. **Jama**, v. 309, n. 13, p. 1351–1352, 2013.

OKUBO, Yoshiko. Bibliometric Indicators and Analysis of Research Systems: Methods and Examples. 1997. Disponível em: <[https://www.oecd-ilibrary.org/science-and-technology/bibliometric-indicators-and-analysis-of-research-systems\\_208277770603](https://www.oecd-ilibrary.org/science-and-technology/bibliometric-indicators-and-analysis-of-research-systems_208277770603)>.

Acesso em: 22 set. 2020.

ORACLE. **O Que é Big Data? | Oracle Brasil**. Disponível em: <<https://www.oracle.com/br/big-data/what-is-big-data.html#link2>>. Acesso em: 26 jun. 2020.

PAUL ZIKOPOULOS, I. B. M.; EATON, Chris; ZIKOPOULOS, Paul. **Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data**. 1st Edition. New York: McGraw-Hill Osborne Media, 2011.

PHILIP CHEN, C. L.; ZHANG, Chun-Yang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. **Information Sciences**, v. 275, p. 314–347, 2014.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. Big data analytics in healthcare: promise and potential. **Health Information Science and Systems**, v. 2, n. 1, p. 3, 2014.

RUSSOM, Philip. Big Data Analytics. **BIG DATA ANALYTICS**, p. 38, .

SANTAFERRARO, John. **EMA White Paper: Creating Business Value With Modernization**. SAP. Disponível em: <<https://www.sap.com/documents/2018/06/6013b53a-0b7d-0010-87a3-c30de2ffd8ff.html>>. Acesso em: 26 jun. 2020.

SANTOS, Raimundo Nonato Macedo dos. Produção científica: por que medir? o que medir? **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 1, n. 1, p. 22, 2004.

SCHOENHERR, Tobias; SPEIER-PERO, Cheri. Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential. **Journal of Business Logistics**, v. 36, n. 1, p. 120–132, 2015.

SEBRAE. CADERNO DE TENDÊNCIAS #2019-2020. Disponível em: <<https://m.sebrae.com.br/Sebrae/Portal%20Sebrae/Anexos/CADERNO%20DE%20TENDENCIAS%202019-2020%20Sebrae%20Abihpec%20vs%20final.pdf>>. Acesso em: 3 nov. 2019.

SHVACHKO, Konstantin; KUANG, Hairong; RADIA, Sanjay; *et al.* The Hadoop Distributed File System. **2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)**, p. 1–10, 2010.

SILVA, Márcia Regina da; HAYASHI, Carlos Roberto Massao; HAYASHI, Maria Cristina Piumbato Innocentini. Análise bibliométrica e cientométrica: desafios para especialistas que

atuam no campo. **InCID: Revista de Ciência da Informação e Documentação**, v. 2, n. 1, p. 110–129, 2011.

SMITH, Thomas M; REECE, James S. The relationship of strategy, fit, productivity, and business performance in a services setting. **Journal of Operations Management**, v. 17, n. 2, p.145-161,1998.

THUSOO, Ashish; SARMA, Joydeep Sen; JAIN, Namit; *et al.* Hive: a warehousing solution over a map-reduce framework. **Proceedings of the VLDB Endowment**, v. 2, n. 2, p. 1626–1629, 2009.

TUCKER, Hank. **Global 2000: as maiores empresas de tecnologia do mundo em 2020**. Forbes Brasil. Disponível em: <<https://forbes.com.br/listas/2020/05/global-2000-as-maiores-empresas-de-tecnologia-do-mundo-em-2020/>>. Acesso em: 5 fev. 2021.

WALLER, Matthew A.; FAWCETT, Stanley E. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. **Journal of Business Logistics**, v. 34, n. 2, p. 77–84, 2013.

WANG, Gang; GUNASEKARAN, Angappa; NGAI, Eric W. T.; *et al.* Big data analytics in logistics and supply chain management: Certain investigations for research and applications. **International Journal of Production Economics**, v. 176, p. 98–110, 2016.

WHITE, Tom. **Hadoop: the definitive guide**. Third edition. Beijing: O’Reilly, 2012.

WU, Xindong; ZHU, Xingquan; WU, Gong-Qing; *et al.* Data Mining with Big Data. p. 26, . ZAHARIA, Matei; CHOWDHURY, Mosharaf; FRANKLIN, Michael J; *et al.* Spark: Cluster Computing with Working Sets. p. 7, .

ZANELLA, Andrea; BUI, Nicola; CASTELLANI, Angelo; *et al.* Internet of Things for Smart Cities. **IEEE Internet of Things Journal**, v. 1, n. 1, p. 22–32, 2014.

ZHONG, Ray Y.; HUANG, George Q.; LAN, Shulin; *et al.* A big data approach for logistics trajectory discovery from RFID-enabled production data. **International Journal of Production Economics**, v. 165, p. 260–272, 2015.

ZHONG, Ray Y.; NEWMAN, Stephen T.; HUANG, George Q.; *et al.* Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. **Computers & Industrial Engineering**, v. 101, p. 572–591, 2016.

**How companies are using big data and analytics | McKinsey**. Disponível em: <<https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-companies-are-using-big-data-and-analytics>>. Acesso em: 12 jul. 2020.

**O que é OLAP? Conceitos Básicos Sobre OLAP**. DevMedia. Disponível em: <<https://www.devmedia.com.br/conceitos-basicos-sobre-olap/12523>>. Acesso em: 23 jan. 2020.