

**UNIVERSIDADE METODISTA DE PIRACICABA**

**FACULDADE DE ENGENHARIA ARQUITETURA E URBANISMO**

**PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

**PROPOSTA DE UM MODELO DE DESCOBERTA DE  
CONHECIMENTO PARA O PROJETO INFORMACIONAL DO  
PROCESSO DE DESENVOLVIMENTO DO PRODUTO**

**EMERSON RABELO**

**ORIENTADOR: PROF. DR. FERNANDO CELSO DE CAMPOS**

**SANTA BÁRBARA D'OESTE**

**2017**

**UNIVERSIDADE METODISTA DE PIRACICABA**

**FACULDADE DE ENGENHARIA ARQUITETURA E URBANISMO**  
**PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

**PROPOSTA DE UM MODELO DE DESCOBERTA DE  
CONHECIMENTO PARA O PROJETO INFORMACIONAL DO  
PROCESSO DE DESENVOLVIMENTO DO PRODUTO**

**EMERSON RABELO**

**ORIENTADOR: PROF. DR. FERNANDO CELSO DE CAMPOS**

Tese de doutorado apresentada no Programa  
de Pós-Graduação em Engenharia de  
Produção da Faculdade de Engenharia,  
Arquitetura e Urbanismo da Universidade  
Metodista de Piracicaba – UNIMEP

**SANTA BÁRBARA D'OESTE**

**2017**

**PROPOSTA DE UM MODELO DE DESCOBERTA DE CONHECIMENTO PARA O  
PROJETO INFORMACIONAL DO PROCESSO DE DESENVOLVIMENTO DO  
PRODUTO**

**Emerson Rabelo**

Tese de Doutorado defendida e aprovada em 14 de agosto de 2017 pela banca examinadora, constituída pelos professores:

Prof. Dr. Fernando Celso de Campos  
PPGEP/UNIMEP

Prof. Dr. Carlos Roberto Camello Lima  
PPGEP/UNIMEP

Prof. Dr. Aparecido dos Reis Coutinho  
PPGEP/UNIMEP

Prof. Dr. José Alcides Gobbo Junior  
FEB - DEP/UNESP - Bauru

Prof. Dr. Edson Walmir Cazarini  
DEP - EESC/USP

## DEDICATÓRIA

*A Deus,  
pela força e ensinamentos.*

*À minha família:  
avós, pais, irmão, primos e ao  
tio Francisco Rabelo (in memorian),  
pelos ensinamentos e exemplos de vida.*

*À minha esposa;  
que com seu amor me  
apoiou nos momentos mais difíceis.*

*Às minhas filhas  
Maria Eduarda e Maria Julia, que  
embora ainda não tem conhecimento disso, foram o  
fator motivacional maior para finalização deste trabalho*

## **AGRADECIMENTOS**

Agradeço a primeiramente a Deus, o que seria de mim sem a fé que tenho nele. Deu-me forças nos momentos mais difíceis e, iluminou e guiou meus passos direcionando-me no caminho da aprendizagem.

A minha mãe, Josefa e ao meu pai Valentim, pela dedicação e amor, sempre mostrando por meio de atitudes o valor da humildade e do trabalho, e a educação que proporcionaram a mim e ao meu irmão com valores morais e éticos. Agradeço também ao meu irmão Jefferson que sempre esteve presente e disposto a me ajudar a qualquer momento.

A Deus, mais uma vez, por ter colocado como esposa na minha vida a Juliana Furlan Rabelo. Com seu apoio nos momentos mais difíceis me fortaleceu e, com certeza sem ela ter acreditado em mim incondicionalmente, nada disso seria possível.

As minhas filhas Maria Eduarda Rabelo e Maria Julia Rabelo, que foram a minha principal motivação.

Ao meu Orientador, Prof. Dr. Fernando Celso de Campos, que tenho grande admiração. Meu agradecimento pelo seu profissionalismo, soluções das dúvidas que me ocorreram ao longo do trabalho e, nunca mediu esforços no apoio para realização deste trabalho, compartilhando o seu conhecimento e sempre com palavras de incentivo.

Nos momentos mais difíceis da vida, surgem pessoas abençoadas por Deus para nos apoiar e, graças ao nosso Senhor, sempre recebi essa bênção. Nesses últimos momentos, o meu primo irmão Heber Rabelo foi uma peça fundamental no apoio e com palavras motivacionais para finalização do presente trabalho.

À todos os professores da UNIMEP que, de forma direta ou indireta, contribuíram para o desenvolvimento deste trabalho, principalmente aos professores, Dr. Aparecido dos Reis Coutinho, com suas sabias palavras de conforto e importantes recomendações apresentadas na qualificação e, ao Prof. Dr. Carlos Roberto Camello que também contribuiu, na qualificação, com sugestões e observações construtivas proporcionando a evolução do presente trabalho.

À Marta Helena T. Bragaglia, pelo apoio e colaboração para o andamento do trabalho e esforços na qualidade do atendimento.

Aos amigos e acadêmicos da UNIMEP pelos momentos de entusiasmos partilhados em conjunto. Aos que me acompanharam nas viagens para UNIMEP, em especial ao professor Narciso A. Frazin pelas suas orações e apoio.

Aos colaboradores e diretores da Indústria Amarelo Manga, que acolheram a presente pesquisa e ofereceram as condições para o seu desenvolvimento, em especial ao Diretor Antonio Rodrigues Gomes Filho. Estendo ainda meus agradecimentos ao Grupo Morena Rosa, em especial às estilistas e a atenção recebida da Barbara Fernandez Pereira do departamento de comunicação e do representante comercial Wagner Menezes.

Ao Instituto Federal de Educação, Ciências e Tecnologia do Paraná - IFPR, pela concessão de afastamento no último ano da pesquisa.

Aos amigos professores e técnicos administrativos do IFPR, em especial do campus de Ivaipora e Astorga.

Aos meus alunos e ex-alunos que, mesmo sem imaginarem, indiretamente colaboram para o meu crescimento pessoal e profissional.

RABELO, E. **Proposta de um Modelo de Descoberta de Conhecimento para Projeto Informacional do Processo de Desenvolvimento do Produto**. 2017, 210p. Tese Doutorado em Engenharia. de Produção. Faculdade de Eng. Arquitetura e Urbanismo, Universidade Metodista de Piracicaba, Santa Bárbara d'Oeste.

## RESUMO

A facilidade e a evolução do acesso tecnológico têm sido responsáveis pela velocidade e pelo volume com que os dados são produzidos. Em consequência, surgem cenários, oportunidades e desafios que favorecem as tomadas de decisão e auxiliam o processo de desenvolvimento do produto (PDP). A literatura recente tem evidenciado a falta de modelos que considerem as características (5V's) de dados complexos para apoiar a descoberta de conhecimento no PDP. A descoberta de conhecimento a partir de dados estruturados é um processo estabelecido; entretanto, ainda está em desenvolvimento para o caso de dados que apresentam pouca ou nenhuma estrutura, razão pela qual seu estudo vem se destacando nos últimos anos. No presente trabalho, o objetivo é propor um modelo conceitual que contribua para o projeto informacional do PDP. Para dar suporte ao modelo proposto, foram discutidas as metodologias tradicionais associadas às demandas do *Big Data*, os modelos de referência para o PDP e a identificação das atividades do projeto informacional. O diferencial apresentado no trabalho é que, no modelo proposto, são considerados todos os tipos de estrutura de dados. Além disso, englobam-se as possibilidades existentes para realizar a adequação de conjunto de dados com características *Big Data* e aplicá-los em soluções tradicionais. O desenvolvimento do modelo proposto e sua aplicação em uma indústria de confecção evidenciaram que esforços empreendidos na compreensão antecipada dos dados podem contribuir para que os dados extraídos sejam menos complexos, tornando o *Big Data* viável para o uso na indústria. Por fim, a aplicação do modelo proposto, em cenário real de produção, gerou conhecimentos novos e úteis no PDP.

**PALAVRAS-CHAVE:** Processo de Desenvolvimento de Produto, Descoberta de Conhecimento, Projeto Informacional, Modelo Conceitual, *Big Data*.

RABELO, Emerson. **Proposal of a Knowledge Discovery Model for Informational Design of the Product Development Process**. 2017, 210p. Tese Doutorado em Engenharia de Produção. Faculdade de Engenharia Arquitetura e Urbanismo, Universidade Metodista de Piracicaba, Santa Bárbara d'Oeste.

### ***ABSTRACT***

Both, facility and technological access evolution have been responsible for the speed and volume according to which data are produced. As a result, scenarios, opportunities and challenges emerge, which favor decision-making and support the product development process (PDP). Recent literature has highlighted the lack of models that consider the complex data characteristics (5V's) to support the knowledge discovery on the PDP. Knowledge discovery based on structured data is an established process; however, it is still under development for the case of data that have little or no structure, and this is the reason why its study has been emphasized in recent years. The present study aims at proposing a conceptual model that contributes to the PDP informational project. In order to support such a model, the traditional methodologies associated with the Big Data demands, the reference models for the PDP, and the identification of the informational project activities were discussed. The differential shown in this study is that in the model suggested hereby all types of data structure are considered. In addition, the existing possibilities to perform the dataset adequacy with Big Data characteristics are included, as well as how to apply them in traditional solutions. The development of the model suggested and its application in a textile industry have shown that efforts undertaken in the early understanding of the data may contribute to the extraction of less complex data, making Big Data feasible for being used in the industry. Finally, the application of the proposed model, in real production scenario, generated new and useful knowledge in the PDP.

**KEYWORDS:** Product Development Process, Knowledge Discovery, Product Development, Informational Project, Conceptual Model, Big Data.



## LISTA DE ABREVIATURAS E SIGLAS

API	–	<i>Application Programming Interface</i>
BD	–	Banco de Dados
BDA-PL	–	<i>Data-Based Analytics for Product Lifecycle</i>
BoL	–	<i>Begin of Life</i>
BSON	–	<i>Binary Structured Object Notation</i>
CAD	–	<i>Computer Aided Design</i>
CAE	–	<i>Computer Aided Engineering</i>
CAPP	–	<i>Computer Aided Process Planning</i>
CC	–	<i>Cloud Computing</i>
CCEVP	–	<i>Cloud Computing Based Effective Virtual Physical</i>
CM	–	<i>Cloud Manufacturing</i>
CRAN	–	<i>Comprehensive R Archive Network</i>
Crisp-DM	–	<i>Cross Industry Standard Process for Data Mining</i>
CRM	–	<i>Customer Relationship Management</i>
CVP	–	Ciclo de Vida do Produto.
DDM	–	<i>Distributed Data Mining</i>
DIP	–	Diagnóstico, Implantação e Perpetuação
DM	–	<i>Data Mining</i>
DW	–	<i>Data Warehouse</i>
EoL	–	<i>End of Life</i>
ERP	–	<i>Enterprise Resource Planning</i>
ETL	–	<i>Extract, Transform and Load</i>
HACE	–	Heterogeneidade, Autônomo, Complexo e Evolução
HDFS	–	<i>Hadoop Distributed File System</i>
HP	–	<i>Hewlett Packard</i>
HQL	–	<i>Hive Query Language</i>
IA	–	Inteligência Artificial
IBM	–	<i>International Business Machines</i>
IDC	–	<i>International Data Corporation</i>

IoT	–	<i>Internet of Things</i>
JSON	–	<i>JavaScript Object Notation</i>
JSS	–	<i>Journal of Statistical Software</i>
KDD	–	<i>Knowledge Discovery in Database</i>
LCA	–	<i>Life Cycle Assessment</i>
LCE	–	<i>Life Cycle Engineering</i>
MD	–	Mineração de Dados
MFW	–	<i>Milan Fashion Week</i>
ML	–	<i>Machine Learning</i>
MoL	–	<i>Midle of Life</i>
NoSQL	–	<i>Not Only Structured Query Language</i>
NYFW	–	<i>New York Fashion Week</i>
PDM	–	<i>Parallel Data Mining</i>
PDP	–	Processo de Desenvolvimento do Produto
PEIDs	–	<i>Product Embedded Information Devices</i>
PLM	–	Product Lifecycle Management
PIB	–	Produto Interno Bruto
RSL	–	Revisão Sistemática da Literatura
SCM	–	Suply Chain Management
SQL	–	Structured Query Language
SGBD	–	Sistema de Gerenciamento em Banco de Dados
TEL	–	Transform, Extract and Load
TMT	–	Técnicas Métodos e Tecnologias

## LISTA DE FIGURAS

FIGURA 1 – VISÃO MACRO DA PESQUISA .....	6
FIGURA 2 – ETAPAS DO PROCESSO KDD .....	10
FIGURA 3 – ELEMENTOS DO PROCESSO DE DESCOBERTA DE CONHECIMENTO .....	10
FIGURA 4 – FASES DO MODELO CRISP-DM .....	15
FIGURA 5 – QUESTÕES CRÍTICAS DE MÁQUINA DE APRENDIZAGEM PARA BIG DATA.....	22
FIGURA 6 – CARACTERÍSTICAS DO BIG DATA, DESAFIOS E SOLUÇÕES. ....	24
FIGURA 7 – DIAGRAMA DO PROCESSO BIG DATA.....	27
FIGURA 8 – FACETAS DO BIG DATA .....	29
FIGURA 9 – ARQUITETURA GLOBAL BASEADA NO CVP – BDA-PL.....	34
FIGURA 10 – SISTEMAS DE INFORMAÇÃO NAS FASES DO CVP.....	37
FIGURA 11 – FONTES INTERNA E EXTERNA DE DADOS.....	39
FIGURA 12 – INFORMAÇÕES DE ENTRADA E SAÍDA PARA CADA FASE DO CVP .....	41
FIGURA 13 – FRAMEWORK DO BIG DATA NO PLM .....	42
FIGURA 14 – PROCESSO DE DESENVOLVIMENTO DE PRODUTO.....	46
FIGURA 15 – PRINCIPAIS INFORMAÇÕES E DEPENDÊNCIAS ENTRE AS ATIVIDADES DA FASE DE PROJETO INFORMACIONAL .....	48
FIGURA 16 – TÉCNICAS GEOMÉTRICAS – A) MATRIZ DE DISPERSÃO DE DADOS; B) GRÁFICO DE DISPERSÃO; C) COORDENADAS PARALELAS.....	50
FIGURA 17 – A) FACES DE CHERNOFF ; B) GRÁFICO DE ESTRELA; C) ARESTA.....	51
FIGURA 18 – PROPOSTA METODOLÓGICA.....	53
FIGURA 19 – FASES DA RSL .....	56
FIGURA 20 – PROCESSO DE DESENVOLVIMENTO DA PESQUISA .....	57
FIGURA 21 – ETAPAS DO PROCESSO DE SELEÇÃO .....	60
FIGURA 22 – FASES DO MODELO PROPOSTO.....	62
FIGURA 23 – FASE I DO MODELO PROPOSTO .....	64
FIGURA 24 – ATIVIDADES DA FASE DE PROJETO INFORMACIONAL E A INSERÇÃO DO MODELO PROPOSTO.....	65
FIGURA 25 – HISTÓRICO E COMPREENSÃO DOS DADOS.....	67
FIGURA 26 – FONTES DE DADOS.....	68
FIGURA 27 – DIAGRAMA DE ATIVIDADES .....	81
FIGURA 28 – ETAPAS DA FASE III.....	84
FIGURA 29 – ANÁLISE DA QUANTIDADE E COMPLEXIDADE DOS DADOS.....	95
FIGURA 30 – FASE IV - DESCOBERTA DE CONHECIMENTO. ....	98
FIGURA 31 – WORD CLOUD.....	106
FIGURA 32 – WORD TREE .....	107
FIGURA 33 – TÉCNICAS DE VISUALIZAÇÃO PARA DADOS TEXTUAIS .....	108
FIGURA 34 – FLUXOS DO CONHECIMENTO .....	111
FIGURA 35 – ARQUITETURA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO PROJETO INFORMACIONAL .....	112
FIGURA 36 – FACETA E SUBFACETAS DA REDE SOCIAL TWITTER.....	119
FIGURA 37 – POSTAGENS DO CONJUNTO NYFW .....	127
FIGURA 38 – POSTAGENS DOS USUÁRIO DO TWITTER REFERENTE AO EVENTO NYFW APÓS AS ATIVIDADES DE PREPARAÇÃO DOS DADOS .....	131

FIGURA 39 – AGRUPAMENTOS (K-MEANS) PARA O CONJUNTO DE DADOS NYFW..	137
FIGURA 40 – AGRUPAMENTOS (K-MEANS) PARA O CONJUNTO DE DADOS MFW....	138
FIGURA 41 – VISUALIZAÇÃO WORD CLOUD PARA OS CONJUNTOS NYFW E MFW .	141
FIGURA 42 – GRUPO DE VISUALIZAÇÃO WORD CLOUD LIGADO POR TERMOS .....	142
FIGURA 43 – VISUALIZAÇÃO WORD TREE DO CONJUNTO DE DADOS MFW .....	144
FIGURA 44 – DIAGRAMA DE ASSOCIAÇÃO DE TERMOS DO CONJUNTO DE DADOS MFW .....	146
FIGURA 45 – DIAGRAMA DA APLICAÇÃO DO MODELO PROPOSTO.....	151

## LISTA DE TABELAS

TABELA 1 – DADOS ESTRUTURADOS VS DADOS NÃO ESTRUTURADOS .....	19
TABELA 2 – VISÃO GERAL DAS TECNOLOGIAS DO BIG DATA .....	25
TABELA 3 – AVALIAÇÃO DAS CARACTERÍSTICAS DOS BD NoSQL .....	93
TABELA 4 – CARACTERÍSTICAS DO MODELO DE DADOS NoSQL .....	95
TABELA 5 – ANÁLISE DAS CARACTERÍSTICAS DAS TÉCNICAS DE VISUALIZAÇÃO.....	104
TABELA 6 – FREQUÊNCIAS DA ASSOCIAÇÃO ENTRE OS TERMOS - NYFW .....	135
TABELA 7 – FREQUÊNCIAS DA ASSOCIAÇÃO ENTRE OS TERMOS - MFW .....	136
TABELA 8 – RESPOSTAS AO QUESTIONÁRIO - INDÚSTRIA AM .....	149
TABELA 9 – RESPOSTAS AO QUESTIONÁRIO - GRUPO MORENA ROSA.....	149

## LISTA DE QUADROS

QUADRO 1 – PALAVRAS CHAVES EMPREGADAS NA PESQUISA.....	58
QUADRO 2 – QUANTITATIVO DE ARTIGOS ENCONTRADOS NA RSL.....	58
QUADRO 3 – FACETA FONTE DE DADOS. ....	71
QUADRO 4 – FORMULÁRIO DO DETALHAMENTO DO CONJUNTO DE DADOS.....	75
QUADRO 5 – COMPARAÇÃO ENTRE SOLUÇÕES TRADICIONAIS E BIG DATA. ....	79
QUADRO 6 – RESUMO DOS ASPECTOS ABORDADOS NA INDÚSTRIA AM.....	117
QUADRO 7 – POSTAGENS EXTRAÍDAS. ....	121
QUADRO 8 – FORMULÁRIO DE DOCUMENTAÇÃO E DETALHAMENTO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS.....	123
QUADRO 9 – FREQUÊNCIA DOS TERMOS DESTACADOS.....	133

## SUMÁRIO

<b>LISTA DE ABREVIATURAS E SIGLAS .....</b>	<b>VIII</b>
<b>LISTA DE FIGURAS.....</b>	<b>X</b>
<b>LISTA DE TABELAS .....</b>	<b>XII</b>
<b>LISTA DE QUADROS.....</b>	<b>XIII</b>
<b>1. INTRODUÇÃO.....</b>	<b>1</b>
1.1. JUSTIFICATIVA E RELEVÂNCIA.....	2
1.2. PROBLEMA DA PESQUISA .....	4
1.3. OBJETIVO GERAL E ESPECÍFICO.....	4
1.4. ESTRUTURA DA TESE.....	5
<b>2. REVISÃO DA LITERATURA.....</b>	<b>8</b>
2.1. DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD) .....	8
2.1.1. MINERAÇÃO DE DADOS .....	11
2.1.2. USO DE MÍDIAS SOCIAIS PARA DESCOBERTA DE CONHECIMENTO.....	13
2.1.3. MODELO CRISP-DM.....	14
2.2. <i>BIG DATA</i> .....	16
2.2.1. MINERAÇÃO DE DADOS E <i>BIG DATA</i> .....	20
2.2.2. APRENDIZAGEM DE MÁQUINA NA DESCOBERTA DE CONHECIMENTO .....	21
2.2.3. QUESTÕES CRÍTICAS SOBRE MÁQUINA DE APRENDIZAGEM PARA <i>BIG DATA</i> .....	21
2.2.4. TECNOLOGIAS RELACIONADAS AO <i>BIG DATA</i> .....	24
2.2.5. FLUXO DE DADOS NO <i>BIG DATA</i> .....	27
2.2.6. ANÁLISE DE FACETAS NO <i>BIG DATA</i> .....	28
2.3. <i>GERENCIAMENTO DO CICLO DE VIDA DE PRODUTO - PLM</i> .....	30
2.4. <i>BIG DATA</i> E PLM .....	33
2.4.1. PRODUÇÃO DE DADOS.....	35
2.4.2. FONTES DE DADOS: INTERNAS E EXTERNAS .....	37
2.4.2.1. Dados de Entrada e Saída Utilizados nas Fases do CVP.....	40
2.5. <i>QUALIDADE DOS DADOS</i> .....	43
2.6. <i>MODELOS DE REFERÊNCIA PARA O PDP</i> .....	44
2.6.1. IDENTIFICAÇÃO DAS ETAPAS DO PROJETO INFORMACIONAL .....	47
2.7. <i>TÉCNICAS DE VISUALIZAÇÃO</i> .....	49
<b>3. MÉTODO DA PESQUISA .....</b>	<b>53</b>
3.1. PROCEDIMENTOS TÉCNICOS .....	54
3.1.1. REVISÃO SISTEMÁTICA DA LITERATURA - RSL.....	55
3.1.1.1. Planejamento.....	56
3.1.1.2. Busca e Base de Dados .....	57
3.1.1.3. Classificação dos Artigos Seleccionados.....	59
3.1.2. SUBSÍDIOS PARA O DESENVOLVIMENTO DO MODELO PROPOSTO.....	60
3.1.2.1. Modelo KDD e Crisp-DM .....	60
3.1.2.2. Qualidade dos Dados .....	60
3.1.2.3. Análise de Facetas.....	61
3.1.2.4. Arquitetura .....	61
3.1.3. APLICAÇÃO DO MODELO PROPOSTO .....	61
<b>4. DESENVOLVIMENTO DO MODELO PROPOSTO .....</b>	<b>62</b>
4.1. FASE I – DIAGNÓSTICO .....	63
4.2. FASE II – HISTÓRICO E COMPREENSÃO DOS DADOS .....	66

4.2.1.	ETAPA 1 - IDENTIFICAÇÃO DAS FONTES DE DADOS .....	67
4.2.2.	ETAPA 2 – FACETAS PARA A FONTE DE DADOS.....	70
4.2.3.	ETAPA 3 – AVALIAÇÃO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS .....	74
4.2.4.	ETAPA 4 – DECISÃO DAS SOLUÇÕES TECNOLÓGICAS .....	78
4.3.	<b>FASE III – SONDAGEM E USO DE TMT .....</b>	<b>83</b>
4.3.1.	ETAPA 1 – PREPARAÇÃO DOS DADOS .....	84
4.3.2.	ETAPA 2 – SOLUÇÃO TECNOLÓGICA DE ARMAZENAMENTO.....	90
4.4.	<b>FASE IV – DESCOBERTA DE CONHECIMENTO .....</b>	<b>98</b>
4.4.1.	ETAPA 1 – ANÁLISE.....	99
4.4.2.	ETAPA 2 – TÉCNICAS DE VISUALIZAÇÃO .....	103
4.5.	ARMAZENAMENTO DO CONHECIMENTO .....	110
4.6.	ARQUITETURA RESULTANTE DO MODELO PROPOSTO .....	111
4.7.	CONSIDERAÇÕES FINAIS DO CAPÍTULO .....	113
<b>5.</b>	<b>APLICAÇÃO DO MODELO PROPOSTO.....</b>	<b>115</b>
5.1.	FASE I – DIAGNÓSTICO .....	116
5.2.	FASE II – HISTÓRICO E COMPREENSÃO DOS DADOS .....	118
5.2.1.	ETAPA 1 – IDENTIFICAÇÃO DAS FONTES DE DADOS .....	118
5.2.2.	ETAPA 2 – FACETAS PARA A FONTE DE DADOS.....	119
5.2.3.	ETAPA 3: AVALIAÇÃO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS .....	120
5.2.4.	ETAPA 4: DECISÃO DAS SOLUÇÕES TECNOLÓGICAS .....	126
5.3.	FASE III – SONDAGEM E USO DE TMT .....	130
5.3.1.	ETAPA 1: PREPARAÇÃO DOS DADOS.....	130
5.3.2.	ETAPA 2: SOLUÇÃO TECNOLÓGICA DE ARMAZENAMENTO .....	131
5.4.	FASE IV – DESCOBERTA DE CONHECIMENTO .....	132
5.4.1.	ETAPA 1: ANÁLISE .....	132
5.4.2.	ETAPA 2: TÉCNICAS DE VISUALIZAÇÃO.....	139
5.5.	ARMAZENAMENTO DO CONHECIMENTO .....	147
<b>6.</b>	<b>CONCLUSÃO.....</b>	<b>153</b>
6.1.	SUGESTÕES PARA TRABALHOS FUTUROS.....	156
	<b>REFERÊNCIAS .....</b>	<b>158</b>
	<b>APÊNDICE A – QUESTIONÁRIO DE DIAGNÓSTICO GERAL .....</b>	<b>172</b>
	<b>APÊNDICE B – FUNÇÃO INICIAR.....</b>	<b>175</b>
	<b>APÊNDICE C – FUNÇÕES LISTAR TERMO E LISTAR USUÁRIO .....</b>	<b>176</b>
	<b>APÊNDICE D – PRÉ-FILTRAGEM, PRÉ-LIMPEZA E PRÉ-TRANSFORMAÇÃO .....</b>	<b>178</b>
	<b>APÊNDICE E – PREPARAÇÃO DOS DADOS PARA ANÁLISE.....</b>	<b>179</b>
	<b>APÊNDICE F – MODELAGEM DE DADOS ESTRUTURADA PARA REDE SOCIAL .....</b>	<b>180</b>
	<b>APÊNDICE G – CONEXÃO COM <i>MONGODB</i> NA LINGUAGEM R. ....</b>	<b>181</b>
	<b>APÊNDICE H – ANÁLISE DE ASSOCIAÇÃO E AGRUPAMENTO .....</b>	<b>182</b>
	<b>APÊNDICE I – TÉCNICAS DE VISUALIZAÇÃO PARA AGRUPAMENTO.....</b>	<b>183</b>
	<b>APÊNDICE J – TÉCNICAS DE VISUALIZAÇÃO .....</b>	<b>184</b>
	<b>APÊNDICE K – QUESTIONÁRIO DE CONHECIMENTOS EXTRAIDOS.....</b>	<b>187</b>
	<b>ANEXO A – TABELA PERIÓDICA PARA TÉCNICAS DE VISUALIZAÇÃO .....</b>	<b>192</b>
	<b>ANEXO B – AMBIENTE PARA O BIG DATA.....</b>	<b>193</b>



## 1. INTRODUÇÃO

Atualmente, em virtude do crescente aumento no volume de informações, vem se instalando um novo cenário, em que, aliados às novas necessidades de consumidores cada vez mais exigentes, novos desafios surgem. As oportunidades geradas por esses novos desafios podem favorecer as tomadas de decisão e auxiliar na gestão do Ciclo de Vida do Produto (CVP). A competitividade e a velocidade dos processos nas empresas exigem que as tomadas de decisão e o desenvolvimento de estratégias sejam realizados com base em conhecimentos úteis e concretos. Assim, os gestores adquirem diferentes visões das mais variadas dimensões na dinâmica da empresa, passando a se interessar pela criação de um diferencial na relação entre empresa/cliente, o que cria a possibilidade de surgirem novas ideias para o desenvolvimento de produtos.

A descoberta de conhecimento por meio de dados estruturados e internos, que já era um possível suporte à decisão, se faz agora por meio de dados não estruturados ou semiestruturados, intitulado *Big Data*. Desde 2008, essa nova forma vem apresentando muitos avanços, que, por sua vez, geram expectativas e preocupações tecnológicas no que se refere a armazenamento, processamento, *frameworks*, entre outros. A literatura recente tem evidenciado a falta de modelos para apoiar o processo de descoberta de conhecimento no desenvolvimento do produto. Em vista disso, o *Big Data* oferece um diferencial competitivo, especialmente quanto a: identificar em tempo real os anseios dos consumidores; evidenciar os defeitos dos produtos; gerar oportunidades para a otimização dos produtos e apontar tendências e hábitos de consumo.

No que se refere à análise de dados, anteriormente ao surgimento do *Big Data*, as empresas e os pesquisadores tinham a atenção voltada para a descoberta de conhecimento em banco de dados (em inglês, *Knowledge Discovery in Database* ou KDD), desde a obtenção (pré-processamento), até o processamento (mineração de dados) e a interpretação dos dados (pós-

processamento). No entanto, dentre essas etapas, o processamento dos dados, conhecido e denominado como Mineração de Dados (MD), apresenta mais destaque no que se refere ao interesse científico. A maioria dos trabalhos realizados sobre MD utiliza dados internos, produzidos pelos sistemas de informação computacional da própria empresa. Nesse caso, a matéria prima (dados) está organizada, estruturada e armazenada em Sistemas de Gerenciamento de Banco de Dados (SGBD). Diferentemente da MD, o *Big Data* faz uso de dados não estruturados e heterogêneos, isto é, dados documentais (textos, planilhas, *slides*, vídeos, fotos, dentre outros) que estão armazenados em lugares distintos e, muitas vezes, desconhecidos.

Com o advento do *Big Data*, surge a oportunidade de se obterem novas visões e dimensões em relação aos processos e produtos da organização e, assim, de se revolucionar a tomada de decisão durante a fabricação e a venda desses produtos. Porém, isso só ocorrerá com a aceitação dessa nova forma, juntamente com a aplicação de esforços para o aprendizado das técnicas, das tarefas e dos métodos que envolvem a descoberta de conhecimento e de toda a filosofia de base dessas atividades. Isso porque não é suficiente dispor dos recursos físicos; é necessário também mudar a cultura e suas práticas. Li *et al.* (2015) constatam que a tarefa no futuro será abordar problemas de como aplicar detalhadamente técnicas avançadas em dados complexos.

As informações geradoras de conhecimento são relevantes para o Processo de Desenvolvimento do Produto (PDP) e, nesse contexto, o *Big Data* altera a visão e o foco sobre a matéria prima existente para a extração do conhecimento. A literatura destaca a elevada quantidade de dados produzidos, além da ausência de trabalhos relatando a aplicação do *Big Data* como solução de apoio ao PDP.

### **1.1. JUSTIFICATIVA E RELEVÂNCIA**

Ao longo do CVP, por meio de diversas ferramentas e fontes com diferentes formatos estruturais e, a cada momento, com mais velocidade, as empresas e

consumidores produzem um volume significativo de dados referentes ao produto. Portanto, como o interesse no tema “descoberta de conhecimento” tem evoluído e se inovado nos últimos anos, as possibilidades de aplicação do *Big Data* são inúmeras. Pesquisa desenvolvida por McAfee e Brynjolfsson (2012) mostra as vantagens adquiridas por empresas que utilizam o *Big Data* e concluem que sua aplicação eleva a produtividade e consequentemente os lucros dessas empresas.

No que se refere ao CVP, dentre os vários objetivos da abordagem denominada “gestão do ciclo de vida do produto”, destaca-se o gerenciamento das informações, cujo propósito é a melhoria dos processos ao longo de todas as suas fases. Essa abordagem contempla métodos e ferramentas de apoio ao PDP, desde a concepção da ideia, até o produto final, incluindo ainda o processo de descarte quando do término de sua vida útil. Para auxiliar o PDP, foram elaborados modelos de referência, incluindo fases de formalização e sistematização dos processos de desenvolvimento. Nesse contexto, o projeto informacional é uma das fases iniciais do modelo de referência, tendo o importante papel de definir as especificações do produto e o detalhamento dos requisitos, conhecimentos esses que serão utilizados como base para as demais atividades do PDP (ROZENFELD *et al.*, 2006; BACK *et al.*, 2008). Como fonte de entrada para essa fase, pode-se utilizar dados externos à empresa, como, por exemplo, as redes sociais, cujos dados podem se mostrar relevantes, se explorados.

Li *et al.* (2015) mostram a aplicabilidade do *Big Data* na indústria e no CVP e enfatizam a necessidade de maiores estudos.

Assim, a escolha por tal temática é justificada pela ausência de métodos e de discussões acadêmicas mais aprofundadas a respeito desses instrumentos de obtenção de conhecimentos para o desenvolvimento de produtos. O que existe atualmente está em formato genérico, isto é, não é específico para o apoio no PDP. Então, as motivações para essa pesquisa advêm da possibilidade de descoberta de conhecimento com dados mal aproveitados.

O interesse pela utilização de dados complexos para a descoberta de conhecimento tem avançado desde 2008, porém com um salto significativo no ano de 2013 e, no momento, ocupa as pautas de discussões em encontros, debates e congressos acadêmicos.

O potencial da exploração de dados para o modelo de PDP e a etapa do projeto informacional que, de acordo com Rozenfeld *et al.* (2006), requer informações mais abrangentes para ser utilizada como base para critérios de avaliação e de tomadas de decisão ao longo do PDP, evidenciam a relevância da presente pesquisa.

## **1.2. PROBLEMA DA PESQUISA**

Da lacuna identificada, surge o problema de pesquisa.

***Como desenvolver um modelo de descoberta de conhecimento que, com o auxílio de soluções tradicionais e do Big Data, possa apoiar o projeto informacional do processo de desenvolvimento do produto?***

## **1.3. OBJETIVO GERAL E ESPECÍFICO**

O objetivo deste trabalho é propor um modelo de descoberta de conhecimento que abranja, do início ao fim, as atividades necessárias para o processo de extração de conhecimentos novos e úteis para auxiliar o projeto informacional.

Para atingir esse objetivo geral, foram definidos os seguintes objetivos específicos:

- identificar as fontes de dados internas, ou seja, geradas pelos processos de produção, e também as possíveis fontes externas presentes na rede mundial de computadores com potencial para apoiar as atividades do PDP;
- identificar, na literatura, as soluções tecnológicas que podem se complementar e, assim, contribuir para o desenvolvimento do modelo proposto;

- demonstrar o funcionamento do modelo por meio de aplicação prática (nível de protótipo – Linguagem R) em cenário real de produção.

#### 1.4. ESTRUTURA DA TESE

A estrutura do trabalho, em seis capítulos, está disposta conforme mostra a Figura 1.

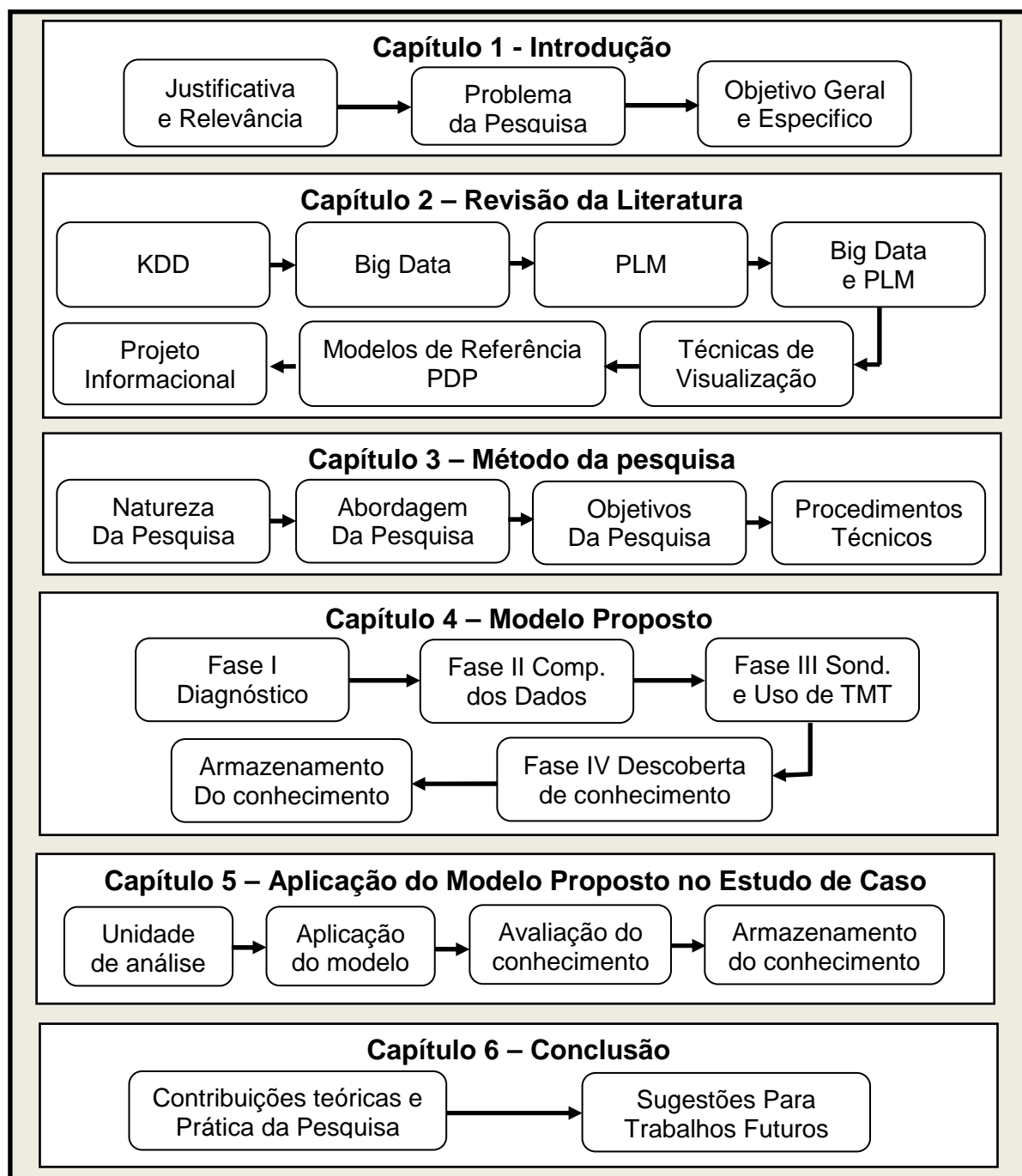
O primeiro capítulo é dedicado à formulação do tema da pesquisa, com destaque para os fatores motivacionais, as justificativas, a problemática da pesquisa, o objetivo geral e os objetivos específicos, além dos aspectos teóricos que justificam a opção temática da pesquisa. Nesse capítulo, discutem-se também a organização e a estruturação do trabalho.

No segundo capítulo, aborda-se primeiramente o processo da descoberta de conhecimento e a apresentação do modelo Crisp-DM (*Cross Industry Standard Process for Data Mining*). Ao longo do desenvolvimento da descoberta de conhecimento, são apresentadas as discussões sobre o *Big Data* e sua definição, com ênfase nas questões críticas e respectivas soluções, assim como as tecnologias relacionadas ao *Big Data*. Inclui-se uma discussão a respeito da relação entre o *Big Data* e a Gestão do CVP, com destaque para a produção de dados internos e externos. Para melhor interpretação dos resultados da descoberta de conhecimento, são discutidas as técnicas de visualização. Apresentam-se ainda dois modelos de referência no cenário nacional para o PDP, os quais se referem ao processo de negócio relacionado à gestão do CVP.

No terceiro capítulo, são expostos os procedimentos metodológicos utilizados nas atividades da pesquisa, com a descrição detalhada do tipo de estudo, o delineamento do roteiro metodológico e, por fim, os procedimentos técnicos.

No quarto capítulo, detalham-se as fases e etapas do modelo proposto, com um relato sobre o armazenamento do conhecimento extraído e sobre a

construção de uma arquitetura para descoberta de conhecimento com base no modelo proposto.



**FIGURA 1 – VISÃO MACRO DA PESQUISA**  
 FONTE: ELABORADA PELO AUTOR

No quinto capítulo, relatam-se os resultados da aplicação do modelo proposto em uma indústria de confecção, na qual as avaliações dos conhecimentos

foram realizadas por seus próprios colaboradores. Os mesmos conhecimentos também foram avaliados por colaboradores de outra indústria de maior porte.

No sexto capítulo apresentam-se as contribuições teóricas e práticas da pesquisa e sugestões de trabalhos futuros.

## 2. REVISÃO DA LITERATURA

Neste capítulo, discute-se o referencial teórico utilizado no trabalho.

Inicialmente, são abordados os assuntos relacionados à descoberta de conhecimento e ao recente tema denominado *Big Data*. Após essa apresentação, discute-se a gestão do ciclo de vida do produto (em inglês, *Product Lifecycle Management* ou PLM) e sua relação com o *Big Data* e, apresentam-se os resultados de um levantamento da produção e das fontes internas e externas de dados relacionadas ao CVP. Na sequência, é descrita a análise de facetas no *Big Data* e, algumas técnicas de visualização para a interpretação dos resultados produzidos pelo modelo proposto. Por fim, são discutidos os modelos de referência de Back *et al.* (2008) e Rozenfeld *et al.* (2008) para o PDP e identificado as atividades do projeto informacional, a qual o modelo proposto pode ser aplicado.

### 2.1. DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Fayyad *et al.* (1996b) dissertaram sobre o elevado ritmo na coleta e no acúmulo de dados e alertaram para a urgente necessidade de uma nova geração computacional e de ferramentas tecnológicas que auxiliem na extração de conhecimentos do crescente volume de dados digitais.

Tal preocupação permanece atual, pois, embora os avanços tecnológicos tenham apoiado as descobertas de conhecimento, estas geraram novas oportunidades e criaram novos desafios. Dessa forma, foram iniciados trabalhos efetivos de mineração de dados para o KDD.

O termo KDD foi formalizado em 1989 para nomear a abordagem destinada a atender aos processos referentes à busca de conhecimento a partir de bases de dados. Fayyad *et al.* (1996b, p. 30) propuseram a definição que se tornou a mais popular na literatura: “*KDD é um processo, de várias etapas, não trivial,*

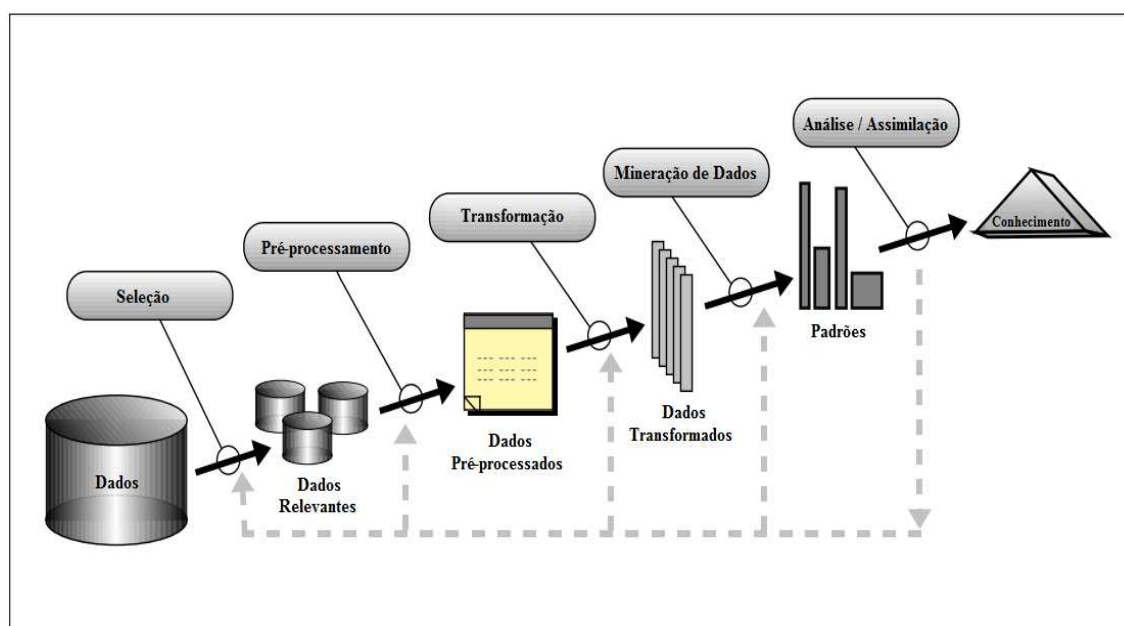


*interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”.*

Duas décadas depois, a análise dessa definição permite que o mesmo propósito seja mantido quando se trata de associá-la às novas demandas do *Big Data*. Hashem *et al.* (2015) argumentam que a natureza do *Big Data* é indistinta e envolve consideráveis processos para identificar e converter os dados em novas ideias. Os termos interativo e iterativo indicam a atuação do homem na realização dos processos; ele é o responsável pela utilização das ferramentas computacionais para buscar, analisar e interpretar os dados.

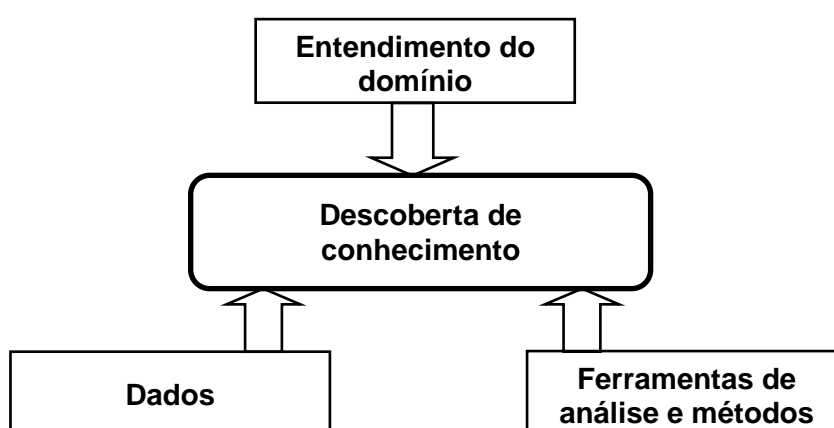
Em suma, o KDD é empregado na identificação de padrões por meio da manipulação de dados (AGRAWAL *et al.*, 1993; BRACHMAN *et al.*, 1996; MITCHELL, 1999). Enquanto Fayyad *et al.* (1996a) asseguram que o processo KDD, como mostrado na Figura 2, é fundado em várias etapas não triviais, Feldens *et al.* (1998) apresentam as etapas iterativas e interativas para a identificação de padrões, sintetizando-as em:

- pré-processamento - atividades que visam gerar uma representação conveniente para os algoritmos de mineração, a partir da base de dados, incluindo a seleção (automática e/ou manual de atributos relevantes), a amostragem e as transformações de representação;
- mineração de dados (MD) - aplicação de algoritmos de mineração aos dados pré-processados, ou seja, busca efetiva por conhecimentos úteis a partir dos dados;
- pós-processamento - seleção e ordenação das descobertas interessantes e dos mapeamentos de representação de conhecimento; é nesta etapa que o especialista em KDD e o especialista no domínio de aplicação avaliam os resultados obtidos e criam novas alternativas para as investigações de dados.



*FIGURA 2 – ETAPAS DO PROCESSO KDD*  
 FONTE: ADAPTADO DE FAYYAD ET AL. 1996A, P.29)

O estudo de cada etapa do processo KDD requer a prática da coleta de dados, do armazenamento, da organização, utilização e desenvolvimento de ferramentas tecnológicas. Isso favorece a compreensão e a aplicação efetiva dos métodos analíticos, bem como o entendimento da estrutura da natureza do problema e do significado dos dados implícitos (BEGOLI e HOREY, 2012). A Figura 3 ilustra os elementos necessários para a obtenção do conhecimento.



*FIGURA 3 – ELEMENTOS DO PROCESSO DE DESCOBERTA DE CONHECIMENTO*  
 FONTE: TRADUZIDO DE BEGOLI E HOREY (2012, PAG. 1)

Em virtude do poder de processamento e armazenamento, a velocidade do avanço tecnológico nos setores de comunicação proporciona um elevado

volume de dados. Ao mesmo tempo, a evolução da computação nas áreas de sistema distribuído e paralelo oferece inúmeras possibilidades de simulação de modelos complexos e de descoberta de conhecimentos. Com a taxa de crescimento e de velocidade das informações, que são processadas, transferidas e compartilhadas na velocidade da luz, surgem inúmeras variáveis que desafiam a ciência, como: volume, velocidade, heterogeneidade e segurança (KHAN *et al.*, 2014).

A forma como os dados são controlados e tratados está sofrendo mudanças, já que, como na nova era o mundo está em transformação, as constantes possibilidades de previsão com base em dados criam desafios que podem alterar nossas instituições e a nós mesmos (MAYER-SCHONBERGER e CUKIER, 2013)

#### **2.1.1. MINERAÇÃO DE DADOS**

Os termos MD e KDD têm duas definições populares: na primeira, os termos são considerados sinônimos; na segunda, a MD é um passo importante do KDD (ZHOU, 2003).

Neste trabalho, assume-se a segunda definição, ou seja, de que MD faz parte do processo KDD, que é uma das etapas de reconhecimento de padrões e regras de dados que estão armazenados em grandes bancos de dados (FELDENS *et al.*, 1998; FAYYAD *et al.*, 1996a; BERRY e LINOFF, 1997). Esses autores, além de corroborar essa definição, argumentam que a mineração de dados, envolvendo diversas áreas, dentre as quais marketing, vendas e suporte, deve fornecer conhecimentos às corporações, apoiando o desenvolvimento de estratégias e a melhoria dos negócios. A MD é composta por um conjunto de tarefas avaliadas por sua capacidade de realizá-las (LAROSE, 2014):

- classificação – função de associar cada registro de um banco de dados em rótulos categóricos; por exemplo, identificar e classificar regiões com mais probabilidade de surgimento de uma determinada doença;

- associação – função de identificar associações entre conjuntos de dados; por exemplo, associar itens de um carrinho de compras;
- regressão – função similar à de classificação; deve-se mapear os registros e fazer uso de valores numéricos; por exemplo, identificar a quantidade de calorias que devem ser gastas por um atleta, baseando-se no tipo de exercícios relacionados à idade, ao gênero e à massa corporal;
- agrupamento – função de agrupar os registros com base em critérios de semelhança; por exemplo, agrupar consumidores com comportamento de compra similar;
- predição - similar às tarefas de classificação; o intuito é identificar valores futuros de uma determinada aplicação e, por exemplo, prever o aumento das vendas com base em atributos dos produtos e consumidores,
- descrição - função para identificar e apresentar de forma concisa os padrões e as tendências revelados pelos dados; por exemplo, identificar as características do consumidor de acordo com a natureza da loja, faixa etária, renda anual, grau de instrução, região em que reside.

Zhou (2003) analisou as definições da MD de três perspectivas: *i)* estatística, *ii)* aprendizado de máquina, *iii)* banco de dados.

Adotando a perspectiva estatística, Hand *et al.* (2001) argumentam que, primeiramente, os resultados devem ser compreensíveis, para que, então, se possam encontrar relacionamentos inesperados.

Cabena *et al.* (1998), da perspectiva de banco de dados, referem-se à interdisciplinaridade para reconhecimento de padrões em grandes bases de dados.

Fayyad *et al.* (1996a), considerando a perspectiva do aprendizado de máquina na definição da MD, descrevem as limitações computacionais para a produção de conhecimentos.

Abordando-se essas três perspectivas, podem-se obter resultados de mineração de dados bem sucedidos (ZHOU, 2003). Em virtude do aumento frenético na massa de dados, surge a abordagem denominada “*Big Data Mining*”, que reforça essa argumentação. Leung *et al.* (2014) e Kumar *et al.* (2013) utilizam a mineração de dados integrada a técnicas de diversas áreas, tais como, computação em nuvem, aprendizagem de máquina, matemática e estatística.

### 2.1.2. USO DE MÍDIAS SOCIAIS PARA DESCOBERTA DE CONHECIMENTO

A proliferação do uso das mídias sociais<sup>1</sup> na sociedade é inegável, bem como os estudos sobre a importância e a atenção que elas vêm ganhando ultimamente na comunidade científica.

Segundo Lahuerta-Otero e Cordero-Gutierrez (2016), a cada segundo, doze novos usuários móveis se juntam às plataformas de redes sociais, isto é, um milhão de pessoas por dia. Diante desses elevados números, as empresas perceberam a importância das redes sociais e passaram a utilizá-las como ferramentas de relacionamento com o cliente no mundo digital. Dentre as plataformas de redes sociais, os *microblogging*<sup>2</sup> destacam-se pelas dinâmicas dos conteúdos produzidos por seus usuários.

*Twitter* é uma plataforma de *microblogging*. Dos 313 milhões de usuários mensais ativos, 82% acessam por meio de dispositivos móveis (TWITTER, 2017). Como a maioria desses usuários produz seus conteúdos sem restrição de privacidade, é possível extrair um elevado volume de postagens e de atributos por meio da API<sup>3</sup> (*Application Programming Interface*) do próprio *Twitter*.

---

<sup>1</sup> Espaços interativos entre usuários em blogs, fóruns, sistemas de mensagem instantânea, wikis, redes sociais (*facebook*, *linkedin*, *twitter*, entre outros) e conteúdos multimídia (*youtube*, *flickr*, entre outros).

<sup>2</sup> Site em que os usuários postam mensagens curtas (aproximadamente 140 caracteres) para visualização por meio de uma rede de pessoas.

<sup>3</sup> Instruções e padrões de programação para acesso a um aplicativo ou software.

A literatura apresenta trabalhos que utilizam as redes sociais para a descoberta de conhecimento. Dentre eles, destaca-se o de Oliveira *et al.* (2017), que trabalham com métodos de análise de sentimento e preveem variáveis para o mercado de ações. Lahuerta-Otero e Cordero-Gutierrez (2016) combinam a teoria dos grafos com a teoria da influência social com o intuito de descobrir as características das postagens dos formadores de opinião no *Twitter*, no que se refere a duas empresas automotivas japonesas (Toyota e Nissan). Amaral e De Pinho (2017) utilizaram a mídia social para descrever e analisar as postagens de parlamentares e de partidos políticos brasileiros e, dessa maneira, mostrar, com base em uma classificação ideológica dos partidos políticos, se essa é uma variável que distingue a adoção e o uso do *Twitter* por parte dos parlamentares do congresso brasileiro.

### **2.1.3. MODELO CRISP-DM**

O modelo Crisp-DM estabelece um conjunto de regras e tarefas para a orientação do processo de descoberta de conhecimento. Esse é um dos modelos de maior aceitação, especialmente para o desenvolvimento de atividades de MD (PIATETSKY, 2014).

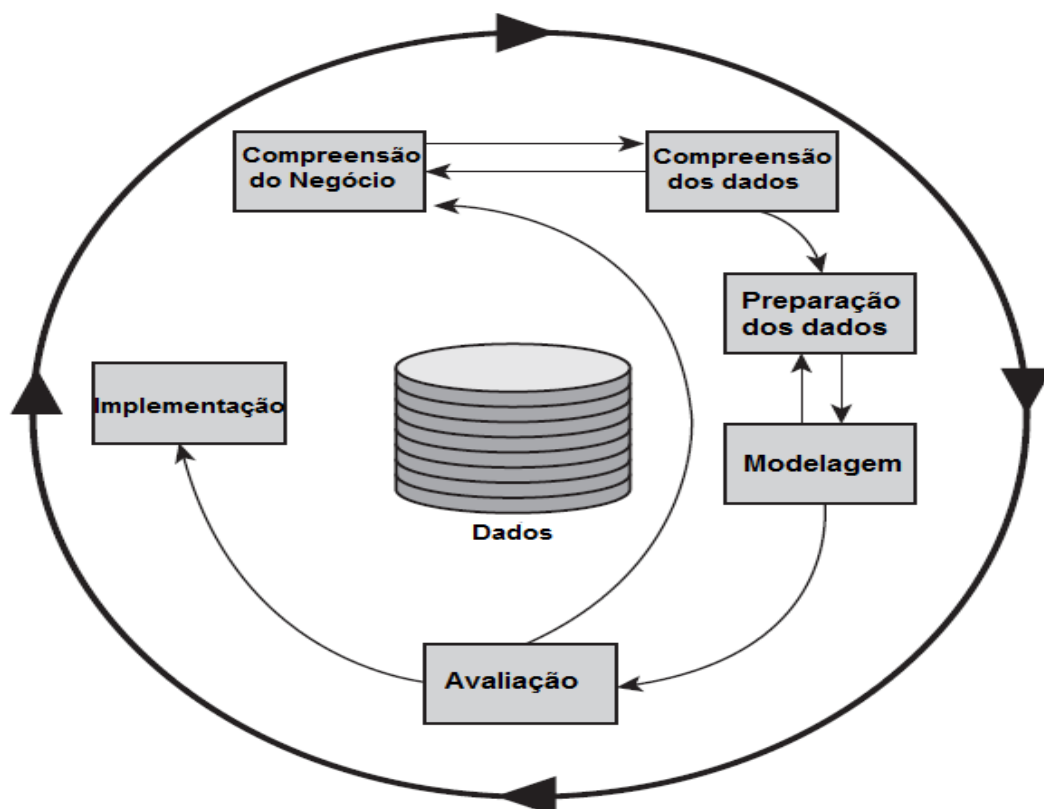
Proposto inicialmente em 1996, o modelo foi criado para melhorar os processos de MD em ambiente real, no sentido de apoiar decisões de negócios (CHAPMAN *et al.*, 2000). O processo de MD compõe-se de seis fases, como ilustra a Figura 4, as quais são descritas em Piatetsky (2014) e Goldschmidt *et al.* (2015).

A fase de compreensão do negócio no modelo Crisp-DM consiste em estudar melhor o processo de negócio e, assim, conhecer o contexto em que o processo de descoberta de conhecimento será aplicado na empresa.

A compreensão dos dados abrange a coleta dos dados, a análise e a avaliação das informações no sentido de seu significado e de sua qualidade; no entanto, o exame é restrito à regularidade técnica dos dados.

Na fase de preparação dos dados, promovem-se a limpeza e a adequação dos dados, de forma a construir o conjunto final.

Na sequência, os dados são aplicados por meio de técnicas de MD, que é a fase de modelagem, cujo foco está na seleção adequada da técnica. Caso seja necessário, essa fase prevê o retorno à fase anterior.



*FIGURA 4 – FASES DO MODELO CRISP-DM*  
*FONTE: TRADUZIDO DE CHAPMAN ET AL. (2000, P.10)*

Com os resultados obtidos nas fases anteriores, realiza-se a fase de avaliação da viabilidade do modelo.

Por fim, na fase de implementação, promove-se o desenvolvimento do modelo, além da produção de relatórios e outras análises.

O modelo Crisp-DM foi desenvolvido para auxiliar no processo de descoberta de conhecimento para dados estruturados. No entanto, esse modelo se encontra desatualizado em relação aos recentes desafios da descoberta de

conhecimento e não atende adequadamente às novas exigências do *Big Data* (ASAMOA e SHARDA, 2015; PIATETSKY, 2014).

## **2.2. BIG DATA**

Com a evolução da comunicação, não apenas foram desenvolvidas novas tecnologias como também aumentou significativamente a quantidade de dados acessíveis. Nesse contexto, os pesquisadores perceberam que a maior parte das informações existentes no mundo se origina de postagens em massa, mas infelizmente esse crescimento não se manteve proporcional às tecnologias de armazenamento (KAISLER *et al.*, 2013).

A sociedade produz e consome um volume significativo de dados, aqui denominados de matéria prima, os quais, com velocidade cada vez maior, são gerados por diversas ferramentas e adquirem diferentes formatos estruturais.

Mcafee e Brynjolfsson (2012) afirmam que, por dia, ocorre um aumento de 2,5 exabytes na produção de dados; Davenport (2014), por sua vez, ressalva que o mundo utilizou 2,8 zetabytes de dados em 2012, mas que apenas 0,5% desses dados foram analisados de alguma forma. Esse autor estima que aproximadamente 25% deles têm valor potencial e reconhece que essa estimativa é modesta quando se considera a quantidade de dados disponíveis.

Dervojeda *et al.* (2013) afirmam que aproximadamente 90% dos dados produzidos no mundo correspondem aos dois últimos anos e que 20% desses dados são numéricos. Essa evidência incentiva a realização de pesquisas sobre a área de MD, inclusive algoritmos de dados quantitativos.

De acordo com Gantz e Reinsel (2012), a previsão é de que, em 2020, serão produzidos mais de 40.000 exabytes (mais de 5.200 gigabytes para cada habitante no mundo). Khan *et al.* (2014) relatam que a expectativa para 2020 é de que 50 bilhões de novos dispositivos sejam conectados à internet e preveem que a produção de dados será 44 vezes maior do que em 2009.



O *Big Data*, mesmo com alguns vestígios de ceticismo, ganhou a atenção da academia. Segundo alguns autores, tais desconfianças desaparecerão com o passar do tempo, já que esse método tem relevância considerável para muitas organizações (KHAN *et al.* 2014; DAVENPORT, 2014; LI *et al.*, 2015).

As definições de *Big Data* na literatura convergem quanto aos seguintes fatos: utilização de diferentes fontes de dados e características como tipo de dados, volume, velocidade e variedade (MANYIKA *et al.*, 2011; IBM, 2011; BEGOLI e HOREY, 2012; MCFEE e BRYNJOLFSSON, 2012; KAISLER *et al.*, 2013; DAVENPORT, 2014; LI *et al.*, 2015; GANTZ e REISEN, 2012). Estendendo a definição, Zikopoulos e Eaton (2011) acrescenta a característica veracidade e Kaisler *et al.* (2013) mencionam as características valor e complexidade. Davenport (2014) agrega venalidade, isto é, a possibilidade de ser vendido.

Kaisler *et al.* (2013) assumem que a definição para o *Big Data* se refere à quantidade de dados que está além da capacidade tecnológica nas tarefas de armazenar, gerenciar e processar de forma eficiente. Gantz e Reinsel (2012) desenvolveram pesquisa para uma das grandes organizações especializadas em tecnologia de informação e desenvolvimento de software dos EUA, denominada IDC (*International Data Corporation*). Afirmam eles que *Big Data* seria:

“..uma nova geração de tecnologias e arquiteturas, concebido para extrair economicamente valor a partir de grande volume de dados, de uma ampla variedade de dados, permitindo uma alta velocidade de captura, descoberta e análise.” (GANTZ e REINSEL, 2012, p.9).

Portanto, nessa definição, o *Big Data* se caracteriza por “V’s”.

### **Volume dos dados**

Demchenko *et al.* (2013) afirmam que o volume é a característica de maior relevância no *Big Data*, cujos requisitos adicionais e específicos (tamanho, escala, quantidade e dimensão dos dados) impõem limites às tecnologias tradicionais de descoberta de conhecimento. O volume dos dados não se restringe ao disponível internamente na organização: compreende também o acesso a dados externos (KAISLER *et al.*, 2013). Complementarmente, Qiu *et*

al. (2016) ponderam que as tendências do desenvolvimento tecnológico propiciam às organizações o armazenamento e a análise dos dados na magnitude de petabyte a exabyte. O estatístico Nate Silver, em uma conferência promovida pela empresa HP (*Hewlett-Packard*) na cidade de Boston, apresentou os problemas que acompanham o *Big Data* no que se refere ao volume (BUTLER, 2015):

- problema relacionado a armazenamento, gerenciamento dos dados e tempo de resposta;
- diferentes visões – resultados imprecisos: quanto maior o volume de dados maiores são as chances de distorção ou as manobras de respostas para resultados de interesse;
- falso positivo – a dificuldade para analisar os dados por completo pode gerar falsas conclusões;
- complexidade – quanto maior o volume de dados, maiores são as dificuldades para se encontrar algo valioso.

### **Velocidade dos dados**

Refere-se à dinâmica de criação, transferência e aglomeração dos dados (KAISLER *et al.*, 2013), as quais ocorrem em “tempo real”. McAfee e Brynjolfsson (2012) afirmam que a velocidade agrega a possibilidade de vantagem competitiva nas atividades de extração e processamento em tempo real.

### **Variedade dos dados**

A análise de dados em formatos não estruturados ou semiestruturados pode ser complexa. Em razão disso, os dados devem ser previamente estruturados (KHAN *et al.* 2014). Consequentemente, diante da possibilidade de tratamento dos dados, os algoritmos tradicionais de MD são capazes de localizar padrões desconhecidos.

Os diferentes formatos de dados, não estruturado e semiestruturado, como textos, imagens, vídeos e áudios, oferecem desafios significativos diante das

exigências impostas pela gestão do armazenamento e pela arquitetura de banco de dados (KAISLER *et al.*, 2013, DEMCHENKO *et al.*, 2013).

A relação entre dados estruturados e não estruturados é descrita por KHAN *et al.* (2014) e ilustrada na Tabela 1.

*TABELA 1 – DADOS ESTRUTURADOS VS DADOS NÃO ESTRUTURADOS*  
*FONTE: TRADUZIDO DE KHAN ET AL. (2014 P. 12)*

	<b>Dados estruturados</b>	<b>Dados não estruturados</b>
<b>Formato</b>	Linhas e colunas	Volumosos objetos binários
<b>Armazenamento</b>	SGBD	Documentos não gerenciados e arquivos não estruturados
<b>Metadados<sup>4</sup></b>	Sintático	Semântico
<b>Ferramentas</b>	Tradicionais para MD	Processamento em lotes

### **Valor dos dados**

De acordo com Kaisler *et al.* (2013), a característica valor está relacionada à sua utilidade no processo de tomada de decisão. Ou seja, essa característica assume o mesmo propósito do termo útil, definido no processo KDD por Fayyad *et al.* 1996b. Complementarmente, Demchenko *et al.* (2013) destacam a relevância do valor e a relação com as características volume e variedade.

### **Veracidade dos dados**

Obviamente, todas as características do *Big Data* são pertinentes, no entanto, tornam-se irrelevantes caso os dados não sejam confiáveis.

Segundo Demchenko *et al.* (2013), isso depende de aspectos como origem, métodos de extração e tratamento de dados, os quais estão associados à confiabilidade dos dados no que diz respeito à segurança da fonte, à integridade e à autenticidade.

<sup>4</sup> Dados que descrevem outros dados.

### 2.2.1. MINERAÇÃO DE DADOS E *BIG DATA*

No contexto do *Big Data* os dados, as lógicas e os algoritmos de MD concentrados em uma única memória não são apropriados para a descoberta de conhecimento. Wu *et al.* (2014) ponderam que a execução das tarefas dos algoritmos tradicionais de MD em um único computador é suficiente para o cumprimento da meta de descoberta de conhecimento, pois eles são projetados para essa finalidade. Diferentemente das soluções tradicionais, o *Big Data* requer novos modelos computacionais, isto é, soluções de computação paralela e mineração coletiva para poder agregar os dados de diferentes fontes.

Para descoberta de conhecimento no *Big Data*, a composição do algoritmo de MD apresenta métodos de divisão e conquista, utilizados em ambiente de programação paralela (JI *et al.*, 2012; LUO *et al.*, 2012). Com interesse em atender às demandas do *Big Data*, surgiram arquiteturas para aumentar o processamento em tempo real, como a plataforma de *software* denominada *Apache Hadoop*, que possui código aberto e é voltada para clusterização e processamento de grande volume de dados (WU *et al.*, 2014).

Os algoritmos de MD se embasam em modelos capazes de sumarizar e realizar previsões. Essas abordagens fazem parte de uma subárea da Inteligência Artificial (IA), denominada aprendizagem de máquina (em inglês, *Machine Learning*), e são dedicadas ao desenvolvimento de técnicas e algoritmos para a construção de sistemas automatizados que utilizam os dados para o seu aprendizado (BRINK *et al.*, 2016). Zhou (2003) destaca a importância da contribuição oferecida pela comunidade científica nas pesquisas de MD utilizando a aprendizagem de máquina.

A IA possibilita às organizações a ampliação de conhecimentos e a melhoria na interação com os clientes. Existem casos em que a IA é capaz de substituir departamentos inteiros, especificamente a subárea de aprendizagem de máquina, tornando possível a execução de tarefas que antigamente eram reservadas a seres humanos (DERVOJEDA *et al.*, 2013).

### **2.2.2. APRENDIZAGEM DE MÁQUINA NA DESCOBERTA DE CONHECIMENTO**

A aprendizagem de máquina é dividida em aprendizagem supervisionada e não supervisionada (BRINK *et al.*, 2016), além da aprendizagem por esforço (ADAM e SMITH 2008).

De acordo com Brink *et al.* (2016), os algoritmos de aprendizado supervisionado são os mais comuns e requerem treinamentos com dados já classificados para induzir um modelo que gere novas classificações. Já, a aprendizagem não supervisionada, não necessita de dados treinados. A aprendizagem por esforço permite o aprendizado por meio do retorno dos resultados recebidos, ou seja, por meio de interações com o ambiente (ADAM e SMITH 2008).

De forma sintetizada, pode-se dizer que tanto a aprendizagem supervisionada, quanto a não supervisionada se concentram na análise de dados, na medida em que a aprendizagem por reforço trata problemas referentes à tomada de decisões.

A aprendizagem de máquina apresenta dificuldades relativas à memória e ao volume de dados; em razão disso, surge a necessidade de estudos e do desenvolvimento eficiente de ferramentas para suprir o elevado volume de dados (QIU *et al.*, 2016; BRINK *et al.*, 2016).

### **2.2.3. QUESTÕES CRÍTICAS SOBRE MÁQUINA DE APRENDIZAGEM PARA *BIG DATA***

Os métodos de aprendizagem de máquina são potencialmente significativos no apoio ao desenvolvimento de algoritmos para descoberta de conhecimento no *Big Data*. Esse argumento é confirmado por Qiu *et al.* (2016), que ressaltam que a aprendizagem de máquina é promissora, mas apresentam algumas questões críticas das técnicas desse tipo de aprendizagem em diferentes perspectivas, como ilustrados na Figura 5.

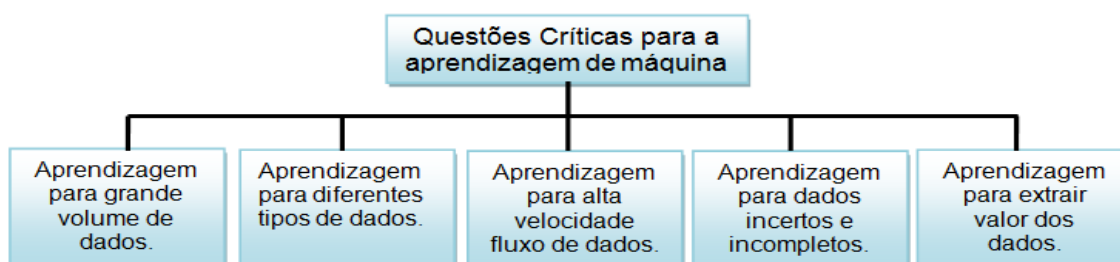


FIGURA 5 – QUESTÕES CRÍTICAS DE MÁQUINA DE APRENDIZAGEM PARA *BIG DATA*  
 FONTE: TRADUZIDO DE QIU ET AL. (2016, P. 6)

Observa-se que essas questões associam-se às definições dadas pelos “V’s”, utilizadas na literatura para definir o *Big Data*.

Em relação ao volume, Qiu *et al.* (2016) argumentam que essa principal característica apresenta grandes desafios para a aprendizagem de máquina. A utilização de *framework* oferece suporte à computação paralela e distribuída e pode ser solução para trabalhar com volume de dados. Goldschmidt *et al.* (2015) discutem duas áreas relacionadas ao *Big Data*: mineração de dados paralela (em inglês, *Parallel Data Mining* ou PDM) e mineração de dados distribuída (em inglês, *Distributed Data Mining* ou DDM). A PDM é ideal para organizações que trabalham com dados centralizados e a DDM é interessante para dados distribuídos.

A velocidade torna-se uma característica emergente no contexto da aprendizagem de máquina, sendo imprescindível na execução de tarefas em um determinado período de tempo para não tornar o conhecimento menos valioso ou inútil (QIU *et al.*, 2016). Quanto ao ponto crítico da velocidade, uma das soluções apresentadas é a da aprendizagem *on-line*. De acordo com Shalev-Shwartz (2012), nesse paradigma de aprendizagem, utilizam-se recursos teóricos e práticos com o objetivo de realizar uma sequência de previsões para tarefas que tornem os dados disponíveis e sequenciais. Essas tarefas são utilizadas para prever dados futuros.

Qui *et al.* (2016) afirmam que a variedade de dados é o que torna o *Big Data* desafiador e que é perceptível o grau de complexidade no contexto da

aprendizagem de máquina. Khan *et al.* (2014) corroboram essa afirmação e sugerem a integração de dados como uma solução.

Para compreender melhor essa sugestão, toma-se aqui o exemplo discutido por Hendler (2014) para integrar dados de diferentes fontes. Nesse exemplo, os EUA e a China têm publicado informações abertas sobre os seus respectivos PIB (Produto Interno Bruto) por vários anos, no entanto, os valores publicados pelos EUA são em dólares e na China *yuan*. Para a integração desses dados, é necessário um terceiro mapeamento que proporcione a integração das fontes, isto é, um intermediador, que, nesse caso, pode ser o sistema *on-line* “*Federal Reserve*” do governo do EUA. Com isso, é possível realizar cálculos matemáticos simples sobre o conjunto de dados extraídos das diferentes fontes, para conversão e comparações dos valores.

O ponto crítico relatado por Qiu *et al.* (2016) é conseguir extrair conhecimentos a partir do alto volume de dados; por exemplo, encontrar comentários úteis de consumidores de uma determinada região sobre um produto de interesse, dentro de um conjunto que integra diversos comentários sobre outro produto realizados por consumidores que moram em outras regiões. Nesse sentido, WU *et al.* (2014) propõem um teorema denominado HACE, que é composto pelos seguintes aspectos: heterogeneidade proveniente de fontes autônomas com controle distribuído e descentralizadas para explorar o complexo e a evolução das relações entre os dados. Este teorema sugere um modelo de processamento de *Big Data*, a partir de uma perspectiva de mineração de dados; trata-se de uma referência orientada para dados que envolvem a agregação baseada na procura de fontes de informação, mineração e análise.

Em teoria, é possível considerar confiáveis os resultados provenientes de dados internos e estruturados, simplesmente pelo fato de que geralmente pertencem a fontes conhecidas. Entretanto, essa realidade não se aplica ao *Big Data*, pois o controle e o gerenciamento dos dados podem não ser conhecidos pela organização, de forma que a precisão e a confiança tornam-se um problema. Qiu *et al.* (2016) argumentam que uma das soluções para essa

questão crítica é a utilização da estatística por meio dos cálculos de médias e variâncias para distribuição amostral.

Por fim, na Figura 6, apresenta-se um resumo das questões críticas do *Big Data* e suas respectivas soluções. No entanto, tal descrição não esgota os desafios e as soluções.

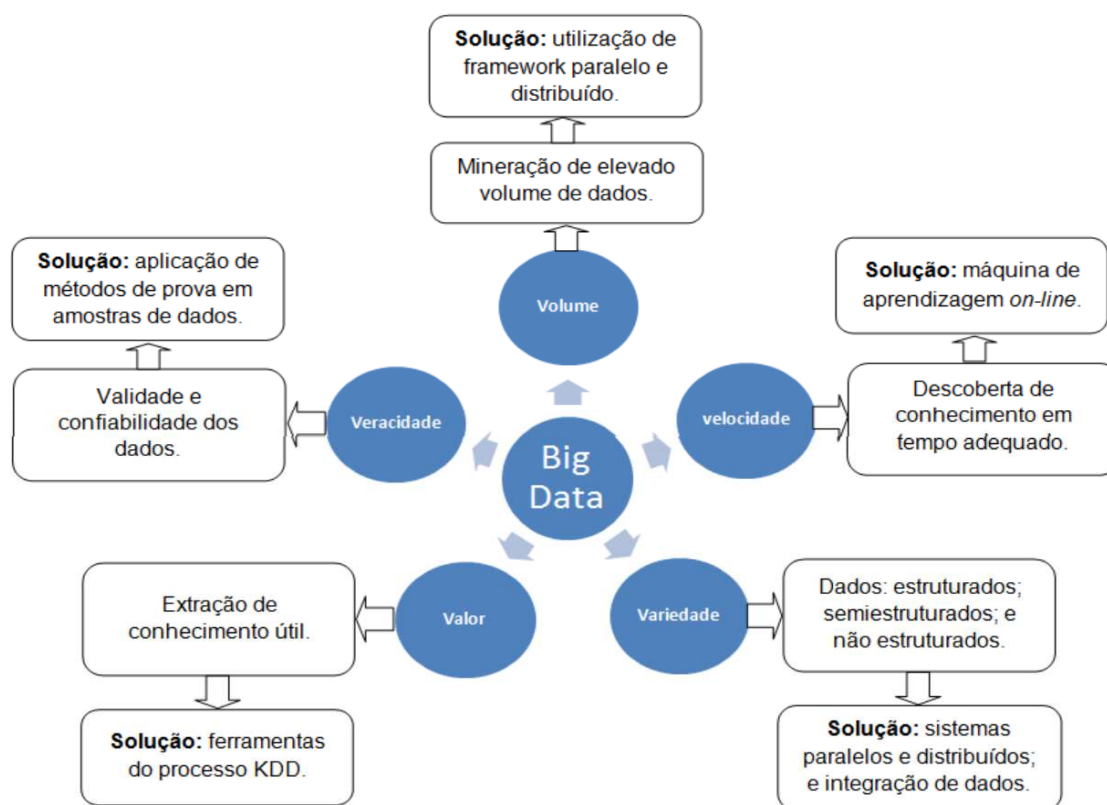


FIGURA 6 – CARACTERÍSTICAS DO *BIG DATA*, DESAFIOS E SOLUÇÕES.

FONTE: ELABORADA PELO AUTOR

#### 2.2.4. TECNOLOGIAS RELACIONADAS AO *BIG DATA*

O *Hadoop*, uma plataforma de código aberto<sup>5</sup> desenvolvida pela fundação Apache<sup>6</sup>, permite analisar e processar grande volume de dados, com suas complexidades, já que agrega a capacidade de manipulação de dados para processamento computacional avançado (DAVENPORT, 2014).

<sup>5</sup> Os usuários têm livre acesso ao código-fonte do software e fazem alterações conforme suas necessidades.

<sup>6</sup> <http://hadoop.apache.org>



Com essa plataforma, é possível criar um ambiente computacional de gerenciamento de dados, cujo ambiente é formado por *clusters* distribuídos com base no modelo *MapReduce* (GOLDSCHMIDT *et al.*, 2015).

A plataforma é considerada um *framework* computacional para sistemas distribuídos. Projetado com mecanismos de tolerância a falhas, favorece a execução paralela de tarefas em sistemas distribuídos (*clusters*), que dispõem de um conjunto de processadores para apoiar a distribuição de subtarefas. As tarefas executadas por esses sistemas geram subconjuntos de resultados, os quais, posteriormente, são unificados em um único conjunto (BENGFORT e KIM, 2016; GOLDSCHMIDT *et al.*, 2015).

Bengfort e Kim (2016) argumentam que o *Hadoop* evoluiu para um ecossistema, que contempla diversas ferramentas para ingerir, disponibilizar, processar e gerenciar os dados. Ferramentas como *Sqoop* e *Kafka* são utilizadas para a ingestão; para o armazenamento, existem tecnologias como *Hive*, *Hbase* e *MongoDB*; para a análise, existem os pacotes *GraphX*, *MLlib* ou *Mahout*.

Pensando em uma visão geral dessas tecnologias e procurando auxiliar seu entendimento, Davenport (2014) apresenta a Tabela 2.

TABELA 2 – VISÃO GERAL DAS TECNOLOGIAS DO BIG DATA  
 FONTE: DAVENPORT (2014, P.112)

<b>Tecnologia</b>	<b>Definição</b>
<i>Hadoop</i>	<i>Software</i> de código aberto para processamento de <i>Big Data</i> em série com servidores paralelos.
<i>MapReduce</i>	<i>Framework</i> arquitetônico que o <i>Hadoop</i> utiliza
Linguagem de <i>script</i>	Linguagem de programação adequada ao <i>Big Data</i>
Aprendizagem de máquina	<i>Software</i> para identificação de modelo adequado para o conjunto de dados
Análise visual	Apresentação dos resultados analíticos em formatos visuais
Processamento de linguagem natural (PLN)	<i>Software</i> para análise de texto, frequência, sentido, dentre outros.

Dois componentes utilizados no desenvolvimento de soluções *Big Data*, envolvendo o armazenamento e o processamento distribuído são o HDFS (*Hadoop Distributed File System*) e o YARN (*Yet Another Resource Negotiator*). O primeiro componente possibilita o armazenamento em *clusters*; o segundo, o gerenciamento dos recursos do *clusters*. Para minimizar o tráfego de dados, esses componentes trabalham simultaneamente, garantindo que os dados estejam em disco local no momento do processamento. A flexibilidade dos componentes permite o desenvolvimento de ferramentas e tecnologias para o *Big Data*, característica que torna o *Hadoop* um ecossistema robusto (BENGFORT E KIM, 2016).

Inserida no ecossistema do *Big Data*, a linguagem R proporciona um ambiente de desenvolvimento integrado para manipulação dos dados e a possibilidade de expansão com a inserção de pacotes (bibliotecas). Essa linguagem é utilizada por cientistas de dados com finalidade de estatística e de análise. Esses pacotes são implementados por desenvolvedores espalhados geograficamente e disponíveis em uma rede de distribuição CRAN<sup>7</sup> (*Comprehensive R Archive Network*).

Para suportar grandes tarefas de processamento e análise de dados, a Linguagem R dispõe do *RHadoop*, um conjunto de pacotes compatíveis com o *framework Hadoop* (*rhdfs*, *rhbase*, *rmr2*, *ravro* e *plyrmr*) (SARNOVSKY *et al.*, 2017).

Davenport (2014), ao refletir sobre a maneira de como o *Big Data* deve ser trabalhado, identifica duas condições. Uma delas pode ser considerada satisfatória, porque o ecossistema é rico em recursos e, como estes pertencem à licença de código aberto, os trabalhos que fazem uso dessas tecnologias têm custo relativamente baixo. A outra condição é insatisfatória, porque o trabalho com as tecnologias *Big Data* não é simples, demandando muita atenção e tempo de todos os envolvidos.

---

<sup>7</sup><https://cran.r-project.org/>

### 2.2.5. FLUXO DE DADOS NO *BIG DATA*

O processamento de dados no ciclo de vida para *Big Data* é diferente do processamento dos dados transacionais. No ambiente tradicional, a arquitetura do fluxo de dados é eficiente do ponto de vista do desempenho da gravação e do processamento. Primeiramente, realiza-se a análise, que leva à descoberta de conhecimento e, conseqüentemente, à criação de modelo de dados e à criação de uma estrutura de banco de dados.

No processo para o *Big Data*, os dados são primeiramente coletados e carregados em plataforma de destino e, posteriormente, aplicados em uma camada de metadados para ser transformados e analisados. Nesse processo, a aplicação em um banco de dados tradicional é ineficiente e, por essa razão, sugere-se uma arquitetura baseada em arquivos com o uso de linguagem de programação (KRISHNAN, 2013). A Figura 7 apresenta os processos referentes à sequência do fluxo de dados para o *Big Data*.

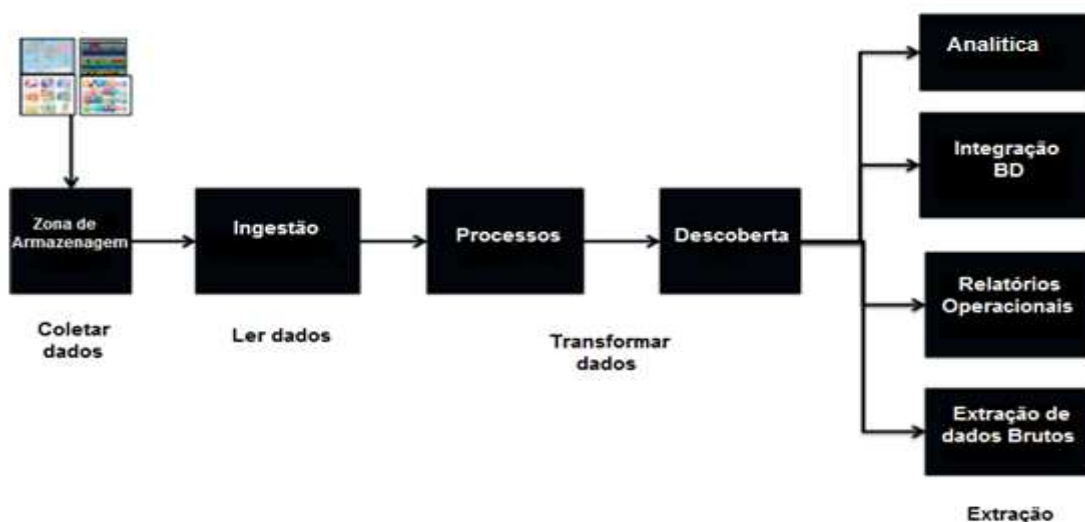


FIGURA 7 – DIAGRAMA DO PROCESSO *BIG DATA*  
 FONTE: TRADUZIDO DE KRISHNAN (2013, P.38)

No diagrama do processo *Big Data*, apresenta-se uma visão geral dos estágios para o processamento do fluxo de dados.

- Coleta – recebimento de dados de diferentes fontes, para posterior armazenamento em um local denominado “zona de armazenamento”. Essa etapa é ideal para modificação nos arquivos,

pois é nessa área que os dados podem ser classificados em estruturados e não estruturados.

- Leitura – carregamento dos dados, aplicação de metadados e preparação para transformação. Esse processo subdivide os dados em pequenas partições.
- Transformação – transformação dos dados por meio da aplicação das regras de negócio e por meio do processamento dos conteúdos. Nesse estágio, são executadas várias atividades de transformação, o que gera resultados intermediários que podem ser armazenados para análises.
- Extração – utilização dos resultados anteriores para análise; integração em *Data Warehouse* (DW); geração de relatórios operacionais; e possibilidades de utilizar técnicas de visualização para saídas que contenham dados brutos.

#### **2.2.6. ANÁLISE DE FACETAS NO BIG DATA**

A análise de faceta foi desenvolvida pelo matemático indiano Shialy Rammarita Ranganathan e publicada no ano de 1930. A faceta é utilizada na classificação e na identificação de características comuns em categorias de um determinado assunto (NETTO, 2016). Esse método é conhecido na área de ciência da informação e tem sido amplamente utilizado como mecanismo em vários sistemas de organização do conhecimento, dentre eles, os sistemas de classificação, taxonomias, incluindo o desenvolvimento de arquiteturas de *sites* e de estruturas de informação visual (SHIRI, 2014).

Como um método utilizado na organização do conhecimento, a análise de facetas favorece o reconhecimento de muitos aspectos de um único assunto (LIMA, 2007). Esses aspectos contribuem para identificar, organizar e detalhar as características de uma entidade, o que envolve a observação, a exploração e a compreensão das fontes do *Big Data*. Segundo Milonas (2011), em virtude dos benefícios da aplicação da análise de facetas, suas classificações têm sido

amplamente exploradas na organização e na recuperação de informações no domínio da *Web*.

Shiri (2014) descreve a aplicação da análise de facetas no mapeamento dos aspectos do *Big Data*, selecionando seis entidades, denominadas facetas: tipo de dados; ambiente; pessoas; operações e atividades; análises e metadados. Essas facetas foram utilizadas na captura de aspectos-chave do *Big Data*. Para cada faceta foram elaboradas as subfacetas, que demonstram seus aspectos específicos.

Para efeito ilustrativo, a Figura 8 apresenta as subfacetas e as instâncias referentes à faceta tipo de dados. Na perspectiva do *Big Data*, essa análise contribui para melhor compreensão dos seus principais componentes.

Para ser bem sucedidos, os processos de muitas empresas que produzem, processam, gerenciam, utilizam e mantêm grande volume de dados dependem de uma clara compreensão da sua complexidade multifacetada (SHIRI, 2014). A análise de faceta realizada por La Barre (2010) fornece uma revisão e apresenta diversas aplicações, como sistemas de recuperação, interfaces, bancos de dados, modelagem de dados e sistemas de busca e navegação.



FIGURA 8 – FACETAS DO BIG DATA  
 FONTE: ADAPTADO E TRADUZIDO DE SHIRI (2014, P. 361).

Nesta tese, utiliza-se a análise de faceta com o intuito de auxiliar a compreensão das fontes de dados e, dessa forma, seu emprego no modelo proposto.

Ao longo do CVP, são relevantes a gestão e a disseminação dos dados sobre o produto. É nesse contexto que entra a perspectiva da gestão dos dados, isto é, diante da necessidade de um gerenciamento eficiente para prover abordagens tecnológicas de sistemas de informações, surge à abordagem estratégica PLM.

### **2.3. GERENCIAMENTO DO CICLO DE VIDA DE PRODUTO - PLM**

Uma das definições mais encontradas na literatura sobre o PLM é a da CIMdata (2012), uma empresa de consultoria que está há mais de três décadas desenvolvendo trabalhos sobre iniciativas de gestão do CVP:

“uma abordagem estratégica da empresa que se aplica em um consistente conjunto de soluções empresariais para apoiar a criação, o gerenciamento, a disseminação e uso da informação relativa à definição do produto na empresa. PLM abrange desde a concepção do produto, até o fim da vida, integrando pessoas, processos, sistemas de negócios e informações. Ela forma espinha dorsal das informações sobre o produto em toda extensão da empresa.” (CIMDATA, 2012, p. 2).

Corallo *et al.* (2013) analisaram a definição proposta pela CIMdata (2012) e concluíram que a complexidade da gestão dos dados é proporcional à extensão do CVP e ao número de componentes do produto. Assim, afirmam que o PLM é uma “espinha dorsal” do produto e que os dados provêm dos meios interno e externo envolvidos no PDP.

Considerando que o PLM é o centro das informações do produto, as quais são produzidas e extraídas de meio internos e externos, observa-se a possibilidade de uma relação entre o PLM e o *Big data*, pois as duas abordagens têm em comum dados complexos e de diferentes fontes.

O PLM surgiu no início do século XXI com a intenção de gerenciar o processo de conhecimento em relação ao mercado, ao *design* de produtos, ao

desenvolvimento de processo, à fabricação de produtos, ao pós venda, aos serviços e à reciclagem de produtos (LI *et al.*, 2015).

As definições apresentadas pela CIMdata (2012) são similares a outras encontradas na literatura, já que apontam para a importância da gestão e do processo de integração dos dados relacionados ao produto. Grives (2006) define o PLM como um método que integra pessoas, processos/práticas e tecnologias, desde a ideia, a fabricação e a implantação até a remoção do produto. De forma semelhante, Saaksvuori e Immonen (2004) definem o PLM como uma gestão do processo e do controle das informações relacionadas ao produto nas diferentes fases do seu ciclo de vida.

Zancul (2009) ressalta os seguintes aspectos para o PLM: *i)* gestão integrada, que oferece suporte aos processos de negócios e apoia a gestão da informação; *ii)* aplicação do início ao fim do CVP para apoiar a colaboração na empresa estendida; *iii)* sua implantação necessita de infraestrutura, especialmente de tecnologia da informação.

Analisando-se as definições, conclui-se que o PLM não é um *software* ou um conjunto de aplicativos, mas sim uma infraestrutura alinhada aos processos e fases do CVP. Zancul (2009) realiza um levantamento da utilização da abordagem do PLM e dos sistemas de informação (Sistemas PLM ou Solução PLM).

As empresas utilizam o PLM como apoio à implantação dos processos transversais e, com isso, as áreas entram em consenso e a definição do escopo dos projetos se torna adequada (VIEIRA *et al.*, 2013).

Para melhor compreender o PLM, é fundamental entender o conceito de processo de negócio. A definição mais utilizada na literatura é a de Davenport (1993), para quem o processo de negócio é um conjunto de atividades estruturadas com o propósito de obter resultados específicos para um mercado ou cliente. Sua ênfase é a configuração do trabalho realizado no interior da organização, isto é, a essência está muito mais em como o trabalho é realizado

do que no produto ou serviço fornecido. Dessa forma, por meio da sequência ordenada nas atividades de trabalho, com começo e fim, é possível claramente identificar as entradas e as saídas.

Rozenfeld *et al.* (2006) explicam que um processo de negócio é um conjunto de atividades realizadas na empresa, associadas a informações manipuladas e com base nos recursos e na organização da empresa. Exemplos típicos de processo de negócio são: *i)* atendimento ao cliente; *ii)* planejamento estratégico; *iii)* desenvolvimento de produto; *iv)* venda de produto.

Em uma das definições do KDD, encontra-se a conquista de novos conhecimentos que sejam úteis (FAYYAD *et al.*, 1996b). Portanto, essa característica, aliada às possibilidades de busca de conhecimento e ao processo de negócio, oferece oportunidades de visualizar diferentes perspectivas do consumidor. De acordo com a definição de De Sordi (2008, p.18) “*os processos de negócios são fluxos de trabalho que atendem a um ou mais objetivos da organização e que proporcionam, sob a ótica do cliente final, agregação de valor*”. Dessa maneira, entende-se que o apoio do KDD ao PLM proporcionará valores ao processo de negócio.

Para Willaert *et al.* (2007), os processos de negócios precisam ser constantemente avaliados, aperfeiçoados e implementados na estrutura organizacional por meio de um quadro de apoio de recursos humanos e sistemas de informação. Já, para Vieira *et al.* (2013), os modernos processos de negócios necessitam de soluções, nas quais a integração do conjunto de informações apoia a otimização no desenvolvimento do produto. Por exemplo, os dados dos sistemas CRM (*Customer Relationship Management*) devem ser utilizados para incorporar as exigências do consumidor de novos produtos. Esses mesmos autores consideram que o PLM, integrado à estratégia de negócios da empresa, assume proporções significativas e efetivas e apresenta claramente os requisitos necessários para a obtenção de resultados.



## 2.4. **BIG DATA E PLM**

Li *et al.* (2015) afirmam que a utilização do *Big Data* no PLM é deficitária e que algumas empresas não exploram essa possibilidade. Argumentam ainda que o *Big Data* permeará toda a cadeia de produção, passando pela otimização do processo de montagem, pelo aumento da produtividade e pela satisfação das necessidades dos clientes. Esses fatores implicam a necessidade de trabalhos sistemáticos e exaustivos no sentido de identificar os benefícios relacionados às áreas do *Big Data* e PLM.

O *Big Data* não atua diretamente no gerenciamento dos dados no contexto do PLM, no entanto, pode utilizar tais dados para aumentar a inteligência e a eficiência no processo. Com base nesse argumento, os autores Li *et al.* (2015) corroboraram a ideia e realizaram uma pesquisa para verificar se o conceito de técnicas do *Big Data* pode ser empregado na fabricação de produtos com a abordagem do PLM. Interessados em utilizar a descoberta de conhecimento para aumentar a eficiência no processo de produção, eles concluíram que o potencial do *Big Data* é significativo e que suas técnicas podem acompanhar todo o CVP.

As empresas, para oferecer o produto adequado, no momento certo e pelo preço justo, depositam sua confiança em estratégias e soluções propostas pelo PLM. Além de utilizar os dados gerados internamente, fazem uso das informações produzidas na rede mundial de computadores para apoiar o PDP. Com o crescimento no volume de dados internos e externos que podem ser aplicados no PLM, surge a necessidade da procura de novos métodos para transformar esses dados em conhecimento (CIMDATA, 2012).

Para o desenvolvimento de seus produtos, as empresas implementam e utilizam sistemas de tecnologia da informação para gerenciar o CVP, como desenho assistido por computador (em inglês, *Computer Aided Design* ou CAD) e ferramentas de gerenciamento. Os dados gerados por essas ferramentas ficam restritos ao seu âmbito, porém, é possível extraí-los e

posteriormente integrá-los e, assim, obter uma colaboração eficaz (MANYIKA *et al.*, 2011).

Quando deseja comprar determinado produto, o consumidor busca opiniões de outros consumidores na rede mundial de computadores, visto que as empresas divulgam apenas os pontos positivos de seus produtos. Essas opiniões e os dados produzidos pelo PLM podem gerar conhecimentos isoladamente. Entretanto, o cenário *Big Data* permite a junção dessas fontes de dados, o que possibilita novos conhecimentos para o PDP.

A integração dos dados produzidos pelo PLM pode ser observada no trabalho de Zhang *et al.* (2017). Para a análise de dados de diferentes fontes do CVP, eles propuseram a arquitetura BDA-PL. Essa arquitetura oferece visões do armazenamento, do processamento de dados do PLM, das tecnologias RFID e de outros sensores e se encontra dividida em estágios, como mostra a Figura 9.

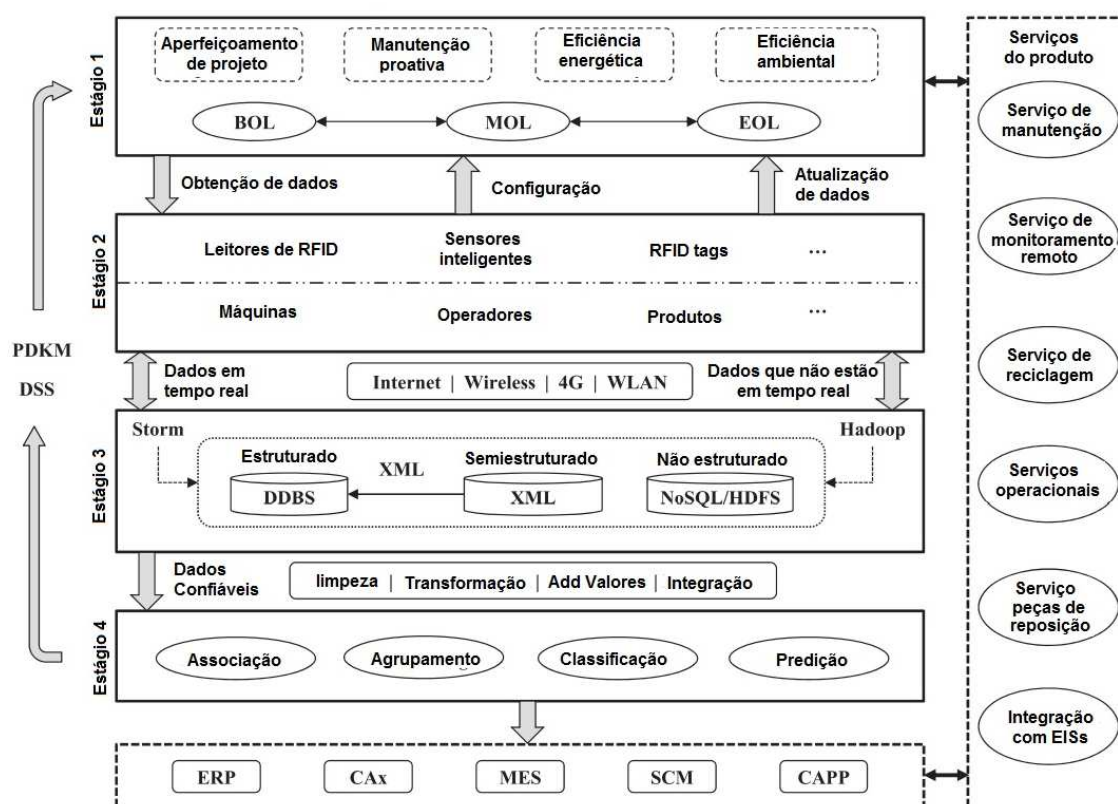


FIGURA 9 – ARQUITETURA GLOBAL BASEADA NO CVP – BDA-PL.  
FONTE: TRADUZIDO DE ZHANG ET AL. 2017, P.630.

- estágio 1 – extração dos dados referentes aos serviços aplicados para o gerenciamento do CVP;
- estágio 2 – aquisição e integração dos dados do estágio 1 com os dados da RFID e sensores inteligentes;
- estágio 3 – armazenamento dos dados;
- estágio 4 – aplicação das tarefas de descoberta de conhecimento.

O processo de limpeza, transformação e integração dos dados ocorre entre os estágio 3 e 4.

Outra pesquisa nesta área é de Zhuang *et al.* (2016), que apresentam uma arquitetura aberta e escalável denominada “*D-Ocean*”, a mesma não está relacionada diretamente com PLM, porém, atua com sistema de gestão de dados não estruturados. A proposta é oferecer uma estrutura de armazenamento unificado, com possibilidades de processamentos incrementais, fornecendo um “*search engine*”, que combina várias consultas compostas.

#### **2.4.1. PRODUÇÃO DE DADOS**

Diferentes sistemas de informação, como CAD (Computer Aided Design), CAPP (*Computer Aided Process Planning*), CAE (*Computer Aided Engineering*), são empregados para gerenciar dados durante o PDP. Como solução para integrar os dados produzidos individualmente por esses sistemas, surge o ERP (*Enterprise Resource Planning*), que é capaz de suprir a necessidade de informações e gerar uma base de dados única para ser utilizada por diversos setores da empresa (ROZENFELD, 2006).

Na construção de um modelo de dados integrado aos aspectos técnicos e de negócios, torna-se necessário utilizar dados relevantes para produzir produtos inovadores. Entretanto, com o aumento da quantidade e da variedade de tipos de dados, eleva-se o grau de dificuldade na compreensão dos projetos em andamento (VIEIRA et al., 2013).

A produção de dados durante o CVP apresenta as seguintes situações: integração de diferentes aspectos (técnico e negócios); aumento no volume de dados; diferentes tipos de dados; complexidade na compreensão. Essas situações combinam com as características apresentadas por Shiri (2014) em relação ao *Big Data*:

- grande conjunto de dados integrados;
- variedade de dados e tipologia;
- meio de armazenamento;
- capacidade de processamento;
- dimensionalidade temporal e espacial;
- heterogeneidade, diversificação, distribuição, complexidade e evolução natural.

Além dos sistemas tecnológicos de informação mencionados, existem outros grupos usados de forma mais abrangente nas empresas, dentre os quais destacam-se os seguintes.

- CRM (*Customer Relationship Management*) - responsável pelas informações referentes aos clientes. Gerencia as informações detalhadas dos canais de comunicação, com interesse na fidelização dos clientes (KOTLER e KELLER, 2012). Tem ainda a função de atender e antecipar as necessidades dos clientes e, para isso, capta as informações existentes sobre eles para consolidá-las, analisá-las e identificar padrões (ROZENFELD *et al.*, 2006). Dessa forma, o CRM apoia as estratégias de negócios voltadas para identificar o cliente, atraí-lo, retê-lo e apoiar seu ciclo de vida (CHANG *et al.*, 2010).
- SCM (*Supply Chain Management*) - gerencia o fluxo de produção na cadeia de suprimentos e integra os processos de negócios junto aos fornecedores. Integra não apenas a área de suprimentos, mas também a demanda de vendas, compras, recebimento, estoque, planejamento de produção e transporte (ROZENFELD *et al.*, 2006).

O SCM facilmente gera dados com volume elevado, variedade e velocidade (WALLER e FAWCETT, 2013).

- PLM (*Product Life cycle Management*) – contempla a criação e a gestão dos dados dos produtos e dos projetos ao longo do CVP. Rozenfed *et al.* (2006) argumentam que o PLM integra todas as soluções tecnológicas relacionadas ao PDP.

Utilizando esses sistemas de informações, Zancul (2009) apresenta quatro macrofases no CVP: desenvolvimento, produção, uso e serviços, e descarte. Cada macrofase é composta por fases mais específicas. A Figura 10 ilustra a utilização de sistema de informação ao longo do CVP.

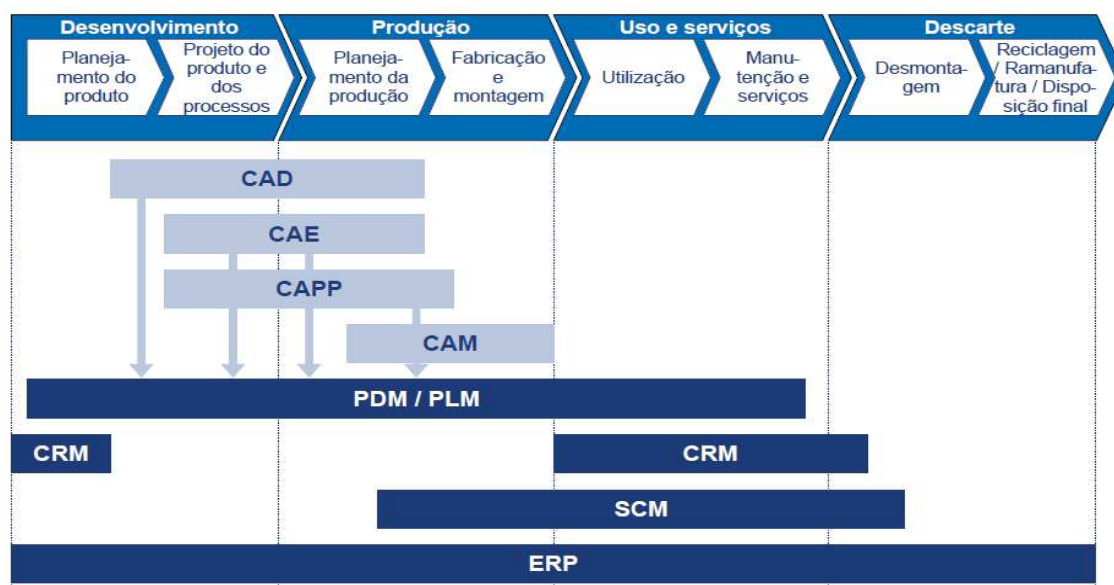


FIGURA 10 – SISTEMAS DE INFORMAÇÃO NAS FASES DO CVP.

FONTE: ZANCUL, (2009, P. 48).

#### 2.4.2. FONTES DE DADOS: INTERNAS E EXTERNAS

Fornecedores de soluções para o PLM estimam que o volume de dados utilizados na produção de um veículo está na casa dos gigabytes. Considerando ainda os dados do processo de engenharia, incluindo todos os projetos em evolução, possivelmente essa quantidade se eleva para a casa dos terabytes. Tais dados estão em formato estruturado e não estruturado, sendo encontrados em servidores próprios da empresa ou externamente em parceiros, fornecedores e clientes (CIMDATA, 2012).

Dessa forma, os dados que podem ser utilizados não se restringem à empresa, ou seja, dados externos também podem ser acessados (KAISLER *et al.*, 2013). Um exemplo é o das informações dos clientes que, além de estar armazenadas em banco de dados internos (CRM), podem ser encontradas nas mídias sociais. De acordo com Kaplan e Haenlein (2010), essas mídias permitem que as empresas estreitem o contato com o consumidor final, oportunizando a redução de custo, se comparadas com ferramentas de comunicação tradicional. Jun *et al.* (2014) preveem mudanças nas atitudes dos consumidores e, realizando comparações com a pesquisa convencional, procuram demonstrar a credibilidade da pesquisa no tráfego de informações. A utilização das mídias sociais não é uma tarefa simples: além de poder exigir uma nova forma de pensamento, tem um potencial que não pode ser desprezado.

Em relação à pesquisa que inclui mídias sociais, Fan e Gordon (2014) argumentam que é necessário que as empresas estejam em sintonia com os desejos dos clientes a fim de antecipar alterações significativas em seus produtos. Não menos relevante é o fato de que as informações de alguns clientes, publicadas nas mídias sociais, podem influenciar positivamente ou negativamente o comportamento de outros clientes. Nesse contexto, o diretor sênior de estratégia de *software* PLM da Siemens, localizada na cidade de Plano no Texas, afirma: *“As próximas gerações de sistema PLM serão plataformas para enxergar a inovação, pelo fato de envolver usuários que fornecem informações necessárias para a tomada de decisão no menor tempo possível”* (MARTIN, 2015, p. 46).

Figura 11 apresenta exemplos de fontes de dados interna e externa à organização que gera o *Big Data* e, ilustra os sistemas de informação inseridos nas macrofases apresentadas por Zancul (2009).

Tripathy *et al.* (2012) utilizam a MD nos *“blogs”* de consumidores para identificar conceitos no desenvolvimento de novos produtos. Essa aplicação é verificada em Carr *et al.* (2015): com uma abordagem inovadora, eles exploram

dados das mídias sociais, considerando as ferramenta não apenas para marketing, mas também para o desenvolvimento do produto.

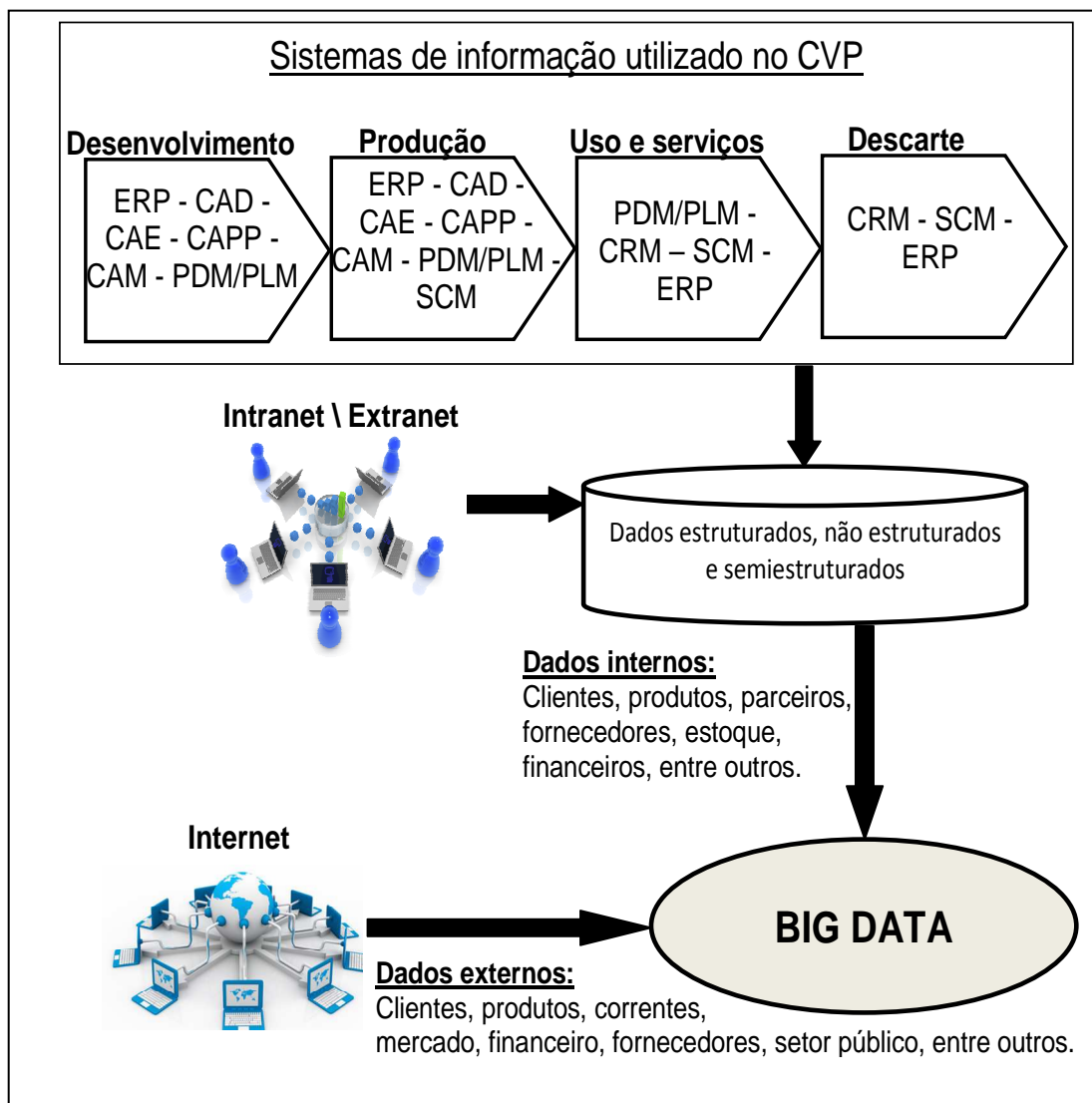


FIGURA 11 – FONTES INTERNA E EXTERNA DE DADOS.

FONTE: ELABORADA PELO AUTOR

O PDP gera grande quantidade de dados, assim como as mídias sociais, mas existem outras fontes que podem melhorar a “experiência com o produto”, como a inserção de instrumentos capazes de capturar dados durante a utilização do produto. Esses dados fornecem conhecimentos relevantes a respeito de falhas nos produtos, necessidades e hábitos dos clientes (CIMDATA, 2012).

Como exemplo de análise da fase de utilização do produto, Bhinge *et al.* (2015), avaliam o ciclo de vida de bateria de celular de íons de lítio por meio de seus próprios dados. Já, Lachmayer *et al.* (2014), trabalham com dados de suspensão de carros de corridas e adicionam dispositivos (sensores) em pontos estratégicos das peças com o intuito de coletar dados. Após a coleta, eles aplicam técnicas e tarefas de MD para extrair conhecimentos que serão utilizados no aprimoramento das novas gerações de produtos.

#### **2.4.2.1. DADOS DE ENTRADA E SAÍDA UTILIZADOS NAS FASES DO CVP**

Li *et al.* (2015) analisam a entrada e a saída de dados para as fases do CVP:

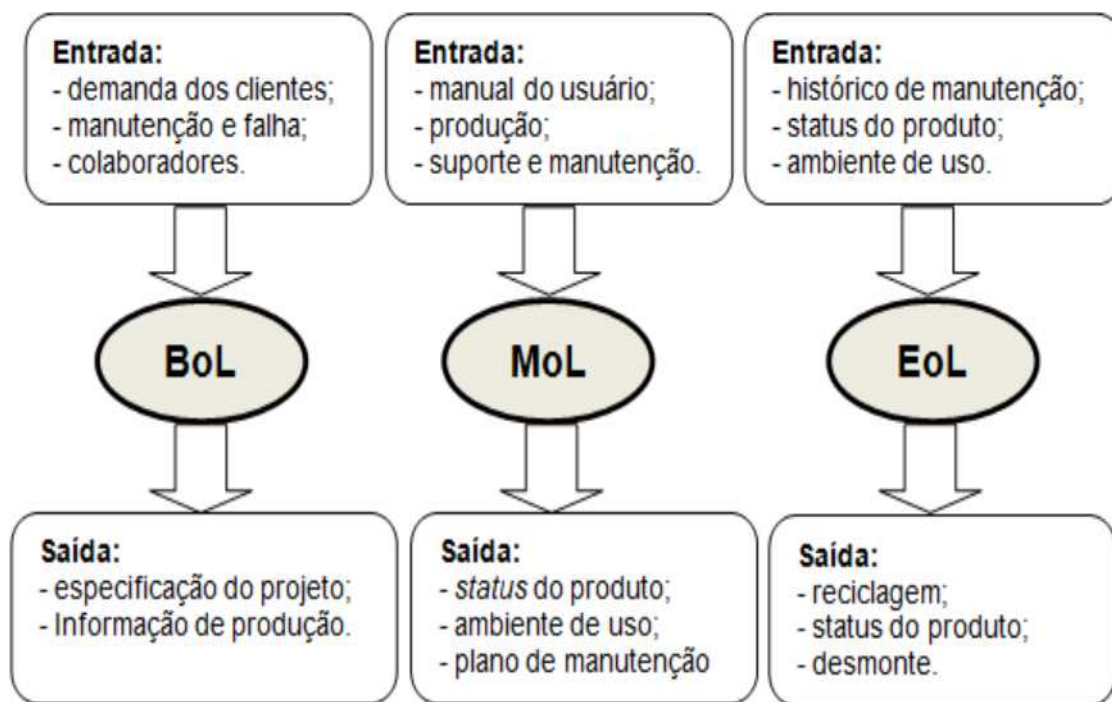
- BoL (*Begin of Life*) – período referente a planejamento e desenvolvimento (projeto e produção);
- MoL (*Midle of Life*) – período referente a distribuição, logística, distribuição, uso e manutenção;
- EoL (*End of Life*) – período referente a reciclagem (produtos abandonados pelos fabricantes e/ou descartados pelos clientes), incluindo logística reversa e remanufatura.

A entrada da fase inicial, *BoL*, demanda dados dos clientes, os quais podem ser encontrados em mídias sociais. Na fase de produção, os dados são gerados continuamente e em tempo real, de forma que sensores podem ser instalados para monitorar e fornecer informações sobre os parâmetros do ambiente, do produto e dos desgastes dos equipamentos.

A fase *MoL*, na qual estão os dados gerados pelo setor logístico, tem como entrada informações de pedidos, as quais podem ser transformadas em arranjos inteligentes. Na manutenção, recorre-se aos dados de entrada do monitoramento do ambiente, os quais podem oferecer instruções de apoio na prevenção de avarias.

Na fase *EoL*, as decisões se baseiam na reciclagem e no descarte do produto. A partir das informações originadas da fase *MoL*, é possível maximizar os valores dos produtos na fase *EoL*.





**FIGURA 12 – INFORMAÇÕES DE ENTRADA E SAÍDA PARA CADA FASE DO CVP**  
**FONTE: ADAPTADO DAS TABELAS DE LI ET AL. (2015, P. 6 E 7).**

A Figura 12 sintetiza as informações de entrada e saída para as fases do CVP.

As entradas e as saídas de informações mostradas na Figura 12 denotam elevado volume de dados. Esse volume apresenta potencial para a extração do conhecimento que pode apoiar o PDP.

O framework desenvolvida por Li *et al.* (2015), ilustrado na Figura 13, apresenta as etapas desenvolvidas na fases BoL, MoL e EoL, o desenvolvimento do modelo proposto tem como foco a fase BoL, especificamente na etapa do projeto de produto que agrega a atividade da tomada de decisão sobre o detalhamento do produto.

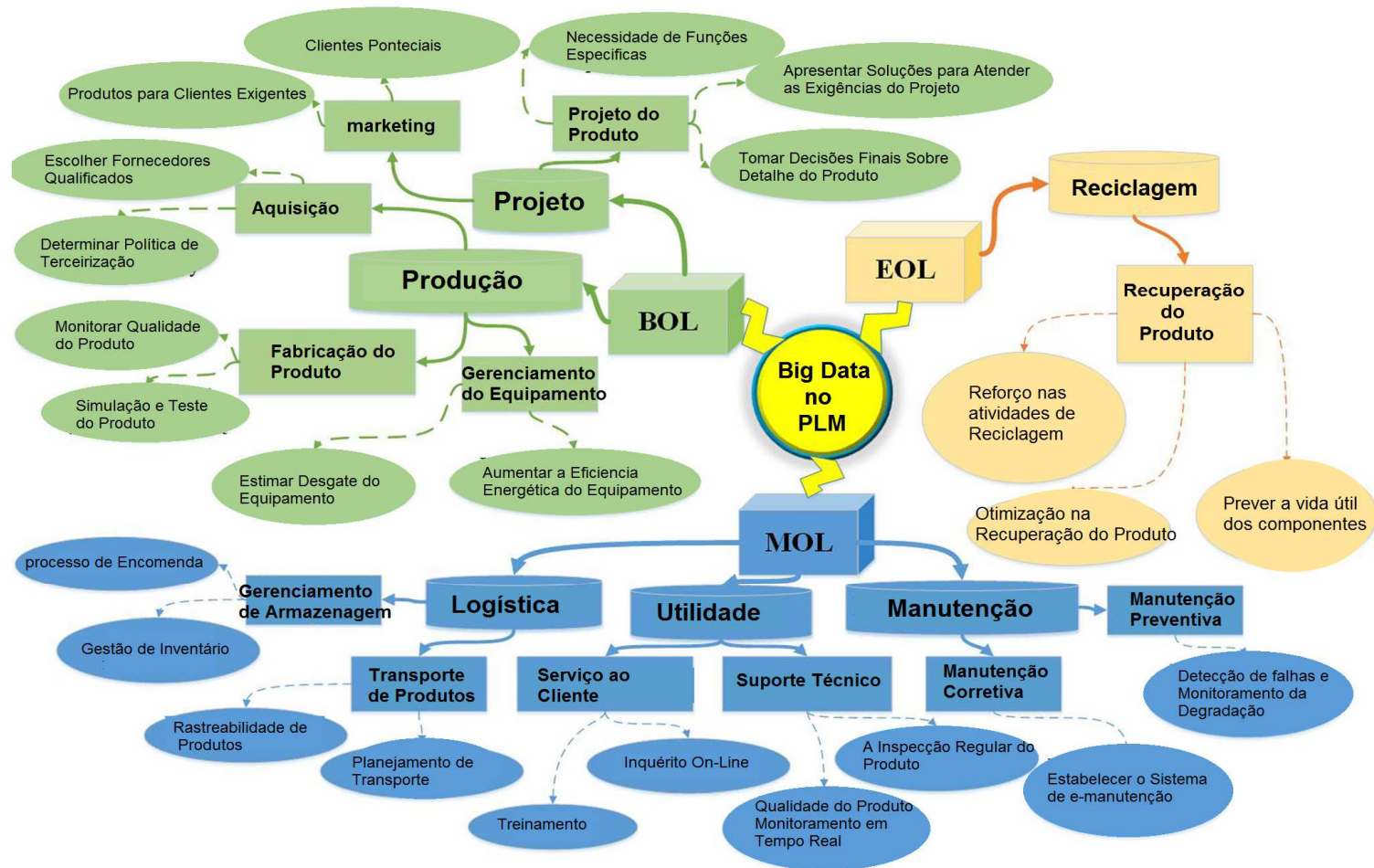


FIGURA 13 – FRAMEWORK DO BIG DATA NO PLM  
 FONTE: TRADUZIDO DE LI ET AL. (2015, P. 8)

## 2.5. QUALIDADE DOS DADOS

Nos últimos anos, é grande a variedade de tipos de dados linguísticos e visuais presentes em mídias sociais e nos sistemas de informação, em que dentre outros, são produzidos textos, músicas, informações de projeto e imagens. Tal variedade vem recebendo a atenção de pesquisadores e profissionais, dando origem a investigações mais avançadas sobre o conceito de qualidade dos dados (BATINI *et al.*, 2015).

Tais pesquisas têm como proposta a análise da composição dos dados em suas diferentes dimensões (WANG, 1998; PIPINO *et al.*, 2002; LEE *et al.*, 2002; OLSON, 2003). Batini *et al.* (2015) corroboram essas linhas de pensamento e destacam que a qualidade de dados é um conceito multifacetado, cujas definições possuem diferentes dimensões. Em razão disso, tais autores têm discutido as características estruturais e específicas dos diferentes tipos de dados. Para Orr (1998), qualidade dos dados não é para garantir que ele seja perfeito, mas sim, que seja suficientemente consistente no apoio da tomadas de decisão.

Merino *et al.* (2016) afirmam que, quanto à qualidade, existem diferentes modelos, mas estes se restringem a avaliar dados regulares, de forma que nenhum deles foi adaptado às novas exigências do *Big Data*. Por isso, propuseram um modelo denominado “3As Data Quality-in-Use model”, cujo objetivo é fornecer indicadores de confiabilidade na etapa de análise de dados no *Big Data*. Nesse modelo de avaliação, são consideradas três características do projeto *Big Data*.

- Adequação contextual – o conjunto de dados<sup>8</sup> deve estar em consonância com o domínio de interesse, independentemente de seu formato (estruturado ou não estruturado), de seu tamanho ou de sua velocidade. Assim, é importante que esses dados sejam relevantes; semanticamente compreensíveis e precisos; que demonstrem

---

<sup>8</sup>Conjunto de dados será tratado no presente trabalho como uma parte da extração de dados de uma fonte.

credibilidade; e, para casos confidenciais, que sejam utilizados pelo mesmo grupo autorizado a desenvolver a análise;

- Adequação temporal - os dados devem estar dentro de um intervalo de tempo apropriado para a análise. Assim, devem ser:
  - concomitantes – devem estar no mesmo intervalo de tempo;
  - atuais – devem ser semelhantes em seus tempos, pois a fusão de dados de diferentes temporalidades pode causar ruídos<sup>9</sup> na análise;
  - oportunos – devem ser devidamente atualizados, e seus períodos de tempo convenientes para análise,
  - frequentes – devem possibilitar a produção de resultados quando há necessidade de análise de tendências.
- Adequação operacional – refere-se às condições em que os dados podem ser analisados e processados por um conjunto de tecnologias, sem que se deixe parte do conjunto de dados fora do processamento e da análise. Na relação custo/benefício e desempenho, devem-se considerar o volume, a velocidade e a variedade dos dados. Dessa forma, o conjunto de dados deve ter:
  - disponibilidade - facilmente recuperável e acessível para a análise;
  - autorização para os fins previstos;
  - representação eficiente, a fim de evitar o desperdício de recursos,
  - possibilidades de auditoria para permitir rastreamentos e possíveis alterações.

## **2.6. MODELOS DE REFERÊNCIA PARA O PDP**

Back *et al.* (2008) e Rozenfeld *et al.* (2006), tratando de processo de negócio relacionado à gestão do CVP com foco no PDP, propõem modelos de referência que são destaques no cenário nacional.

---

<sup>9</sup> Conjunto de dados inconsistente com o restante dos dados existentes. Exemplo: erro de digitação ou medições, dados incompletos, corrompidos ou distorcidos, entre outros.

O modelo de referência proposto por Back *et al.* (2008) é composto por três macrofases: *i)* planejamento do projeto; *ii)* elaboração do projeto do produto; *iii)* implementação do lote piloto. Essas macrofases se decompõem em oito fases: *i)* planejamento do projeto; *ii)* projeto informacional; *iii)* projeto conceitual; *iv)* projeto preliminar; *v)* projeto detalhado; *vi)* preparação para produção; *vii)* lançamento; *viii)* validação.

O modelo de referência de Rozenfeld *et al.* (2006), mostrado em alto nível de abstração na Figura 16, é composto por macrofases, fases, atividades e tarefas. Esse modelo, com previsão de pré-projeto, subdivide-se em planejamento estratégico de produtos e planejamento de projetos.

O planejamento estratégico de produtos tem como objetivo gerar um plano que contenha o portfólio dos produtos e a especificação das características e metas de novos produtos. Seus principais participantes são os membros da diretoria e os gerentes funcionais. O planejamento do produto orienta o projeto, apresentando a viabilidade de seu desenvolvimento. Ou seja, é utilizado como orientação para a macrofase de desenvolvimento, decomposta nas seguintes fases.

- Projeto informacional - etapa de identificação das necessidades dos clientes e dos requisitos dos produtos e também de levantamento de informações dos produtos concorrentes.
- Projeto conceitual - desenvolvimento de atividades de busca, criação, representação e seleção de soluções para o problema do projeto.
- Projeto detalhado – seu objetivo é a finalização e o desenvolvimento das especificações do produto. Essa fase corresponde à concepção do produto e tem como resultados as configurações finais, a especificação dos componentes, os desenhos finais com tolerância, o protótipo funcional e o projeto dos recursos necessários.
- Preparação da produção - homologação do produto e treinamento do pessoal.

- Lançamento – estabelecimento dos processos de vendas, treinamento da força de vendas, distribuição, assistência e atendimento aos clientes.

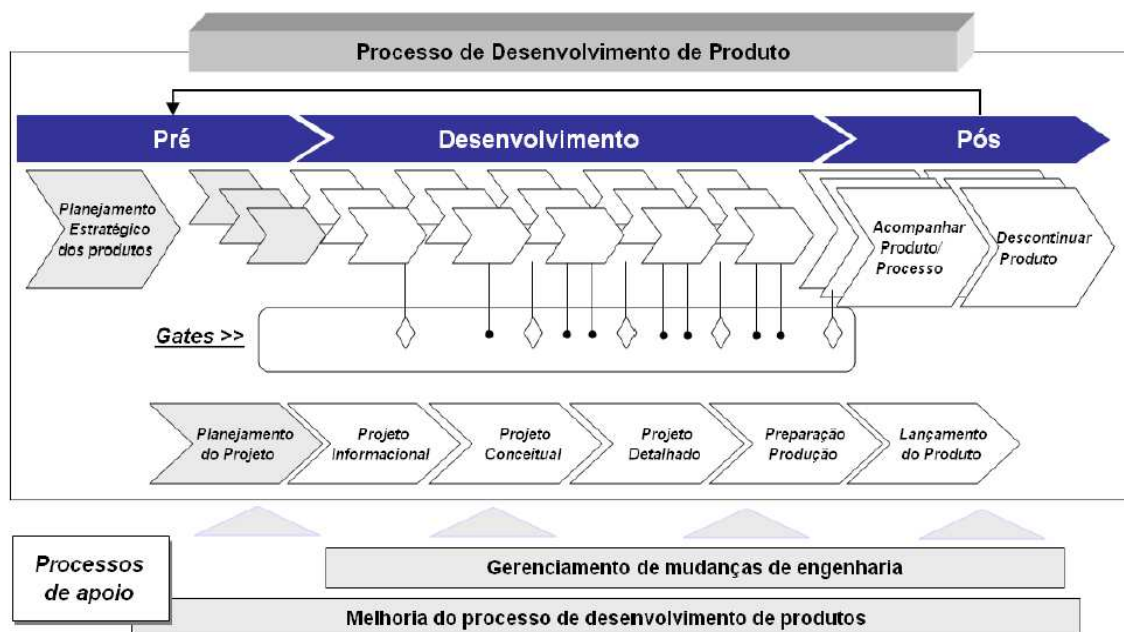


FIGURA 14 – PROCESSO DE DESENVOLVIMENTO DE PRODUTO  
 FONTE: ROZENFELD ET AL. (2006, P. 44)

O pós-desenvolvimento é subdividido em acompanhamento do produto/processo e sua possível descontinuidade. Nessa fase, as ações de acompanhamento do produto podem gerar informações que indicam melhorias no produto ou a descontinuidade de sua produção.

Rozenfeld *et al.* (2006) argumentam que o entendimento e a visão de cada pessoa participante no PDP resultam na criação de diferentes visões e linguagens próprias, o que pode gerar ineficiência e dificuldade na integração e na comunicação. Esse modelo de referência tem o propósito de criar um ponto de vista comum entre os participantes do processo. A ideia é que o modelo de referência seja genérico, o que favorece sua adaptação às diferentes realidades encontradas nas empresas.

Zancul (2009) avalia os modelos de referência do processo de negócio, os quais são resultado de projetos de pesquisa e de trabalhos padronizados

realizados nas indústrias (modelo unificado do PDP, projeto TFB-57, projeto PLM4KMU e modelo da empresa IDS Scheer).

### **2.6.1. IDENTIFICAÇÃO DAS ETAPAS DO PROJETO INFORMACIONAL**

Este trabalho se restringe ao projeto informacional, fase da identificação dos requisitos e das necessidades dos clientes. Essa primeira fase da macrofase do desenvolvimento tem como objetivos auxiliar as tomadas de decisão e melhorar os elementos obtidos na fase anterior. Atende, portanto, a dois propósitos, que determinam o sucesso das fases subsequentes: direcionar as etapas seguintes do desenvolvimento do produto e utilizar os preceitos construídos para a tomada de decisão ao longo do CVP.

Back *et al.* (2009) apresentam uma lista das atividades do processo de planejamento do produto. O modelo proposto pode sugerir conhecimentos novos e úteis nas seguintes atividades.

- Análise do consumidor – identificar as necessidades e os desejos dos consumidores.
- Análise dos concorrentes – analisar os produtos concorrentes e comparar aos produtos da empresa.
- Descrição dos requisitos – identificar os requisitos do mercado e utilizar informações dos consumidores, dos produtos da própria empresa (novos ou velhos), dos produtos concorrentes, das patentes, dos requisitos de diferentes mercados, da legislação e das normas.
- Integração dos conhecimentos – agregar novos conhecimentos ao processo.

Segundo Rozenfeld *et al.* (2006), no modelo de referência, a fase do projeto informacional possui o objetivo de utilizar informações coletadas nas fases de planejamento e em outras fontes de dados, ou seja, desenvolver um conjunto de informações para ser utilizadas como base para etapas posteriores, inclusive para a definição de critérios de avaliação e tomadas de decisão.

A Figura 17 mostra, por meio dos retângulos, as atividades do projeto informacional.

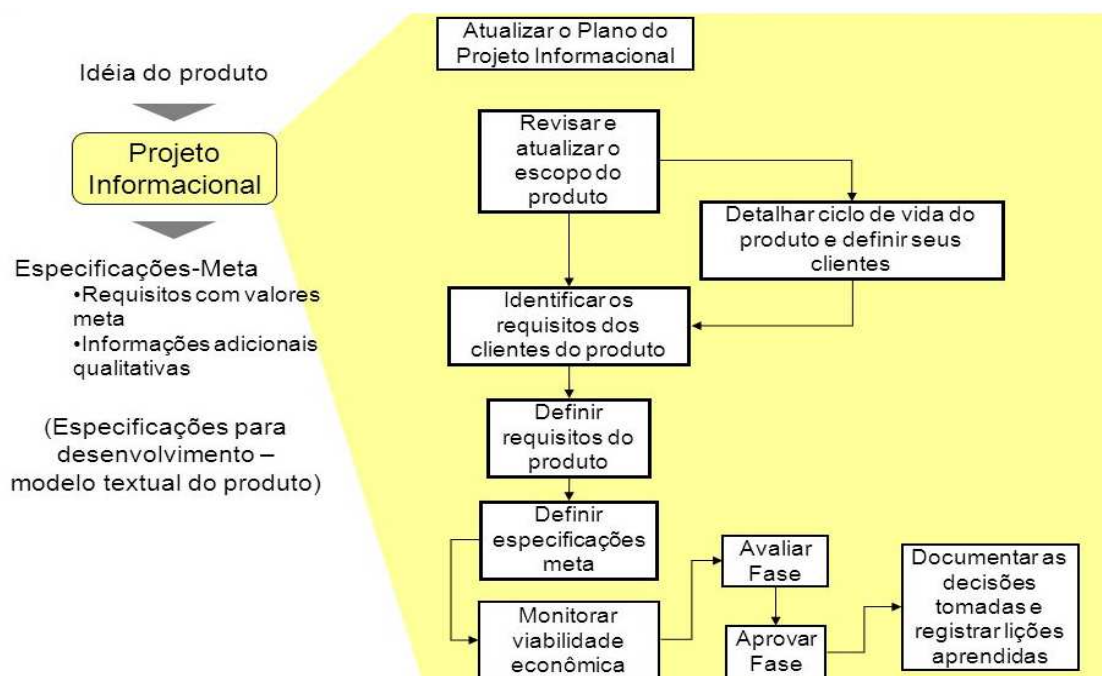


FIGURA 15 – PRINCIPAIS INFORMAÇÕES E DEPENDÊNCIAS ENTRE AS ATIVIDADES DA FASE DE PROJETO INFORMACIONAL

FONTE: ADAPTADO DE ROZENFELD ET AL. (2006, P. 212).

Na atividade de revisão e atualização do escopo do produto, para estudar problemas relacionados ao projeto, procura-se reunir o maior volume possível de informações, tais como: *i)* tipo de produto; *ii)* tipo de projeto; *iii)* volume planejado de fabricação; *iv)* desejos explícitos expostos pelos clientes; *v)* restrições ao projeto e ao produto. Nessa atividade, podem ser necessários mais conhecimentos referentes aos produtos da empresa, além de estudos de produtos similares, concorrentes, tecnologias e métodos de fabricação disponíveis.

A atividade de identificação dos requisitos dos produtos envolve a coleta das necessidades dos clientes em cada fase do CVP. Para isso, ROZENFELD *et al.* (2006) e BACK *et al.* (2008) recomendam a utilização de algumas ferramentas, tais como: QFD (*Quality Function Deployment*)<sup>10</sup>, questionários,

<sup>10</sup> Possibilita estabelecer relação entre as necessidades dos clientes e os requisitos do projeto e auxilia a equipe a buscar o consenso nas diferentes definições do produto.



entrevistas, *checklist*, *brainstorming*, diagrama de afinidades. Logo, o modelo proposto nesta tese torna-se uma ferramenta complementar, podendo gerar conhecimentos provenientes de diversas fontes de dados. Fato relevante é que os resultados (informações) das ferramentas sugeridas podem ser utilizados como matéria prima a ser aplicada no modelo proposto.

Na atividade de definição dos requisitos do produto, as informações da “voz dos clientes” podem ser associadas às características do produto. Dessa forma, o modelo proposto pode ser utilizado para ampliar a “voz dos clientes” e, conseqüentemente, descobrir novas características para o produto.

Portanto, a utilização do modelo proposto no projeto informacional pode resultar em novos conhecimentos, que serão úteis nas atividades que exigem melhor compreensão do produto.

## **2.7. TÉCNICAS DE VISUALIZAÇÃO**

As técnicas de visualização de informação podem ser utilizadas como mecanismos para auxiliar na compreensão dos resultados da descoberta de conhecimento. Além de transmitir conhecimentos, essas técnicas possuem potencial para receber comandos, tais como: *i)* controlar a quantidade de dados na tela; *ii)* alterar a representação da visualização, *iii)* ajustar escalas. Lengler e Eppler (2007) utilizaram a seguinte definição para técnica de visualização: “...uma representação gráfica que mostra informações de uma forma que é propício para a aquisição de conhecimento, desenvolvendo uma compreensão elaborada...”.

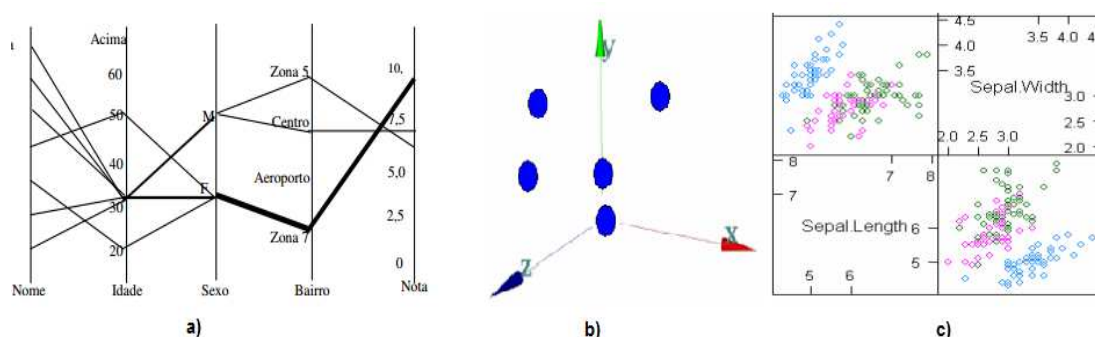
Zhong *et al.* (2016) aplicaram técnicas de visualização na área do *Big Data*, utilizando dados dos dispositivos RFID (*Radio-Frequency IDentification*) para apoiar a manufatura nas nuvens (em inglês, *Cloud Manufacturing* ou CM). Esse trabalho desenvolve um modelo denominado “*RFID-Cuboid*”, cuja proposta é realizar a organização e o encadeamento dos dados originados do RFID, sugerindo, com isso, uma visualização logística do chão de fábrica para

demonstrar o desempenho no progresso da produção e o comportamento da logística.

Keim e Kriegel (1996) descrevem técnicas de visualização de informação multidimensional, as quais são agrupadas em categorias geométricas, iconográficas, hierárquicas e orientadas a pixel.

As técnicas de projeções geométricas geram projeções bidimensionais e tridimensionais em base de dados multidimensionais e revelam informações de interesse da descoberta de conhecimento. A Figura 14 ilustra exemplos das técnicas geométricas:

- a) coordenadas paralelas - representam todos os atributos em uma mesma visualização e permitem realizar interpretações visuais entre os atributos (RABELO *et al.*, 2008);
- b) gráfico de dispersão tridimensional - eficiente para determinar a existência de relações, padrões ou tendências entre atributos. Essa visualização permite a inserção de propriedades visuais (cor, tamanho, forma, orientação, etc.), aumentando o número de atributos que podem ser representados;
- c) matriz de dispersão de dados - permite visualizar o relacionamento entre os atributos. Essa técnica projeta os atributos em pares, gerando células associadas a dois atributos que são mapeados pelo eixo x (linha horizontal) e eixo y (linha vertical).



**FIGURA 16 – TÉCNICAS GEOMÉTRICAS – A) MATRIZ DE DISPERSÃO DE DADOS; B) GRÁFICO DE DISPERSÃO; C) COORDENADAS PARALELAS**  
**FONTE: RABELO ET AL. (2008).**

As técnicas iconográficas são utilizadas para representações multidimensionais e podem ser compostas por características geométricas (forma, tamanho e orientação) e características de aparência (cor e textura). Essas características podem ser associadas aos atributos em análise. A Figura 15 ilustra algumas técnicas iconográficas:

- a) faces de Chernoff - associa as características de uma face humana (forma da boca, cabelos, olhos, orelha, etc.) aos atributos dos dados (CHERNOFF, 1973);
- b) gráfico de estrela - utiliza um círculo como referência e, em seu centro, são projetadas linhas que representam os atributos que emanam como raios que formam uma estrela (NASCIMENTO e FERREIRA, 2005);
- c) figura de arestas - consiste em segmentos de linhas denominados ramos e possui três parâmetros (ângulo, intensidade e comprimento) que podem ser utilizados para representar os atributos de dados.

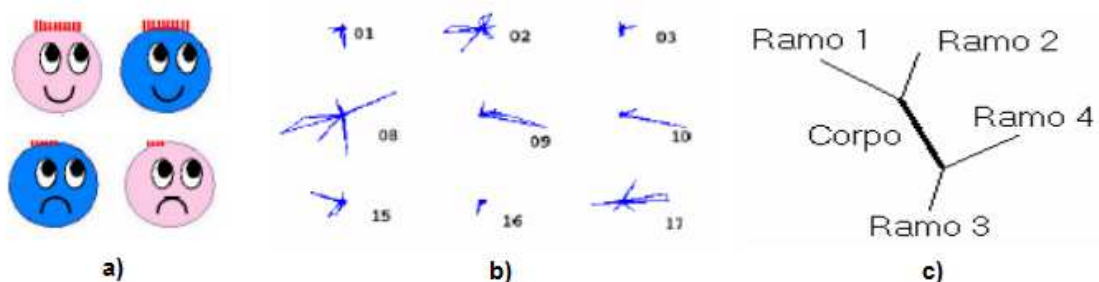


FIGURA 17 – A) FACES DE CHERNOFF ; B) GRÁFICO DE ESTRELA; C) ARESTA.  
 FONTE: ADAPTADO – NASCIMENTO E FERREIRA (2005, P. 1268 E 1301).

Lengler e Eppler (2007) elaboraram um método de seleção de técnicas de visualização de uma perspectiva sistemática, embasada na lógica da tabela periódica dos elementos químicos, como mostra o Anexo A. Nessa pesquisa, encontraram aproximadamente cento e sessenta métodos visuais, que foram reduzidos a cem métodos por meio de alguns critérios de seleção: *i)* o método deve ser documentado; *ii)* ter sido previamente aplicado na vida real; *iii)* estar apto a representar conhecimento intensivo; *iv)* ter sido aplicado por não-especialistas, *v)* ter sido previamente avaliado.

As técnicas de visualização de dados são organizadas na tabela, apresentada no Anexo A, em seis grupos: *i)* quantitativas representadas por gráficos de linha, pizza, barra ou áreas; *ii)* informações, que representam visualmente os dados e ampliam a cognição (conhecimento e percepção); *iii)* conceitual, que representa os relacionamentos dos elementos com capacidade para representar conceitos qualitativos; *iv)* metafóricas, que representam metáforas visuais; *v)* estratégicas, que representam as estratégias nas empresas; *vi)* compostas, que contemplam as visualizações de dois ou mais grupos.

Além de serem organizadas em grupos, as técnicas de visualização são classificadas em dimensões: *i)* complexidade da visualização; *ii)* área de aplicação; *iii)* ponto de visão; *iv)* tipo de ajuda esperada, *v)* tipo de informação. As linhas da tabela representam sua complexidade (quanto mais direita, maior sua complexidade). Em relação às cores das letras, o preto representa as informações estruturadas; o azul, as informações de processo. Na posição central, acima de cada quadro, está representado o ponto de visão. Essa tabela apresenta avaliações atualizadas e está disponível também em versão interativa *online*<sup>11</sup>, na qual, posicionando-se o cursor do mouse sobre cada método, aparecerá uma figura que o representa.

---

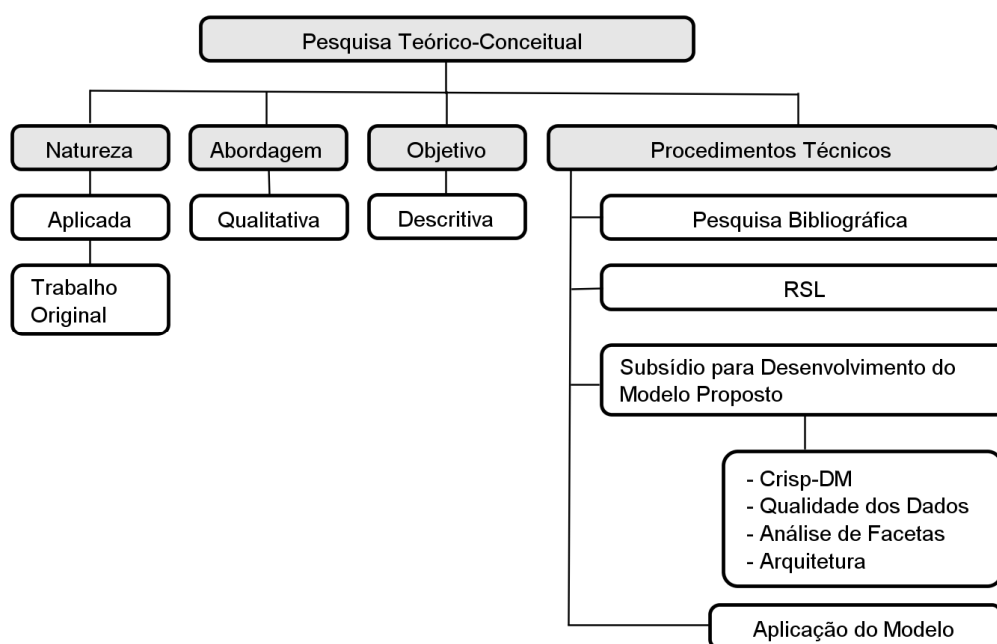
<sup>11</sup>[www.visualliteracy.org/periodic\\_table/periodic\\_table.html](http://www.visualliteracy.org/periodic_table/periodic_table.html)

### 3. MÉTODO DA PESQUISA

Neste capítulo, descreve-se a abordagem adotada para a pesquisa.

Declara-se, inicialmente, a opção pela metodologia teórico-conceitual. Essa abordagem tem o objetivo de elaborar um cenário sobre um tema pré-definido para identificar, desenvolver e aprimorar determinada área do conhecimento (MIGUEL, 2007).

A Figura 18 mostra os aspectos relacionados à natureza, à abordagem, aos objetivos e aos procedimentos técnicos adotados.



**FIGURA 18 – PROPOSTA METODOLÓGICA**  
**FORTE: ELABORADA PELO AUTOR**

Para Wazlawick (2010), um trabalho original corresponde à busca de um conhecimento novo com base em observações e teorias construídas para explicá-lo. No caso deste trabalho, cada fase do modelo proposto foi construída com base em observações e teorias, e teve como finalidade gerar novos conhecimentos. A pesquisa pode ser caracterizada como aplicada, pois atende

ao interesse prático de apoiar processos ou produtos conforme às necessidades do mercado (TURRIONI e MELLO, 2012).

De acordo com Miguel *et al.* (2011), a diferença entre a pesquisa quantitativa e a qualitativa é que a primeira tem sua essência nos elementos do objeto de estudo, ao passo que a qualitativa se preocupa com os processos. Martins (2010) corrobora esse argumento ao afirmar que o foco da abordagem qualitativa são os processos e seus significados. Logo, a abordagem qualitativa pressupõe o delineamento do contexto da pesquisa, não é muito estruturada, tem múltiplas fontes de evidências e atribui importância à realidade organizacional.

Quanto aos objetivos, a pesquisa é classificada como descritiva. Seu intuito é equilibrar as perspectivas acadêmicas e as industriais, proporcionar maior familiaridade com o problema e aprimorar ideias, descobertas ou confirmações de intuições. A pesquisa descritiva visa descrever as características de determinado fenômeno e possibilita o uso de técnicas padronizadas para a observação sistemática (TURRIONI e MELLO, 2012). Da mesma forma, Cervo *et al.* (2007) referem-se à pesquisa descritiva como o ato de observar, registrar, analisar e correlacionar os fatos sem interferências. Para Miguel *et al.* (2011), a pesquisa empírica descritiva direciona para uma compreensão real dos processos, pois tem como preocupação o desenvolvimento de modelos, por meio dos quais se podem descrever as relações causais existentes na realidade.

### **3.1. PROCEDIMENTOS TÉCNICOS**

Neste item, apresentam-se os procedimentos técnicos e as diretrizes para a condução e o desenvolvimento da pesquisa. A base é a pesquisa bibliográfica, por meio da qual o cientista pode encontrar o que há de disponível sobre determinado assunto na literatura (MARCONI e LAKATOS, 1996).

Gil (2009) menciona a pesquisa bibliográfica como um elemento importante para o delineamento do tema, identificando dois grandes grupos: fontes de

“papel” e dados fornecidos por pessoas. Como o escopo deste trabalho é o de uma modelagem conceitual, o método de pesquisa adotado é classificado como teórico-conceitual. Assim, o desenvolvimento do modelo proposto foi embasado em uma ampla revisão bibliográfica e na compilação de conceitos provenientes de artigos de diversos autores.

A pesquisa bibliográfica foi realizada em dois momentos: uma Revisão Sistemática da Literatura (RSL), detalhada na seção 2.1.1, seguida de uma revisão bibliográfica. Para essa revisão bibliográfica, foi utilizado o Portal de Periódico da Capes, no período de 2008 a 2016, com as seguintes palavras chaves: *Knowledge Discovery in Database; Big Data; Product Life Cycle Management; Data Mining; Visualization Techniques; Machine Learning; Social Media e Quality of Data*.

### 3.1.1. REVISÃO SISTEMÁTICA DA LITERATURA - RSL

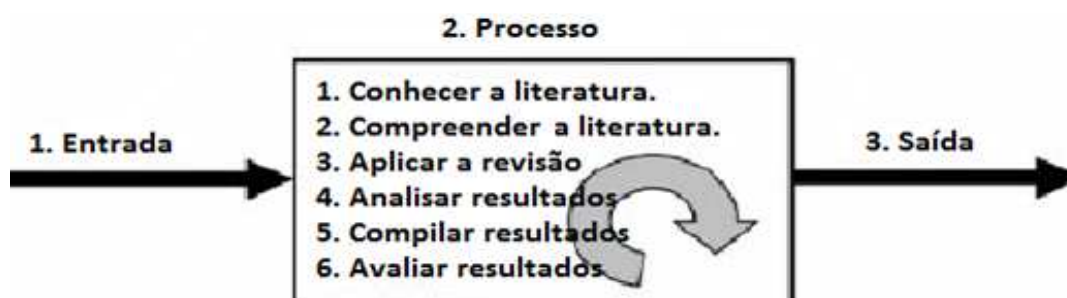
Foi realizada no ano de 2014 e publicada em 2016 uma avaliação Bibliométrica para a Descoberta de Conhecimento na Gestão do Conhecimento (RABELO e CAMPOS, 2016). Essa avaliação forneceu ideias e direcionou a RSL, cujo objetivo foi identificar trabalhos publicados até meados de 2016 a respeito de temas específicos da utilização do *Big Data* no CVP. Esta fase da pesquisa foi norteada pela seguinte questão: **“O que há de pesquisas disponíveis na literatura sobre a área de análise de dados e descoberta de conhecimento, e que utilizam o Big Data para apoiar as fases contempladas no CVP?”**.

Por meio dessa revisão, foram identificadas lacunas no que se refere a pesquisas e aplicações efetivas, bem como à utilização do elevado e heterogêneo volume de informações (texto, vídeo, imagens e som) que poderiam levar à descoberta de conhecimento e apoiar as fases do CVP (introdução, crescimento, maturidade e declínio).

A RSL é uma ferramenta para o mapeamento de publicações científicas sobre temas específicos, apoiando a construção de ideias e a otimização do

conhecimento para embasar a teoria científica sobre tais temas. Levy e Ellis (2006) tratam a RSL como um processo de coleta, conhecimento, compreensão, análise, sintetização e avaliação de um conjunto de artigos científicos. Da mesma forma, Kitchenham e Charters (2007) afirmam que, para atingir os objetivos da RSL, é preciso identificar, avaliar e interpretar os trabalhos realizados sobre o tema de interesse.

Na Figura 19, apresentam-se as três fases propostas por Levy e Ellis (2006) para caracterizar uma RSL. A fase de entrada seria a da seleção dos trabalhos preliminares, isto é, do material relacionado ao assunto de interesse, e também do plano de condução da RSL, que é um documento destinado à descrição das etapas e ferramentas a serem utilizadas na fase de processo. A fase de saída seria a dos relatórios e das sínteses dos resultados.



**FIGURA 19 – FASES DA RSL**  
 FONTE: TRADUZIDO DE LEVY E ELLIS (2006, P. 182).

### 3.1.1.1. PLANEJAMENTO

Portanto, de acordo com Levy e Ellis (2006), de maneira mais abrangente, a RSL busca:

- apoiar a compreensão do assunto de interesse, a descoberta da ausência de pesquisa e a existência das lacunas;
- solidificar a fundamentação teórica;
- prover evidência;
- apoiar as justificativas da pesquisa;
- contribuir para a estrutura da pesquisa e dos objetivos.



Com base nas fases e etapas sugeridas por Levy e Ellis (2006) e Kitchenham e Charters (2007), apresentam-se, na Figura 20, os processos utilizados na pesquisa.



**FIGURA 20 – PROCESSO DE DESENVOLVIMENTO DA PESQUISA**  
**FONTE: ADAPTADO DE LEVY E ELLIS (2006) E KITCHENHAM E CHARTERS (2007).**

### 3.1.1.2. BUSCA E BASE DE DADOS

Por meio de palavras chaves, buscou-se identificar os trabalhos científicos para responder a questão norteadora desta RSL. Inicialmente, foi realizada uma pesquisa geral com diversas palavras; na sequência, optou-se pela busca avançada, utilizando-se o operador lógico “and”. O Quadro 1 ilustra os termos utilizados na pesquisa.

Para a realização da RSL, foram utilizadas diferentes bases de dados das áreas de engenharias, administração, computação e outras:

- *ACM Digital Library*<sup>12</sup>;
- *ScienceDirect*<sup>13</sup>;
- *Engineering Village*<sup>14</sup>;
- *IEEEExplore*<sup>15</sup>;
- *Scopus*<sup>16</sup>.

<sup>12</sup> <http://dl.acm.org/>

<sup>13</sup> <http://www.sciencedirect.com/>

<sup>14</sup> <http://www.engineeringvillage.com/>

<sup>15</sup> <http://ieeexplore.ieee.org/>

<sup>16</sup> <http://www.scopus.com/>

QUADRO 1 – PALAVRAS CHAVES EMPREGADAS NA PESQUISA.

Pesquisa	Palavras chaves	Opção de busca
geral	<ul style="list-style-type: none"> <li>• <i>big data in product life cycle</i>;</li> <li>• <i>knowledge discovery in product life cycle</i></li> <li>• <i>data mining in product life cycle</i></li> <li>• <i>social media in product life cycle</i></li> </ul>	todo o texto do artigo
avançada	<ul style="list-style-type: none"> <li>• <i>“big data” <u>and</u> “life cycle”</i></li> <li>• <i>“knowledge discovery” <u>and</u> “product life cycle”</i></li> <li>• <i>“big data” <u>and</u> “product life cycle”</i></li> <li>• <i>“mining information system” <u>and</u> “product life cycle”</i></li> <li>• <i>“knowledge discovery” <u>and</u> “product life cycle”</i></li> </ul>	palavras chave

FONTE: ELABORADO PELO AUTOR.

O Quadro 2 apresenta os resultados da pesquisa para cada base de dados com as respectivas palavras chaves.

QUADRO 2 – QUANTITATIVO DE ARTIGOS ENCONTRADOS NA RSL.

Termos da busca	ACM		Science Direct		Eng. Village		IEEE		Scopus	
	Quant.	Desde	Quant.	Desde	Quant.	Desde	Quant.	Desde	Quant.	Desde
Big Data	4406	2013	1156	2012	8382	2006	5394	2011	10714	2007
Big data and product Life Cycle	27	2014	2	2014	13	2013	2	2014	12	2013
Knowledge Discovery and Product Life Cycle	9	2002	0	-	6	2008	8	1998	9	2006
Data Mining	30799	1993	35969	1996	59801	1996	73404	1990	86650	2007
Data Mining and product Life Cycle	27	2009	5	2001	44	2006	40	1998	56	2004
Social media and product Life Cycle	10	2014	0	-	6	2010	5	2003	7	2010

FONTE: ELABORADO PELO AUTOR.

### 3.1.1.3. CLASSIFICAÇÃO DOS ARTIGOS SELECIONADOS

Por meio de um estudo preliminar, os artigos selecionados foram organizados e agrupados em três categorias, conforme objetivo contido na questão norteadora desta RSL:

- categoria A – aborda diretamente o tema da pesquisa com foco na questão (*Big Data* e CVP);
- categoria B – aborda diretamente o tema da pesquisa, mas não considera a questão como foco principal
- categoria C – aborda indiretamente o tema da pesquisa em questão com foco em outros temas correlatos.

Paralelamente às buscas, realizou-se a seleção dos artigos que aparentemente demonstravam evidências significativas, isto é, que estavam em consonância com a questão principal dessa RSL. Assim, a interpretação do artigo deveria abordar, de alguma forma, a utilização do conhecimento extraído por meio de dados (estruturados ou não estruturados) e sua aplicação em algumas ou todas as fases do CVP. Ou seja, deveria conter evidências de que tinham sido utilizados dados internos ou externos à empresa e, posteriormente, aplicados na descoberta de conhecimento por meio de análise, mineração e visualização dos dados. Em suma, deveria demonstrar que tais dados teriam tido utilidade em alguma fase no processo do CVP.

Muito embora não mencionassem os termos *Big Data* e CVP, destacado na questão da seção 2.1.1, alguns dos artigos apresentaram ideias úteis que poderiam ser aplicadas no contexto da pesquisa. Em razão disso, os artigos foram classificados em categorias (A, B e C).

Em resumo, a Figura 21 ilustra as etapas para a seleção e a classificação dos artigos. Na etapa um, ocorreram a seleção dos bancos de dados e a formatação da busca. Na etapa dois, foi realizada a pré-seleção com base no título, nas palavras chaves e nos resumos. Por fim, na etapa três, foi identificado se o artigo atendia à RSL.

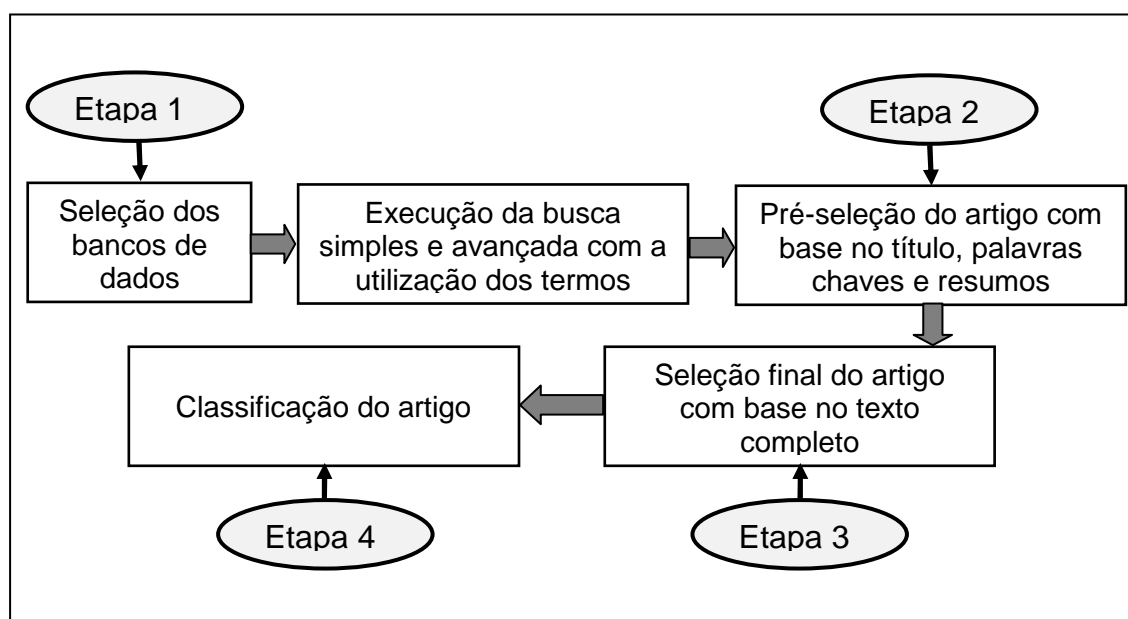


FIGURA 21 – ETAPAS DO PROCESSO DE SELEÇÃO  
 FONTE: ELABORADA PELO AUTOR

### 3.1.2. SUBSÍDIOS PARA O DESENVOLVIMENTO DO MODELO PROPOSTO

Aqui serão descritos alguns modelos e técnicas utilizados como subsídio para o desenvolvimento do modelo proposto.

#### 3.1.2.1. MODELO KDD E CRISP-DM

A base do modelo proposto nesta tese foi extraída de ideias provenientes dos modelos KDD e Crisp-DM. O modelo proposto compreende as diferentes estruturas de dados de fontes distintas, podendo, portanto, atender às novas demandas na área do *Big Data*.

#### 3.1.2.2. QUALIDADE DOS DADOS

As dimensões sobre a qualidade de dados encontradas na literatura (WANG, 1998; OLSON, 2003 ; BATINI *et al.*, 2015) e o modelo “3As Data Quality-in-Use model”, de Merilo *et al.* (2015) foram utilizados como subsídio para o desenvolvimento da Fase II do modelo proposto, descrito no Capítulo 4.

### **3.1.2.3. ANÁLISE DE FACETAS**

Utiliza-se a análise de facetas nesta tese no desenvolvimento das atividades relacionadas com a qualidade e as características das fontes de dados, apresentadas na Fase II do modelo proposto, descrito no Capítulo 4.

### **3.1.2.4. ARQUITETURA**

O objetivo da arquitetura desenvolvida nesta tese, além de auxiliar no processo de descoberta de conhecimento, é fornecer apoio em todo o desdobramento das fases desenvolvidas para o modelo proposto e apresentadas no Capítulo 4.

Como o modelo proposto agrega as soluções tradicionais e as do *Big Data*, faz-se necessário que a estrutura da arquitetura utilizada como apoio a essa modelagem contemple ambas as soluções e, de forma clara e concisa, descreva as fases do modelo proposto em camadas. Dessa forma, como embasamento para o desenvolvimento dessa arquitetura foram analisadas as arquiteturas apresentadas na literatura (ZHUANG *et al.*, 2016; ZHANG *et al.*, 2017); WU *et al.*, 2014).

### **3.1.3. APLICAÇÃO DO MODELO PROPOSTO**

Para avaliar a utilização e identificar possíveis aperfeiçoamentos, apresenta-se, no Capítulo 5, a aplicação do modelo proposto em ambiente real em uma empresa de indústria e comércio de vestuário, roupas e acessórios femininos, que acompanha as tendências de moda. Denominada Amarelo Manga (INDÚSTRIA AM), a indústria atua no mercado há aproximadamente sete anos, e está situada na cidade de Maringá, no estado do Paraná.

#### 4. DESENVOLVIMENTO DO MODELO PROPOSTO

O Crisp-DM não está sendo atualizado para as demandas do *Big Data* e tampouco para a ciência de dados moderna (PIATETSKY, 2014). Nesse contexto, O modelo proposto da presente tese tem características genéricas, o que torna possível realizar o processo de descoberta de conhecimento do início ao fim e atender às demandas do *Big Data* sem deixar de considerar os métodos tradicionais de MD.

Diferentemente das arquiteturas e modelos encontrados na literatura, o modelo proposto apresenta uma abordagem detalhada por meio de fases, etapas e atividades para a execução do processo de descoberta de conhecimento, além da sua especificidade no projeto informacional. O modelo proposto também prevê a execução antecipada de atividades de ETL na modificação de conjunto de dados, para aplicação em soluções tradicionais.

Atualmente, existem várias opções tecnológicas para o *Big Data*, como mostram Turck e Hao (2016) no Anexo B. Essa diversidade de escolhas, conseqüentemente, gera indecisões. Assim, este trabalho propõe a construção de um modelo composto por quatro fases, como mostra a Figura 22, nas quais se buscam organizar as atividades de descoberta de conhecimento, além de auxiliar a compreensão das soluções tecnológicas necessárias.

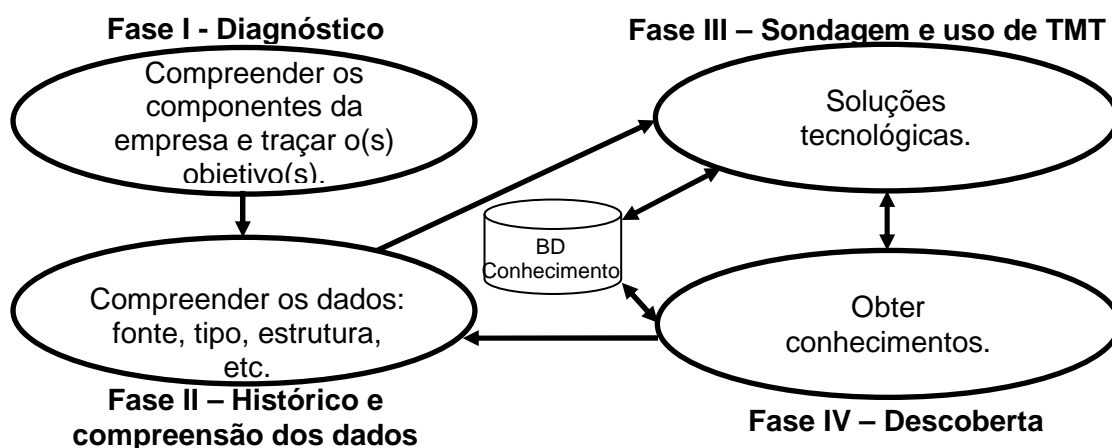


FIGURA 22 – FASES DO MODELO PROPOSTO  
 FONTE: ELABORADA PELO AUTOR.

Krishnan (2013), em relação ao ciclo de vida do processamento de dados, diferencia o modelo tradicional e *Big Data*. Para o modelo tradicional, analisam-se os dados para determinar um conjunto de requisitos, o que permite realizar descobertas e obter a modelagem dos dados. Para o *Big Data*, os dados são extraídos e armazenados em uma plataforma, à qual, em seguida, é aplicada uma camada de metadados. Dessa forma, surge uma estrutura de dados que permite transformação e análise.

Portanto, o modelo proposto carrega em sua essência as etapas definidas no KDD e atende às demandas do *Big Data*, ou seja, atende aos processos tradicionais e *Big Data*.

Cada fase do modelo proposto busca respostas ordenadas para as questões que aos poucos vão desenvolvendo concepções que propiciam maior amadurecimento e, prevê soluções tecnológicas para as possíveis dificuldades encontradas nas etapas dos processos de manipulação dos dados. Dessa maneira, ao final da aplicação do modelo, espera-se adquirir ampla visão e diferentes perspectivas tecnológicas para promover a descoberta de conhecimento.

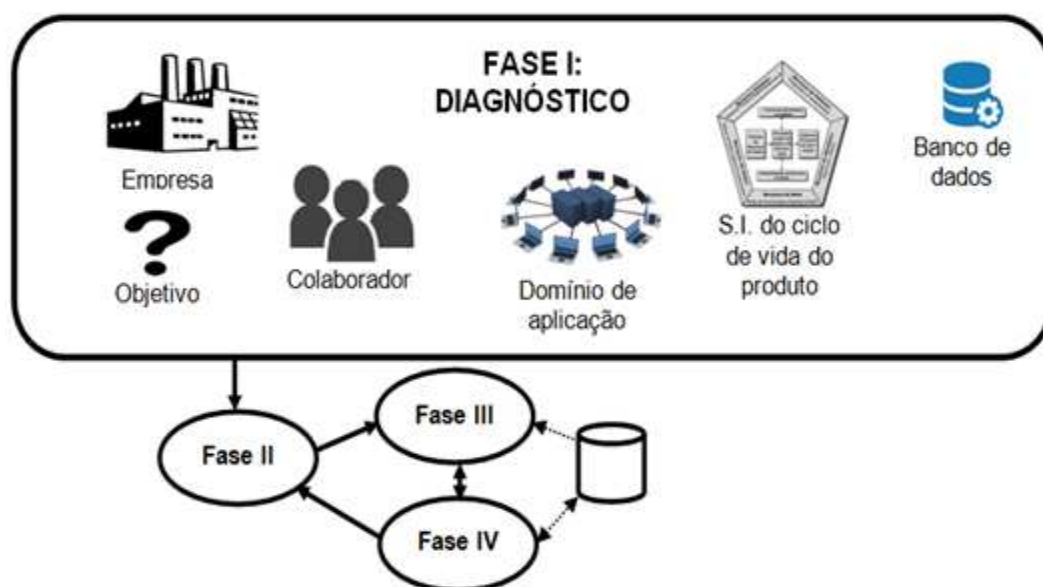
A primeira fase, “Diagnóstico”, corresponde às atividades que permitem compreender a empresa, isto é, obter familiaridade no processo de negócio e no domínio de aplicação. A segunda, “Histórico e compreensão dos dados”, refere-se ao levantamento das fontes de dados e suas características, cujo foco é a qualidade da informação. A terceira, “Sondagem e uso de TMT (Técnicas, Métodos e Tecnologias)”, é composta pelo levantamento de técnicas, métodos e tecnologias utilizadas atualmente. A quarta, “Descoberta”, caracteriza-se pela análise dos dados e pelo uso de técnicas de visualização na descoberta de conhecimento.

#### **4.1. FASE I – DIAGNÓSTICO**

Nesta fase, discutem-se quais conhecimentos são relevantes para o projeto informacional no PDP, assim como os questionamentos que podem surgir das atividades do processo de planejamento do produto.

Seus detalhes, ilustrados na Figura 23, têm como objetivo conhecer o contexto em que o trabalho de descoberta de conhecimento está inserido e contemplam as seguintes atividades:

- identificação do domínio de aplicação da empresa;
- conhecimento dos colaboradores;
- verificação dos sistemas de informação e banco de dados;
- pesquisa a respeito dos equipamentos de *hardware* existentes;
- levantamento da atual infraestrutura tecnológica;
- levantamento das necessidades de dados referente ao produto;
- levantamento de outras informações que permitam conhecer o ambiente da empresa.



*FIGURA 23 – FASE I DO MODELO PROPOSTO*  
*FONTE: ELABORADA PELO AUTOR*

Nesta fase ocorre a construção do objetivo da descoberta de conhecimento, ou seja, da diretriz que será adotada na realização da pesquisa. Com base nos modelos de referência para o PDP, destaca-se, na fase do projeto



informacional, a obtenção de informações relevantes referentes ao produto. Ratificando essa perspectiva, Li *et al.* (2014) concluem que, na era do *Big Data* e com o desenvolvimento generalizado da rede mundial de computadores, o comportamento e as necessidades dos clientes podem ser analisados e considerados nos projetos de produtos novos ou existentes.

A Figura 24 indica, por meio das linhas tracejadas, a aplicação do modelo proposto nas atividades do projeto informacional.

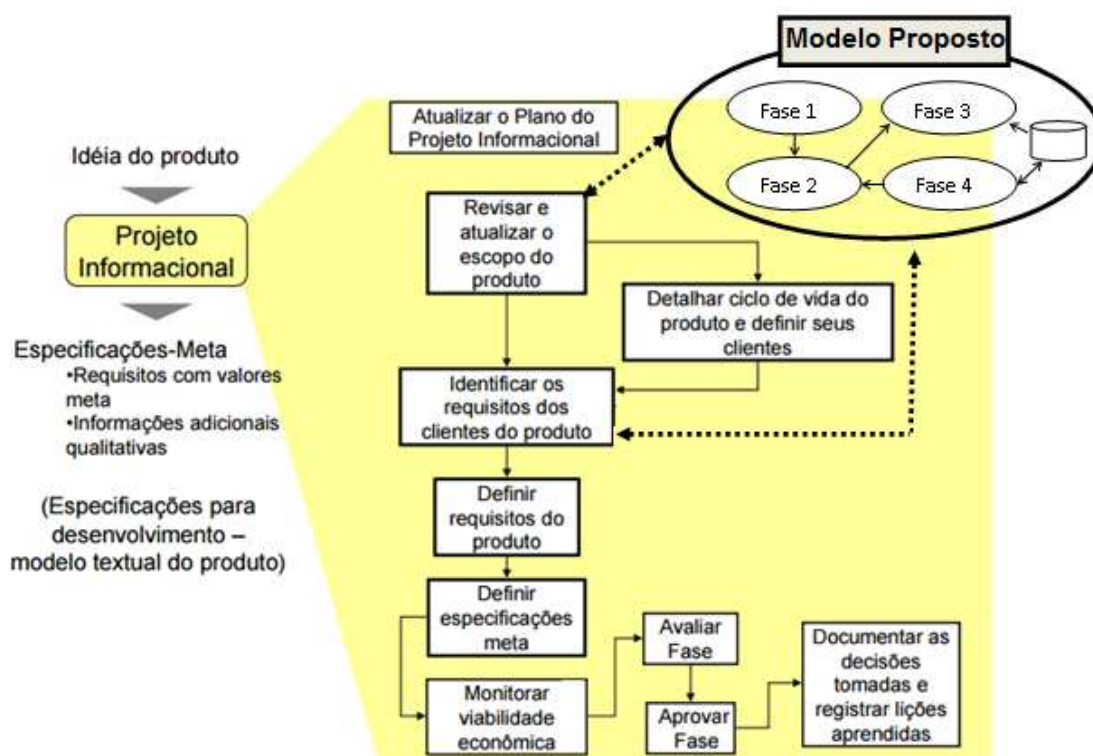


FIGURA 24 – ATIVIDADES DA FASE DE PROJETO INFORMACIONAL E A INSERÇÃO DO MODELO PROPOSTO

FONTE: ADAPTADO DE ROZENFELD ET AL. (2006, P. 212).

Com referência ao projeto informacional, foi elaborada uma lista com alguns pontos que podem pautar a definição do objetivo desta fase:

- produtos concorrentes e similares;
- novos nichos de clientes em potencial;
- necessidades dos clientes;
- requisitos dos clientes sobre o produto;
- requisitos do produto;

- ideias para novos produtos.

Ressalta-se que a construção dos objetivos da descoberta de conhecimento deve ser discutida juntamente com os colaboradores do domínio de aplicação.

#### **4.2. FASE II – HISTÓRICO E COMPREENSÃO DOS DADOS**

Esta fase tem como objetivo conhecer os dados já produzidos, ou seja, o conhecimento que foi explicitado e registrado. Seguindo o objetivo traçado na primeira fase, é possível manter o foco no que se deseja e, dessa forma, pesquisar e identificar as fontes interna e externa de dados.

As tarefas que podem ser realizadas com base nas fontes de dados são:

- levantamento dos tipos e da estrutura de dados;
- identificação do propósito com que foram gerados;
- pesquisa a respeito da qualidade e da necessidade de limpeza, transformação e filtragem;
- identificação do volume e dos demais atributos existentes.

No modelo proposto, sugere-se a aplicação da análise de facetas, cujo intuito é compreender melhor os componentes chaves e as características das fontes de dados. Para tanto, é fundamental estudar previamente a fonte de dados que gerará o conhecimento, uma vez que é esse estudo que direcionará a escolha de soluções *Big Data* e/ou tradicionais para o armazenamento, processamento e análise dos dados. Para o cenário *Big Data*, o modelo proposto neste capítulo utiliza como base o diagrama de processo apresentado por Krishnan (2013), detalhado na Seção 3.3.5.

Para um melhor entendimento, a segunda fase foi dividida em etapas, como mostra a Figura 25:

- identificação das fontes interna e externa de dados;
- criação de facetas;
- avaliação das características dos dados;
- decisão quanto à solução tecnológica de armazenamento.

Na etapa de decisão quanto à solução tecnológica de armazenamento, o modelo proposto sugere as tarefas de pré-filtragem, pré-limpeza e pré-transformação dos dados, buscando proporcionar sua qualidade e criando condições para a decisão quanto à solução a ser utilizada.

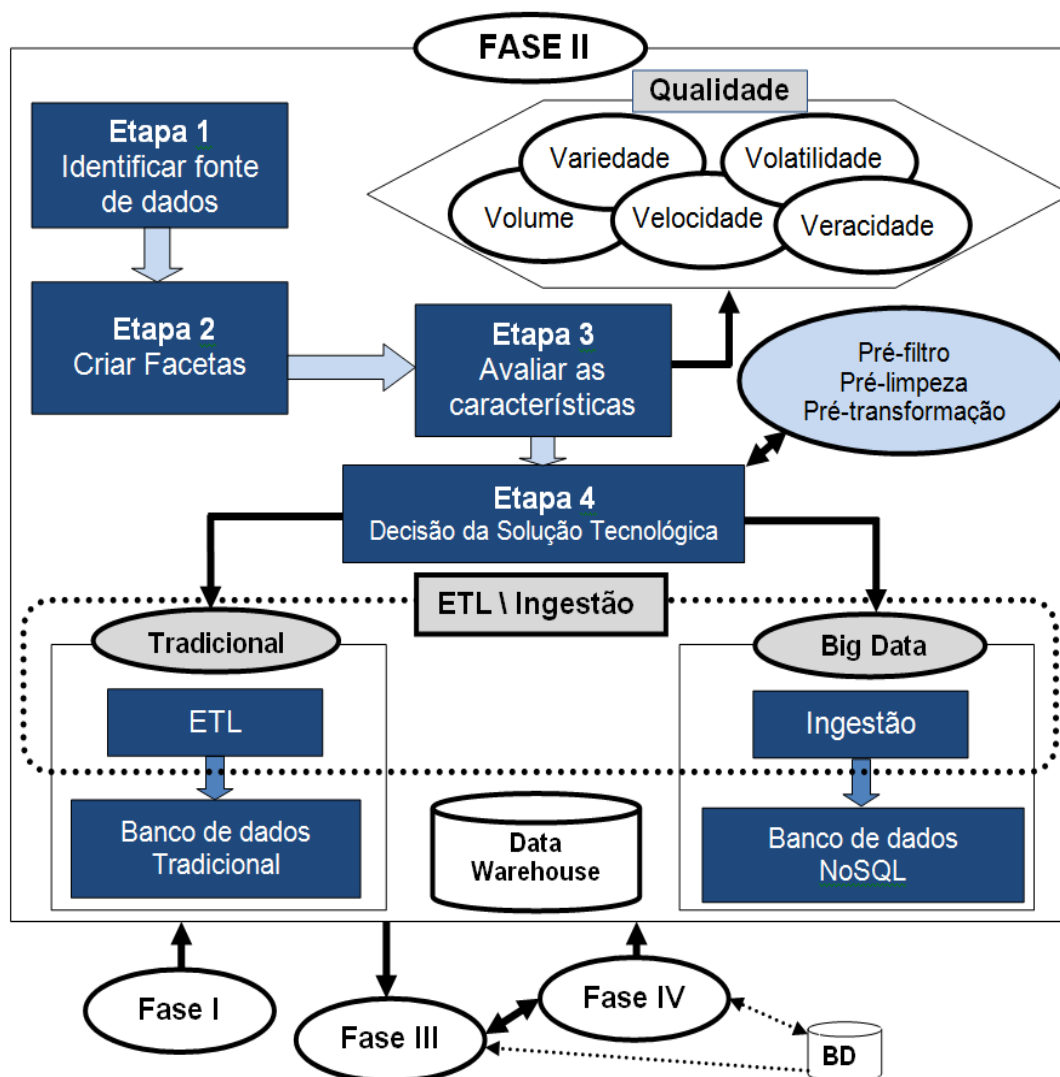


FIGURA 25 – HISTÓRICO E COMPREENSÃO DOS DADOS  
FONTE: ELABORADA PELO AUTOR

#### 4.2.1. ETAPA 1 - IDENTIFICAÇÃO DAS FONTES DE DADOS

Os trabalhos encontrados na literatura evidenciaram ou utilizaram as mais diversas fontes de dados aplicadas no processo de descoberta de conhecimento e tais fontes podem ser utilizadas no projeto informacional (FAN e GORDON, 2014; LACHMAYER *et al.*, 2014; CARR *et al.*, 2015; Li *et al.*,

2015; ZHANG *et al.*, 2017). Nesta subseção, por meio da Figura 26, mostra-se uma visão geral da integração dessas fontes de dados.

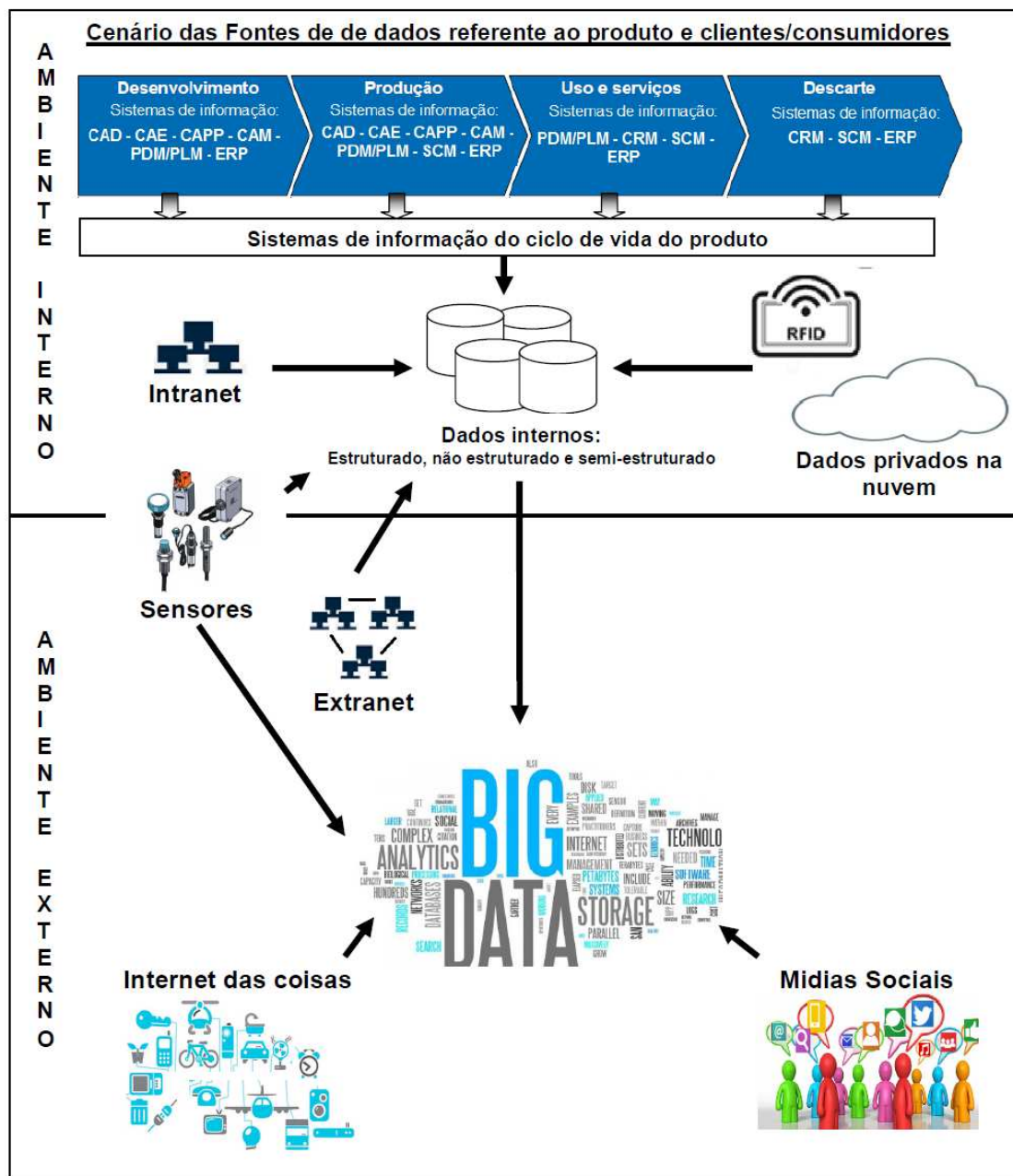


FIGURA 26 – FONTES DE DADOS  
FONTE: ELABORADA PELO AUTOR

Neste trabalho, referente ao armazenamento e à produção dos dados, as fontes estão divididas em ambiente interno e externo. No ambiente interno estão os dados produzidos e armazenados na própria empresa e que geralmente são de acesso exclusivo. No ambiente externo estão os produzidos

e armazenados em recursos tecnológicos que não pertencem à empresa (mídias sociais, fóruns, jornais, entre outros). Conforme ilustra a Figura 26 as fontes são:

- sistemas de informação - apoiam o processo de desenvolvimento do produto (CAD, CAE, CAPP, CAM, PDM/PLM, ERP, SCM);
- intranet - rede local de computadores, circunscrita ao limite interno da empresa;
- RFID - tecnologia utilizada para rastrear, identificar e gerenciar produtos;
- internet das coisas (em inglês, *Internet of Things* ou IoT) - conecta os produtos do dia-a-dia, como aparelhos eletrodomésticos, meios de transporte e máquinas industriais;
- sensores - geram dados referentes à utilização do produto;
- mídias sociais - postagens das necessidades e opiniões dos clientes quanto a produtos, tendências e inovação;
- extranet - rede privada que partilha informações com fornecedores, clientes, parceiros ou até mesmo entre pessoas da própria empresa em locais geograficamente distintos.

As tecnologias emergentes, como sensores, RFID e comunicação em rede, favorecem o desenvolvimento de uma nova área de conhecimento denominada IoT. Essa tecnologia oferece uma estrutura para conectar dispositivos eletrônicos utilizados no dia-a-dia e gerar fluxo e troca de informações em tempo real. Os envolvidos (*stakeholders*<sup>17</sup>) nos processos ao longo do CVP podem obter dados com a utilização de dispositivos de informações de produtos embutidos PEIDs (*product embedded Information devices*), implantados nos recursos da produção e nas principais peças e componentes que fazem parte do produto (ZHANG *et al.*, 2017). A maioria das aplicações relacionada à IoT gera um elevado volume de dados que, por meio de análise, modelagem e transformação, levam à produção de conhecimento. As tecnologias digitais atendem às necessidades de armazenamento, mas as

---

<sup>17</sup> Pessoa ou grupo que tem papel direto ou indireto na gestão e nos resultados da organização.

dificuldades na velocidade da descoberta de conhecimento e na gestão de dados permanecem (MISHRA, 2015).

As fontes de dados existentes em mídias sociais possibilitam o acompanhamento do “comportamento do consumidor”. Segundo Kotler e Keller (2012), esse acompanhamento é definido como o estudo da conduta do indivíduo em relação à seleção, à compra, à utilização e ao descarte do produto, tendo em vista a satisfação de suas necessidades e seus desejos. Tal conduta pode ser influenciada por fatores culturais, sociais e pessoais.

Nessa mesma perspectiva, as oportunidades podem estar disponíveis em mídias sociais, por meio das quais é possível acompanhar as opiniões dos consumidores, suas alterações de sentimentos e a percepção em relação aos produtos (CARR *et al.*, 2015). Algumas ferramentas para esse acompanhamento são desenvolvidas com métodos linguísticos computacionais, como processamento de linguagem natural e outros métodos de análise de texto em qualquer nível de granularidade<sup>18</sup>.

Com base nas informações extraídas das mídias sociais quanto a pessoas e produtos, é possível: *i)* prever o movimento dos mercados de ações; *ii)* identificar tendências do mercado e *iii)* analisar defeitos dos produtos (FAN e GORDON, 2014). Os métodos utilizados nessas análises incidem sobre as menções do consumidor aos produtos e consistem em gerar listas de pontos positivos e negativos e analisar a frequência dos termos mencionados. Os resultados gerados podem ser estruturados e armazenados em banco de dados relacionais, que favorecem processos de descoberta de conhecimento por meio de métodos tradicionais.

#### **4.2.2. ETAPA 2 – FACETAS PARA A FONTE DE DADOS**

Para melhor entender as características das fontes de dados, o modelo proposto desenvolve a análise de faceta. Shiri (2014) utiliza a análise de faceta

---

<sup>18</sup> Nível de detalhe contido nas unidades de dados. Quanto menor é o nível de detalhes, maior é o nível de granularidade e, conseqüentemente, maiores são as dificuldades de gerenciamento.

para categorizar termos e conceitos para o *Big Data* e, assim, facilitar a compreensão e a investigação. Diferentemente de Shiri, a análise de faceta desenvolvida neste modelo visa à compreensão da fonte de dados e das características relacionadas à qualidade dos dados, conforme apresentado na literatura (OLSON, 2003; BATINI *et al.*, 2015; MERINO *et al.*, 2016).

Merino *et al.* (2016) argumentam que um dos desafios de se trabalhar com volume de dados é sua qualidade, cujo papel é decisivo na confiabilidade da entrada de dados, já que níveis indesejados na qualidade podem gerar e espalhar erros imperceptíveis ao longo do processo de descoberta de conhecimento. Afirmando também que existem diferentes modelos de qualidade de dados cuja intenção é avaliar somente dados regulares. Portanto, nenhum modelo foi adaptado para o *Big Data*. Com o intuito de ampliar esse cenário, a análise de faceta empregada neste modelo tem como objetivo compreender dados estruturados e não estruturados.

Posteriormente à escolha da faceta para a fonte de dados, foram realizados estudos que levantaram e identificaram suas subfacetas, levando à definição das características, conforme apresentado no Quadro 3.

QUADRO 3 – FACETA FONTE DE DADOS.

Faceta	Sub-Facet	Características	Descrição
Fonte de dados	Qualidade	Acuracidade	Medida de quão próximo o dado se aproxima da realidade
		Disponibilidade em tempo	Grau de atualização
		Relevância	Utilidade dos dados
		Completeness	Amplitude e profundidade
		Compreensibilidade	Facilidade de entendimento
		Confiabilidade	Indica quão correta os dados estão
		Abrangência	Aspectos relevantes da realidade
		Redundância	Redução dos recursos informativos
	Acesso	Público	Dados - (licença livre)
		Privado	Dados - (licença proprietário)
	Finalidade	Gestão	Sistemas de informação (operacional, gerencial, executivo, entre outros)
		Controle	
		Operação	
		Entretenimento	Mídias \ canais \ jogos
		Pesquisa	Artigos \ livros \ apostila
		Informativo	Fórum \ jornais \ blogs \ revistas
	Estrutura	Estruturado	Exemplo: banco de dados
		Semi estruturado	Exemplo: XML \ RDF \ OWL

	Volume	Não estruturado	Exemplo: Vídeos\ textos\ fotos
		Elevado	A empresa não comporta armazenamento interno
		Médio	Comporta o armazenamento interno com restrição
	Natureza	Baixo	Permite armazenamento interno
		Quantitativos	Discreto
			Contínuo
		Qualitativo	Nominal
			Ordinal
	Usuário	Pesquisador	
		Estudante	
		Consumidor	
		Colaborador	
		Empresa\ indústria \ comércio	

*FONTE: ELABORADO PELO AUTOR*

A faceta da fonte de dados descrita no Quadro 3 está representada por suas subfacetas e respectivas características.

- Qualidade – dimensionalidade apresentada na literatura:
  - acuracidade - medida da proximidade dos dados em relação à realidade; para Olson (2003), essa característica é uma das mais relevantes e, caso o nível não seja satisfatório, não importa o significado das outras características;
  - disponibilidade – o quanto atualizado o dado está para ser utilizado no momento necessário;
  - relevância – grau de utilidade inserido no contexto;
  - completude - indica a dimensionalidade e a profundidade esperadas, isto é, indica o nível de presença dos dados necessários para atender à demanda do negócio;
  - compreensibilidade e confiabilidade - grau de facilidade de entendimento e confiança;
  - abrangência - capacidade de representar os aspectos relevantes da realidade,
  - redundância - refere-se à capacidade de representar com o uso mínimo de recursos informativos.
- Acesso – referente às características públicas e privadas:



- pública – dados de acesso livres, isto é, os proprietários não cobram por sua utilização, a exemplo dos dados de portais governamentais, como IBGE, ou até mesmo de fóruns de discussão abertos ao público,
  - privada - dados pertencentes à própria empresa, a exemplo dos bancos de dados dos sistemas de informação (CRM, ERP, SCM, entre outros).
- Finalidade – origem e o objetivo para qual foi desenvolvido.
- Estrutura - organização do armazenamento. A literatura apresenta de forma explícita três tipos: *i)* dados estruturados, geralmente pertencentes a um sistema de gerenciamento de banco de dados; *ii)* dados não estruturados, normalmente codificados em linguagem natural; *iii)* dados semiestruturados tem uma estrutura que o autor descreve (MERILO *et al.*, 2015; BATINI *et al.*, 2015);
- Volume - quantidade de dados existentes na fonte tendo em vista sua grandeza (gigabyte, terabyte, petabyte, etc.). Por convenção, foram adotados três níveis: alto, médio e baixo, os quais correspondem à possibilidade de armazenamento interno na empresa:
  - elevado - a quantidade de dados é maior que os recursos de armazenamento disponíveis na empresa;
  - médio - a empresa suporta o armazenamento, porém não garante o processamento e/ou gerenciamento,
  - baixo – as demandas são supridas pelos recursos de armazenamento e processamento.
- Natureza - refere-se a dados estruturados, quantitativos (discreto ou contínuo) e qualitativos (nominal ou ordinal):
  - quantitativos - representados por valores numéricos que podem ser discretos ou contínuos: os discretos referem-se a contagens, por exemplo, o número de compra de um determinado produto por mês; os contínuos são determinados por escala, por exemplo, área, volume, peso e velocidade;

- qualitativos - utilizados para atribuir rótulos e identificar os atributos: dados nominais não apresentam ordenação, permitindo operações de igualdade ou de diferença, por exemplo, a respeito do estado civil (casado, divorciado, solteiro, viúvo); dados ordinais possibilitam ordenação, por exemplo, dos níveis de escolaridade, de temperatura (frio, morno e quente).
- Usuários - pessoas que utilizam ou produzem dados.

O detalhamento proporcionado pela construção da “faceta” da fonte de dados cria possibilidades de apoio às atividades das próximas etapas e fases do modelo proposto.

Evidentemente, os trabalhos relacionados à descoberta de conhecimento têm como objeto central estudar os dados. Dessa forma, a compreensão dos dados em diferentes perspectivas pode evitar possíveis transtornos no processo de descoberta de conhecimento.

#### **4.2.3. ETAPA 3 – AVALIAÇÃO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS**

Depois da definição da fonte, é estabelecido o conjunto de dados, para, então, serem avaliadas suas características.

Sabe-se que, em qualquer tipo de arquitetura projetada para o gerenciamento de dados do *Big Data*, devem ser consideradas as características referentes aos “3V’s” (volume, velocidade e variedade). Além dos “3V’s”, este trabalho adota as características veracidade e valor, formando-se os “5V’s”, conforme foi discutido na Seção 3.3. Ressalta-se que a preocupação é a descoberta de conhecimento para o projeto informacional, independentemente do ambiente (*Big Data* ou tradicional).

Davenport (2014) refere-se a algumas diferenças entre *Big Data* e análise tradicional. No que se refere ao volume dos dados, o autor afirma que o volume

do *Big Data* está acima de cem terabytes, com fluxo constante de dados; o do tradicional é de dezenas de terabytes ou menos, com “pool” estático de dados.

Para a elaboração do formulário representado no Quadro 4, além dos “5V’s”, consideram-se as características referentes à subfaceta qualidade, descritas no Quadro 3.

QUADRO 4 – FORMULÁRIO DO DETALHAMENTO DO CONJUNTO DE DADOS.

Detalhamento do conjunto de dados				
<b>1. Objetivo da descoberta de conhecimento</b>				
<b>2. Informações do conjunto de dados</b>				
<b>2.1 - Fonte(s):</b>				
<b>2.2 - Formato(s) (variedade):</b> ( ) Estruturado ( ) Não Estruturado ( ) Semiestruturado				
<b>2.3 - Atualização (alimentação da fonte de dados):</b> ( ) Tempo real ( ) Diário ( ) Semanal ( ) Mensal ( ) Anual ( ) Outras: _____				
<b>2.4 - Intervalo de tempo - (período inicial e final):</b>				
<b>2.5 - Gerador (autor da fonte):</b>				
<b>2.6 - Observações:</b>				
<b>3. Acompanhamento das características do conjunto de dados</b>				
Características	Avaliação			Observações
	1ª	2ª	3ª	
<b>3.1 - Credibilidade</b>				
<b>3.2 - Veracidade</b>				
<b>3.3 - Imparcialidade</b>				
<b>3.4 - Utilidade</b>				
<b>3.5 - Atualidade</b>				
<b>3.6 - Complexidade</b>				
<b>3.7 - Objetividade</b>				
<b>3.8 - Inconsistência</b>				
<b>3.9 - Apresenta qualidade mínima?</b>	Sim ( )	Sim ( )	Sim ( )	
	Não ( )	Não ( )	Não ( )	
<b>Legenda:</b> (1) muito baixa- (2) baixa - (3) média - (4) alta - (5) muito alta				
<b>4. Lista de recursos tecnológicos</b>				
<b>4.1 - Hardware:</b>				
<b>4.2 - Disponibilidade de armazenamento:</b>				

FONTE: ELABORADO PELO AUTOR

Para a proposição das características do conjunto de dados, é imprescindível atentar para o objetivo previamente traçado na Fase I e para a identificação das fontes de dados, descrita na Etapa 1 da presente fase. Levantamentos relacionados à qualidade, juntamente com as informações fornecidas pela análise de facetas, asseguram a aplicação desse formulário.

Para explicar melhor os itens do formulário mostrado no Quadro 4, destacam-se:

- Item 1 – resumo do objetivo da descoberta de conhecimento traçado na Fase I;
- Item 2 – informações do conjunto de dados:
  - Item 2.1 - nome da fonte que gerou o conjunto de dados;
  - Item 2.2 – conforme a literatura, os dados são de três tipos: *i)* estruturados, semiestruturados e não estruturados (MERILO *et al.*, 2015; BATINI *et al.*, 2015); Esse item pode ser modificado no decorrer do processo de descoberta de conhecimento, pois existe a possibilidade de a estrutura ser modificada;
  - Item 2.3 - tempo em que ocorrem as atualizações da fonte de dados, sendo relevante considerar a dimensão da disponibilidade apresentada por Olson (2003), além dos atributos de concomitância, atualidade e oportunidade descritos por Merilo *et al.* (2015);
  - Item 2.4 - refere-se ao período necessário para a descoberta de conhecimento;
  - Item 2.5 – refere-se aos responsáveis pela produção do conteúdo, os quais podem ser: a própria organização, empresas concorrentes, pessoas físicas, ONGs, governo, entre outras. Indiretamente, esse campo fornece indícios do grau de credibilidade da fonte.

O Item 3 refere-se ao acompanhamento das características do conjunto de dados, cuja avaliação é realizada em conjunto com os colaboradores do

domínio de aplicação. Trata-se de uma medida complexa, subjetiva e individual, podendo variar de fonte para fonte. A avaliação e a mensuração do conjunto de dados estabelecem até que ponto eles podem ser utilizados na descoberta de conhecimento; para isso, é necessário atentar para as propriedades de tais características.

- Item 3.1 – credibilidade: mede o nível de autenticidade dos dados, de forma a se ter certeza de que eles correspondem a uma medida real ou de que foram obtidos por métodos de aquisição apropriados;
- Item 3.2 – veracidade: mede a incerteza ou a imprecisão dos dados, de forma a avaliar a confiança por eles proporcionada, isto é, se os dados ou os processos utilizados são confiáveis, e assim identificar sua importância;
- Item 3.3 - imparcialidade: mede o nível de influência da fonte dos dados, avaliando sua isenção na intervenção de seus dados.
- Item 3.4 – utilidade: mede o propósito da fonte de dados, avaliando sua pertinência em relação ao objetivo da descoberta de conhecimento; em alguns casos, essa característica pode ser estabelecida no final do processo de descoberta de conhecimento;
- Item 3.5 - atualidade - mede o grau de atualização de um conjunto de dados, em correspondência com o objetivo traçado na Fase I;
- Item 3.6 - complexidade - mede o grau de compreensibilidade do conjunto de dados, avaliando quão complexa é a estrutura dos dados e se há necessidade de ferramentas intermediárias para torná-la mais simples e viável;
- Item 3.7 - objetividade - mede o grau de assertividade dos dados, avaliando a conformidade dos dados em relação ao objetivo do processo de descoberta de conhecimento;
- item 3.8 - inconsistência - mede o grau de ruídos existentes nos dados, avaliando dados sem sentido e que não podem ser entendidos ou interpretados corretamente.

As avaliações assumem pontuações que variam de 1 a 5, sendo que 1 representa intensidade muito baixa e 5, muito alta. Como a análise é individualizada, a intensidade atribuída às características assume significado distinto, isto é, a relevância da intensidade independe da pontuação designada para as características. Por exemplo, o ideal é que a intensidade da característica inconsistência seja próxima do nível 1 e a da credibilidade, próxima do nível 5.

Por convenção, a “qualidade mínima”, apresentada no item 3.9, indica se todas as características possuem pontuação próxima do nível de intensidade desejado. Com base nisso, pode-se considerar o conjunto de dados habilitado para o processo de descoberta de conhecimento. A “qualidade mínima” será utilizada nas fases seguintes do modelo proposto.

A avaliação inicial do conjunto de dados mantém sua natureza bruta; no entanto, ao longo das fases do modelo proposto, são executadas atividades que alteram esse conjunto, o que implica a necessidade de uma nova avaliação das características.

#### **4.2.4. ETAPA 4 – DECISÃO DAS SOLUÇÕES TECNOLÓGICAS**

Esta etapa visa o processamento do conjunto de dados bruto e à seleção das soluções tecnológicas. Nesse processamento, busca-se melhorar a qualidade do conjunto de dados em função da solução a ser utilizada. A estrutura do conjunto de dados pode ser alterada para ser suportada pelas soluções tradicionais e/ou *Big Data*, sendo assim, um conjunto de dados com características sustentadas por *Big Data* pode sofrer alterações para se adequar as soluções tradicionais. As diferenças entre as dimensões dessas soluções são apresentadas no Quadro 5.

Em razão das diferenças entre os dados para a descoberta de conhecimento, que pode utilizar soluções tradicionais ou *Big Data*, Soares (2013) recomenda a adoção das seguintes práticas para garantir a qualidade dos dados:

- melhorar a qualidade de dados estruturados por meio do potencial que os dados semiestruturados ou não estruturados podem oferecer,
- realizar limpeza no volume dos dados antes ou após o armazenamento.

QUADRO 5 – COMPARAÇÃO ENTRE SOLUÇÕES TRADICIONAIS E BIG DATA.

Dimensões	Tradicional	<i>Big Data</i>
Processamento	Orientado em lote.	Orientado em lote e/ou em tempo real.
Formato	Estruturados.	Estruturados, semiestruturados e não estruturados.
Níveis de confiança	Os dados precisam estar em um BD e em boas condições para análise.	Os dados precisam ser filtrados para eliminar os ruídos. A qualidade dos dados pode impedir a análise.
Limpeza dos dados	Os relacionamentos dos dados são identificáveis.  Os dados estão limpos antes de ser carregados no BD.	Os relacionamentos dos dados podem não ser identificados.  O volume e a velocidade podem exigir métodos de análise em fluxo, dessa forma, a limpeza pode reduzir requisitos de armazenamento.
Localização de análise	Os dados são direcionados para o processamento.	Os processos e a análise são direcionados para os dados.
Gerenciamento	Administração de um alto percentual dos dados.	Administração de um percentual reduzido de dados, em razão de seu volume e de sua velocidade.

FONTE: ADAPTADO E TRADUZIDO DE SOARES (2013, P. 49)

Para realizar o processamento do conjunto de dados e definir a solução tecnológica adequada, foram utilizadas como base as tarefas de ingestão e ETL (*Extract, Transform and Load*), como mostra a área em destaque da Figura 25, representada pelo retângulo com linhas tracejadas e bordas arredondadas.

A ingestão referente ao *Big Data*, para garantir o armazenamento e o processamento dos dados, relaciona a carga dos dados não estruturados a ferramentas. A ingestão pode ser contínua ou não e, como mostra o Quadro 5, apresentar processamento em tempo real, em lote ou ambos.

O processo ETL é empregado na etapa inicial do KDD e, por meio das funções relacionadas à extração, transformação e carga dos dados, auxilia na

descoberta de conhecimento. De acordo com Guo *et al.* (2015), a ETL é utilizada na construção de um DW, sendo considerada um processo complexo que demanda muito tempo no apoio à construção da estrutura de armazenamento. Em vista disso, os autores apresentam uma abordagem denominada TEL (*Transform, Extract and Load*), cuja diferença com a abordagem ETL é que o processo de transformação é realizado antes da extração, com o uso de tabelas virtuais e não da área de teste ou do banco de dados temporário.

Na ETL, a atividade de extração envolve a seleção de fontes com formatos de arquivos distintos e, na transformação e na limpeza, em alguns momentos, é considerada a qualidade dos dados. Geralmente, essas atividades envolvem alguns procedimentos, como *i)* remoção de duplicidades; *ii)* controle da violação de integridade; *iii)* filtragem com base em expressões regulares; *iv)* classificação; *v)* agrupamento; *vi)* aplicação de funções internas. No processo de ingestão, a atividade de carga detém a tarefa de direcionar os dados para um banco de dados operacional (BANSAL e KAGEMANN, 2015).

No modelo proposto, prevê-se que, antes de entrar efetivamente nos processos de ETL ou de ingestão dos dados, os quais serão focados na parte destinada ao armazenamento e retomados na Fase III, sejam aplicadas as tarefas denominadas pré-filtragem, pré-limpeza e pré-transformação, as quais, além de apoiar na decisão da solução tecnológica, proporcionam melhorias no conjunto de dados.

A compreensão da qualidade, das características do conjunto de dados e das informações referentes aos recursos tecnológicos disponíveis na empresa possibilita o emprego do diagrama de atividades apresentado na Figura 27.

Esse diagrama utiliza como referência os 3V's" (volume, velocidade e variedade) para realizar modificações no conjunto de dados. O intuito é melhorar a qualidade e não afastar nenhuma possibilidade de aplicação desse conjunto em soluções tradicionais. Tais soluções estão sendo utilizadas há mais tempo, se comparadas às soluções *Big Data*. Portanto, suas técnicas



foram testadas, modificadas e atualizadas, o que proporciona maior confiabilidade em sua utilização.

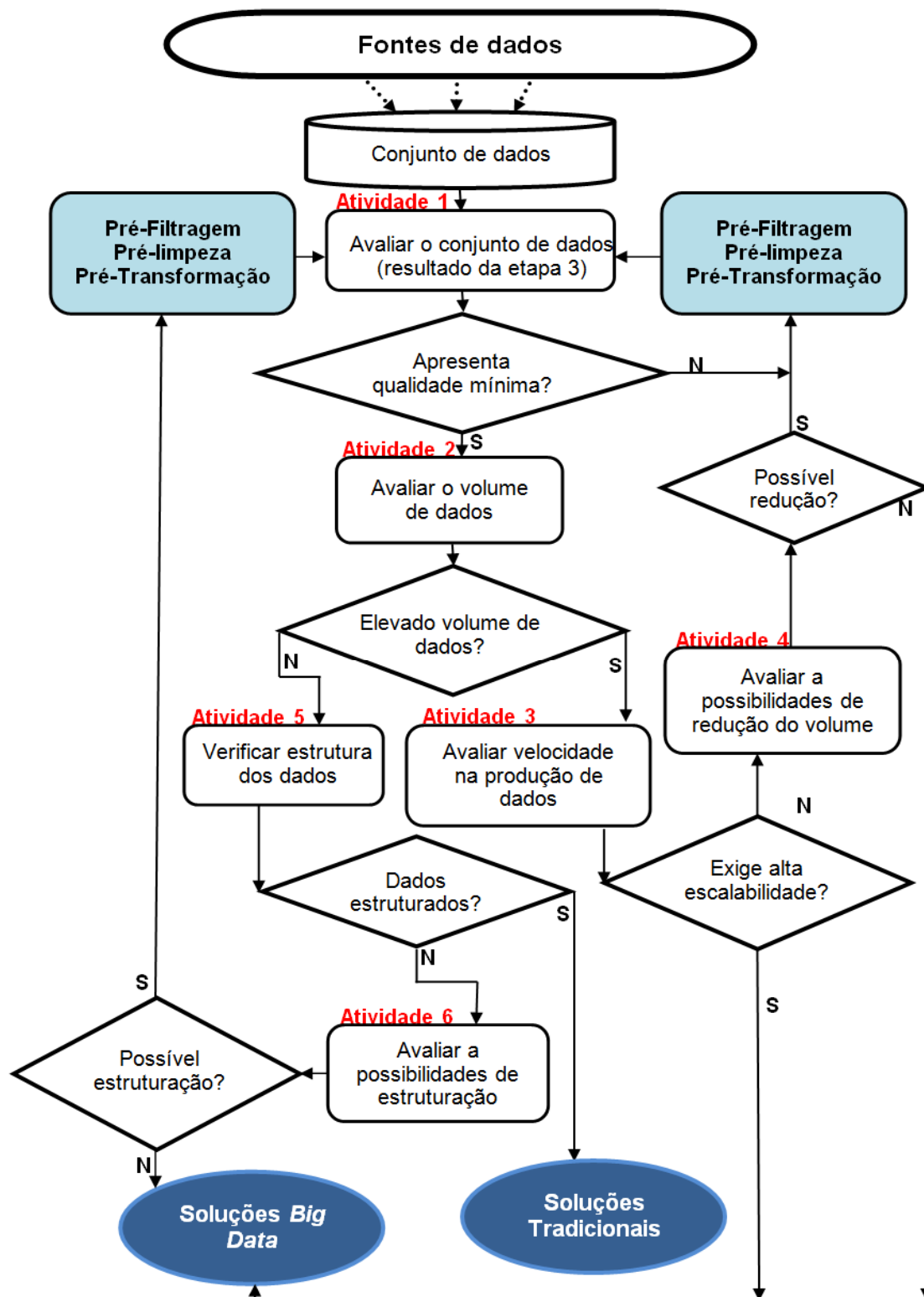


FIGURA 27 – DIAGRAMA DE ATIVIDADES  
FONTE: ELABORADA PELO AUTOR

### **Atividade 1**

A avaliação do conjunto de dados deriva da execução da Etapa 3, cujos resultados são descritos no formulário de detalhamento do conjunto de dados apresentado na Quadro 4. Uma vez que são executadas as tarefas de pré-filtragem, pré-limpeza, e pré-transformação, faz-se necessário retornar à Etapa 3 para realizar nova avaliação do conjunto de dados, principalmente das características descritas no item 3 do formulário.

### **Atividade 2**

Verificada a qualidade mínima exigida no conjunto de dados, confronta-se o volume desse conjunto com a disponibilidade de recursos da empresa, isto é, avalia-se a capacidade da empresa para suportar a demanda de armazenamento e processamento.

### **Atividade 3**

Essa atividade avalia a velocidade com que os dados são produzidos. Em alguns casos, para manter o equilíbrio e a uniformidade na manipulação do crescimento dos dados, são necessárias ferramentas de alta escalabilidade<sup>19</sup>. Para esses casos, as soluções *Big Data* oferecem ferramentas, como o *Hadoop* que, por meio de nós de *clusters*, utiliza a computação distribuída para manter a alta escalabilidade.

### **Atividade 4**

Para conjunto de dados com elevado volume e que não exige alta escalabilidade, avalia-se a possibilidade de redução do conjunto de dados. Caso seja possível, aplicam-se as tarefas de pré-filtragem, pré-limpeza, e pré-transformação. Algumas das possibilidades proporcionadas por essas tarefas são:

- converter vídeos em áudios, visando redução do volume de dados;

---

<sup>19</sup> Facilidade de reconfiguração nos casos em que seja necessário utilizar mais recursos computacionais

- converter áudios em texto, pois arquivos de textos são menores e menos dispendiosos;
- utilizar somente os arquivos de metadados;
- filtrar os dados de interesse e adicioná-los a um banco de dados estruturado, modelado previamente de acordo com os dados filtrados.

Nos casos em que o volume dos dados é elevado e não permite redução, indica-se a utilização das soluções *Big Data*.

### **Atividade 5**

Nesta atividade, em que se avalia a estrutura do conjunto, preenchida no item 2.2 do formulário ilustrado no Quadro 4, o volume do conjunto de dados não é problema para o armazenamento e o processamento.

Para casos estruturados indicam-se as soluções tradicionais. Para outras estruturas, avalia-se a possibilidade de sua modificação, conforme a próxima atividade.

### **Atividade 6**

Para situações em que o conjunto de dados é semiestruturado ou não estruturado, aplicam-se as tarefas de pré-filtragem, pré-limpeza, e pré-transformação para verificar a possibilidade de torná-lo estruturado e, assim, possibilitar condições para a utilização de soluções tradicionais.

## **4.3. FASE III – SONDAGEM E USO DE TMT**

Esta fase é composta por duas etapas: preparação dos dados e avaliação das soluções tecnológicas para armazenamento, como mostra a Figura 28.

Uma vez compreendido o conjunto de dados, conforme as fases anteriores, as etapas de preparação dos dados e de identificação das possibilidades de soluções de armazenamento serão realizadas com menos dúvidas e incertezas.

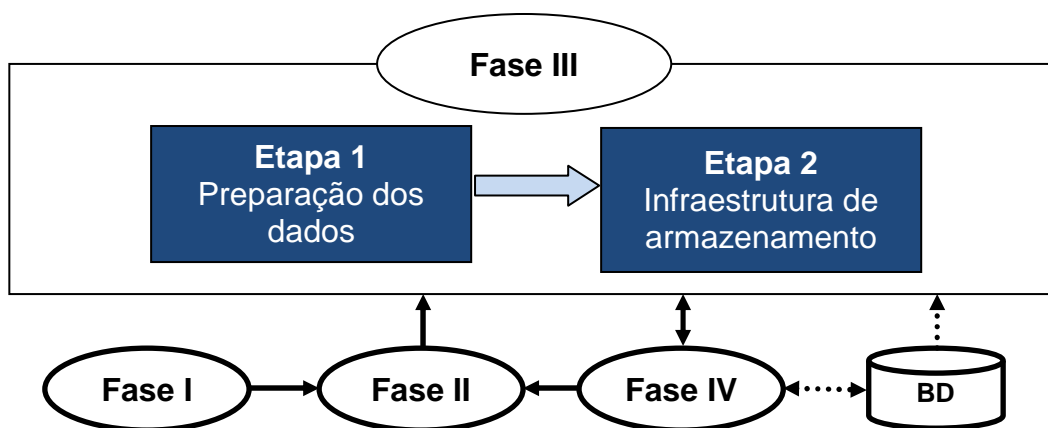


FIGURA 28 – ETAPAS DA FASE III  
 FONTE: ELABORADA PELO AUTOR

Esta fase evidencia as possíveis e recentes soluções tecnológicas que dão suporte às características dos conteúdos do conjunto de dados. Em particular, as ferramentas de código aberto que, são essenciais para pesquisas relacionadas ao *Big Data*. Até mesmo grandes corporações como a IBM<sup>20</sup> (*International Business Machines*) estão aderindo às soluções de código aberto para o *Big Data*.

#### 4.3.1. ETAPA 1 – PREPARAÇÃO DOS DADOS

A Etapa 4 da Fase II atende à necessidade de limpeza, filtragem e transformação apresentadas na ETL com o intuito de melhorar a qualidade dos dados e auxiliar na decisão de soluções tradicionais ou *Big Data*. Diferentemente disso, a Etapa 1 da Fase III, embora atenda às mesmas necessidades, tem a finalidade de preparar o conjunto de dados para o armazenamento e/ou para a análise de descoberta de conhecimento.

O pré-processamento é um dos processos operacionais do KDD e tem como objetivo a preparação dos dados que serão aplicados nos algoritmos de MD (FAYYAD *et al.*, 1996a; GOLDSCHMIDT *et al.*, 2015). As atividades de ETL possuem o mesmo objetivo, no entanto, geralmente são aplicadas no armazenamento em um DW (SALAKI *et al.*, 2016; GUO *et al.*, 2015; BANSAL e KAGEMANN, 2015).

<sup>20</sup> <https://www.ibm.com/br-pt/marketplace/biginsights>

No *Big Data*, a preparação dos dados é denominada de ingestão e, embora teoricamente execute o mesmo papel do pré-processamento do KDD, possui ferramentas específicas para transformar e utilizar os dados no *Big Data*. A ferramenta *Sqoop* desempenha essas funções quando transfere dados de banco de dados relacionais para o *Hadoop* (BENGFORT e KIM, 2016).

Enquanto na arquitetura de Zhang *et al.* (2017), ilustrada na Figura 9, a preparação dos dados (limpeza, transformação e integração) ocorre entre os estágios de armazenamento dos dados brutos e as tarefas de descoberta de conhecimento, no modelo proposto, a preparação é realizada anteriormente ao efetivo armazenamento. Como as soluções tradicionais já estão consolidadas e são menos custosas, a preparação antecipada do conjunto de dados pode direcionar seu uso por essas soluções.

No modelo proposto, as tarefas de limpeza, filtragem e transformação têm como objetivo amenizar problemas referentes a valores ausentes e ruidosos, remover dados indesejados e mudar a natureza ou a grandeza dos dados. Na transformação, pode haver necessidade de conversão e de caracterização dos dados, uma vez que os algoritmos de classificação trabalham com dados discretos.

Dentre as dificuldades encontradas na preparação dos dados, destaca-se o tratamento de dados ausentes. Como solução desse problema, o teorema HACE, de Wu *et al.* (2014) propõe um modelo em que cada item de dado é representado como uma distribuição amostral do conjunto de dados e não como dado isolado; dessa forma, os parâmetros estimados são obtidos por esse modelo. Como exemplo, os autores relatam o emprego de média e variância no algoritmo *NaiveBayes* para o desenvolvimento de um modelo classificador.

Como os conteúdos das fontes de dados podem ser diversificados, observa-se que, muitas vezes, os trabalhos de preparação dos dados requerem procedimentos específicos. Inclusive, há circunstâncias em que os envolvidos com os trabalhos de descoberta de conhecimento desenvolvem suas próprias

ferramentas para realizar a preparação dos dados. Davenport (2014) argumenta que, para trabalhar com elevado volume de dados, são necessárias pessoas com habilidade para: *i)* desenvolver algoritmos para transformar dados não estruturados em estruturados; *ii)* analisar dados; *iii)* interpretar os resultados; *iv)* orientar a execução dos conhecimentos adquiridos.

### **Soluções tradicionais**

Comercialmente, existem ferramentas proprietárias e livres, as quais auxiliam na preparação dos dados. Dentre as proprietárias, têm-se: *Microsoft SQL Server Integration Services*; *IBM Infosphere*; e *Oracle Warehouse Builder*. Para as ferramentas livres (*open source*<sup>21</sup>), destacam-se: *CloverETL*; *Talend Open Studio*; e *Pentaho Data Integration (Kettle)*. As últimas são desenvolvidas em linguagem de programação Java, cuja licença permite que sejam estudadas, modificadas e distribuídas gratuitamente independentemente da finalidade. Nesta tese, o interesse é voltado para pesquisas referentes às ferramentas livres de código aberto.

- *CloverETL* - apresenta uma interface gráfica que permite fácil acesso aos principais comandos de manipulação dos dados, além de transferir os conteúdos dos dados de diferentes fontes.
- *Talend Open Studio* - permite a manipulação dos dados por meio de interface gráfica e contém componentes para integrar, sincronizar e migrar dados de diversas fontes de dados.
- *Pentaho Data Integration* - conhecida também como *Kettle*, essa ferramenta contempla as camadas física, de negócios, e de visualização, sendo orientada para metadados. Além de realizar as tarefas da ETL em ambientes de DW para soluções tradicionais, permite: *i)* transferência de dados entre aplicações e banco de dados; *ii)* análise dos dados; *iii)* limpeza dos dados; *iv)* exportação e importação de dados entre os banco de dados.

---

<sup>21</sup> Código fonte de um *software* livre que pode ser modificado e adaptado para diferentes finalidades.

Em relação às soluções tradicionais, o modelo proposto se atém à integração dos conjuntos de dados de diferentes BD (Banco de Dados) relacionais, os quais são utilizados no armazenamento dos dados oriundos de sistemas de informação empregados na gestão do CVP, como CRM, PLM, ERP.

Essa integração torna-se evidente no *Seabase*, um BD em nuvem que, em sua arquitetura, permite compor diferentes bancos, como *MySQL*, *Oracle*, *SQL Server*, *PostgreSQL* e *Firebird*. Embora haja essa integração, o *Seabase* não é considerado um conjunto de banco de dados, pois não há armazenamento de dados persistente, apenas o armazenamento dos metadados necessários. Essa abordagem é promissora para ETL, já que reúne múltiplas fontes de dados relacionais e, por meio da virtualização dos dados, se constitui como uma abordagem unificada. Essa virtualização, envolvendo um processo de extração de elevado volume de dados, permite, por meio de camadas abstratas, que o conteúdo permaneça em seu local de origem.

Nesse contexto, baseando-se no *Seabase*, Guo *et al.* (2015) realizam a transformação dos dados antes mesmo da extração. Para tanto, empregam o conceito da virtualização para agregar ao processo de transformação uma camada virtual do modelo CCEVP (*Cloud Computing Basead Effective-Virtual-Physical*).

Na mesma linha de raciocínio, Basal e Kagemann (2015) propõem a utilização de relações semânticas, considerando-as importantes para a integração dos dados de fontes heterogêneas. A aplicação semântica dos atributos do conjunto de dados, realizada pela ETL, permite uma integração mais detalhada dos dados. Os autores recomendam a utilização e a inserção da semântica em sistema de código aberto, destacando como exemplo a ferramenta *CloverETL*.

Para fontes com elevado volume de dados, cuja migração entre diferentes repositórios se mostra inviável, uma possível solução pode ser a incorporação da proposta de Guo *et al.* (2015). Esse modelo busca realizar previamente a transformação, antes mesmo da extração.

Considerando as dificuldades das fontes cujo volume inviabiliza a migração de dados entre diferentes repositórios, uma das soluções pode ser a junção da proposta de Guo *et al.* (2015), no sentido de realizar a tarefa de transformação dos dados anteriormente à extração, com a de Basal e Kagemann (2015), que adicionam o contexto da semântica na tarefa de transformação. Essa junção, aliada à possibilidade de utilização das técnicas de virtualização, pode gerar, tanto para a solução tradicional quanto para o *Big Data*, uma ferramenta poderosa para o processo de preparação dos dados.

Nas circunstâncias em que a demanda de velocidade na geração dos dados e a variedade do conteúdo dos conjuntos de dados sejam frequentes, torna-se necessário que a capacidade das tarefas de preparação dos dados evolua proporcionalmente a essas demandas. Esses casos dificilmente são suportados pelas soluções tradicionais de armazenamento.

Em vista disso, para o processamento de elevado volume de dados, propõe-se a utilização da plataforma de computação distribuída denominada *Hadoop*, orientada para a tarefa de ingestão de dados.

### **Soluções Big Data**

A ingestão dos dados empregada nas soluções *Big Data* é realizada para obter conteúdos das fontes e exportá-los para BD pertencente à família NoSQL (*Not Only Structured Query Language*), que será discutida na próxima etapa da presente fase. Essa família de BD oferece suporte para dados estruturados, semiestruturados e não estruturados.

Os dados estruturados, adquiridos por sistemas de informação empregados na gestão do CVP, geralmente são armazenados em diferentes BD relacionais. No entanto, para ser utilizados em ambientes *Hadoop*, é necessário que, antes, esses dados sejam carregados em um dos repositórios de dados da família NoSQL. Dessa forma, a ingestão proporciona a integração e a interação dos dados aos conteúdos de outras fontes, os quais, posteriormente, podem ser acessados por meio da aplicação *MapReduce*.



A realização de script ou de atividades manuais torna-se inviável para a realização da tarefa de ingestão de dados com volume elevado, pois isso demanda um tempo considerável. Nesse cenário, a ferramenta *Apache Sqoop* (BENGFORT e KIM, 2016) permite realizar a tarefa de importação ou exportação dos dados entre o *Hadoop* e diversos serviços de armazenamento (MATTMANN *et al.*, 2014).

Para as fontes que dispõem de dados não estruturados, como *logs*<sup>22</sup> de sistemas computacionais, uma das soluções para a ingestão é a ferramenta *Apache Flume*, desenvolvida para coletar e agregar grande quantidade de dados de *logs*. No entanto, tal ferramenta não se restringe aos *logs*, permitindo também a ingestão dos conteúdos originados de diversas fontes, como mídias sociais; tráfego de redes; e-mails; dados de sensores.

Outra solução para a ingestão de dados é a ferramenta *Apache Storm*, que possui plataforma escalável e distribuída e realiza o processamento e a análise do conjunto de dados em tempo real. Possui fácil comunicação com o *Twitter* e é executada continuamente ao longo do fluxo dos dados de entrada. A literatura relaciona as aplicações dessa ferramenta com o *microblogging twitter* (KARUNARATNE *et al.*, 2017; TOSHNIWAL *et al.*, 2014). Outra de suas vantagens é o emprego da semântica no processamento de dados e nos algoritmos de aprendizagem de máquina, o que permite a coleta das fontes de dados em *call-centers*, mídias sociais e sistemas de chat.

Para a ingestão, é fundamental o cuidado com a qualidade dos dados, o que requer dos sistemas que trabalham com volume elevado de dados a capacidade de corrigir erros. Pensando nisso, Wang *et al.* (2016) propõem o protótipo *Cleanix* para realizar a limpeza dos dados, carregar os arquivos no HDFS (*Hadoop Distributed File System*), para, em seguida, aplicar as seguintes tarefas: *i)* detecção e correção de valores anormais; *ii)* preenchimento dos dados incompletos; *iii)* correção dos dados duplicados; *iv)* resolução de conflitos.

---

<sup>22</sup> Expressão empregada para reproduzir, geralmente em forma de texto, o processo de registro de eventos de sistema computacional.

#### 4.3.2. ETAPA 2 – SOLUÇÃO TECNOLÓGICA DE ARMAZENAMENTO

Os procedimentos de MD exigem unidades de computação intensivas para analisar e comparar os dados. Para realizar pequenas tarefas de mineração, um computador de pequeno porte é suficiente, no entanto, quando as aplicações envolvem volume elevado de dados, são necessários mais computadores para atender à demanda (WU *et al.*, 2014).

Davenport (2014) define a solução tecnológica do *Big Data* como um conjunto de funções que determinam o desempenho do processamento, incluindo a capacidade de manipular, integrar e administrar os dados.

Os conjuntos de dados atendidos por soluções tradicionais dispõem de algumas características, como dados estruturados, processamento e armazenamento suportado internamente pela empresa, baixa escalabilidade. Conjuntos condizentes com a solução *Big Data* precisam de mineração paralela entre o sistema de memória compartilhada, no qual a memória física global de um sistema é igualmente acessível por todos os processadores, e o sistema de memória distribuída, que consiste em múltiplos processamentos independentes com memórias dedicadas.

Embasados nesses sistemas de memória, Karunaratne *et al.* (2017) implementaram, por meio da ferramenta *Apache Storm*, versões de algoritmos distribuídos que rodam paralelamente em duas arquiteturas distribuídas e realizam o agrupamento em fluxo de dados. A arquitetura centralizada propõe que cada tarefa paralela e individual mantenha um conjunto completo de pequenos agrupamentos, enquanto, na arquitetura descentralizada, cada tarefa mantém uma imagem global dos agrupamentos, a qual é comum a todos os recursos.

Embora essas arquiteturas não apresentem diferenças significativas de desempenho, Karunaratne *et al.* (2017) argumentam que, se o objetivo for a estabilidade dos dados proporcionada pela transferência entre os períodos de sincronização, fica evidente que o melhor emprego é o da arquitetura

descentralizada. Os algoritmos aplicados nessas arquiteturas foram os do ambiente de computação nas nuvens (em inglês, *Cloud Computing*).

Ultimamente esse ambiente tem se mostrado viável, uma vez que dispõe de capacidade virtual ilimitada para processamento e armazenamento de volume elevado de dados, além de oferecer diversos recursos de *hardware* e *software* (GOLDSCHIDT *et al.*, 2015; BOTTA *et al.*, 2016).

Considerando que, nesse ambiente, a utilização de memória e de processamento ocorre na nuvem, as fontes de dados originadas na internet, como dados da IoT, podem ser manipuladas virtualmente, sem necessidade de extrair um conjunto de dados para armazenamento local.

Já o conjunto de dados estruturado, preparado na Etapa 1 da presente fase, pode ser direcionado para o processo de MD ou para armazenamento em solução tecnológica tradicional, como, DW ou BD relacionais. No caso do conjunto de dados com características *Big Data*, os dados podem ser analisados concomitantemente com o fluxo ou passar por um processo de ingestão para armazenamento em BD NoSQL, os quais se contrapõem aos relacionais por não utilizar *SQL*<sup>23</sup> (*Structured Query Language*).

Dentre os BD relacionais, que apresentam escalabilidade, flexibilidade e disponibilidade, destacam-se: *MySQL*; *Oracle*; *SQL Server*; *PostgreSQL* e *Firebird*. Enquanto que para os BD NoSQL: *Redis*; *Cassandra*; *MongoDB*; *Riak*; *CouchDB*; *OrientDB*; *SimpleDB*; *Neo4J*; *DEX*; *Titan*; e *CouchDb*.

A família de BD NoSQL possui técnicas e finalidades diferentes das dos BD relacionais. Estes possuem menos diferenças entre si, se comparados à família dos BD NoSQL (POKORNÝ, 2015; CHANDRA, 2015). Dessa forma, é importante conhecer os modelos de dados utilizados pelos BD pertencentes à família NoSQL. São eles: valor-chave; orientado a coluna; documento e grafo (TOSHNIWAL *et al.*, 2014; LOURENÇO *et al.*, 2015; POKORNÝ, 2015; CHANDRA, 2015).

---

<sup>23</sup> linguagem padrão de gerenciamento de dados que interage com os principais bancos de dados baseados no modelo relacional.

Como os BD relacionais são consolidados e amplamente discutidos na comunidade científica e tecnológica, a presente tese se concentrará em modelos de BD pertencentes à família NoSQL. Por ser objeto de estudo e para melhor entendimento desses modelos, os mesmos serão descritos e avaliados para posteriormente serem associados às fontes de dados que podem apoiar o PDP no projeto informacional.

Valor-chave é um modelo utilizado para trabalhar com BD simples e de fácil operação, como *REDIS* e *SimpleDB*, utilizando tabela *hash*. Todo acesso ao valor é realizado por meio de uma chave única e contínua, pois o BD desconhece esse valor. Esse modelo oferece agilidade no armazenamento dos dados que necessitam de rápido acesso, como sessões de usuários e *logs*. Nesta tese, o interesse se concentra no BD *REDIS*, que se destaca por possuir licença livre e código aberto, além de apoiar as fontes de dados com origem em arquivos com informações referentes a sessões dos usuários de comércio eletrônico, como *login*, produtos e interesses do consumidor. Uma característica considerada como ponto forte do *REDIS* é a possibilidade de realizar pesquisas rápidas.

O modelo orientado a coluna utiliza chaves mapeadas para valores, agrupados em colunas. Esse modelo permite o armazenamento de conjuntos com volume elevado de dados, como, por exemplo, os originados de sensores, RFID e até mesmo os disponíveis em mídias sociais. Existem diversos BD que operam esse modelo, como *Cassandra*, *Hbase* e *Riak*, em cujas características se destaca a alocação adequada no armazenamento dos dados.

O BD *Cassandra* possui licença livre e código aberto e, conforme avaliado por Lourenço *et al.* (2015), apresenta excelentes pontuações na maioria dos atributos mencionados na Tabela 3. Essa tabela apresenta a avaliação dos BD *Cassandra* e *Hbase*, que pertencem ao modelo orientado a colunas, e *MongoDb* e *CouchDb*, pertencentes ao modelo documento.

TABELA 3 – AVALIAÇÃO DAS CARACTERÍSTICAS DOS BD NoSQL

Características	Cassandra	CouchDB	Hbase	MongoDB
Disponibilidade	5	5	2	2
Consistência	5	4	3	5
Durabilidade	4	2	4	4
Manutenção	3	4	2	3
Desempenho de leitura	2	3	2	5
Desempenho de escrita	5	2	4	2
Confiabilidade	4	4	3	5
Robustez	4	3	1	3
Escalabilidade	5	2	5	2

**Legenda:** (5) Excelente (4) Bom (3) Médio (2) Ruim (1) Péssimo (x) Desconhecido

FONTE: ADAPTADO E TRADUZIDO DE LOURENÇO ET AL. (2015, P. 19)

As características dos BD avaliadas na Tabela 3 são:

- disponibilidade - falhas e tempo ininterrupto de operação dos sistemas;
- consistência – desempenho das transações;
- durabilidade – se os dados validados foram gravados no disco após uma operação bem sucedida (essa característica está correlacionada à consistência, pois, se um sistema apresenta inconsistência, sua durabilidade estará comprometida);
- manutenção – facilidade nas operações de atualização, reparação e depuração;
- desempenho – execução das operações de leitura e escrita;
- confiabilidade – probabilidade de ausência de falhas no sistema por um determinado período de tempo;
- robustez – capacidade de lidar com erros durante a execução;
- escalabilidade – capacidade de lidar com o crescimento das cargas de trabalho.

O modelo orientado a documento possui estrutura flexível e não está vinculado à disponibilidade de colunas previamente definidas para manipular documentos, que, geralmente, são apresentados nas extensões XML

(*eXtensible Markup Language*), JSON (*JavaScript Object Notation*), BSON (*Binary Structured Object Notation*) e estão armazenados em conjunto de valores no campo valor chave. Alguns BD que utilizam esse modelo são *MongoDB*, *OrientDB* e *CouchDB*.

Muito embora pertença à família NoSQL, o *MongoDB* oferece algumas das funcionalidades dos banco tradicionais e realiza operações de escrita e leitura eficientes. A aplicação do *MongoDB* na descoberta de conhecimento para auxiliar o projeto informacional é adequada para situações que necessitam de intensivas leituras, que são relevantes para o armazenamento e a recuperação dos dados oriundos dos sistemas de informação empregados na gestão do CVP.

Além de possuir licença livre e código aberto, o *MongoDB* apresenta, como características positivas, linguagens de consulta e funções de agregação consideráveis e não tão específicas como as disponíveis em outros BD NoSQL. Embora sejam utilizadas para diferentes finalidades, tais características se destacam no armazenamento temporário.

No modelo orientado a grafos, diferentemente dos outros BD da família NoSQL, os dados não são armazenados em linhas e colunas, mas sim em estruturas de grafos, estabelecendo relações entre os objetos de um determinado conjunto. As características positivas desse modelo são os relacionamentos e os algoritmos que buscam os melhores percursos dentre as possibilidades disponíveis nos grafos. Alguns dos BD que utilizam esse modelo são: *Neo4J*; *DEX*; *Titan* (*BerkeleyDB* e *CassandraDB*); e *OrientDB*. Nesse grupo, o BD *Neo4J* destaca-se na comunidade científica.

As características relacionadas aos modelos empregados pelos BD da família NoSQL são avaliadas por Chandra (2015), como mostra a Tabela 4.

TABELA 4 – CARACTERÍSTICAS DO MODELO DE DADOS NoSQL

Modelo	Desempenho	Escalabilidade	Flexibilidade	Complexidade
Valor-chave	Alta	Alta	Alta	Baixa
Orientado a Coluna	Alta	Alta	Moderada	Baixa
Documento	Alta	Variável (alto)	Alta	Baixa
Grafos	Variável	Variável	Alta	Alta

FONTE: TRADUZIDO E ADAPTADO DE CHANDRA, (2015, P16)

Referindo-se aos BD do modelo orientado a coluna, Chandra (2015) e Lourenço *et al.* (2015) apresentaram avaliações convergentes. Entretanto, para o modelo orientado a documento, não houve consenso na avaliação da característica escalabilidade. Enquanto Chandra (2015) mostra, na Tabela 4, que os BD desse modelo possuem escalabilidade variável propensa a alta, Lourenço *et al.* (2015) avaliam que os BD *CouchDb* e *MongoDb* apresentam baixa escalabilidade, como mostra a Tabela 3. Essa divergência na avaliação dos autores pode ter como causa a particularidade na escolha dos parâmetros de configuração da característica.

Outra característica dos modelos empregados nos BD da família NoSQL e destacada por Chandra (2015) foi a complexidade que, como mostra a Figura 29, foi avaliada em função da quantidade de dados.

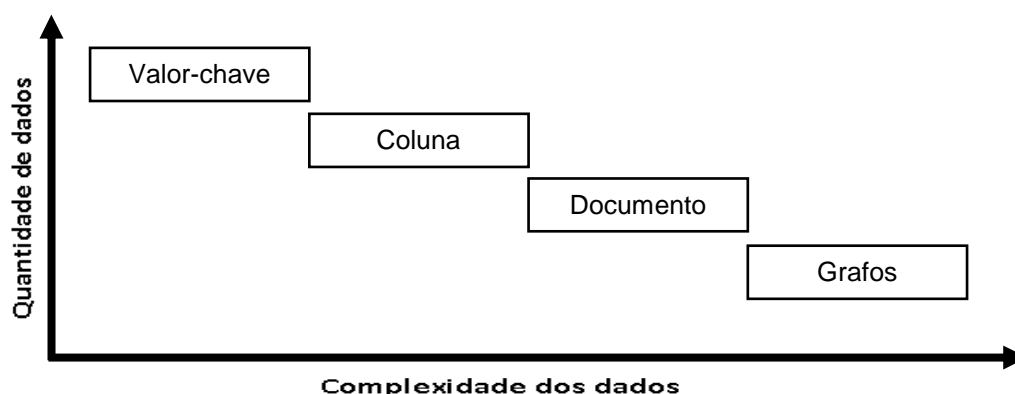


FIGURA 29 – ANÁLISE DA QUANTIDADE E COMPLEXIDADE DOS DADOS RELACIONADOS COM OS MODELOS DE BD NoSQL  
FONTE: ADAPTADO DE CHANDRA, (2015, P. 18)

Observa-se, na Figura 29, que o modelo orientado a valor-chave é indicado para volume elevado de dados com baixa complexidade; já o modelo orientado a grafos é adequado para trabalhar com dados mais complexos, porém com volume reduzido. O *MongoDB*, por ser orientado a documento, apresenta equilíbrio entre a complexidade e a quantidade de dados.

As avaliações e os estudos realizados nesta etapa quanto aos modelos empregados nos BD NoSQL tornam possível fazer recomendações de armazenamento para as fontes de dados que foram previamente identificadas na Etapa 1 da Fase II do modelo proposto, tendo em vista a realização de análises para a descoberta de conhecimento para o projeto informacional.

Para as fontes de dados originadas de sistemas de informação, como CAD, CAE, CAPP, CAM, SCM, ERP, CRM, recomenda-se a utilização de um BD pertencente ao modelo documento, como o *MongoDB*, que, embora apresente divergência em relação à característica escalabilidade, demonstra eficiência nas características de desempenho de leitura, flexibilidade, consistência e confiabilidade. Em relação às fontes de dados, geralmente os sistemas de informações utilizam BD relacionais e, apesar de pertencer à família que se contrapõe aos BD relacionais, o *MongoDB* agrega suas funcionalidades.

As fontes de dados oriundas das redes de computadores (*intranet* ou *extranet*) dispõem de arquivos em diversos formatos, como planilhas, gráficos, fotos, textos, dados de sistemas legados. Além de apresentar características positivas para as fontes originadas dos sistemas de informação, o *MongoDB* se mostra útil por sua atuação adequada em relação aos dados complexos, tornando-se necessário para as fontes que contêm dados com formatos diferenciados. Nos casos em que a complexidades dos dados não é tida como requisito principal, outra opção a se considerar pode ser a dos BD pertencentes ao modelo orientado a valor-chave. O BD *Redis* suporta volume elevado de dados, apresenta avaliações altas para as características desempenho, escalabilidade e flexibilidade, além de permitir avaliações dos dados em tempo real.



As fontes de dados provenientes da IoT (*RFID*, sensores e outros) precisam de BD com alta escalabilidade. O BD pertencente ao modelo orientado a coluna atende a esse requisito. Como a escalabilidade é aliada às avaliações positivas das características disponibilidade, consistência, desempenho na escrita e baixa complexidade, sugere-se a utilização do BD *Cassandra*.

As mídias sociais são fontes que englobam dados oportunos relacionados aos consumidores e produtos. Embora não seja uma regra, a utilização do *MongoDB* pode ser útil para identificar possíveis formas de associar os conteúdos dos dados existentes nessas fontes, aos conteúdos dos BD dos sistemas de informação utilizados pelo CRM. Isso porque o *MongoDB* oferece suporte e funcionalidades tanto para dados estruturados quanto para dados não estruturados. Em razão da alta consistência e do desempenho do *MongoDB*, outra possibilidade é sua utilização como um BD transitório.

Como o BD *Cassandra*, orientado a coluna, tem apresentado avaliações favoráveis nas características imprescindíveis para o armazenamento de dados extraídos em fluxo, alternativamente pode ser útil para o trabalho com fontes oriundas de mídias sociais.

Os BD orientados a grafos permitem armazenamentos que favorecem a identificação das relações entre usuários de mídias sociais e a descoberta dos formadores de opinião. Um dos BD que se destaca é o Neo4j, que, na avaliação de Jouili e Vansteenbergh (2013) com os outros BD, como DEX, *Titan* e *OrientDB*, superou todos os outros, independentemente da carga ou dos parâmetros utilizados na avaliação.

Em razão das diferentes particularidades da análise requerida e das características próprias de cada conjunto de dados, a aplicação dos conteúdos das fontes não está limitada aos BD sugeridos.

#### 4.4. FASE IV – DESCOBERTA DE CONHECIMENTO

Na Fase IV, da descoberta de conhecimento, são apresentadas as possibilidades para a realização da análise do conjunto de dados, visando estabelecer a técnica adequada para visualizar tanto os resultados provenientes dessas análises quanto os dados brutos.

Esta fase está dividida em duas etapas, como ilustra a Figura 30. A Etapa 1 refere-se às soluções de análise, ao passo que, na Etapa 2, descrevem-se algumas técnicas de visualização empregadas para auxiliar a interpretação dos resultados.

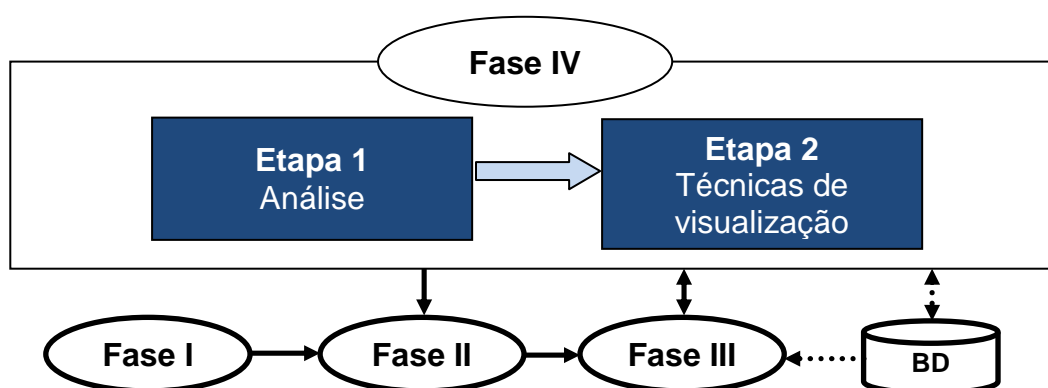


FIGURA 30 – FASE IV - DESCOBERTA DE CONHECIMENTO.

FONTE: ELABORADA PELO AUTOR

A execução dessa fase tem a finalidade de produzir resultados que possam produzir conhecimentos novos e úteis. Isso implica uma seleção adequada da fonte de dados e a realização do processo de busca de conhecimento alinhada aos objetivos definidos na Fase I do modelo proposto (diagnóstico). Espera-se, assim, alcançar resultados que contribuam para o projeto informacional. Evidentemente, existe a possibilidade de o resultado obtido não corresponder ao objetivo previamente definido, mas estar em consonância com as diretrizes apresentadas na definição do KDD, quais sejam, “...*identificação de padrões compreensíveis, válidos, novos e potencialmente úteis* ...” (FAYYAD *et al.* 1996a, p. 30). Dessa forma, poderia atender a outros objetivos que não foram previamente estabelecidos.

#### 4.4.1. ETAPA 1 – ANÁLISE

Nesta etapa, realiza-se a análise do conjunto de dados com a aplicação de métodos e técnicas, tendo em vista a busca de conhecimento.

Anteriormente à análise, convém atentar para a observação e para a proposição de Wu *et al.* (2014) a respeito das análises realizadas com a utilização de dados parciais. Considerando que a visão fragmentada dos dados pode induzir a tomadas de decisão tendenciosas e exemplificando com a parábola do elefante e dos cegos<sup>24</sup>, os autores sugerem a aplicação de técnicas para fundir os dados e, posteriormente, por meio de cálculos, determinar o grau de relevância estabelecido pela correlação das diferentes fontes de dados, de forma a assegurar que todas sejam analisadas em conjunto.

Outra observação a ser considerada refere-se à análise que utiliza dados históricos com o objetivo de gerar previsões. Neste caso, Tyagi *et al.* (2015) propõem o método de quantificação do grau de continuidade, no qual são utilizadas as ocorrências do passado para identificar comportamentos semelhantes e, assim, realizar previsões.

Quando o assunto é a análise para a descoberta de conhecimento aplicada a dados estruturados, imediatamente imagina-se MD. Esse pensamento decorre do fato de a MD pertencer a uma das etapas do processo de descoberta de conhecimento. A MD emprega, por meio dos relacionamentos e padrões existentes entre os dados, tarefas e técnicas para a obtenção de conhecimento. As tarefas empregadas na MD, descritas na seção 3.2.1, são: classificação, associação, regressão, agrupamento e predição. Para cada uma dessas tarefas existem técnicas, representadas por algoritmos já consolidados e eficientes que buscam atender às demandas relacionadas a dados estruturados.

---

<sup>24</sup> História dos homens cegos tentando descrever um elefante. Cada um tenta descrever a parte que conseguiu captar pelo tato, cuja limitação impõe a cada homem uma determinada região do corpo do elefante. Assim, surgem diferentes perspectivas, segundo as quais o elefante seria uma corda, um cavalo, uma árvore ou uma parede.

Segundo Wu *et al.* (2008), os algoritmos mais utilizados pela comunidade científica em soluções tradicionais de MD são: C4.5, *k-means*, *vector machines*, *Apriori*, EM, *PageRank*, *AdaBoost*, kNN, *Naive Bayes* e *CART*. Importante ressaltar que, dentre os algoritmos empregados nas tarefas de agrupamento de dados, destaca-se o *Kmeans*, que é utilizado há mais de meio século (CUI *et al.*, 2014).

Conceitualmente, o *Big Data* apresenta aspectos que dificultam a utilização dos algoritmos construídos em soluções tradicionais de MD. Entretanto, esse cenário vem mudando nos últimos anos, uma vez que a literatura tem apresentado recursos e técnicas para aperfeiçoá-los, como a computação paralela e a distribuída.

Esse aperfeiçoamento pode ser visto em Cui *et al.* (2014), que apresentam resposta para o processamento de dados em larga escala. Utilizando o algoritmo de agrupamento *Kmeans*, os autores propuseram um modelo de roteiro para processamento e empregaram o *MapReduce*, cujo intuito é eliminar dependências de interações para obter processamento com alta performance. Conclusivamente, o emprego desse método gerou eficiência e robustez.

Como mencionado, existem possibilidades de aprimoramento desses algoritmos tradicionais e, conseqüentemente, muitas outras pesquisas serão realizadas para atender à demanda por soluções *Big Data*, incluindo dados originados de fontes múltiplas.

A característica heterogeneidade dos dados de fontes múltiplas apresenta diferenças essenciais na descoberta de conhecimento em fonte única e em fontes múltiplas (WU *et al.*, 2014). As soluções *Big Data* dispõem de recursos para trabalhar essa heterogeneidade dos dados, além de possibilitar a integração com as soluções tradicionais já testadas e consolidadas.

Para fontes que necessitam de análise nos dados brutos e volumosos, a plataforma *Hadoop* oferece ferramentas para que tais dados sejam

armazenados no HDFS sem a necessidade de realizar a ETL. Por meio dessa plataforma, pode-se fazer uma análise exploratória e criar protótipos para a MD, além do que ela amplia o cenário para atender novos tipos de dados e realizar novas análises (BENGFORT e KIM, 2016).

O *framework* para armazenamento de dados, denominado *Hive*, oferece um código intuitivo e um dialeto próximo ao SQL para a análise com o uso do *Hadoop*, o que simplifica o trabalho dos participantes no projeto de descoberta de conhecimento. Esse dialeto, denominado HQL<sup>25</sup> (*Hive Query Language*), permite criar tabelas e realizar a carga dos dados e, dessa forma, realizar análises como se fosse em banco de dados tradicionais. Por meio desse mecanismo, torna-se possível extrair os resultados e gerar novos conjuntos de dados para então submetê-los ao processo de MD tradicional.

Dentre as diversas opções de linguagens de programação e análise para dados estruturados e não estruturados, têm-se *Java*, *Python*, *C++* e *R*. Pesquisa realizada pela IEEE<sup>26</sup> *Spectrum's* por Cass (2016) apresenta um ranking das linguagens de programação mais populares do ano de 2016. Nessa lista, aparece a linguagem *R*, que está na quinta posição do ranking e vem se expandindo nos últimos anos, em decorrência do crescimento da demanda por soluções *Big Data*. Seguindo essa linha de pesquisa, Fox e Leanage (2016) realizaram um levantamento dos trabalhos publicados no JSS (*Journal of Statistical Software*) entre os anos de 1996 e 2016 e destacaram que, dentre as linguagens, a *R* é a que proporcionou maior contribuição científica.

Atualmente, a linguagem *R* dispõe em seu repositório de 10.816 pacotes, que estão disponíveis em sua rede de distribuição CRAN (*Comprehensive R Archive Network*). Esses pacotes atuam com diversas finalidades, como *i*) manipulação de conjunto de dados com diferentes estruturas, *ii*) tarefas de MD, *iii*) técnicas de visualização e *iii*) conexão e interface com as soluções *Big Data*.

---

<sup>25</sup> Linguagem de consulta semelhante ao SQL.

<sup>26</sup> Institute of Electrical and Electronics Engineers.

A escolha das soluções tecnológicas (tradicional ou *Big Data*) para a análise dos dados está diretamente relacionada às fontes de dados, uma vez que são os próprios dados que estabelecem as regras para a seleção de métodos e técnicas.

Em relação às fontes oriundas de mídias sociais, as análises realizadas podem gerar conhecimentos referentes à opinião do consumidor. Para Rozenfeld *et al.* (2006), essa opinião definirá os requisitos do produto; da mesma forma, para Back *et al.* (2008), o importante é entender o desejo dos consumidores e participantes do PDP, visto que a qualidade do produto está diretamente relacionada à satisfação, à necessidade e ao desejo do consumidor. Além de apoiar o PDP, Fan e Gordon (2014) argumentam que as análises das fontes originadas em mídias sociais geram conhecimentos referentes aos fornecedores e concorrentes, além de ajudar a compreender o ambiente do negócio.

Dentre as técnicas empregadas na análise de fontes de dados produzidas por mídias sociais, estão o cálculo de frequência dos termos e a análise de texto proporcionada pela combinação de métodos estatísticos e de processamento de linguagem natural (PLN). Sapountzi e Psannis (2016) definem as principais técnicas, que são: *i)* análise das conexões sociais, *ii)* análise de sentimento, *iii)* análise de tendência.

Algumas ferramentas destinadas à análise de dados possuem códigos abertos, assim como os pacotes da linguagem R, os quais apresentam como vantagem a possibilidade de se realizarem alterações em seu código, contribuindo, dessa forma, para os casos que necessitam de análise personalizada.

O intuito nesta fase não foi descrever detalhes técnicos das possibilidades de análise de dados, mas apenas discutir ferramentas, técnicas métodos promissores. Isso porque o modelo proposto não está condicionado a nenhuma solução tecnologia específica.

#### 4.4.2. ETAPA 2 – TÉCNICAS DE VISUALIZAÇÃO

As técnicas de visualização de informações são utilizadas para simplificar a tarefa de interpretação dos resultados obtidos por meio dos algoritmos empregados no processo de descoberta de conhecimento. Essas técnicas baseiam-se na capacidade humana de percepção para analisar eventos complexos, permitindo reconhecer o que é útil e ao mesmo tempo desconsiderar o que não é de interesse (RABELO *et al.*, 2008). Além de representar os resultados extraídos pelos algoritmos, essas técnicas podem ser empregadas na visualização dos dados brutos.

As discussões sobre as técnicas de visualização de dados implicam diversas abordagens, envolvendo conceitos, modelos, processos e técnicas propriamente ditas. Neste trabalho, não são abordados conceitos, modelos e processos; o foco é refletir sobre algumas técnicas e em como sua utilização poderá auxiliar na seleção da técnica apropriada para visualizar os conhecimentos extraídos, já que ainda existem dificuldades na escolha das técnicas de visualização para a descoberta de conhecimento (RABELO e CAMPOS, 2014).

Assim como a qualidade dos dados influencia o processo de descoberta de conhecimento, a seleção inadequada da técnica de visualização também o faz, podendo gerar resultados incorretos ou inconclusivos. Para evitar esses resultados, faz-se necessário conhecer as características das técnicas de visualização.

Na época em que não se cogitava a possibilidade de extração de conhecimento no *Big Data*, Lengler e Eppler (2007) desenvolveram um esquema em formato de “tabela periódica”, como apresentado no Anexo A. Nesse esquema, os “elementos químicos” foram substituídos por técnicas de visualização, que são classificadas por:

- tipo de visualização – dados, informação, conceitual, metafórica, estratégica e composta;

- dimensões – complexa, área de aplicação, ponto de visão e tipo de informação representada.

Mais detalhes referentes à “tabela periódica” de Lengler e Eppler (2007) estão descritos na seção 3.5.

Dentre as diversas técnicas de visualização, Rabelo *et al.* (2008) avaliaram, com base em suas características, as técnicas pertencentes aos grupos de projeção geométrica e iconográfica, conforme apresentada na Tabela 5. Essa avaliação fornece subsídios para apoiar a seleção da técnica adequada para a extração de conhecimento.

**TABELA 5 – ANÁLISE DAS CARACTERÍSTICAS DAS TÉCNICAS DE VISUALIZAÇÃO**

Características	Técnica de projeção geométrica			Técnicas iconográficas		
	Matriz de dispersão	Dispersão de dados 3D	Coordenadas Paralelas	Star Glyphs	Figuras de arestas	Faces de Chernoff
Escalabilidade	5	2	5	1	5	1
Dimensionalidade	4	3	4	5	3	3
TIPO DE DADOS	Qualitativo nominal	3	3	1	1	1
	Qualitativo ordinal	4	4	5	5	5
	Quantitativo discreto	5	5	5	5	5
	Quantitativo contínuo	5	5	5	5	5
Interação	5	5	5	5	5	5
Interpretabilidade	5	4	4	3	3	4
Relacionamento	5	3	3	1	1	1
Correlação	5	3	3	x	x	x

*FONTE: TRADUZIDO DE RABELO ET AL. (2008, P 1232)*

Para cada uma das técnicas de visualização descritas na Tabela 5, foram apresentadas as características e suas respectivas avaliações de aplicabilidade, as quais variam de 1 a 5, sendo que 1 corresponde a não satisfaz e 5 a satisfaz completamente.



A característica escalabilidade refere-se à quantidade de registros que podem ser apresentados simultaneamente em técnicas de visualização. Dependendo da técnica, esse elevado número de dados pode gerar resultados em grau de desordem considerável. Algumas técnicas de projeção geométrica, como matriz de dispersão e coordenadas paralelas, demonstraram eficácia na identificação de padrões com elevado número de dados já tratados por algoritmos de MD. As técnicas iconográficas, como *star glyphs* e faces de *chernoff*, são limitadas, o que não ocorre com a técnica da figura de arestas, por meio da qual se realiza a identificação de padrões em quantidade elevada de dados, representados por figuras gráficas posicionadas na forma de texturas.

A característica dimensionalidade corresponde à capacidade que a técnica possui para representar a quantidade de atributos do conjunto de dados, sem causar poluição visual ou “borrões” indecifráveis. As coordenadas paralelas são úteis para conjunto de dados com vários atributos, entretanto, seu limite de representação está condicionado ao tamanho da tela. Outra visualização capaz de representar a dimensionalidade são as matrizes de dispersão em 2D ou 3D, as quais permitem utilizar figuras gráficas com formas e cores para representar os atributos.

As técnicas iconográficas de visualização têm seus atributos representados por meio de identificadores visuais perceptíveis, tais como: cor, forma e textura. Como exemplo dessa técnica, têm-se as faces de *Chernoff* que relaciona os atributos do conjunto de dados às características existentes em uma face humana. Embora essa técnica seja útil para a exibição de dados multidimensionais, ela não transmite os valores em sua integralidade. A técnica *Star Glyphs*, se comparada às faces de *Chernof*, permite representar os dados com maior número de atributos. Do grupo da técnica de visualização iconográfica faz parte a figura de aresta que, apesar de apresentar alta escalabilidade, possui restrição em termos de dimensionalidade.

Todas as técnicas de visualização avaliadas na Tabela 5 são satisfeitas completamente pelas características dos dados quantitativos (discretos e

contínuos). Enquanto que para as técnicas iconográficas, as avaliações das características referentes aos dados qualitativos nominais não permitem boa representatividade.

A característica interação traduz o diálogo entre usuário e tecnologia e permite realizar alterações nas visualizações para demonstrar conhecimentos que podem estar ocultos. As abordagens de filtragem interativa e zoom interativo se aplicam a essas visualizações. Para essa característica, todas as técnicas apresentadas no Tabela 5 obtiveram pontuação máxima.

A característica interpretabilidade refere-se à facilidade para extrair conhecimento por meio das técnicas de visualização. A matriz de dispersão satisfaz completamente essa característica. A técnica de coordenadas paralelas satisfaz parcialmente a interpretabilidade, pois evidencia a facilidade na identificação das correlações entre atributos, a qual projeta em padrões bidimensionais os relacionamentos entre os atributos do conjunto de dados. A técnica de faces de *Chernoff* apresenta razoável interpretabilidade, pois possibilita uma visualização rápida e compacta de diversas figuras gráficas simultaneamente.

Para representar o conjunto de dados não estruturados, destaca-se a técnica de visualização nuvem de palavra (em inglês, *Word Cloud* ou WC). Essa visualização identifica a frequência com que os termos ocorrem em um texto. Carr *et al.* (2015) utilizam essa visualização, ilustrada na Figura 31, para obter conhecimentos referentes aos consumidores de café.



FIGURA 31 – WORD CLOUD  
 FONTE: CAR ET AL. (2015, P. 360)

Outra técnica que permite a visualização de dados não estruturados é a árvore de palavras (em inglês, *Word Tree* ou WT), ilustrada na Figura 32. Essa técnica mostra os termos e as frases de um conjunto de dados em formato de texto. A busca por um termo específico retorna diferentes frases relacionadas a esse termo. Dessa forma, é possível identificar, por meio de ramificações, a ligação entre um termo pré-definido a outros termos fornecidos no texto.



**FIGURA 32 – WORD TREE**  
**FONTE: ELABORADA PELO AUTOR**

A técnica da *word tree* é a versão gráfica do método “palavra chave em contexto”. Wattenberg e Viégas (2008) apresentam as características da técnica da *word tree*: *i)* detecta repetições; *ii)* torna óbvia a estrutura natural da árvore de contexto ; *iii)* proporciona facilidade na exploração do contexto. Essas características e o desempenho proporcionado por essa técnica podem ser úteis para o processo de descoberta de conhecimento, em particular, para as fontes de dados oriundas de mídias sociais. Como desvantagem dessa técnica, destaca-se a necessidade de executar atividades exaustivas de interação e iteração.

Além das técnicas já mencionadas para trabalhar com dados não estruturados, Henderson e Segal (2013) propõem um *framework*, incluindo outras técnicas, mostradas na Figura 33, as quais ampliam as possibilidades de visualização por meio de técnicas utilizadas para dados textuais.



**FIGURA 33 – TÉCNICAS DE VISUALIZAÇÃO PARA DADOS TEXTUAIS**  
**FONTE: TRADUZIDO DE HENDERSON E SEGAL, (2013, P. 56)**

A Figura 33 apresenta essas técnicas dispostas em um gráfico bidimensional, em que o eixo vertical das ordenadas representa o nível de exibição, isto é, as possibilidades de escolha em relação ao objetivo desejado na exploração de palavras, frases ou tema/narrativa. O eixo das abscissas representa a possibilidade de escolha da técnica pretendida, de acordo com o grau de complexidade. Esse grau de complexidade refere-se ao nível de habilidade necessária para desenvolver a visualização em cada técnica.

As avaliações das técnicas de visualização apresentadas na literatura servem para orientar a escolha da técnica adequada, conforme as características existentes em cada conjunto de dados. As avaliações realizadas por Rabelo *et al.* (2008) e a “tabela periódica”, proposta por Lengler e Eppler (2007), apoiam a seleção da técnica de visualização utilizada em dados originados dos sistemas de informação (CAD, CAE, CAPP, CAM, SCM, ERP e CRM) para

auxiliar a gestão do CVP e, também para auxiliar a escolha da técnica para visualizar as fontes em diferentes formatos, como planilhas, textos e dados de sistemas legado.

Existem eventos em que os dados originados dos dispositivos utilizados na IoT (RFID, sensores, entre outros) necessitam de técnicas que suportam visualização em tempo real de uma quantidade elevada de dados. Para esses casos, as características de escalabilidade e de dimensionalidade são fundamentais e, nesse quesito, as técnicas de matriz de dispersão, coordenadas paralelas e figuras de arestas são adequadas.

Os dados provenientes de mídias sociais necessitam de técnicas cujas características representem os dados qualitativos ordinais e/ou nominais: as técnicas pertencentes ao grupo geométrico satisfazem parcialmente por essas características. Para os dados ordinais, as técnicas iconográficas apresentam avaliações satisfatórias, entretanto, as faces de *Chernof* e *Star Glyphs* possuem avaliações que não satisfazem a característica escalabilidade, o que pode dificultar sua utilização.

Entre as técnicas que contemplam dados textuais, a visualização *word cloud* depende da frequência com que o termo aparece no conjunto de dados e não apresenta contexto, isto é, não demonstra conotações positivas ou negativas e nem conexão com termos correlatos (WATTENBERG E VIÉGAS 2008; HENDERSON E SEGAL, 2013). A falta de contexto apresentada na *word cloud* pode ser suprida com a utilização paralela da técnica *word tree*.

Ao explorar a palavra-chave nas frases da visualização da *word tree*, podem-se identificar consistências e padrões e até mesmo observar como as palavras são utilizadas em contextos diferentes (HENDERSON E SEGAL, 2013). Essa argumentação sugere a aplicação dessa técnica nas atividades de pré-limpeza e pré-filtragem para que sejam obtidas visões diferentes de um mesmo conjunto de dados, conforme o que está exposto na Fase II da Etapa 4 do modelo proposto.

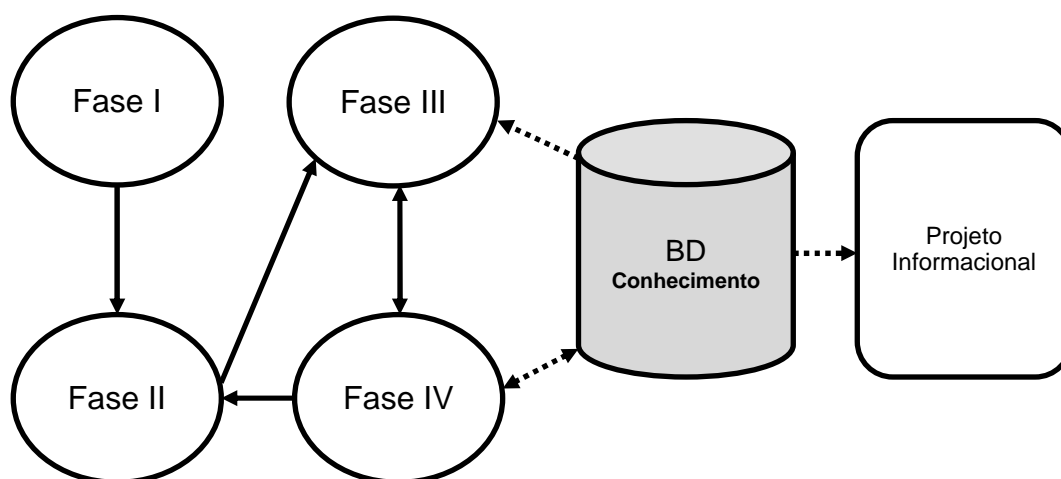
A *word cloud* pode ser utilizada para acompanhar a quantidade de vezes que o produto ou a empresa são citados em uma mídia social. O acompanhamento diário, semanal e mensal da frequência em que determinado termo aparece nas mídias sociais pode evidenciar alguma anormalidade não detectada pela empresa e, assim, motivar investigações para descobrir a causa geradora dessa anormalidade. A técnica também pode ser empregada como termômetro para mostrar o interesse dos consumidores por determinado produto.

É preciso esclarecer aqui que não foram esgotadas todas as técnicas de visualização e que o intuito nesta fase foi discutir algumas delas e mostrar sua importância para a descoberta de conhecimento. O modelo proposto não está condicionado a nenhuma técnica de visualização, mas demonstra sua importância no processo.

#### **4.5. ARMAZENAMENTO DO CONHECIMENTO**

Depois de extraídos, os conhecimentos devem ser armazenados e posteriormente avaliados e interpretados, em conjunto com os colaboradores do domínio de aplicação. Conforme o objetivo planejado na Fase I, constatando-se que são novos e úteis, tais conhecimentos são adicionados ao projeto informacional para auxiliar o PDP.

A partir desses conhecimentos armazenados, novas fontes de dados podem surgir. As setas tracejadas na Figura 34 ilustram o direcionamento do fluxo do conhecimento extraído. A seta que interliga o BD à Fase III indica que o conhecimento já armazenado pode ser agregado a outros conjuntos de dados e, juntamente com estes, ser direcionado para a Fase IV para ser analisado. A seta bidirecional que interliga o BD à Fase IV mostra que os conhecimentos armazenados podem retornar a essa fase para auxiliar complementarmente tanto a análise da descoberta de conhecimento quanto à comparação com conhecimentos prévios.



*FIGURA 34 – FLUXOS DO CONHECIMENTO*  
*FONTE: ELABORADA PELO AUTOR*

#### **4.6. ARQUITETURA RESULTANTE DO MODELO PROPOSTO**

O desenvolvimento do modelo proposto resultou em uma arquitetura de processos que contribui para a geração dos conhecimentos a ser empregados no projeto informacional, como mostra a Figura 35. Tal arquitetura, que oferece suporte para dados estruturados e não estruturados (tradicionais e *Big Data*), é motivada por pesquisas recentes na literatura (WU *et al.*, 2014; ZHUANG *et al.*, 2016; ZHANG *et al.*, 2017). Como ilustra a Figura 35, suas camadas estão diretamente relacionadas às fases II, III e IV desenvolvidas no modelo proposto.

As camadas incluem as seguintes abordagens:

- fontes de dados;
- manipulação de dados e seleção de infraestrutura tecnológica;
- tarefas para auxiliar a descoberta de conhecimento;
- algoritmos para MD;
- soluções tecnológicas;
- técnicas de visualização e MD.

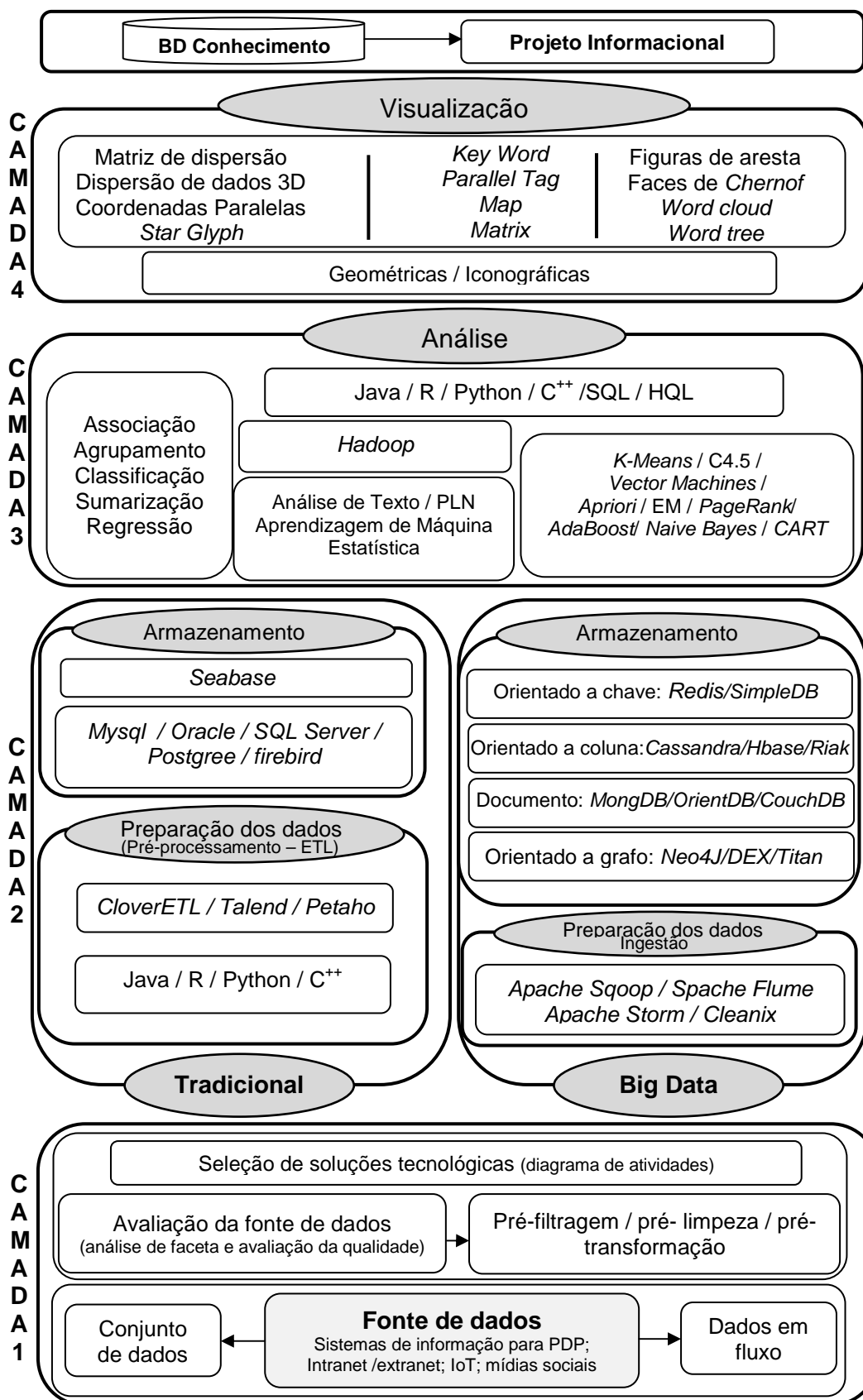


FIGURA 35 – ARQUITETURA PROPOSTA PARA DESCOBERTA DE CONHECIMENTO NO PROJETO INFORMACIONAL  
 FONTE: ELABORADA PELO AUTOR



#### 4.7. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Na presente pesquisa, propôs-se o desenvolvimento de um modelo conceitual que, com o auxílio de métodos utilizados no processo de descoberta de conhecimento tradicional e *Big Data*, apoiasse a execução do projeto informacional no PDP. Nesse modelo, foram criadas fases e etapas, nas quais se realizam atividades que contribuem para organizar esforços em uma área tão vasta quanto à descoberta de conhecimento em dados.

Para atingir o objetivo da pesquisa, primeiramente foram discutidas as metodologias de descoberta de conhecimento tradicional associadas às novas demandas do *Big Data*. Depois, foram analisados os modelos de referência do PDP, com base nos quais foram identificadas as atividades do projeto informacional. Tais atividades, realizadas para atualizar o escopo do produto e identificar seus requisitos, foram auxiliadas pelo modelo proposto.

No desenvolvimento do modelo proposto, antes de efetivamente serem trabalhados os dados, foi elaborada a Fase I que preconiza a compreensão da empresa e do seu domínio de aplicação. Ademais, essa fase tem o importante papel de traçar o objetivo que dará início às atividades descritas no modelo.

Sabe-se que os dados são fator determinante e fundamental nos processos de negócio de uma organização, da mesma forma que sua qualidade. Em face disso, o modelo proposto atende as organizações manufatureiras, cujo interesse é, por meio dos recursos de dados, reforçar as decisões que serão utilizadas no PDP. Com essa finalidade, na Fase II do modelo proposto, voltada aos dados e suas respectivas fontes, empregou-se a análise de faceta e se adotaram procedimentos para verificar, trabalhar, acompanhar e garantir essa qualidade.

Com base nos resultados extraídos, na fase final do modelo proposto, utilizaram-se técnicas de visualização. Estas são um instrumento eficiente para a exploração de conhecimentos e, por isso, no modelo, foi proposto que, antes

de aplicar tais técnicas de visualização, fossem compreendidas suas características.

## 5. APLICAÇÃO DO MODELO PROPOSTO

Para ilustrar sua aplicabilidade, o modelo proposto no Capítulo 5 foi empregado em uma empresa de confecção industrial, denominada “AM - Amarelo Manga<sup>27</sup>”, em atividade há aproximadamente oito anos. Com matriz situada na cidade de Maringá, no estado do Paraná, essa indústria dispõe fisicamente de oito lojas, um escritório e centro de distribuição. Virtualmente, possui uma loja, que se encontra em reestruturação no momento. A indústria desenvolve parcerias com oito facções especializadas em confecções, além de comprar roupas prontas, fabricadas por outros fornecedores.

No ano de 2016, a INDÚSTRIA AM contratou consultorias, encarregando-as de aplicar a metodologia DIP (Diagnóstico, Implantação e Perpetuação). Inicialmente, os setores da indústria foram diagnosticados detalhadamente, em busca da identificação de aspectos positivos e negativos em seu nível organizacional. Dentre os vários aspectos levantados para cada setor, observa-se a necessidade de aprimoramento das análises de tendências de moda no setor produtivo da indústria.

Para tomar decisões quanto ao desenvolvimento de novos produtos, a referida indústria utiliza informações relacionadas às tendências de moda feminina, como cores, estilos, temas, marcas e *designer*, e também a personalidades em destaque nas mídias.

O desenvolvimento dos produtos ocorre sazonalmente, sendo a produção dos diferentes períodos denominada de “coleção”. Em reuniões realizadas com os colaboradores da INDÚSTRIA AM, os mesmos relataram que alguns dos indicadores de tendências utilizados como parâmetro para a produção dessas coleções são os desfiles internacionais de moda. De acordo com o diretor, os eventos internacionais influenciam as tendências do ano seguinte no Brasil.

---

<sup>27</sup> <http://mundoam.com/>

Na INDÚSTRIA AM, o ciclo de vida da coleção é de no mínimo 12 meses, contados desde a fase de criação da coleção e de desenvolvimento de seus produtos, até a do lançamento no mercado.

Segundo seu diretor, decisões não assertivas resultaram em prejuízos. Caso fossem obtidas mais informações sobre as tendências e os temas da moda, algumas dessas decisões poderiam ter sido mais assertivas e menos prejudiciais. Portanto, o modelo proposto foi aplicado com o objetivo de fornecer conhecimentos adicionais que pudessem ajudar na assertividade das decisões.

Os resultados da aplicação do modelo proposto, além de terem sido avaliados pelos colaboradores da referida indústria, também foram avaliados por colaboradores da indústria de roupa do grupo Morena Rosa. Essa indústria, composta por diversas marcas, está localizada na cidade de Cianorte no estado do Paraná, atendendo boa parte do Brasil.

### **5.1. FASE I – DIAGNÓSTICO**

Como o objetivo desta fase é discutir quais informações são relevantes para o desenvolvimento do produto, foram realizadas entrevistas, visitas técnicas de acompanhamento das atividades e reuniões com os colaboradores da INDÚSTRIA AM. Os contatos ocorreram, em média, de duas a quatro vezes por mês, entre os meses de novembro de 2016 e maio de 2017. Além das entrevistas, visitas e reuniões, foram realizados contatos telefônicos e troca de mensagens eletrônicas.

A INDÚSTRIA AM dispõe de uma equipe de estilistas que, apoiada pelo setor de *marketing* da indústria, realiza levantamentos e coletas dos elementos referentes aos temas e tendências da moda. Para tanto, a equipe acompanha os eventos de moda, as revistas, os portais eletrônicos, os comentários e as opiniões de especialistas e até mesmo as influências sociais.

Tais influências sociais têm se destacado ultimamente em razão do surgimento de pessoas que, nos canais de comunicação disponíveis na internet, partilham opiniões, ideias, conceitos e críticas com seus seguidores. Para *Halvorsen et al.* (2013), os comentários de moda produzidos e divulgados nesses canais influenciam o comportamento de compra e também estimulam o consumo dos seus seguidores. Assim, os estilistas e o departamento de marketing, desenvolvem trabalhos de pesquisa para identificar esses influenciadores e seus comportamentos no mercado da moda.

De acordo com os colaboradores da INDÚSTRIA AM, os eventos de moda internacional, especialmente os ocorridos na Europa e nos EUA, repercutem no Brasil com atraso médio de um ano. Atualmente, a equipe de criação está trabalhando no desenvolvimento da coleção outono-inverno para o próximo ano (2018) e, para isso, inspira-se em eventos de moda ocorridos durante o ano de 2017.

O Quadro 6 apresenta um resumo dos principais aspectos da Fase I do modelo proposto aplicado na INDÚSTRIA AM.

**QUADRO 6 – RESUMO DOS ASPECTOS ABORDADOS NA INDÚSTRIA AM.**

<b>Aspectos FASE I</b>	<b>Questionamentos Pertinentes</b>	<b>Respostas aos Questionamentos</b>
Atividades da empresa	<ul style="list-style-type: none"> <li>- O que a empresa produz?</li> <li>- Quais são as principais atividades da empresa?</li> </ul>	<ul style="list-style-type: none"> <li>- Confecção de artigos de moda casual feminina.</li> <li>- Desenvolvimento e comercialização de produtos.</li> </ul>
Componentes da Empresa	<ul style="list-style-type: none"> <li>- Possui diversidade de produtos?</li> <li>- Como são desenvolvidos esses produtos?</li> <li>- Qual o mercado alvo?</li> </ul>	<ul style="list-style-type: none"> <li>- Diversidade razoável de produtos.</li> <li>- Os produtos são desenvolvidos por estilistas próprios que utilizam eventos de moda, revistas técnicas, matérias e entrevistas publicadas em mídia sociais para trabalhar as novas coleções, desde a concepção do projeto das vestimentas até sua fabricação.</li> <li>- Em geral atende o público feminino,</li> </ul>

		faixa etária adolescente-jovem, atuação no varejo por meio de sua rede de lojas. A maioria das lojas estão presentes em <i>shopping centers</i> .
Objetivos para utilização assertiva do modelo proposto	<p>- O que se buscará em nível de conhecimento?</p> <p>- Qual a importância desse conhecimento para o desenvolvimento de novos produtos (coleção)?</p>	<p>- A ideia é buscar conhecimentos sobre os dois maiores eventos de moda do mundo, o de Nova Iorque, nos Estados Unidos, e o de Milão, na Itália, identificando detalhes e nuances do que pode se tornar tendência a ser explorada.</p> <p>- A importância desse tipo de conhecimento é auxiliar a geração de novas possibilidades e no desenvolvimento de aspectos diferenciais e de detalhes atrativos na nova coleção.</p>

FONTE: ELABORADO PELO AUTOR

## 5.2. FASE II – HISTÓRICO E COMPREENSÃO DOS DADOS

Com a finalidade de compreender os dados disponíveis, levando em conta o objetivo definido na fase anterior, esta fase foi desenvolvida nas seguintes etapas.

### 5.2.1. ETAPA 1 – IDENTIFICAÇÃO DAS FONTES DE DADOS

Para identificação das possíveis fontes de dados, foram realizadas reuniões, nas quais os colaboradores expuseram os procedimentos realizados para o levantamento das informações que subsidiavam as tomadas de decisão durante o desenvolvimento da coleção. Os destaques incidiram sobre a obtenção de dados nas mídias sociais, cujos usuários produzem comentários e discussões sobre eventos de moda. Nesses ambientes, sem limite geográfico, diferentes usuários se conectam para compartilhar conteúdos e trocar experiências.

Nesse caso específico, tendo em vista sua objetividade e a possibilidade de vincular termos específicos a eventos de moda, dentre as mídias sociais disponíveis, optou-se pela utilização de conjuntos de dados originados da rede social *Twitter*. Os conjuntos de dados são formados por postagens de usuários: *i)* que compartilham fotos e vídeos; *ii)* que geram e opinam sobre determinado tema; *ii) broadcasters*<sup>28</sup> *que postam conteúdos já publicados*; *iii)* que divulgam e promovem seus produtos.

### 5.2.2. ETAPA 2 – FACETAS PARA A FONTE DE DADOS

Nesta etapa, destinada a apresentar uma visão mais ampla da fonte de dados, atenta-se para as características, qualidade, acesso, finalidade, estrutura, volume, natureza e usuário.

Para melhor compreensão da fonte de dados, realiza-se o desenvolvimento de “faceta” e “subfacetas”, as quais representam, de forma generalizada, as características da fonte de dados. A Figura 36 ilustra as subfacetas derivadas da rede social *twitter* (faceta), as quais foram desenvolvidas juntamente com os colaboradores da INDÚSTRIA AM.

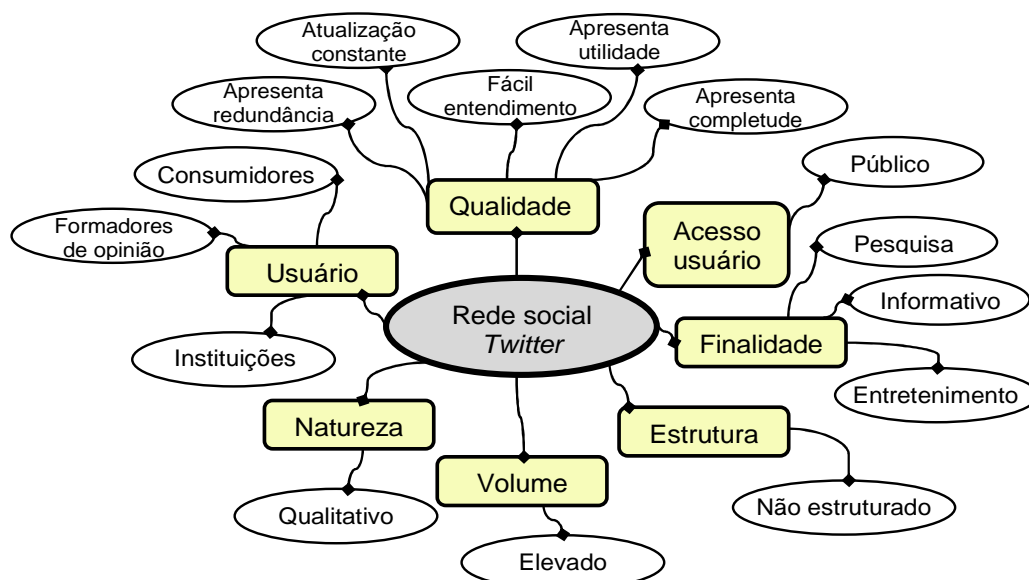


FIGURA 36 – FACETA E SUBFACETAS DA REDE SOCIAL TWITTER  
FONTE: ELABORADA PELO AUTOR

<sup>28</sup> Método de transferência de mensagem para todos os receptores simultaneamente.

A discussão gerada durante a elaboração da Figura 36 proporcionou aos colaboradores o entendimento das características e aproximou-os do processo de descoberta de conhecimento.

### **5.2.3. ETAPA 3: AVALIAÇÃO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS**

Esta etapa, que corresponde a uma avaliação mais detalhada das características dos dados, foi realizada por meio de um formulário, cuja finalidade foi auxiliar no processo de descoberta de conhecimento, conforme mostrado no Quadro 4.

O formulário se inicia com a questão do objetivo da descoberta de conhecimento, ou seja, do propósito da obtenção de informações referentes ao produto em relação ao consumidor, além de outras informações relevantes do produto. Rozenfeld *et al.* (2006) e Back *et al.* (2008) destacam essas atividades e outras ferramentas e métodos convencionais para prover a aquisição de informações referente ao produto.

Sequencialmente, vem à necessidade de conhecer melhor o conjunto de dados para que seja possível realizar a extração de conhecimento. Neste caso, após a análise, constatou-se que havia a possibilidade de se realizar essa extração.

Os conjuntos de dados foram extraídos da fonte originada do *Twitter* com auxílio da linguagem R, que foi útil para a manipulação desses dados. O primeiro passo para a extração foi desenvolver um cadastro de aplicação<sup>29</sup> no *Twitter*, o qual gerou a chave e a senha de acesso. O pacote para linguagem R, denominado “*ROAuth*” (GENTRY e LANG, 2015), contém um protocolo aberto<sup>30</sup> que permite que aplicações de terceiros acessem recursos de usuários. Para a conexão com o *Twitter*, foi desenvolvida a função “iniciar”, mostrada no Apêndice B, e, para o credenciamento do acesso ao *Twitter*, foi utilizado o pacote “*ROAuth*”. Para a extração dos conjuntos de dados, foram utilizadas as funções do pacote *TwitterR* (GENTRY, 2016).

---

<sup>29</sup> <http://apps.twitter.com/>

<sup>30</sup> <http://oauth.net/>



Para referenciar as mensagens enviadas, o *Twitter* utiliza o símbolo *hashtag* (#), acompanhado do termo-chave; o símbolo *arroba* (@) é usado para identificar os usuários. O Quadro 7 apresenta os termos-chaves vinculados aos eventos de moda realizados em Nova Iorque e em Milão.

O evento *New York Fashion Week* (NYFW) ocorreu entre os dias 09 e 17 de fevereiro de 2017. Para a identificação dos comentários a esse respeito, foram utilizados os termos-chaves “#NYFW” e “#NYFW17”. Para o evento *Milan Fashion Week* (MFW), realizado entre os dias 22 e 28 de fevereiro de 2017, os termos-chaves foram “#milanfashionweek”, “#mfw17” e “#mfw”.

O Quadro 7 apresenta a quantidade de postagens extraídas para cada um dos termos e os parâmetros de busca utilizados no processo de extração.

QUADRO 7 – POSTAGENS EXTRAÍDAS.

TERMOS	Quant. de postagens	Período de extração	Idioma	Conjunto de Dados
#NYFW	9.478	09/02 a 25/02	Inglês	NYFW
#NYFW17	554	09/02 a 25/02	Inglês	NYFW
#MFW	20.225	22/02 a 07/03	Inglês	MFW
#MFW17	882	22/02 a 07/03	Inglês	MFW
#MILANFASHION	4.619	22/02 a 07/03	Inglês	MFW

FONTE: ELABORADO PELO AUTOR

Por convenção, serão utilizados NYFW e MFW para representar o evento e o respectivo conjunto de dados. Por motivo de privacidade, os nomes das pessoas e das marcas serão substituídos por termos genéricos, como *brand*, *designer* e *model*, e, para diferenciá-los entre si, os mesmos serão seguidos de hífen e uma letra do alfabeto.

Para simplificar a busca e a agregação das postagens, foram desenvolvidas, por meio da linguagem R, as funções “listartermo” e “listarusuario”, apresentadas no Apêndice C.

Os dados extraídos foram temporariamente alocados em variáveis do tipo “*list*”<sup>31</sup>, disponibilizadas na linguagem R, e seu armazenamento ocorreu em um computador pessoal. Importante ressaltar que, para conjuntos de dados com volumes superiores a 20 MB, algumas funções apresentaram tempo de processamento significativo (aproximadamente 180 minutos).

Após a extração e a realização de uma breve análise dos conjuntos de dados, foi possível obter informações para responder ao formulário de detalhamento. A primeira avaliação das características do conjunto de dados é mostrada no Quadro 8. Salienta-se que essa avaliação pode ser alterada durante o processo de descoberta de conhecimento, ou seja, a manipulação do conjunto de dados pode, naturalmente, prover maior clareza na avaliação dessas características.

Algumas das características mencionadas no Quadro 8 foram atribuídas com base nas discussões realizadas com os colaboradores da Indústria AM; outras, de nível técnico, foram definidas com base em argumentações e na experiência, ou seja, na prática.

#### Credibilidade - Os geradores da fonte possuem credibilidade?

- Sabe-se que os dados podem ser gerados por todos os usuários ativos nas redes sociais; porém, é necessário avaliar a circunstância em que esses dados são gerados. Mesmo em se tratando de eventos reconhecidos e conceituados na área da moda, o volume elevado de dados gerados torna impossível garantir a credibilidade do conteúdo que cada usuário gera. Contudo, no trabalho de descoberta de conhecimento, é possível destacar conteúdos que tiveram repercussões, o que contribui para a avaliação da credibilidade do conjunto de dados. Nessa primeira avaliação, subjetiva, atribuiu-se ao conjunto de dados credibilidade média.

---

<sup>31</sup> Na linguagem R, o “*list*” é um objeto que pode armazenar diferentes tamanhos e tipos de dados.

**QUADRO 8 – FORMULÁRIO DE DOCUMENTAÇÃO E DETALHAMENTO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS.**

Detalhamento do conjunto de dados				
<b>1. Objetivo da descoberta de conhecimento</b>				
Obter conhecimentos referentes a produtos por meio de comentários em mídias sociais de eventos de referência em moda.				
<b>2. Informações do conjunto de dados</b>				
<b>2.1 - Fonte(s):</b> mídias sociais				
<b>2.2 - Formato(s) (variedade):</b> ( ) Estruturado ( x ) Não Estruturado ( ) Semiestruturado				
<b>2.3 - Atualização (alimentação da fonte de dados):</b> ( x ) Tempo real ( x ) Diário ( ) Semanal ( ) Mensal ( ) Anual ( ) Outras: _____				
<b>2.4 - Intervalo de tempo - (período inicial e final):</b> 29 de fevereiro a 07 de março de 2017.				
<b>2.5 - Gerador (autor da fonte):</b> usuários da rede social				
<b>2.6 - Observações:</b> termos-chaves de busca (NYFW; NYFW17; MFW; MFW17; MILANFASHION).				
<b>3. Acompanhamento das características do conjunto de dados</b>				
Características	Avaliação			Observações
	1ª	2ª	3ª	
<b>3.1 - Credibilidade</b>	3	3		
<b>3.2 - Veracidade</b>	3	3		
<b>3.3 - Imparcialidade</b>	2	2		
<b>3.4 - Utilidade</b>	4	4		
<b>3.5 - Atualidade</b>	4	4		
<b>3.6 - Complexidade</b>	4	3		
<b>3.7 - Objetividade</b>	5	5		
<b>3.8 – Inconsistência</b>	4	3		
<b>3.9 - Apresenta qualidade mínima?</b>	Sim ( )	Sim ( X )	Sim ( )	
	Não ( X )	Não ( )	Não ( )	
<b>Legenda:</b> (1) muito baixa- (2) baixa - (3) média - (4) alta - (5) muito alta				
<b>4. Lista de recursos tecnológicos</b>				
<b>4.1 - Hardware:</b> Computadores de baixo porte				
<b>4.2 - Disponibilidade de armazenamento:</b> Em torno de 10 Terabytes				

*FONTE: ELABORADO PELO AUTOR.*

Veracidade - Os dados gerados pelas fontes correspondem à realidade dos fatos ou podem ser oriundos de um evento sem muita importância, ocorrido localmente? O conteúdo dos dados é obtido diretamente da fonte ou existem possibilidades de os mesmos terem sido manipulados?

- A avaliação dessa característica é semelhante à da credibilidade. O conjunto de dados é referente a um evento importante da

indústria da moda, mas a maneira como esses dados foram gerados não favorece o controle para avaliar a veracidade dos comentários, das opiniões e dos sentimentos postados. Nesse caso, tendo em vista a importância e a proporção do evento de moda, o resultado da avaliação da veracidade foi médio, com a ressalva de que a avaliação pode sofrer alterações de acordo com o resultado da descoberta de conhecimento. Os conteúdos (postagens) são oriundos dos usuários da fonte de dados e não foram manipulados.

Imparcialidade - A fonte de dados é isenta de influência?

- O conjunto de dados em questão não é isento de influência, mas é afetado por ela de maneira diferente da de outros conjuntos de dados em que as consequências podem ser negativas para o produto. Assim, no caso da moda, as influências das mídias sociais, por envolver propagandas e opiniões, são relevantes, pois podem se tornar tendências ou gerar temas para as distintas coleções.
- A imparcialidade é um ponto importante para o processo de descoberta de conhecimento. Como as influências sobre as fontes podem indicar tendências de moda, a avaliação a respeito da imparcialidade do conjunto de dados foi baixa. Observa-se que, nesse contexto, faz-se necessária uma averiguação, por meio de estudo de campo ou em outras fontes de dados, de forma a garantir que a imparcialidade do conjunto de dados seja comprovada.

Utilidade – Anteriormente ao levantamento de dados da fonte, é possível verificar sua utilidade para a descoberta de conhecimento?

- Para responder a esse questionamento, realiza-se a análise prévia dos atributos do conjunto de dados. Além das postagens com os comentários dos usuários de redes sociais, apresentam-

se outras informações, tais como: marcação de uma postagem como favorito, data de criação da postagem, identificação do usuário, quantidade de compartilhamento, quantidade de *retwetter*, localização do usuário (latitude e longitude), idioma e endereço de URL<sup>32</sup> (*Uniform Resource Locator*). Inicialmente, o conjunto de dados apresentou utilidade, mas essa avaliação será confirmada ao longo do processo de descoberta de conhecimento, ao final do qual os dados poderão ser considerados fortemente úteis.

*Atualizada* - O conteúdo da fonte de dados encontra-se atualizado?

- O conjunto de dados foi extraído justamente no período em que ocorreram os eventos, portanto, está atualizado.

*Inconsistência* - O conteúdo da fonte de dados apresenta ruídos?

- Na primeira análise dos dados, foram constatados muitos ruídos, como erro de digitação, termos e caracteres irrelevantes.

*Complexidade* - O conteúdo da fonte de dados apresenta estrutura complexa e necessita de ferramentas intermediárias para transformá-la?

- As postagens do conjunto de dados avaliado são textuais, portanto, em formato não estruturado. Isso as faz ser consideradas complexas quando comparadas com formatos não estruturados, o que implica que necessitam de técnicas e ferramentas para ser tratadas.

*Objetividade* - O conjunto de dados corresponde ao objetivo do processo de descoberta de conhecimento?

- O conjunto de dados contém postagens relacionadas aos eventos de moda analisados e foi extraído exatamente no período em que esses eventos estavam ocorrendo. Portanto, apresenta

---

<sup>32</sup> Endereço de um recurso (foto, vídeo) disponível em uma rede, com a finalidade de ser referenciado e prover acesso aos usuários.

evidências claras de que corresponde ao objetivo traçado no processo de descoberta de conhecimento.

Para completar o preenchimento do formulário apresentado no Quadro 8, foi realizado o levantamento dos recursos tecnológicos (*hardware e software*) disponíveis na INDÚSTRIA AM. Em relação ao sistema de informação para atender às necessidades gerenciais, tais recursos são terceirizados; já, as demandas momentâneas de hardware, necessárias para o funcionamento da indústria, são supridas por computadores de baixo porte.

#### **5.2.4. ETAPA 4: DECISÃO DAS SOLUÇÕES TECNOLÓGICAS**

Nesta etapa, desenvolve-se um maior detalhamento em relação aos dados e com base nisso é possível então realizar a aplicação prévia das tarefas de filtragem, limpeza e transformação. Um estudo para avaliar a maneira como os dados podem ser trabalhados direciona as modificações a ser efetuadas na estrutura bruta do conjunto de dados.

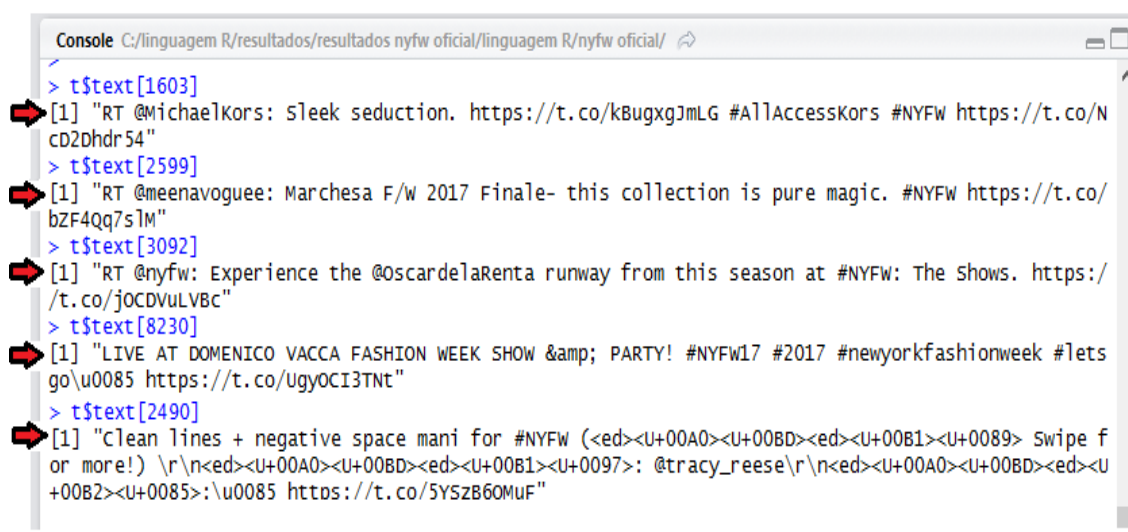
De posse das informações do conjunto de dados, aplica-se o diagrama, ilustrado na Figura 27, para melhorar o conjunto de dados e selecionar a solução adequada para então realizar a descoberta de conhecimento.

O processo de extração da fonte de dados pode gerar conjuntos que demandam grande esforço computacional, o que compromete o tempo de processamento do conjunto. Nesse caso, foram utilizados filtros que, no processo de extração, produziram conjuntos de dados favoráveis ao armazenamento em disco local.

A avaliação e a manipulação do conjunto de dados para definir a escolha da solução adequada passaram pelas seguintes atividades.

Atividade 1 - avaliação do conjunto de dados: a análise prévia realizada no conjunto de dados permitiu avaliar as características apresentadas no Quadro 8, indicadas a seguir:

- A credibilidade e veracidade apresentaram dúvidas, que podem ser dirimidas no decorrer ou no final do processo de descoberta de conhecimento.
- A imparcialidade detectada foi baixa, porque a área de moda é naturalmente influenciável por pessoas e organizações ligadas a essa área.
- A utilidade, atualidade e objetividade demonstradas no conjunto foram aceitáveis.
- Em relação à complexidade e à inconsistência, a Figura 37 mostra claramente que as postagens mencionadas devem ser tratadas e melhoradas, pois, além de não dispor de estruturas definidas, apresentam ruídos, tais como: URL; caracteres especiais; termos sem utilidade e/ou significado; frases e termos em caixa alta e caixa baixa.



```

Console C:/linguagem R/resultados/resultados nyfw oficial/linguagem R/nyfw oficial/
> t$text[1603]
[1] "RT @MichaelKors: sleek seduction. https://t.co/kBugxgJmLG #AllAccessKors #NYFW https://t.co/N
cd2Dhnr54"
> t$text[2599]
[1] "RT @meenavoguee: Marchesa F/W 2017 Finale- this collection is pure magic. #NYFW https://t.co/
bZF4Qq7s1M"
> t$text[3092]
[1] "RT @nyfw: Experience the @OscarDeLaRenta runway from this season at #NYFW: The Shows. https:/
/t.co/jOCdVvLVbc"
> t$text[8230]
[1] "LIVE AT DOMENICO VACCA FASHION WEEK SHOW & PARTY! #NYFW17 #2017 #newyorkfashionweek #lets
go\u0085 https://t.co/ugyOCi3Tnt"
> t$text[2490]
[1] "Clean lines + negative space mani for #NYFW (<ed><U+00A0><U+00BD><ed><U+00B1><U+0089> Swipe f
or more!) \r\n<ed><U+00A0><U+00BD><ed><U+00B1><U+0097>: @tracy_reese\r\n<ed><U+00A0><U+00BD><ed><U
+00B2><U+0085>:\u0085 https://t.co/5YSzB60MuF"

```

**FIGURA 37 – POSTAGENS DO CONJUNTO NYFW**  
**FONTE: ELABORADA PELO AUTOR.**

Embora os processos de ETL e de ingestão sejam utilizados na Fase III para o armazenamento e/ou análise de descoberta de conhecimento, nesta fase, para reduzir a complexidade e a inconsistência apresentadas no conjunto de dados, foram realizadas as tarefas de pré-limpeza, pré-filtragem e pré-transformação. Por meio dessas tarefas, manipulou-se antecipadamente o conjunto para obter melhor desempenho no processamento. Conjuntos de dados com formatos

textuais permitem análises léxicas; nesse caso, para a aplicação dessas tarefas nos conjuntos NYFW e MFW, foi utilizada a linguagem R, especificamente o pacote “*tm*”, empregado na mineração de texto (FEINERE e HORNIK, 2017). Os códigos utilizados no processamento e na manipulação do conjunto de dados são apresentados no Apêndice D.

Os conjuntos de dados foram armazenados em variáveis “*list*”, semelhantes a vetor, porém seus elementos assumem diferentes tipos de dados. Para a utilização das funções do pacote “*tm*”, faz-se necessário transformar essa variável em uma estrutura de dados, denominada “*corpus*”, em que cada postagem do *twitter* se torna um documento, formado por uma cadeia de caracteres.

Observando-se que os conjuntos NYFW e MFW, dispostos na estrutura “*corpus*”, não atenderam ao quesito “qualidade mínima”, foram aplicadas as seguintes tarefas: *i*) transformação da estrutura dos dados; *ii*) remoção de URL’s; *iii*) remoção das pontuações ; e *iv*) remoção de números decimais. Após essas manipulações, retornou-se à Etapa 3 para uma segunda avaliação das características do conjunto de dados, preenchidas no Quadro 8. Essa avaliação indicou aperfeiçoamento nos níveis de complexidade e inconsistência.

Atividade 2 - avaliação do volume dos dados: caso a quantidade de dados a ser armazenados seja elevada, apresenta-se a necessidade de redução de seu volume. Notadamente, a fonte de dados *twitter* dispõe de grande variedade e elevado volume de dados; no entanto para essa aplicação, os conjuntos NYFW e MFW, extraídos com o auxílio da linguagem R, não apresentaram volumes elevados.

Atividade 3 - avaliação da velocidade na produção dos dados: como a extração dos conjuntos NYFW e MFW foi realizada após os eventos, tais conjuntos apresentaram velocidade estática. Caso essa extração precisasse ser realizada durante o evento de moda, os dados seriam gerados continuamente (*streaming* de dados) e, nesse caso, como as postagens ocorreriam em tempo real, não seria possível precisar previamente o volume de dados a ser extraído. Assim,



inicialmente, optar-se-ia pela utilização de sistemas com alta escalabilidade, disponíveis nas soluções *Big Data*. Embora não tenha sido necessária para esta aplicação, a linguagem R dispõe de pacotes para se trabalhar com computação paralela e distribuída, como *rpvm* e *Rmpi*.

Atividade 4 - avaliação da possibilidade de redução do volume: Os conjuntos de dados NYFW e MFW não apresentaram volume elevado e, por esse motivo, não necessitaram de sistemas com alta escalabilidade. Embora não tenha havido necessidade de redução efetiva no volume desses conjuntos, o fato de os mesmos terem sido tratados por meio de filtragens e limpezas, gerou diminuição em seu volume inicial.

Atividade 5 - verificação da estrutura dos dados: como os conjuntos de dados NYFW e MFW foram dispostos em linhas e colunas. Nesse caso, a linha representa os registros, que contêm as postagens, e as colunas, os atributos dessas postagens. Dessa forma, ao considerar o conjunto de dados como um todo, isto é, postagens e atributos, tal conjunto assume formato estruturado. Entretanto, caso o interesse se restrinja apenas às postagens, em formato de texto escrito e linguagem natural, o conjunto se torna não estruturado.

Atividade 6 - avaliação das possibilidades de estruturação: ao realizar a avaliação dos conjuntos NYFW e MFW, busca-se descobrir as possibilidades de torná-los normalizados<sup>33</sup>, de forma que sejam utilizados em armazenamento tradicional. Nesse caso, foi elaborada uma modelagem de dados relacional, mostrada no Apêndice F. Essa modelagem é embasada na estrutura dos conjuntos de dados e permite realizar consultas SQL e aplicação de técnicas de MD tradicionais. No entanto, quando se consideram somente as postagens para realizar a descoberta de conhecimento, indica-se o BD da família NoSQL, que, além de atender aos cenários de análises textuais, possibilita a inclusão de dados advindos de outras fontes e com estruturas diferentes.

---

<sup>33</sup> Processo para aplicar regras a todas as tabelas do banco de dados, com o objetivo de evitar falhas no projeto, como redundância de dados e mistura de diferentes assuntos em uma mesma tabela.

Mediante a execução das atividades do diagrama, ilustradas na Figura 27, foram aplicadas funções para a manipulação dos dados. As tarefas de pré-filtragem, pré-limpeza e pré-transformação foram realizadas com a utilização de pacotes da linguagem R, diminuindo assim os esforços empreendidos no desenvolvimento de novas ferramentas.

As necessidades de manipulação reveladas nesta fase foram supridas pela linguagem R. Embora algumas funções, disponibilizadas nos pacotes dessa linguagem, tenham apresentado lentidão no processamento dos conjuntos NYFW e MFW no mesmo ambiente, esse obstáculo foi superado com a utilização do BD da família NoSQL em conjunto com a linguagem R, o que resultou em agilidade e flexibilidade para o processo de descoberta de conhecimento.

Diante do exposto, para os conjuntos de dados NYFW e MFW, fica evidente a escolha da solução tecnológica durante a realização desta etapa. Em termos de manipulação, a solução adotada foi a tradicional, uma vez que a linguagem R possui uma variedade de pacotes limitados a essa tecnologia. Já, para o armazenamento, a solução utilizada foi a *Big Data*, por meio de BD NoSQL, descrito na Etapa 2 da Fase III.

### **5.3. FASE III – SONDAÇÃO E USO DE TMT**

#### **5.3.1. ETAPA 1: PREPARAÇÃO DOS DADOS**

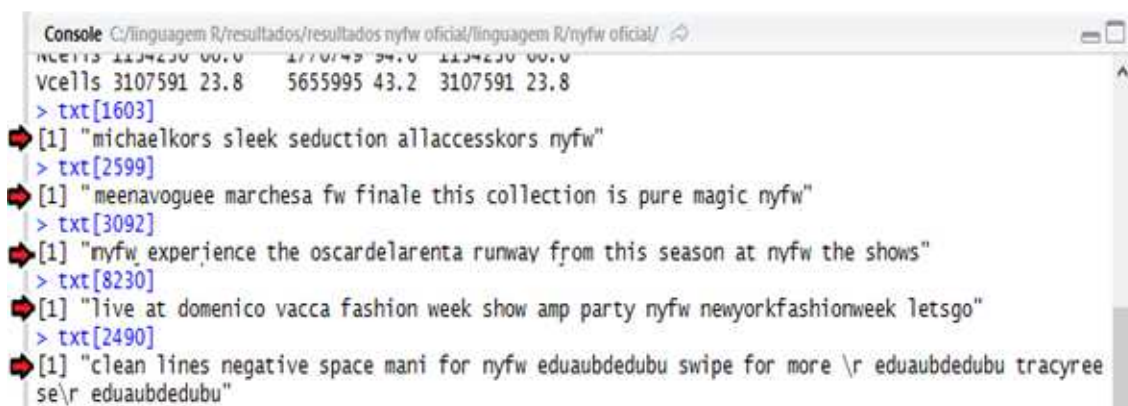
Esta etapa tem como objetivo preparar os dados e, assim, com foco no armazenamento e/ou mineração, possibilitar a efetiva descoberta de conhecimento. A seção 5.3.1 apresenta as ferramentas e os métodos apresentados na literatura.

Nessa aplicação, o processo de descoberta de conhecimento realizado nos conjuntos NYFW e MFW evidenciará os termos conforme as frequências com que aparecem nas postagens. Para tanto, foi necessário realizar a preparação desses dados, por meio das seguintes tarefas:

- remoção de termos sem relevância por intermédio do pacote “*tm*”. Além da lista de termos de acordo com o idioma, disponibilizada por esse pacote, foi elaborada outra lista com termos específicos;
- padronização de todos os caracteres em caixa baixa;
- remoção de espaços em branco;
- redução dos termos aos seus radicais, como *paraded* para *parade* e *walked* para *walk*.

Os códigos desenvolvidos para a execução dessas tarefas estão descritos no Apêndice E.

A Figura 38 apresenta exemplos de postagens depois de realizada a preparação do conjunto de dados, conforme mostrado na Figura 37. Observa-se que a limpeza realizada gerou junções de caracteres que formam palavras indefinidas, como, “*eduaubdedubu*”, as quais foram eliminadas para não prejudicar a análise.



```

Console C:/linguagem R/resultados/resultados nyfw oficial/linguagem R/nyfw oficial/
ncells 1134230 00.0 1770749 94.0 1134230 00.0
vcells 3107591 23.8 5655995 43.2 3107591 23.8
> txt[1603]
[1] "michaelkors sleek seduction allaccesskors nyfw"
> txt[2599]
[1] "meenavoguee marchesa fw finale this collection is pure magic nyfw"
> txt[3092]
[1] "nyfw experience the oscar delarenta runway from this season at nyfw the shows"
> txt[8230]
[1] "live at domenico vacca fashion week show amp party nyfw newyorkfashionweek lets go"
> txt[2490]
[1] "clean lines negative space mani for nyfw eduaubdedubu swipe for more \r eduaubdedubu tracyree
se\r eduaubdedubu"

```

**FIGURA 38 – POSTAGENS DOS USUÁRIO DO TWITTER REFERENTE AO EVENTO NYFW APÓS AS ATIVIDADES DE PREPARAÇÃO DOS DADOS**  
**FONTE: ELABORADA PELO AUTOR**

### 5.3.2. ETAPA 2: SOLUÇÃO TECNOLÓGICA DE ARMAZENAMENTO

Na seção 5.3.2, foram discutidas as soluções tecnológicas de armazenamento tradicional e *Big Data* e avaliados os atributos e as características dos BD. Tais avaliações possibilitaram que a utilização de BD NoSQL fosse orientada de acordo com as fontes de dados úteis no PDP.

Portanto, para os conjuntos de dados originados das mídias sociais, os bancos de dados recomendados foram o *MongoDB* e *Cassandra*. Para trabalhar com a linguagem R, foram desenvolvidos os pacotes “*mongolite*” do autor Ooms (2017) e “*RCassandra*” do autor Urbanek (2015), os quais fornecem interfaces de comunicação e manipulação de dados entre os bancos e a linguagem R.

O processamento com a linguagem R do conjunto de dados NYFW e MFW foi limitado pela RAM (*Random Access Memory*) do computador. Dessa forma, para melhorar o processamento, foi criado um ambiente de armazenamento na linguagem R conectado ao *MongoDB*<sup>34</sup> por meio do pacote *mongolite*, que, além de possibilitar a interface com a linguagem, fornece suporte para indexação, *map-reduce* e funções *streaming*, essencial para conjunto de dados que exige alta escalabilidade. O código de conexão e armazenamento está descrito no Apêndice G.

#### **5.4. FASE IV – DESCOBERTA DE CONHECIMENTO**

##### **5.4.1. ETAPA 1: ANÁLISE**

Nesta etapa é realizada a busca do conhecimento e verificado se o conjunto de dados está apto à aplicação das tarefas de associação e agrupamento.

Sapountzi e Psannis (2016) destacam que a unidade mais básica da estrutura linguística pode ser a palavra, argumentando que as abordagens comumente utilizadas nos conteúdos textuais são a linguística, a semântica, a estatística ou a combinação das três.

O fato de os conjuntos NYFW e MFW apresentarem diferentes atributos dá margem às mais diversificadas análises; no entanto, nesta aplicação, o foco esteve nas postagens de usuários do *twitter*. Foram utilizados algoritmos com cálculos de ocorrência dos termos e, posteriormente, aplicadas ações para associar e/ou agrupar os termos que apresentaram maior relevância<sup>35</sup>. Para

---

<sup>34</sup> [www.mongodb.com/download-versão3.4.2](http://www.mongodb.com/download-versão3.4.2) atualizada em 02.01.2017.

<sup>35</sup> A frequência com que o termo aparece no conjunto de dados.

isso, as postagens contidas nos conjuntos de dados, dispostas na estrutura “corpus”, foram transformadas em estrutura vetorial.

Após o cálculo de frequência, os termos em destaque foram selecionados e apresentados para a equipe de colaboradores da INDÚSTRIA AM. Nessa reunião, a análise dos termos destacados permitiu identificar *i)* nome de estilistas, *designers*, atrizes e modelos; *ii)* estação do ano; *iii)* estilo de roupa; *iv)* termos referentes às possíveis tendências; *v)* usuários; *vi)* portais de moda e *vii)* termos aparentemente inúteis para a descoberta de conhecimento.

No Quadro 9, apresentam-se os termos destacados e suas respectivas frequências nos conjuntos NYFW e MFW .

QUADRO 9 – FREQUÊNCIA DOS TERMOS DESTACADOS

#NYFW				#MFW			
Termo	Freq (%)	Termo	Freq (%)	Termo	Freq (%)	Termo	Freq (%)
runway	8,67	thank	2,27	fw	26,23	brand-c	4,38
fw	7,34	day	2,21	fashion	17,09	designer-h	4,15
designer-y	7,28	see	2,21	show	16,09	designer-b	3,87
style	7,24	design	2,18	milan	12,24	fashionweek	3,63
model-x	6,07	hq	2,18	fall	10,96	see	3,54
fall	5,94	hair	2,14	backstage	9,12	brand-x	3,42
week	5,55	today	2,08	brand-b	8,43	milanfashionweek	3,42
fashionweek	5,29	backstag	2,01	look	7,72	brand-j	3,40
designer-x	5,23	favorit	2,01	model-g	6,95	model-b	3,38
model	4,81	get	1,92	style	6,90	today	3,32
love	4,16	meenavogue	1,82	collection	6,78	brand-d	3,12
street	4,12	pic	1,82	latest	6,64	best	3,03
collect	4,06	time	1,82	walk	6,40	brand-e	2,85
look	3,70	check	1,79	winter	6,09	love	2,75
beauti	3,35	via	1,79	week	5,85	trend	2,46
designer-w	3,35	top	1,75	new	5,62	meenavoguee	2,32
accesdes-y	3,15	like	1,69	designer-c	5,56	day	2,28
latest	2,99	one	1,69	street	5,15	brand-f	2,26
amp	2,70	coach	1,62	hadidnews	5,03	women	2,26
nyc	2,57	just	1,62	runway	4,85	blogger-c	2,20
season	2,31	now	1,62	oscar	4,60	brand-g	2,12
trend	2,31	take	1,62	thank	4,52	model-e	2,02

FONTE: ELABORADO PELO AUTOR

O cálculo da frequência desses termos foi realizado de acordo com o número de postagens do conjunto de dados. Isto é, a frequência de determinado termo

no conjunto de dados correspondeu à divisão da frequência absoluta<sup>36</sup> desse termo pelo total de postagens do conjunto de dados. Tem-se, assim,

$$Fd(x) = \frac{Fa(x)}{Np},$$

em que  $F_d$ ,  $F_a$  e  $N_p$  representam a frequência do termo em relação ao número de postagens, a frequência absoluta e o número total de postagens, respectivamente.

Por meio da análise das frequências mencionadas no Quadro 9, foram identificados termos escritos de diferentes maneiras, porém com significados semelhantes, a exemplo das roupas produzidas pelo estilista Michael Dantas (nome fictício) que foram postadas com os termos "michaeldantas", "michael", "dantas" e "md". Para solucionar os problemas decorrentes dessas diferentes postagens e proceder à análise, os conjuntos de dados foram submetidos à Etapa 1 da Fase III, já que o modelo proposto prevê retorno a fases anteriores para melhorar a qualidade dos dados.

Em relação aos termos destacados, conforme Quadro 9, foi utilizada a função "*findAssocs*", pertencente ao pacote "*tm*", para associá-los a outros termos. A avaliação das associações entre os termos é realizada individualmente e, para melhorar esse processo, foram empregadas as técnicas de visualização descritas na próxima etapa desta fase.

A Tabela 6 ilustra exemplos de associação entre os termos pesquisados no conjunto NYFW e suas frequências em relação aos termos associados. Com base nessa tabela, podem-se realizar algumas observações.

- Associação 1 - o termo "*model-x*" refere-se a uma modelo e atriz indiana e está associado aos termos "*designer-x*" e ao "*designer-y*", que se referem a desenhistas norte-americanos. Essa associação indica que as postagens referentes ao desfile da "*model-x*" com roupas do "*designer-x*" tiveram mais destaque, uma vez que

---

<sup>36</sup> Representa o número de vezes que determinado termo aparece no conjunto de dados.

apresentam maior frequência. Além disso, evidencia a associação do termo “*model-x*” com termo “*atriz-x*”, que representa uma atriz norte-americana de ascendência brasileira.

TABELA 6 – FREQUÊNCIAS DA ASSOCIAÇÃO ENTRE OS TERMOS - NYFW

Associação	Termo pesquisado No conjunto de dados NYFW		Termos associados	
	Termos	Frequência	Termos	Frequência
1	“ <i>model-x</i> ”	6,0%	“ <i>designer-x</i> ” “ <i>designer-y</i> ” “ <i>atriz-x</i> ”	63% 29% 38%
2	“ <i>style</i> ”	7,2%	“ <i>street</i> ”	63%
3	“ <i>designer-x</i> ”	5,2%	“ <i>acessdes-y</i> ” “ <i>model-x</i> ”	61% 38%
4	“ <i>favorit</i> ”	2,0%	“ <i>college</i> ” “ <i>color</i> ”	57% 57%
5	“ <i>designer-w</i> ”	3,3%	“ <i>model-y</i> ”	62%
6	“ <i>season</i> ”	2,3%	“ <i>yeezi</i> ”	31%

FONTE: ELABORADO PELO AUTOR

- Associação 2 - o termo “*style*” está associado ao termo “*street*”, indicando que o “estilo de rua” se destacou mais do que outros estilos.
- Associação 3 - o termo “*designer-x*” se destacou no desfile quando associado aos acessórios (“*acessdes-y*”) e ao termo “*model-x*”, como ocorreu na associação 1.
- Associação 4 - o termo “*favorit*” (favorito) destacou-se quando associado aos termos “*college*” e “*color*”, evidenciando o favoritismo das cores colegiais.
- Associação 5 - o termo “*designer-w*” refere-se a um desenhista de moda norte-americano e se destacou quando associado ao termo “*model-y*”, que representa uma modelo e atriz residente no Japão.
- Associação 6 - o termo “*season*” destacou-se quando associado ao termo “*yeezi*”, referente a uma nova linha de tênis.

Na Tabela 7, mostram-se exemplos de associação entre os termos pesquisados no conjunto MFW e suas frequências em relação aos termos associados

TABELA 7 – FREQUÊNCIAS DA ASSOCIAÇÃO ENTRE OS TERMOS - MFW

Associação	Termo pesquisado no conjunto de dados MFW		Termos associados	
	Termos	Frequência	Termo	Frequência
1	"fw"	26,2%	"model-g"	35%
2	"style"	6,9%	"street"	73%
3	"winter"	6,3%	"theimonation"	55%
4	"collection"	6,7%	"theimonation"	40%
5	"Walk"	6,4%	"model-e"	35%
6	"new"	5,6%	"model-e"	48%
7	"Model-e"	2,0%	"Brand-t"	44%
			"designer-s"	39%
8	"designer-c"	5,5%	"model-b"	36%
9	"trend"	2,4%	"rainbow"	35%
			"moeztali"	36%

FONTE: ELABORADO PELO AUTOR

Com base na Tabela 7, podem ser realizadas algumas observações.

- Associação 1 - o termo "fw" (abreviação de "fashionweek" em português "semana da moda") destacou-se quando associado ao termo "model-g", indicando que as postagens referentes à modelo norte americana se sobressaíram na semana da moda.
- Associação 2 - o termo "style", assim como no evento de moda NYFW, destacou-se quando associado ao termo "street", evidenciando o "estilo de rua".
- Associação 3 e 4 - os termos "winter" (inverno) e "collection" (coleção) destacaram-se quando associados ao termo "theimonation", que se refere a um "blog"<sup>37</sup> alternativo, aparentemente desvinculado de marcas.
- Associação 5 e 6 - os termos "walk" (andar) e "new" (novo) destacaram-se quando associados ao termo "model-e", referente a uma modelo sueca;
- Associação 7 - o termo "model-e" destacou-se quando associado aos termos "brand-t" e "designer-s".
- Associação 8 - o termo "designer-c", referente a um desenhista italiano, quando associado ao termo "model-b", referente a uma modelo americana, indica o destaque das postagens que

<sup>37</sup> <http://imonation.com/about/>



relatavam que a modelo desfilou com coleções do referido desenhista.

- Associação 9 - o termo “*trend*” (tendência) associou-se aos termos “*rainbow*” (arco íris) e “*moeeztali*”. Isso indica o destaque das postagens que relacionavam a tendência a arco íris e o termo “*moeeztali*”, referente a um aplicativo desenvolvido para celulares, utilizado para seguir *blogueiros* relacionados à moda.

Continuando com a análise dos conjuntos NYFW e MFW, especificamente quanto ao procedimento de agrupamento, foram gerados seis grupos (*cluster*), cada qual contendo seis termos, como ilustram as Figuras 39 e 40.



```

Console C:/linguagem R/ambiente total/mfw e nyfw/mfw e nyfw
+ }
Agrupamento 1: love fashionweek look model runway amp
Agrupamento 2: runway designer-y model-x fw designer-x model
Agrupamento 3: fall collect runway week designer-y amp
Agrupamento 4: street style fashionweek fw week designer-w
Agrupamento 5: week designer-w fw runway fashionweek model
Agrupamento 6: style street look runway fw week
>

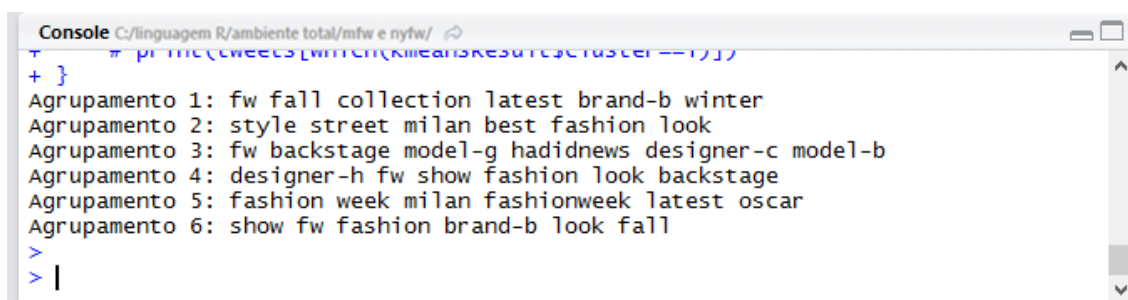
```

FIGURA 39 – AGRUPAMENTOS (K-MEANS) PARA O CONJUNTO DE DADOS NYFW  
FONTE: ELABORADA PELO AUTOR

A interpretação da Figura 39, com seus respectivos termos, resultou em algumas observações sobre os agrupamentos.

- Agrupamento 1 - aparentemente não apresentou evidências.
- Agrupamento 2 - os termos destacados nesse agrupamento e confirmados em breve pesquisa levaram à conclusão de que, no desfile realizado em Nova Iorque, a modelo representada pelo termo “*model-x*” utilizou um vestido azul marinho da coleção do desenhista representado pelo termo “*designer-y*”. Porém, em outro dia, no mesmo evento, essa modelo desfilou com um vestido branco, uma capa emparelhada pertencentes à coleção do desenhista representado pelo termo “*designer-x*”.

- Agrupamento 3 - o agrupamento realizado entre os termos indica que a coleção de outono está relacionada ao desenhista representado pelo termo “*designer-y*”.
- Agrupamento 4 – indica associação entre a expressão estilo de rua, representada pelo termos “*style*” e “*street*”, e o desenhista representado pelo termo “*designer-w*”.
- Agrupamento 5 - aparentemente não apresentou evidências.
- Agrupamento 6 - evidencia que o visual do estilo de rua se sobressaiu nos desfiles ocorridos na semana da moda.



```

Console C:/linguagem R/ambiente total/mfw e nyfw/
+ }
+ }
Agrupamento 1: fw fall collection latest brand-b winter
Agrupamento 2: style street milan best fashion look
Agrupamento 3: fw backstage model-g hadidnews designer-c model-b
Agrupamento 4: designer-h fw show fashion look backstage
Agrupamento 5: fashion week milan fashionweek latest oscar
Agrupamento 6: show fw fashion brand-b look fall
>
> |

```

FIGURA 40 – AGRUPAMENTOS (K-MEANS) PARA O CONJUNTO DE DADOS MFW  
 FONTE: ELABORADA PELO AUTOR

A interpretação da Figura 40, com seus respectivos termos, resultou nas seguintes verificações.

- Agrupamento 1 - identifica semelhança entre os termos “*fw*”, “*collection*” (coleção), “*latest*” (recente) e a marca representada pelo termo “*brand-b*”, ou seja, indica que a marca citada e sua respectiva coleção se destacaram no evento.
- Agrupamento 2 - evidencia as postagens relacionadas ao estilo de rua. Após rápida pesquisa, buscando pelos termos destacados nesse agrupamento, foi confirmado que portais de moda ressaltaram o estilo de rua;
- Agrupamento 3 - identifica os relacionamentos entre os termos “*model-g*” e “*model-b*”, quando associados ao “*designer-c*”. Após rápida pesquisa, descobriu-se que as modelos, representadas pelos termos “*model-g*” e “*model-b*”, são irmãs e desfilaram

vestindo roupas da coleção do desenhista representado pelo termo “*designer-c*”.

- Agrupamento 4 – destaca a associação entre o termo “*designer-h*” e os termos “*look*” (visual) e “*backstage*” (bastidores), o que denota o visual das roupas desse desenhista nos bastidores do evento.
- Agrupamento 5 - aparentemente não apresentou evidências.
- Agrupamento 6 - mostra a associação entre os termos referentes ao evento propriamente dito e a marca representada pelo termo “*brand-b*” e o visual de outono representado pelos termos “*look*” e “*fall*”.

Os destaques gerados pelos processos de associação e agrupamento apenas demonstraram os padrões de dependência e similaridade entre os termos. Isso permitiu a construção de supostas evidências, mas estas foram analisadas e complementadas com novas pesquisas, de forma a confirmar ou não seu aproveitamento como conhecimento novo e útil. A utilização de diferentes parâmetros nas tarefas de análise pode produzir novos e diferentes resultados, mas, em razão da limitação do escopo desta aplicação, não foram esgotadas outras possibilidades de associação e agrupamento.

Os códigos desenvolvidos para transformação do conjunto de dados em estrutura vetorial, para os cálculos de frequência e associação e agrupamento entre os termos estão descritos no Apêndice H.

#### 5.4.2. ETAPA 2: TÉCNICAS DE VISUALIZAÇÃO

As técnicas de visualização, cujo objetivo é facilitar o entendimento dos resultados gerados, proporcionam maior interação com os dados. Para a escolha da técnica de visualização adequada, além de conhecer suas características, faz-se necessário entender os conjuntos de dados, que, nesta aplicação, são NYFW e MFW.

No que se refere aos conjuntos extraídos, seu conhecimento foi assimilado ao longo da aplicação das fases do modelo proposto. Já, para o conhecimento das técnicas de visualização textual, foram utilizadas as avaliações existentes na literatura (RABELO *et al.*, 2008; HENDERSON E SEGAL, 2013).

A Tabela 5, descrita na seção 4.4.2, cujo intuito é auxiliar na escolha da visualização adequada, mostra as avaliações das técnicas de visualização em relação às características apresentadas em cada conjunto de dados. Analisando-se os conjuntos NYFW e MFW, verifica-se que os dados neles contidos apresentam características qualitativas e quantitativas; entretanto, ao reduzir os conjuntos, mantendo apenas as postagens (texto), são mantidas somente as características qualitativas.

A primeira análise gráfica dos conjuntos teve como objetivo calcular a frequência dos termos neles apresentados e, para isso, foi preestabelecido um limiar, de forma que fossem identificados apenas os termos cujas frequências fossem superiores a ele. Esse limiar é importante porque soluciona problemas de escalabilidade e dimensionalidade nas visualizações geradas.

O objetivo das técnicas de visualização é evidenciar e ordenar os termos mais frequentes nos conjuntos de dados. Dessa forma, a característica dos dados contidos nos conjuntos torna-se qualitativa ordinal. Com base nas avaliações realizadas por Rabelo *et al.* (2008), considerou-se que as técnicas iconográficas, quando comparadas às técnicas geométricas, mostraram-se mais adequadas para os conjuntos NYFW e MFW.

Os gráficos de barras ou a pizza são bastante úteis para representar frequências de uma maneira geral, no entanto, quando o objetivo é representar a quantidade de termos, tais mecanismos são limitados. Para suprir essa limitação, a avaliação de Henderson e Segal (2013), ilustrada na Figura 33, apresenta, como sugestões, técnicas de visualização e suas respectivas complexidades de utilização para dados qualitativos. Para a visualização dos termos contidos nos conjuntos NYFW e MFW, foram utilizadas as técnicas *word cloud*, *word tree* e o diagrama de associação. A técnica com maior

popularidade e utilizada em diversas áreas é a *word cloud*, que, além de apresentar menor grau de complexidade em sua utilização, permite a visualização de maior quantidade de termos, quando comparada aos gráficos de barras ou à pizza.

A Figura 41, apresentando a técnica de visualização *word cloud*, evidencia os termos que aparecem com mais frequência nos conjuntos NYFW e MFW. Dessa maneira, os colaboradores da INDÚSTRIA AM puderam visualizar graficamente, para cada um dos conjuntos avaliados, estilistas, marcas, modelos, atrizes e outros termos relacionados ao mundo da moda.

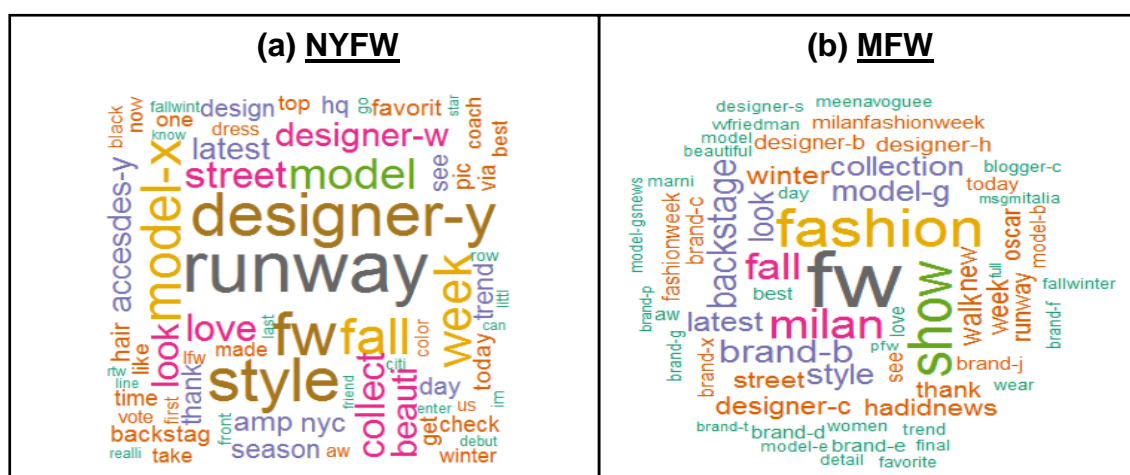


FIGURA 41 – VISUALIZAÇÃO WORD CLOUD PARA OS CONJUNTOS NYFW E MFW  
 FONTE: ELABORADA PELO AUTOR

Durante a reunião com os colaboradores da INDÚSTRIA AM para apresentação dos termos identificados nos conjuntos NYFW e MFW, surgiu a ideia de utilizar os termos destacados para gerar outros conjuntos.

Dessa ideia, aliada à necessidade de se visualizar simultaneamente os termos destacados nesses conjuntos, foi elaborada a Figura 42, que apresenta uma sequência de visualizações *word cloud* interligadas por termos previamente selecionados. Inicialmente, foi gerada a primeira visualização com a junção dos termos “*street*” e “*style*”, disposta mais à esquerda da Figura 42. Na sequência, foram selecionados os termos “*fashionblogger*” e “*downeaststyle*”, destacados por retângulos com bordas e setas de cor vermelha, para gerar a visualização

dos conjuntos e, assim, sucessivamente, de acordo com a importância do termo a ser escolhido.

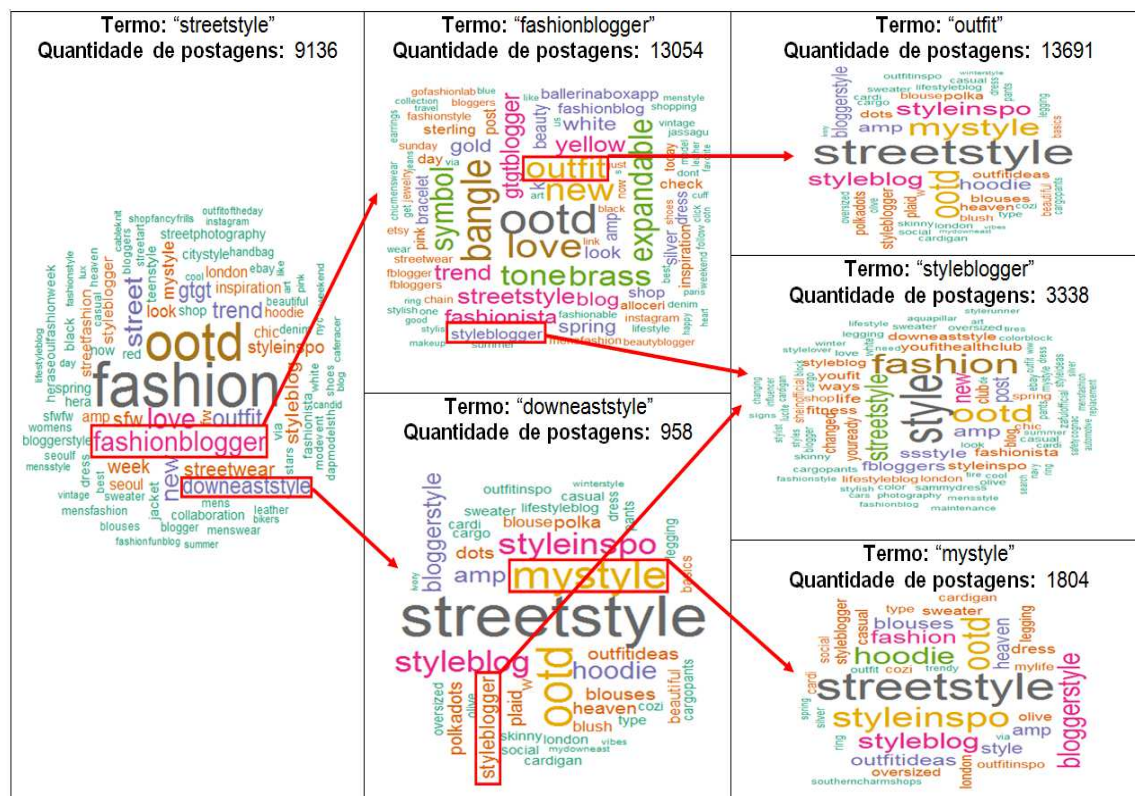


FIGURA 42 – GRUPO DE VISUALIZAÇÃO WORD CLOUD LIGADO POR TERMOS  
FONTE: ELABORADA PELO AUTOR

A dinâmica do processo empregado para gerar a sequência de visualização, mostrada na Figura 42, eleva o volume de dados e, consequentemente, amplia a velocidade na produção, coleta e processamento desses dados. Nesse caso, para a execução da sequência de visualização, torna-se necessária a criação de uma ferramenta com escalabilidade e desempenho satisfatório.

A utilização do modelo proposto, além de possibilitar o entendimento dos processos que geram essas visualizações, permitiu desenvolver um ambiente tecnológico para o processo de descoberta de conhecimento, que compreende: i) roteiro para execução de cada etapa desse processo; ii) linguagem de programação, incluindo as bibliotecas; iii) *script* de programação e, iv), soluções para armazenamento. Tal ambiente proporcionou mais agilidade na realização dos processos que geraram as visualizações *word cloud*. Para efeito

de ilustração, as visualizações geradas limitaram-se a alguns termos destacados na Figura 42. Embora tenha ocorrido limitação na quantidade dos conjuntos extraídos e em suas respectivas visualizações, alguns detalhes foram destacados pelos colaboradores da INDÚSTRIA AM.

- A visualização do conjunto de dados “*streetstyle*” destacou o termo “*fashionblogger*”, que faz referência a “*bloggers*” de moda; nesse subconjunto, destacaram-se os termos “*yellow*” (amarelo) e “*White*” (branco). O termo “*downtownstyle*” refere-se a uma loja sediada nos Estados Unidos que, mantém um *blog* para discutir as tendências da moda e comercializar seus produtos via comércio eletrônico.
- A visualização do subconjunto de dados “*downtownstyle*” destacou os termos “*polka*” (bolinhas), “*plaid*” (xadrex) e “*sweater*” (suéter).
- Na visualização do subconjunto de dados “*outfit*” (roupa) os termos destacados foram, “*polka*” (bolinha), “*hoodie*” (moletom com capuz) e *plaid* (xadrex).
- A visualização do subconjunto de dados “*mystyle*” (meu estilo) evidenciou os termos “*hoodie*” (moletom com capuz) e “*styleinspo*”, que aparece em outras visualizações e se refere a um portal on-line de moda, cujo objetivo é inspirar o mundo da moda.

Os colaboradores da INDÚSTRIA AM enfatizaram ainda a importância de se conhecer e acompanhar *blogs* relevantes para o mundo da moda. A rede mundial de computadores tem conseguido juntar em comunidades pessoas geograficamente distantes e com interesses comuns, quebrando tabus relacionados às tendências do mundo da moda.

Os termos identificados no conjunto MFW e apresentados aos colaboradores da INDÚSTRIA AM foram abordados novamente por outra técnica de visualização, a *word tree*. Essa visualização apresenta postagens, contidas no conjunto MFW, as quais se ramificam a partir de um termo raiz pré-definido. De



acordo com a avaliação de Henderson e Segal (2013), ilustrada na Figura 33, a complexidade de sua utilização é média.

A Figura 43 ilustra a ramificação gerada pela técnica *word tree* para o termo “*fall*” (outono), que representa a raiz da árvore. Essa técnica é interativa e permite que o usuário passe o mouse sobre o termo desejado e, com um clique, abra as postagens para realizar sua leitura.

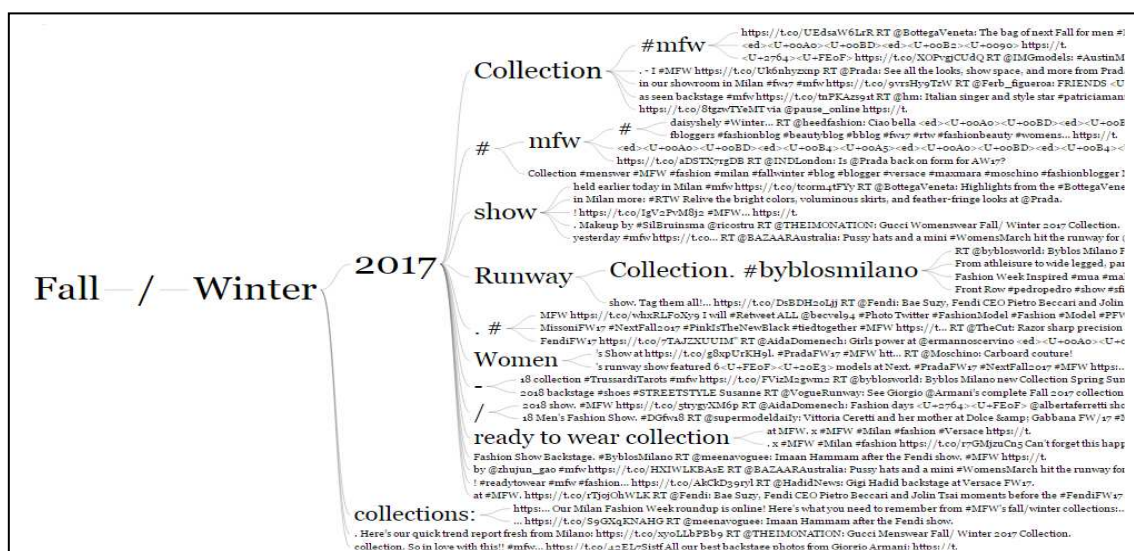


FIGURA 43 – VISUALIZAÇÃO WORD TREE<sup>38</sup> DO CONJUNTO DE DADOS MFW  
 FONTE: ELABORADO PELO AUTOR

Interessante ressaltar que, nessa técnica, é possível visualizar os termos que derivaram toda a sequência de outros termos que compõem a ramificação. Outra possibilidade importante dessa técnica é o acompanhamento visual das postagens do conjunto de dados em seu formato bruto, o que pode ser útil para a avaliação das características dos dados, descrita na Etapa 3 da Fase II do modelo proposto.

Especialmente no mundo da moda, em grande parte, as postagens realizadas no *twitter* possuem fotos ou vídeos referentes a um endereço virtual, ou seja, URL (*Uniform Resource Locator*). Considerando a importância das redes sociais para o trabalho dos colaboradores da INDÚSTRIA AM, particularmente os responsáveis pela criação das coleções, entende-se que a técnica de

<sup>38</sup><https://www.jasondavies.com/wordtree/>



visualização *word tree* ajuda na identificação das postagens e de suas respectivas *URLs*.

Ao definir sequência de termos “*fall*”; “/”; “*winter*” e “2017” para gerar a visualização do resultado com a interação da técnica *word tree*, é possível obter como exemplo, as seguintes postagens e *URL's* para o conjunto de dados MFW:

#### Postagem 1

*“Highlights from the Fall/Winter 2017 show held earlier today in Milan #mfw*  
***https://t.co/tcorm4tFYy*”**

(Tradução: Destaques do outono/inverno 2017 hoje a mostra realizada mais cedo em Milão #mfw)

#### Postagem 2

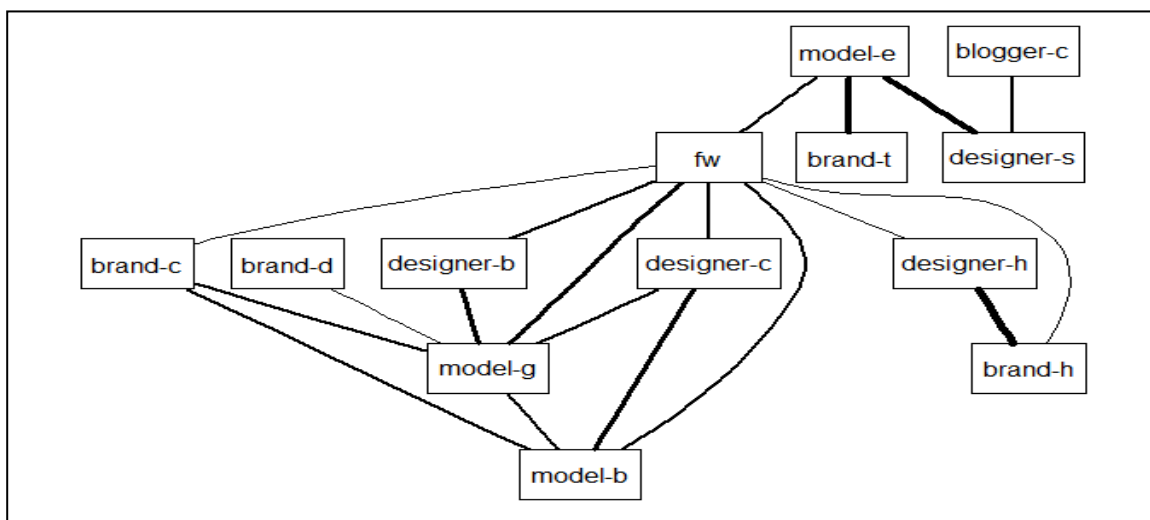
*The look of the brand-f Fall/Winter 2017 collection as seen backstage #mfw*  
***https://t.co/tnPKAzs91t***

(tradução: O visual da coleção *brand-f* outono/Inverno 2017 visto nos bastidores #mfw.

Outra possibilidade é a utilização das técnicas de visualização *word cloud* e *word tree* em conjunto, isto é, a primeira identifica os termos mais frequentes no conjunto de dados e a segunda, as postagens, as fotos e os vídeos que deram origem a esses termos.

Quando o objetivo está na relação entre os termos, a técnica utilizada é a ilustrada na Figura 44, que mostra a associação entre os termos referentes a *designers*, marcas e modelos identificados no conjunto MFW. Nessa visualização, os termos, representados por retângulos, são interligados por linhas de diversas espessuras: quanto mais espessas, maior é o grau de associação entre os termos representado por elas.

A interpretação da Figura 44 mostra claramente a associação das modelos aos designers e marcas. Por meio da análise realizada na primeira etapa da presente fase, identificou-se que, entre os termos do conjunto MFW, a maior frequência foi para o termo “*model-g*”.



**FIGURA 44 – DIAGRAMA DE ASSOCIAÇÃO DE TERMOS DO CONJUNTO DE DADOS MFW**  
**FONTE: ELABORADO PELO AUTOR**

De posse dessa informação, foi realizada uma pesquisa sobre a atuação dessa modelo na semana do evento em Milão, constatando-se que ela desfilou para diferentes marcas e *designers*, que, nesta aplicação, são representados pelos termos: “*brand-u*”, “*designer-b*”, “*brand-x*”, “*brand-c*”, “*designer-c*” e “*brand-d*”. Observa-se que, nas associações ilustradas na Figura 44, os termos “*brand-u*” e “*brand-x*” estão ausentes, o que leva a pressupor que a modelo, representada pelo termo “*model-g*”, não se destacou tanto quando foi relacionada às marcas representadas pelos termos “*brand-u*” e “*brand-x*”. Todavia, o termo “*model-g*”, quando relacionado aos termos “*designer-b*”, “*designer-c*” e “*brand-d*”, destacou-se com frequência de 29%, 19% e 16%, respectivamente.

Os colaboradores da INDÚSTRIA AM afirmaram que os conhecimentos extraídos dessas visualizações são evidências que auxiliam no direcionamento das pesquisas para a criação de novas coleções.

Quando o interesse está na identificação da similaridade existente entre os termos, além da *Word Cloud*, são indicadas outras técnicas de visualização, como a dendograma e a matriz de distância, apresentadas no Apêndice I.

Na técnica de visualização dendograma, os termos são decompostos e aninhados em vários níveis de particionamento, tornando possível a

representação de grupos formados desde o início do processo de análise. Já, a matriz de distância, utiliza sua bidimensionalidade e o contraste de cor para representar a distância entre pares, ou seja, entre pares de termos.

Para apresentar os agrupamentos dos termos identificados aos colaboradores da INDÚSTRIA AM, foi utilizada a técnica “*word cloud*”, por meio da qual é gerada uma visualização para cada grupo. Além de mostrar a relevância desses termos, a técnica permite visualizar sua presença em vários outros agrupamentos.

Para os códigos geradores das visualizações apresentadas nesta aplicação foram utilizados pacotes desenvolvidos e disponibilizados no *Ecran*, os quais se encontram descritos no Apêndice J.

A aplicação do modelo proposto demonstrou a importância da participação dos especialistas do domínio de aplicação, isto é, dos colaboradores da INDÚSTRIA AM. Por se tratar de um processo iterativo e interativo, as sugestões e os direcionamentos relatados por esses especialistas reduzem os passos e o tempo de processamento.

## **5.5. ARMAZENAMENTO DO CONHECIMENTO**

O modelo proposto prevê o armazenamento dos conhecimentos extraídos. No entanto, previamente ao armazenamento, foi necessário confrontá-los com o objetivo estabelecido na fase inicial. Depois de avaliada essa concordância, os prováveis conhecimentos extraídos foram organizados em formato de questionário, como mostra o Apêndice K. Esse questionário apresenta, para cada um dos conjuntos NYFW e MFW, nove prováveis conhecimentos, além de outros cinco relacionados aos termos “*street*” e “*style*”.

O questionário foi respondido pelos colaboradores da INDÚSTRIA AM e do grupo Morena Rosa. A finalidade foi avaliar a probabilidade de esses conhecimentos se tornarem, de fato, conhecimentos novos e úteis para auxiliar no processo de desenvolvimento da coleção, conforme a definição de Fayyad

*et al.* (1996) de que os conhecimentos extraídos devem ser válidos, novos e potencialmente úteis.

Com base na pesquisa empírica realizada na INDÚSTRIA AM, por meio de observação direta e de depoimentos de seus colaboradores, foi possível tecer algumas considerações em relação aos conhecimentos extraídos.

- As evidências relacionadas aos eventos de moda, descritas no questionário, auxiliam o direcionamento das pesquisas para o desenvolvimento da coleção.
- Os colaboradores enfatizaram a importância de se realizar o processo de extração de conhecimento logo após o evento, uma vez que, quanto antes esse conhecimento esteja disponível, maior a contribuição proporcionada ao PDP. Nesse caso, a execução sequencial das atividades, por meio das etapas e fases do modelo proposto, proporciona um ambiente favorável para a utilização e o desenvolvimento de ferramentas computacionais que, por sua vez, agilizam a realização dessas atividades.
- A associação dos termos “*trend*” e “*rainbow*” mencionada no questionário descrito no Apêndice K foi confirmada com decisões previamente tomadas pelos colaboradores da INDÚSTRIA AM.
- Embora a frequente associação entre os termos “estilo” e “rua”, em princípio, não tenha causado surpresa nos colaboradores, sua associação a outros termos destacados pode indicar tendências.
- Diante da elevada quantidade de marcas, modelos e *designers*, os colaboradores destacaram a importância de se acompanhar e avaliar as evidências das postagens publicadas nas redes sociais.

Tais considerações são relevantes e podem até mesmo auxiliar na execução das atividades descritas nas fases e etapas do modelo proposto, isto é, embasar os parâmetros definidos na realização das tarefas, na análise e na seleção das técnicas de visualização.

Na Tabela 8, estão sintetizadas as respostas fornecidas pelos colaboradores da INDÚSTRIA AM para o questionário mostrado no Apêndice K.

*TABELA 8 – RESPOSTAS AO QUESTIONÁRIO - INDÚSTRIA AM*

<b>Conjunto de dados</b>	<b>Quantidade de conhecimentos</b>	<b>Novo</b>	<b>Útil</b>	<b>Novo e útil</b>
NYFW	9	5	7	5
MFW	9	3	7	3
<i>street e style</i>	5	3	5	3

*FONTE: ELABORADO PELO AUTOR*

Para que se tornem efetivamente válidos para auxiliar o PDP, os conhecimentos extraídos devem ser simultaneamente novos e úteis para todos os colaboradores diretamente envolvidos na criação da coleção. Em relação a esse requisito, a Tabela 8 mostra que, dentre as respostas fornecidas pelos colaboradores da INDÚSTRIA AM para os questionamentos referentes ao conjunto NYFW, cinco atestaram positivamente. Dentre as nove respostas referentes aos questionamentos do conjunto MFW, apenas três foram positivas. O mesmo número de respostas positivas foi dado para os questionamentos referentes aos termos “*street*” e “*style*”.

Além dos conhecimentos extraídos terem sido avaliados pelos colaboradores da INDÚSTRIA AM, com intuito de ampliar a relevância desses conhecimentos, foi realizada também uma avaliação por colaboradores do grupo Morena Rosa, incluindo a coordenadora de estilo de uma das marcas de destaque do grupo. Essas avaliações estão sintetizadas na Tabela 9.

*TABELA 9 – RESPOSTAS AO QUESTIONÁRIO - GRUPO MORENA ROSA*

<b>Conjunto de dados</b>	<b>Quantidade de conhecimentos</b>	<b>Novo</b>	<b>Útil</b>	<b>Novo e útil</b>
NYFW	9	5	5	3
MFW	9	4	8	4
<i>street e style</i>	5	2	5	2

*FONTE: ELABORADO PELO AUTOR*

A relevância dos conhecimentos extraídos é inerente às avaliações realizadas individualmente pelas indústrias estudadas. No entanto, quando se tratou de avaliar se concomitantemente os conhecimentos eram novos e úteis, tanto por

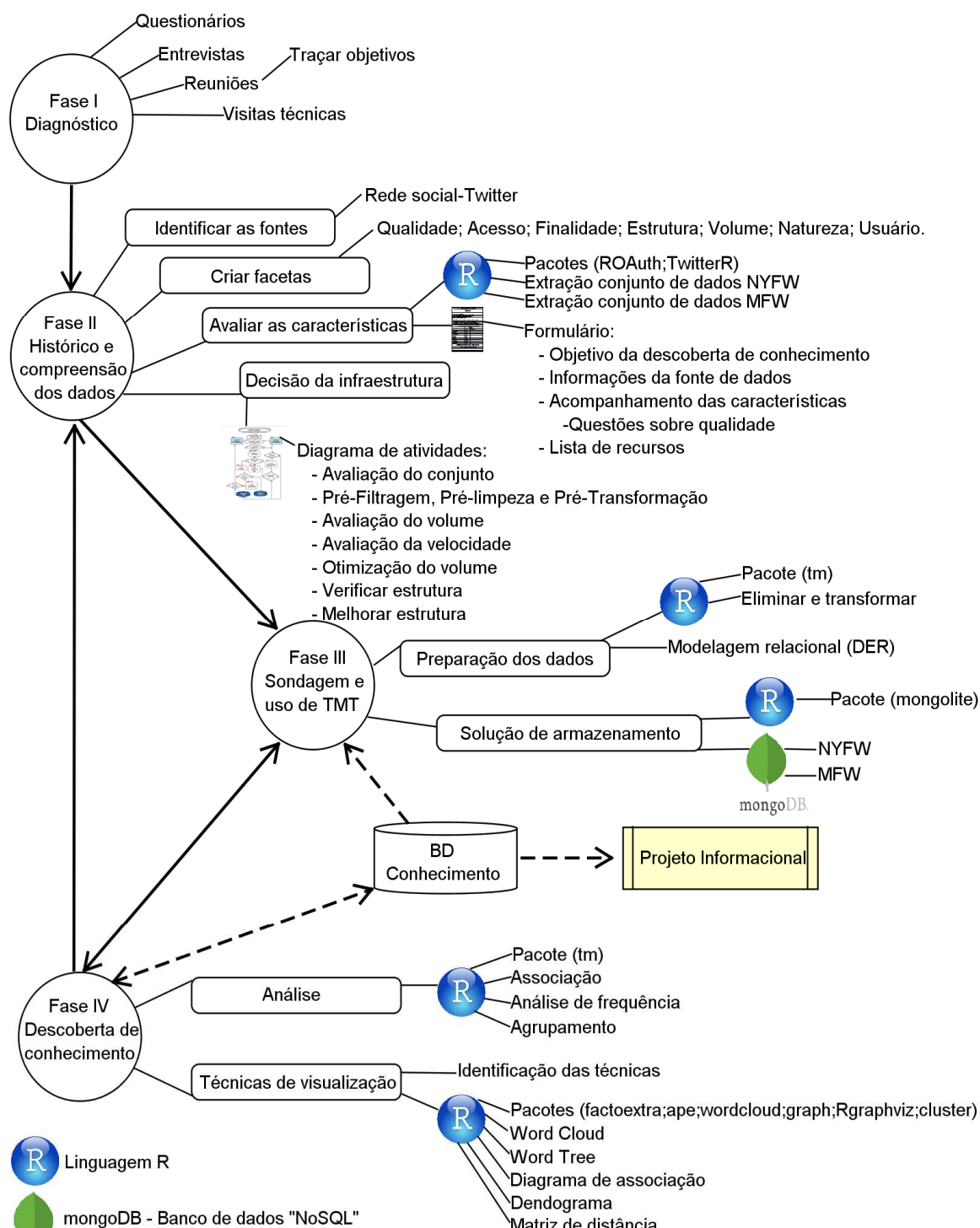
parte dos colaboradores da INDÚSTRIA AM quanto do grupo Morena Rosa, observou-se que houve consenso na avaliação de alguns conhecimentos, mais especificamente 34% dos conhecimentos extraídos. Ao considerar que os conhecimentos foram avaliados por especialistas da moda, esperava-se uma porcentagem menor para a característica relacionada à novidade.

Em virtude da natureza exploratória das análises realizadas na Fase IV, naturalmente surgem indagações em relação aos questionamentos atestados como novos e úteis. Nesse caso, faz-se necessário um estudo mais detalhado para entender a origem e os motivos que levaram essas postagens a se destacar no *twitter*.

No processo de descoberta de conhecimento, a importância está no direcionamento que o conhecimento extraído pode fornecer para ampliar o campo de visão e gerar novas concepções para o desenvolvimento da coleção. Dessa maneira, os conhecimentos extraídos podem não ser a solução para o objetivo traçado, mas podem ser evidências que levam à sua consecução.

Apurada a relevância dos conhecimentos extraídos, o modelo proposto prevê o seu armazenamento em um banco de dados para que sejam utilizados e reutilizados, quando necessário.

Destaca-se que o roteiro gerado em cada aplicação do modelo proposto é intrínseco às ferramentas utilizadas para atingir o objetivo inicialmente traçado no domínio de aplicação. Especificamente nesta aplicação, foram utilizados recursos característicos da linguagem de programação R para realizar a extração, a manipulação, o tratamento dos dados e a aplicação das tarefas de associação e agrupamento. Foi utilizado também o banco de dados mongoDB, da família NoSQL, para o armazenamento dos dados. Para efeito de ilustração, foi elaborado um diagrama, apresentado na Figura 45, o qual sintetiza as fases e etapas do modelo proposto e os recursos desenvolvidos ao longo de toda a aplicação.



**FIGURA 45 – DIAGRAMA DA APLICAÇÃO DO MODELO PROPOSTO**  
**FONTE: ELABORADA PELO AUTOR**

Apesar de o modelo proposto ter sido aplicado em indústria de confecção, o mesmo pode ser utilizado por organização manufatureira que tenha interesse em utilizar dados no processo de descoberta de conhecimento para auxiliar o PDP.



## 6. CONCLUSÃO

*Big Data* é um tema em expansão atualmente e ainda tem muito a evoluir. Conforme pesquisas realizadas neste trabalho, constatou-se que, essa expansão produziu alarde por suas vantagens e complexidades.

A elevada quantidade de dados disponibilizada nos mais diferentes formatos em ambientes digitais, aliada às possibilidades de coleta e análise desses dados proporcionada pelas ferramentas tecnológicas, fomenta o processo de extração de conhecimentos. Embora exista uma grande variedade de ferramentas, essa diversidade não tem significado sem a utilização de métodos para direcionar a escolha das tecnologias a serem utilizadas.

Sendo assim, a partir do questionamento de como desenvolver um modelo de descoberta de conhecimento que, com o auxílio de soluções tradicionais e do *Big Data*, pudesse apoiar o projeto informacional do processo de desenvolvimento do produto, propôs-se um modelo que abrangesse do início ao fim, as atividades necessárias desse processo.

Quanto à escolha das tecnologias, o modelo proposto considerou as características dos dados, prevendo a análise da estrutura do conjunto de dados e também as possibilidades de reestruturação ou transformação da estrutura. Além dessa vantagem, o modelo proposto não se limitou às tecnologias existentes, especialmente tendo em vista suas possíveis futuras evoluções.

As soluções tradicionais, que atendem às demandas de dados estruturados, dispõem de um conjunto aprimorado de ferramentas, métodos e técnicas, de forma que sua utilização se torna menos custosa no que diz respeito ao processamento, tempo e empenho. Diante disso, o modelo proposto engloba todas as possibilidades existentes para adequar o conjunto de dados com características *Big Data* à aplicação em soluções tradicionais.

A partir da concepção e construção do modelo proposto englobando as atividades necessárias para o processo de extração de conhecimentos no projeto informacional, foi possível realizar os objetivos especificados na presente tese. A fase II do modelo proposto cumpre com o primeiro objetivo. Avalia o histórico e a compreensão dos dados, desde a identificação de sua origem ao apoio na decisão de soluções tecnológicas.

A fase III e IV propõe a sondagem de TMT para soluções tradicionais e *Big Data* e, possibilidades tecnológicas para análise do conjunto de dados, dessa forma, cumprem com o segundo objetivo específico. Por fim, o terceiro objetivo específico foi atendido pela ilustração da aplicabilidade do modelo proposto em um cenário real.

O diferencial do modelo proposto está no fato de considerar todos os tipos de estrutura de dados. A descoberta de conhecimento obtida a partir de dados estruturados é um processo estabelecido na comunidade científica e tecnológica. Entretanto, para o caso de dados que apresentam pouca ou nenhuma estrutura, esse processo de descoberta ainda está em desenvolvimento.

Com o desenvolvimento e a aplicação do modelo proposto, foi possível evidenciar que esforços empreendidos na compreensão antecipada dos dados podem ocasionar redução da complexidade dos dados extraídos e tornar o *Big Data* viável para uso na indústria.

Outro diferencial do modelo proposto é o emprego das tarefas de pré-limpeza, pré-filtragem e pré-transformação para realizar a reestruturação ou a transformação do conjunto de dados. A execução dessas tarefas no conjunto de dados implica o tratamento sistemático das características de volume, velocidade, variedade e qualidade.

Em relação à originalidade do trabalho, a mesma é reiterada pela sistematização das fases contidas no modelo proposto para descoberta de

conhecimento. Ao longo do desenvolvimento dessas fases e da aplicação desse modelo, chegou-se a algumas constatações:

- A avaliação de ambientes digitais com elevado volume de dados, que são alimentados continuamente por diferentes fontes de dados, impõe ampla visão na compreensão dos benefícios que os dados originados desses ambientes podem oferecer para o processo de descoberta de conhecimento.
- O uso do modelo proposto mostrou-se adequado na aproximação dos colaboradores da indústria ao processo de descoberta de conhecimento. Essa constatação ocorre principalmente na Fase II, cuja finalidade é avaliar e compreender o conjunto de dados e sua respectiva fonte.
- A utilização do modelo proposto apóia a construção de um roteiro específico para cada objetivo traçado no processo de descoberta de conhecimento e, assim, oportuniza a criação de ambientes tecnológicos, por meio da utilização e/ou desenvolvimento de ferramentas.
- Embora tenha sido destinado especificamente a atender às atividades do projeto informacional, o modelo proposto apresenta condições para auxiliar outras áreas.

O desenvolvimento do modelo proposto neste estudo reforça a ideia de que, embora as soluções tradicionais e *Big Data* disponham de ferramentas para realizar a análise e dar suporte ao tratamento de dados, o fator humano é imprescindível para sua manipulação. Outro entendimento é que o valor dos dados não está vinculado ao seu volume, portanto, o modelo proposto emprega tarefas que dissociam o volume do valor dos dados, isto é, a possibilidade de redução no volume sem prejudicar o seu valor no processo de descoberta de conhecimento.

A principal contribuição acadêmica desta tese é a construção de um modelo conceitual para conduzir o processo de descoberta de conhecimento e colaborar com o PDP. Esse processo foi idealizado por meio da combinação teórica entre os métodos tradicionais e o *Big Data*.

Uma segunda contribuição acadêmica é a criação de um arcabouço teórico referente aos dados e suas respectivas fontes, no qual ficou estabelecido: a análise de faceta, as avaliações das características dos dados e as atividades que avaliam e modificam estrutura de conjunto de dados.

Além das possibilidades de se aplicar o modelo proposto em outros temas de pesquisa, a terceira contribuição incide sobre as discussões geradas na área acadêmica em relação aos métodos tradicionais e *Big Data*.

Em relação às contribuições empresariais ou práticas, especificamente na indústria, o modelo proposto auxilia os responsáveis pelo PDP, orienta os profissionais responsáveis pela execução do processo de descoberta de conhecimento e colabora com o *know-how* para a produção.

Uma segunda contribuição empresarial ou prática do modelo proposto é aproximar os colaboradores do processo de descoberta de conhecimento. Uma terceira contribuição é proporcionar condições para gerar diferentes perspectivas na cultura estratégica no PDP, por meio da utilização dos conhecimentos extraídos com o auxílio do modelo proposto.

Por fim, a utilização do modelo proposto fornece subsídios que promove apoio no direcionamento do desenvolvimento de ferramentas tecnológicas, essenciais na agilidade da extração do conhecimento.

## **6.1. SUGESTÕES PARA TRABALHOS FUTUROS**

Uma das recomendações para a continuidade desta pesquisa é a análise das incertezas e do interesse do setor industrial pela implantação do processo de descoberta de conhecimento. Isso vale também para as indústrias de pequeno porte, uma vez que, nesta tese, foram destacadas ferramentas tecnológicas gratuitas para a utilização e/ou desenvolvimento de outras ferramentas que contribuam, do início ao fim, para a aplicação do modelo proposto. Dessa forma, o maior investimento será concentrado nos profissionais.

Para outros trabalhos futuros, apontam-se ainda algumas sugestões ou possibilidades:

- A criação de ambientes compostos por ferramentas tecnológicas, com interface amigável e baseadas no modelo proposto, os quais possam ser utilizados pelos colaboradores da indústria. O diretor da INDÚSTRIA AM demonstrou interesse e possibilidade real de criação de um projeto de desenvolvimento para esse ambiente.
- Acompanhamento dos resultados proporcionados pela aplicação do modelo proposto na indústria, por um período de tempo mais longo.
- Desenvolvimento de uma ferramenta de análise e de comparação entre os resultados gerados pelo modelo proposto e os conhecimentos anteriormente extraídos e armazenados no BD conhecimento, de forma a gerar previsões de tendências de moda.
- Aplicação em outros tipos de segmentos industriais para verificar a flexibilidade do modelo proposto.
- Utilização e adaptação do modelo proposto como ferramenta de apoio e aperfeiçoamento do BD empregado no sistema CRM, ou seja, para extrair os conhecimentos individuais dos clientes, contidos nas diversas fontes de dados externas, e armazená-los nesse BD.
- Elaboração e desenvolvimento de um modelo de BD específico para o modelo proposto que tenha alto desempenho nas operações de álgebra booleana e relacional, além de suportar diferentes estruturas e volumes de dados;
- Adaptação do modelo proposto para apoiar outras fases do modelo de referência para o PDP, a exemplo da extração de conhecimentos referentes ao lançamento e ao acompanhamento de produtos.

## REFERÊNCIAS

ADAM, B.; SMITH, I. F.. Reinforcement learning for structural control. **Journal of Computing in Civil Engineering**, v. 22, n. 2, p. 133-139, 2008.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A.. Mining association rules between sets of items in large databases. In: **Acm sigmod record**. ACM, v. 22, n. 2, p. 207-216. 1993.

AMARAL, M. S.; DE PINHO, J. A. G.. Ideologias partidárias em 140 caracteres: uso do Twitter pelos parlamentares brasileiros. **Revista de Administração Pública**, v. 51, n. 5 , 2017.

ASAMOAHA, D.; SHARDA, R.. Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare. In: **Twenty-first Americas Conference on Information Systems**, Puerto Rico, 2015.

BACK, N.; OGLIARI, A.; DIAS, A.; DA SILVA, J. C.. Projeto integrado de produtos: planejamento, concepção e modelagem. **Barueri – SP, Malone**, 2008.

BANSAL, S. K.; KAGEMANN, S.. Integrating Big Data: A Semantic Extract-Transform-Load Framework. **IEEE Computer**, v. 48, n. 3, p. 42-50, 2015.

BATINI, C.; RULA, A.; SCANNAPIECO, M.; VISCUSI, G.. From Data Quality to Big Data Quality. **Journal of Database Management (JDM)**, v. 26, n. 1, p. 60-82, 2015.

BEGOLI, E.; HOREY, J.. Design principles for effective knowledge discovery from big data. In: **Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 joint working IEEE/IFIP conference on**. IEEE, 2012. p. 215-218.

BENGFORT, B.; KIM, J.. **Data Analytics with Hadoop: An Introduction for Data Scientists**. O'Reilly Media, Inc., 2016.

BERRY, M. J.; LINOFF, G.. **Data mining techniques: for marketing, sales, and customer support**. NY, USA, John Wiley & Sons, Inc., 1997.

BHINGE, R.; SRINIVASAN, A.; ROBINSON, S.; DORNFELD, D.. Data-intensive Life Cycle Assessment (DILCA) for Deteriorating Products. **Procedia CIRP**, v. 29, p. 396-401, 2015.

BOTTA, A.; DONATO, W.; PERSICO, V.; PESCAPÉ, A.. Integration of cloud computing and internet of things: a survey. **Future Generation Computer Systems**, v. 56, p. 684-700, 2016.

BRACHMAN, R.; KHABAZA, T.; KLOESGEN, W.; PIATETSKY-SHAPIRO, G.; SIMOUDIS, E.. Mining business databases. **Communications of the ACM**, v. 39, n. 11, p. 42-48, 1996.

BRINK, H.; RICHARDS, J.; FETHEROLF, M.. **Real-world machine learning**. Manning Publications Co, 2016.

BUTLER, B.. **CLOUD CHRONICLES- 5 Problems with big data**. 2015. Disponível em: <<http://www.networkworld.com/article/2973963/big-data-business-intelligence/5-problems-with-big-data.html>>. Acesso em 23 maio 2015.

CABENA, P.; STADLER, H.; ZANASI, V.. **Discovering data mining: from concept to implementation**. Upper Saddle River, NJ, USA, Prentice-Hall, Inc., 1998.

CARR, J.; DECRETON, L.; QIN, W.; ROJAS, B.; ROSSOCHACKI, T.; YANG, Y. W.. Social media in product development. **Food Quality and Preference**, v. 40, p. 354-364, 2015.

CASS, S. **The 2016 Top Programming Languages**. IEEE Spectrum, 2016. Disponível em: <<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>>. Acesso em 01 nov. 2016.

CERVO, A. L. BERVIAN, P. A.; DA SILVA, R.. **Metodologia científica**. 6 ed. São Paulo - SP: Pearson Prentice Hall, 2007.

CHANDRA, D. G.. BASE analysis of NoSQL database. **Future Generation Computer Systems**, v. 52, p. 13-21, 2015.

CHANG, W.; PARK, J. E.; CHAIY, S.. How does CRM technology transform into organizational performance? A mediating role of marketing capability. **Journal of Business Research**, v. 63, n. 8, p. 849-855, 2010.

CHAPMAN, P.; CLINTON, J.; KERBER R.; KHABAZA, T.; REINARTZ T.; SHEARER C.; WIRTH, R.. **CRISP-DM 1.0 Step-by-step data mining guide**. The CRISP-DM Consortium, 2000.

CHERNOFF, H.. The use of faces to represent points in k-dimensional space graphically. **Journal of the American Statistical Association**, v. 68, n. 342, p. 361-368, 1973.

CIMDATA. **Product Lifecycle Management and the Data Deluge. PLM and the Data Deluge Transforming Data to Enhance Performance**. 2012. Disponível em: <[www.cimdata.com](http://www.cimdata.com)>. Acesso em: 03.05.2015.

CORALLO, A.; LATINO, M. E.; LAZOI, M.; LETTERA, S.; MARRA, M.; VERARDI, S.. Defining product lifecycle management: A journey across features, definitions, and concepts. **ISRN Industrial Engineering**, v. 2013, 2013.

CUI, X.; ZHU, P.; YANG, X.; LI, K.; JI, C.. Optimized big data K-means clustering using MapReduce. **The Journal of Super computing**, v. 70, n. 3, p. 1249-1259, 2014.

DAVENPORT, T. H.. **Process Innovation: reengineering work through information technology**. Harvard Business Review Press, 1993.

DAVENPORT, T. H.. **Big Data no trabalho: derrubando mitos e descobrindo oportunidades**. São Paulo, Campus, 2014.

DE SORDI, J. O.. **Gestão de Processos: uma abordagem da moderna administração**. 2.ed., São Paulo, Saraiva, 2008.



DEMCHENKO, Y.; GROSSO, P.; DE LAAT, C.; MEMBREY, P.. Addressing big data issues in scientific data infrastructure. In: **Collaboration Technologies and Systems (CTS), 2013 International Conference on**. IEEE, p. 48-55. 2013.

DERVOJEDA, K.; VERZIJL, D.; NAGTEGAAL, F.; LENGTON, M.; ROUWMAAT E.. Big data: Artificial intelligence. **Business Innovation Observatory**, Netherlands. European Union. Case study 9. p. 1-15, 2013.

FAN, W.; GORDON, M. D.. The power of social media analytics. **Communications of the ACM**, v. 57, n. 6, p. 74-81, 2014.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996a.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996b.

Feldens, M. A.; Moraes, R.; Pavan, A.; Castilho, J.. Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization. In: Conferencia Latino americana de Informatica (CLEI' 98), XXIV1998, Quito, Ecuador. **Proceedings...** [S.l.]: PUCE-XEROX, 1998. v.2, p.935-947.

FEINERE I.; HORNIK, K.. "tm": Text Mining Package. R package. **Version 0.7-1**, 2017. Disponível em: <<https://cran.r-project.org/web/packages/tm/index.html>>. Acesso em 10 mar. 2017.

FOX, J.; LEANAGE, A.. R and the Journal of Statistical Software. **Journal of Statistical Software**, v. 73, n. 2, p. 1-13, 2016.

GANTZ, J.; REINSEL, D.. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. **IDC iView: IDC Analyze the future**, v. 2007, p. 1-16, 2012.

GENTRY, J.; LANG, D.. R Interface For OAuth. **Version 0.9.6**, 2015. Disponível em: <<https://cran.r-project.org/web/packages/ROAuth/index.html>> Acesso em 10 jan. 2017.

GENTRY, J.. R Based Twitter Client. **Version 1.1.9**, 2016. Disponível em: <<https://cran.r-project.org/web/packages/twitteR/twitteR>>. Acesso em 14 jan. 2017.

GIL, A. C.. **Como elaborar projetos de pesquisa**. 4ed. São Paulo: Atlas, 2009.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E.. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro, Elsevier, 2015.

GUO, S. S.; YUAN, Z. M.; SUN, A. B.; YUE Q.. A New ETL Approach Based on Data Virtualization. **Journal of Computer Science and Technology**, v. 30, n. 2, p. 311-323, 2015.

HALVORSEN, K.; HOFFMANN, J.; COSTE-MANIÈRE, I.; STANKEVICIUTE, R.. Can fashion blogs function as a marketing tool to influence consumer behavior? Evidence from Norway. **Journal of Global Fashion Marketing**, v. 4, n. 3, p. 211-224, 2013.

HAND, D. J.; MANNILA, H.; SMYTH, P.. **Principles of data mining**. Cambridge, MA, USA, MIT press, 2001.

HASHEM, I. A. T.; YAQOOB, I.; ANUAR, N. B.; MOKHTAR, S.; GANI, A.; KHAN, S. U.. The rise of “big data” on cloud computing: Review and open research issues. **Information Systems**, v. 47, p. 98-115, 2015.

HENDERSON, S.; SEGAL, E. H.. Visualizing qualitative data in evaluation research. **New Directions for Evaluation**, v. 2013, n. 139, p. 53-71, 2013.

HENDLER, J.. Data integration for heterogenous datasets. **Big data**, v. 2, n. 4, p. 205-215, 2014.

IBM - **Bringing big data to the enterprise**, 2011. Disponível em: < <https://www-01.ibm.com/software/sg/data/bigdata/enterprise.html> > Acesso em: 05.04.2015.

JI, C.; LI, Y.; QIU, W.; JIN, Y.; XU, Y.; AWADA, U.; QU, W.. Big data processing: Big challenges and opportunities. **Journal of Interconnection Networks**, v. 13, n. 03 e 04, p. 1250009, 2012.

JOUILI, S.; VANSTEENBERGHE, V.. An empirical comparison of graph databases. In: **Social Computing (SocialCom)**, 2013 International Conference. IEEE, 2013. p. 708-715.

JUN, S. P.; PARK, D. H.; YEOM, J.. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. **Technological Forecasting and Social Change**, v. 86, p. 237-253, 2014.

KAISLER, S.; ARMOUR, F.; ESPINOSA, J. A.; MONEY, W.. Big data: issues and challenges moving forward. In: **System Sciences (HICSS)**, 2013 46th Hawaii International Conference on. IEEE, 2013. p. 995-1004.

KAPLAN, A. M.; HAENLEIN, M.. Users of the world, unite! The challenges and opportunities of Social Media. **Business Horizons**, v. 53, n. 1, p. 59-68, 2010.

KARUNARATNE, P.; KARUNASEKERA, S.; HARWOOD, A.. Distributed stream clustering using micro-clusters on Apache Storm. **Journal of Parallel and Distributed Computing**, v. 108, p. 74-84, 2017.

KEIM, D. A.; KRIEGER, H. P. Visualization techniques for mining large databases: A comparison. **IEEE Transactions on knowledge and data engineering**, v. 8, n. 6, p. 923-938, 1996.

KHAN, N.; YAQOOB, I.; HASHEM, I. A. T.; INAYAT, Z.; MAHMOUD A. W. K.; ALAM, M.; GANI, A.. Big Data: survey, technologies, opportunities, and challenges. **The Scientific World Journal. Hindawi Publishing Corporation**. v. 2014, p. 1-18, 2014.

KITCHENHAM, B.; CHARTES. S.. Guidelines for performing systematic literature reviews in software engineering. **Technical report, EBSE-2007-01, School of Computer Science and Mathematics**. Keele University, 2007.

KOTLER, P.; KELLER, K.. **Administração de Marketing**, 14 ed., São Paulo, Pearson Prentice Hall, 2012.

KRISHNAN, K.. **Data warehousing in the age of big data**. Newnes, Ringgold Inc., Reference and research Book News, v. 28, 2013.

KUMAR, A.; NIU, F.; RÉ, C.. Hazy: making it easier to build and maintain big-data analytics. **Communications of the ACM**, v. 56, n. 3, p. 40-49, 2013.

LA BARRE, K.; Facet analysis. **Annual Review of Information Science and Technology**, v. 44, n. 1, p. 243-284, 2010.

LACHMAYER, R., MOZGOVA, I., SAUTHOFF, B., & GOTTWALD, P.. Evolutionary Approach for an Optimized Analysis of Product Life Cycle Data. **Procedia Technology**, v. 15, p. 359-368, 2014.

LAHUERTA-OTERO, E.; CORDERO-GUTIÉRREZ, R.. Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. **Computers in Human Behavior**, v. 64, p. 575-583, 2016.

LAROSE, D. T. **Discovering knowledge in data: an introduction to data mining**. New Jersey, John Wiley & Sons, 2014.

LEE, Y. W.; STRONG, D. M.; KAHN, B. K.; WANG, R. Y.. AIMQ: a methodology for information quality assessment. **Information & management**, v. 40, n. 2, p. 133-146, 2002.

LEGLER, R.; EPPLER, M. J. Towards a periodic table of visualization methods for management. In: **IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007)**, Clearwater, Florida, USA. 2007.

LEUNG, C. K. S.; MACKINNON, R. K.; JIANG, F.. Reducing the search space for big data mining for interesting patterns from uncertain data. In: **2014 IEEE International Congress on Big Data**. IEEE, p. 315-322, 2014.

LEVY, Y.; ELLIS, T. J. A.. systems approach to conduct an effective literature review in support of information systems research. **Informing Science: International Journal of an Emerging Transdiscipline**, v. 9, n. 1, p. 181-212, 2006.

Li, J.; Tao, F.; Cheng, Y.; Zhao, L.. Big Data in product lifecycle management. **The International Journal of Advanced Manufacturing Technology**, v. 81, n. 1-4, p. 667-684, 2015.

LIMA, G. A. B. O.. A análise facetada na modelagem conceitual para organização hipertextual de documentos acadêmicos: sua aplicação no prototipo MHTX (mapa hipertextual). **Informação & Sociedade**, v. 17, n. 1, 2007.

LOURENÇO, J. R.; CABRAL, B.; CARREIRO, P.; VIEIRA, M.; BERNARDINO, J.. Choosing the right NoSQL database for the job: a quality attribute evaluation. **Journal of Big Data**, v. 2, n. 1, p. 1-26, 2015.

LUO, D.; DING, C.; HUANG, H.. Parallelization with multiplicative algorithms for big data mining. In: **Data Mining (ICDM), 2012 IEEE 12th International Conference on**. IEEE. p. 489-498, 2012.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H.. Big data: The next frontier for innovation, competition, and productivity. Global Institute, **McKinsey & Company**, 2011.

MARCONI, M. A.; LAKATOS, E. M.. **Técnicas de pesquisa: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados**. 7 ed. São Paulo: Atlas, 2010.

MARTIN, J.. Data, Data Everywhere. **Mechanical Engineering**, v. 137, n. 7, p. 46, 2015.

MATTMANN, C. A.; WALISER, D.; KIM, J.; GOODALE, C.; HART, A.; RAMIREZ, P.; LOIKITH, P.. Cloud computing and virtualization within the regional climate model and evaluation system. **Earth Science Informatics**, v. 7, n. 1, p. 1-12, 2014.

MAYER-SCHONBERGER, V.; CUKIER, K.. **Big Data, Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana**. Elsevier, 2013.

MCAFEE, A.; BRYNJOLFSSON, E.. Big data: the management revolution. **Harvard Business Review**, v. 90, n. 10, p. 61-67, 2012.

MERINO, J.; CABALLERO, I.; RIVAS, B.; SERRANO, M.; PIATTINI, M.. A data quality in use model for big data. **Future Generation Computer Systems**, v. 63, p. 123-130, 2016.

MIGUEL, P. A. C.. **Estudo de caso na engenharia de produção: estrutura e recomendações para sua condução**. São Paulo, Produção, V. 17, 216-229. 2007.

MIGUEL, P. A. C.; FLEURY A.; MELLO H. P. C.; NAKANO N. D.; LIMA P. E.; TURRIONI B. J.; HO L. L.; MORABITO R.; MARTINS A. R.; SOUSA R.; COSTA E. G. S.; PUREZA V.. Metodologia de pesquisa em engenharia de produção e gestão de operações. **Editora Elsevier – Campus**, Rio de Janeiro, 2 ed. 2011.

MILONAS, E.. Wittgenstein and web facets. **NASKO**, v. 3, n. 1, p. 33-40, 2011.

MISHRA, N.; LIN, C. C.; CHANG, H. T.. A cognitive adopted framework for IoT big-data management and knowledge discovery prospective. **International Journal of Distributed Sensor Networks**, v. 11, p. 1-12, 2015.

MITCHELL, T. M. Machine learning and data mining. **Communications of the ACM**, v. 42, n. 11, p. 30-36, 1999.

NASCIMENTO, H.; FERREIRA, C.. Visualização de Informações – uma abordagem prática. In: **XXV Congresso da Sociedade Brasileira de Computação. XXIV JAI**. UNISINOS, S. Leopoldo–RS. p. 1262-1312, 2005.

NETTO, C. M.; LIMA, G. A. B. O.; JÚNIOR, I. P.. An Application of Facet Analysis Theory and Concept Maps for Faceted Search in a Domain Ontology: Preliminary Studies. **Knowledge Organization**, v. 43, n. 4, p. 254-264, 2016.

OLSON, J. E.. **Data Quality: the accuracy dimension**. Morgan Kaufmann, San Francisco - EUA, 2003.

OOMS J.. Fast and Simple 'MongoDB' Client for R. R Package **Version 1.1** disponível em: <<https://cran.r-project.org/web/packages/mongolite/index.html>> Acesso em 12 abril 2017.

ORR, K.. Data quality and systems theory. **Communications of the ACM**, v. 41, n. 2, p. 66-71, 1998.

PIATETSKY G.. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. **KDD News** 2014. Disponível em: <<http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>>. Acesso em: 06 set. 2015.

PIPINO, L. L.; LEE, Y; W.; WANG, R. Y.. Data quality assessment. **Communications of the ACM**, v. 45, n. 4, p. 211-218, 2002.

POKORNÝ, J.. Database technologies in the world of big data. In: **Proceedings of the 16th International Conference on Computer Systems and Technologies**, ACM, p. 1-12, 2015.

QIU, J.; WU, Q.; DING, G.; XU, Y.; FENG, S.. A survey of machine learning for big data processing. **EURASIP Journal on Advances in Signal Processing**, v. 2016, n. 1, p. 1-16, 2016.

Rabelo, E.; Dias, M. M.; Franco, C.; Pacheco, R. C.. Information Visualization: Which Is the Most Appropriate Technique to Represent Data Mining Results?.

In: **Computational Intelligence for Modelling Control & Automation, 2008 International Conference on**. IEEE, p. 1228-1233, 2008.

RABELO, E.; CAMPOS, C. F.. Big Data e KDD: Novas Descobertas. In: **Enegep, XXXIV Encontro Nacional de Engenharia de Produção**, 2014.

RABELO, E.; CAMPOS, F. C. Descoberta de Conhecimento na Gestão do Conhecimento: Uma análise bibliométrica. **Espacios**. vol. 37, n. 12, p. 1-10, 2016.

ROZENFELD H.; FORCELLINI A. F.; AMARAL C. D.; TOLEDO C. J.; SILVA L. S.; ALLIPRANDINI H. D.; SCALICE K. R.. **Gestão de Desenvolvimento de Produto uma Referencia para a Melhoria do Processo**. São Paulo, Saraiva, 2006.

SAAKSVUORI, A.; IMMONEN, A.. **Product lifecycle management**. , Springer, Berlin, 2004.

SALAKI, R. J.; WAWORUNTU, J.; TANGKAWAROW, I. R. H. T. Extract transformation loading from OLTP to OLAP data using pentaho data integration. In: **IOP Conference Series: Materials Science and Engineering**. IOP Publishing, Vol. 12, n.1, p.1-8, 2016.

SAPOUNTZI, A.; PSANNIS, K. E.. Social networking data analysis tools & challenges. **Future Generation Computer Systems**, Elsevier, p.1-21, 2016

SARNOVSKY, M.; BUTKA, P.; HUZVAROVA, A.. Twitter data analysis and visualizations using the R language on top of the Hadoop platform. In: **Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 15th International Symposium on**. IEEE, p. 327-332, 2017.

SHALEV-SHWARTZ, S.. Online learning and online convex optimization. **Foundations and Trends in Machine Learning**, v. 4, n. 2, p. 107-194, 2012.

SHIRI, A.. Making Sense of Big Data: A Facet Analysis Approach. **Knowledge Organization**, v. 41, n. 5, p. 357- 368, 2014.



SOARES, S.. **IBM InfoSphere: A Platform for Big Data Governance and Process Data Governance**. USA, MC Press, 2013.

TOSHNIWAL, A.; TANEJA, S.; SHUKLA, A.; RAMASAMY, K.; PATEL, J. M.; KULKARNI, S.; BHAGAT, N.. **Storm@ twitter**. In: **Proceedings of the 2014 ACM SIGMOD - International Conference on Management of Data**. ACM, p. 147-156, 2014.

TRIPATHY, A.; RAINA, S.; MASCARENHAS, R.; PANGOTRA, S.; RODRIGUES, R.. Extracting new product ideas from consumer blogs. In: **Communication, Information & Computing Technology (ICCICT), 2012 International Conference on**. IEEE, p. 1-6, 2012.

TURRIONI, J. B.; MELLO, C. H. P. Metodologia de pesquisa em engenharia de produção. **Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Itajubá. UNIFEI**, p. 191, 2012.

TURCK, M.; HAO, J.. **Is Big Data Still a Thing? (The 2016 Big Data Landscape** (version 2.0). Disponível em: <<http://mattturck.com/2016/02/01/big-data-landscape/>>. acesso em: 04 maio 2016.

TYAGI, A. K.; SREENATH N.. Mining Big Data to Predicting Future. **Journal of Engineering Research and Applications**. v. 5, n. 3 (Part -2), p.14-21, 2015.

TWITTER. (2017). **About the company**. Disponível em: <https://about.twitter.com/pt/company>. Acesso em: 14 jan. 2017.

URBANEK S.. R/Cassandra interface. R Package **Version 0.1-3**. Disponível em: <<https://cran.r-project.org/web/packages/RCassandra/RCassandra.pdf>> Acesso em 20/02/2017.

VIEIRA, D.; DEBAECKER, D.; BOURAS, A.. **Gestão de Projeto de Produto: Baseada na Metodologia Product Lifecycle Management (PLM)**. Rio de Janeiro, Elsevier, 2013.

WALLER, M. A.; FAWCETT, S. 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, R. Y.. A product perspective on total data quality management. **Communications of the ACM**, v. 41, n. 2, p. 58-65, 1998.

WANG, H.; LI, M.; BU, Y.; LI, J.; GAO, H.; ZHANG, J.. Cleanix: a Parallel Big Data Cleaning System. **ACM SIGMOD Record**, v. 44, n. 4, p. 35-40, 2016.

WATTENBERG, M.; VIÉGAS, F. B.. The word tree, an interactive visual concordance. **IEEE transactions on visualization and computer graphics**, v. 14, n. 6, p. 1221-1228, 2008.

WAZLAWICK, R. S.. Uma Reflexão sobre a Pesquisa em Ciência da Computação à Luz da Classificação das Ciências e do Método Científico. **Revista de Sistemas de Informação da FSMA**, v. 6, p. 3–10, 2010.

WILLAERT, P.; VAN DEN BERGH, J.; WILLEMS, J.; DESCHOOLMEESTER, D.. The process-oriented organisation: a holistic view developing a framework for business process orientation maturity. **Business Process Management**. Springer Berlin Heidelberg, p. 1-15, 2007.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H. ; ZHOU, Z. H.. Top 10 algorithms in data mining. **Knowledge and information systems**, v. 14, n. 1, p. 1-37, 2008.

WU, X.; ZHU, X.; WU, G. Q.; DING, W.. Data mining with big data. **IEEE transactions on knowledge and data engineering**, v. 26, n. 1, p. 97-107, 2014.

ZANCUL, S. E.. Gestão do Ciclo de Vida de Produtos: seleção de sistemas PLM com base em modelos de referencia -Tese Doutorado. **Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos**, 2009.

ZHANG, Y.; REN, S.; LIU, Y.; SI, S.. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. **Journal of Cleaner Production**, v. 142, p. 626-641, 2017.

ZHONG, R. Y.; LAN, S.; XU, C.; DAI, Q.; HUANG, G. Q.. Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing. **The International Journal of Advanced Manufacturing Technology**, v. 84, n. 1-4, p. 5-16, 2016.

ZHOU, Z. H.. Three perspectives of data mining. **Artificial intelligence**. Essex, UK, v. 143, n. 1, p. 139–146, 2003.

ZHUANG, Y.; WANG, Y.; SHAO, J.; CHEN, L.; LU, W.; SUN, J.; WU, J.. D-Ocean: an unstructured data management system for data ocean environment. **Frontiers of Computer Science**, v. 10, n. 2, p. 353-369, 2016.

ZIKOPOULOS, P.; EATON, C.. **Understanding big data: Analytics for enterprise class hadoop and streaming data**. McGraw-Hill Osborne Media, USA, 2011.

## APÊNCIDE A – QUESTIONÁRIO DE DIAGNÓSTICO GERAL

### **OBJETIVO:**

O presente questionário tem como objetivo, levantar dados da empresa para propor um modelo para descoberta de conhecimento no apoio do processo de desenvolvimento do produto, na etapa do projeto informacional.

-----

Dados do Entrevistado

Nome \_\_\_\_\_

Empresa \_\_\_\_\_

Cargo/Função \_\_\_\_\_

Data \_\_\_\_/\_\_\_\_/\_\_\_\_

Visão Geral e Reconhecimento da Empresa
---

1.Data de criação da empresa: \_\_\_\_\_

2. Quantitativo Físico

Quantidade de lojas( ) escritório(s) ( ) Centro de Distribuição ( )

3.Quais as parcerias que a empresa mantém atualmente?

\_\_\_\_\_

\_\_\_\_\_

4.Quantidade de facções de roupas?

\_\_\_\_\_

\_\_\_\_\_

5.Quais Sistemas de Informações a empresa Utiliza?

\_\_\_\_\_

\_\_\_\_\_

6. A Empresa utiliza algum método que utilize tratamento de dados presenciais em lojas?

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

7. Tem algum painel comparativo (dashboard) de análise de ranqueamento das lojas, ou produtos, ou algo afim a isso? Que ferramenta de TI utiliza nesse caso?

---

---

**Produtos – características e desenvolvimentos**

8.A empresa já realizou, ou realiza algum método para tomada de decisão na produção de novos produtos? Caso Afirmativo, como é realizado? Se negativo, como ocorrem às atividades atreladas à decisão no desenvolvimento de novos produtos?

---

---

---

9.As tomadas de decisão sobre a fabricação de novos produtos são na maioria assertivas? Caso positivo, saberia dizer qual o maior motivo que se deve a esse fato? Caso negativo, qual a frequência que ocorrem as decisões incorretas e teria algum motivo, o prejuízo acarreta algum impacto significativo a empresa?

---

---

---

---

10. Como faz a análise de tendências de mercado? Utiliza alguma ferramenta? Algum método?

---

---

**Produtos – marketing, vendas e relações com clientes**

11.De que forma é realizado o marketing da empresa?

---

---

12.Existe algum acompanhamento da imagem dos produtos nas redes sociais?

---

---

13. Existe algum acompanhamento e controle sobre as reclamações dos clientes?

---

---

**14.**Qual(is) a(s) maior(es) dificuldade(s) na(s) vendas on-line ?

---

---

**15.** A empresa desenvolve algum programa de fidelidade?

---

---

**16.** A empresa tem algum sistema de pesquisa/enquetes sobre seus produtos?

---

---

Produtos – processo de formação e implementação de estratégias
--

**17.**Empresa possui um processo de comunicação e motivação para que todos os colaboradores participem do processo de formação e implementação de estratégias?

---

---

**18.** Pretende diversificar o ramo de atividade da empresa?

---

---

**19.** Pretende produzir outros produtos?

---

---

**20.** Realiza alguma espécie de benchmarking via internet? Utiliza algum site específico? Já utilizou alguma ferramenta dedicada com recursos *Big Data*?

---

---

## APÊNDICE B – FUNÇÃO INICIAR

```
iniciar<- function()  
{  
  library(ROAuth)  
  chave_aplicacao<- "xxxxxxxxxxxxxxxxxxxxxxxxxxxx"  
  chave_secreta<- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"  
  token<- "xxxxxxxxxxxxxxxx-xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"  
  token_chave_secreta<- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"  
  setup_twitter_oauth(chave_aplicacao, chave_secreta, token, token_chave_secreta)  
}
```

## APÊNDICE C – FUNÇÕES LISTAR TERMO E LISTAR USUÁRIO

### Função Listar termo

```

listartermo = function (quantidade,datainicio,idioma)
{
  tryCatch(remove(m,i,s,n,lista,r1,r2,x,v), error = function(e){warning('desconsiderar
erro de limpeza das variaveis'); NA})
  i<-1
  print('digite a busca: ')
  s <- readLines(n = 1)
  lista<- s
  while (s != "0")
  {
    print('digite a busca: ')
    s <- readLines(n = 1)
    if (s != "0")
    {
      Lista<-c(lista,s)
      i<-i+1
    }
  }
  m<-lista
  for(a in 1:i)
  {
    if (a==1)
    {
      r1<- searchTwitter(m[a], quantidade, since=datainicio, lang=idioma)
    }
    else
    {
      r2<- searchTwitter(m[a], quantidade, since=datainicio, lang=idioma)
      r1<- c(r1,r2)
    }
  }
  return(r1)
}

```



**Função listar usuário**

```

listarusuario = function (quantidade)
{
  i<-1
  s <- readLines(n = 1)
  lista<-s
  while (s != "0")
  {
    print('digite a busca: ')
    s <- readLines(n = 1)
    if (s != "0")
    {
      lista<-c(lista,s)
      i<-i+1
    }
  }
  m<-lista
  for(a in 1:i)
  {
    if (a==1)
    {
      r1<- userTimeline(m[a],quantidade)
      print('buscar do termo:')
      print(m[a])
    }
    else
    {
      r2<- userTimeline(m[a],quantidade)
      r1<- c(r1,r2)
    }
  }
  return(r1)
}

```

## APÊNDICE D – PRÉ-FILTRAGEM, PRÉ-LIMPEZA E PRÉ-TRANSFORMAÇÃO

### **Funções:**

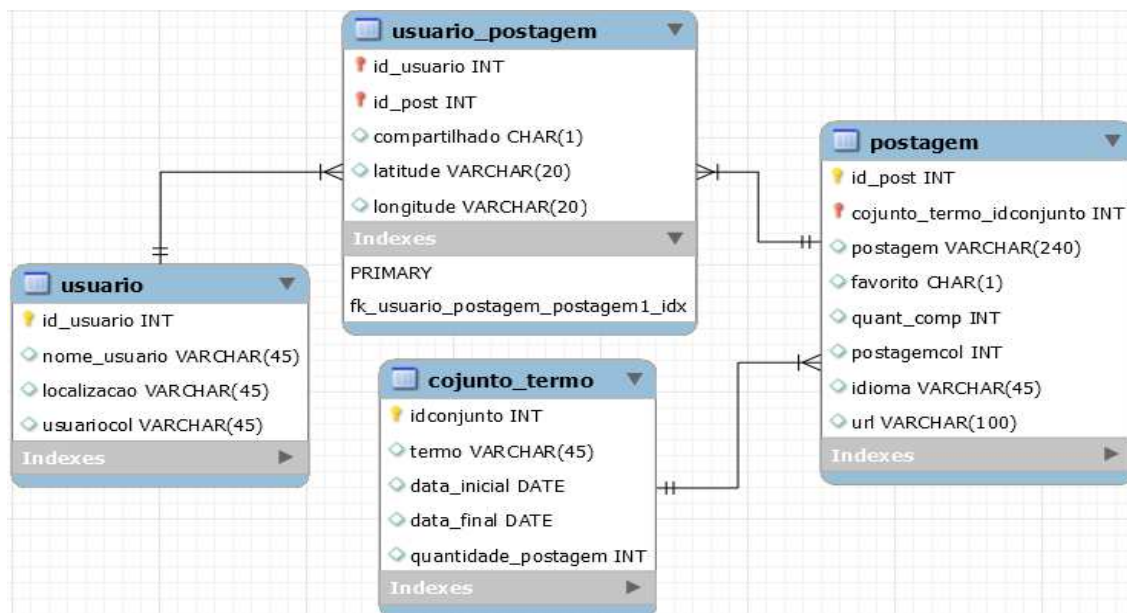
```
>library(twitteR)
>listartermos("1000000","2017-02-09")
+ #nyfw# termo para busca
+ #nyfw2017 # termo para busca
+ 0          # finaliza busca

> postagens <-sapply(t, function(x) x$getText()) # filtrar apenas as postagens
>library(tm) #
> corpus<-Corpus(VectorSource(postagens)) # alterar estrutura
> corpus<- tm_map(corpus, content_transformer(stringi::stri_trans_tolower))
#modificação das postagens para caixa baixa
> remover <- function(x) gsub("http[^\s:]*", "", x) # função para remover URL
> corpus <- tm_map(corpus, content_transformer(rem_url)) # remoção de URL
> corpus <- tm_map(corpus, removePunctuation) # remover pontuações,
> corpus <- tm_map(corpus, removeNumbers) # remover números
```

## APÊNDICE E – PREPARAÇÃO DOS DADOS PARA ANÁLISE.

```
> corpus <- tm_map(corpus, removeWords, c(stopwords("english")
c("nyfw", "fashion", "show", "shows", "new", "york", "awwww", "f", "h", "la", "s", "u", "w", "
x"))) # remoção de termos irrelevantes que foram encontrados no conjunto de dados
> corpus <- tm_map(corpus, function(x) iconv(enc2utf8(x), sub = "byte"))
> corpus <- tm_map(corpus, content_transformer(function(x) iconv( enc2utf8(x),
sub = "bytes")))) # remoção de espaços em brancos.
> corpus <- tm_map(corpus, content_transformer (stringi::stri_trans_tolower))
# padronizar caixa baixa
> corpus <- tm_map(corpus, stripWhitespace) # remoção de espaços em brancos.
> corpus <- tm_map(corpus, stemDocument) # redução dos termos
```

## APÊNDICE F – MODELAGEM DE DADOS ESTRUTURADA PARA REDE SOCIAL



### Descrição dos atributos das tabelas:

Tabela grupo_termo:	
Obs: Armazena o conjunto de dados para cada busca efetuada	
Nome do campo	Descrição
Id_grupo_termo	Identificação do grupo (PK)
termo	Termos utilizados para extração dos dados
data_inicial	Data inicial da extração dos dados
data_final	Data final da extração dos dados
Quant_postagem	Quantidade de postagem do grupo

Tabela usuario:	
Id_usuario	Campo chave para identificação do usuário
nome_usuario	Nome do usuário
Localizacao	Localização do usuário

Tabela postagem	
Id_post	Identificação da postagem (PK - <i>auto increment</i> )
Id_grupo_termo	Grupo a que pertence a postagem (FK)
postagem	Postagem (comentários do usuário)
favorito	Indica se a postagem é favorita
quant_comp	Quantidade de compartilhamento da referida postagem
idioma	Idioma que foi escrito a postagem
url	Endereço de fotos e vídeos ligados a postagem

Tabela usuario_postagem:	
Id_usuario	Identificação do usuário da tabela usuário (FK)
Id_post	Identificação da postagem da tabela postagem (FK)
compartilhado	Identificação se a postagem é original ou compartilhada
latitude	Localização do usuário no momento da postagem
longitude	Localização do usuário no momento da postagem

**APÊNDICE G – CONEXÃO COM *MONGODB* NA LINGUAGEM R.**

```
library(mongolite)
library(jsonlite)
t<- listartermo('100000','2017-02-09','en',)
tweets.df <- do.call("rbind",lapply(t,as.data.frame))
c = mongo(collection = " amarelomanga", db = "nyfw", url =
"mongodb://localhost", verbose = TRUE, options = ssl_options())
c$insert(fromJSON(t))
```

## APÊNDICE H – ANÁLISE DE ASSOCIAÇÃO E AGRUPAMENTO

```

#transformar em estrutura vetorial
>vtd <- TermDocumentMatrix(corpus_tr_ve, control = list(wordLengths = c(1,
Inf)))

>#frequência e associação
>fterms <- findFreqTerms(vtd, lowfreq = xx)
>tfreq <- rowSums(as.matrix(vtd))
>tfreq <- subset(tfreq, tfreq >= xx)
>df <- data.frame(term = names(tfreq), freq = tfreq)

>findAssocs(vtd, "opec", 0.xx)

#agrupamento
>term1 <- removeSparseTerms(vtd, sparse = 0.xx)
>mat1 <- as.matrix(term1)
# cluster terms
>distMatrix <- dist(scale(mat1))

>mat2 <- t(mat1)
>km <- 6 # number of clusters
>kmeansResult <- kmeans(mat2, km)
>round(kmeansResult$centers, digits = 6) # cluster centers

>for (i in 1:k) {
  cat(paste("Agrupamento ", i, ": ", sep = ""))
  rk <- sort(kmeansResult$centers[i, ], decreasing = T)
  cat(names(rk)[1:6], "\n")
}

```

(a) **Dendrograma**



## APÊNDICE J – TÉCNICAS DE VISUALIZAÇÃO

```
#nuvens de palavras
>library(wordcloud)
>mat <- as.matrix(vtd)
>word.freq <- sort(rowSums(mat), decreasing = T)
>pal <- brewer.pal(9, "BuGn")[-(1:4)]
>wordcloud(words = names(word.freq), freq = word.freq, min.freq = xx,
            , random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
```

```
#Associação
>library(graph)
>library(Rgraphviz)
>frequencia_termos <- findFreqTerms(vtd, lowfreq = xx)
#frequencia_termos_intervalo <- findFreqTerms(vtm, lowfreq = xx, highfreq
=xx)
>plot(tdm, term = frequencia_termos, corThreshold = 0.1, weighting = T)
```

### **Visualização Dendograma**

```
> library(cluster)
> library(factoextra)
> tdm <- TermDocumentMatrix(corpus, control = list(wordLengths = c(1, Inf)))
> tdm <- removeSparseTerms(tdm, sparse = 0.95)
> m <- as.matrix(tdm)
> d1 <- na.omit(m)
> d2 <- dist(d1, method = "euclidean")
> r <- hclust(d, method = "ward.D2" )
> plot(r, cex = 0.5, hang = 0)
```



**Visualização Dendograma não enraizado**

```

> library(ape)
> d <- dist(scale(m2), method = "euclidean")
> h1 <- hclust(dd, method = "ward.D2")
> plot(h)
> h2 <- as.dendrogram(h1)
> plot(as.phylo(h2), type = "unrooted", cex = 0.4)

```

**Visualização de matriz de distância**

```

> library("factoextra")
> tdm <- TermDocumentMatrix(corpus_tr_ve, control = list(wordLengths = c(1,
Inf)))
> tdm_aux <- removeSparseTerms(tdm, sparse = 0.xx)
> ma <- as.matrix(tdm_aux)
> distancia_ma <- dist(scale(ma))
> d <- get_dist(distancia_ma, stand = TRUE, method = "pearson")
> viz_dist(d, gradient = list(low = "#000000", mid = "#808080", high =
"#DCDCDC"))

```

**Visualização Word Cloud para cada grupo**

```

> library(wordcloud)
> tdm_aux <- removeSparseTerms(tdm, sparse = 0.xx)
> m_aux1 <- as.matrix(tdm_aux)
> m_aux2 <- t(m_aux1)
> k <- 6
> kmeans_Result <- kmeans(m_aux2, k)
> round(kmeans_Result$centers, digits = k)
> par(mfrow=c(2,3))
> for(j in 1:k){
>   km = corpus_tr[which(kmeans_Result$cluster==j)]

```

```

> t = TermDocumentMatrix(km, control = list(wordLengths = c(3, Inf)))
> m = as.matrix(t)
> frequencia = sort(rowSums(m1), decreasing = T)
> wordcloud(words = names(frequencia), freq = word.freqa, min.freq = 40)) }

```

-----

### ***Função para transferir dados do tipo corpus para um TXT***

```

transferir = function (corpus)
{
  txt <- strwrap(corpus_fun[[1]]$content)
  q1 <- length(corpus_fun)
  for(x in 2:q1)
  {
    temp <- strwrap(corpus_fun[[x]]$content, 500)
    txt <- c(txt, temp)
  }
  return(txt)
}

txt1 <- transferir(corpus_tr)
write.table(txt1, "c:\\temp\\textor\\txt1.txt", row.names = TRUE)

```

## APÊNDICE K – QUESTIONÁRIO DE CONHECIMENTOS EXTRAÍDOS

### Questionário dos conhecimentos extraídos por meio do modelo proposto

Nome: \_\_\_\_\_ Função: \_\_\_\_\_

Empresa: \_\_\_\_\_

#### Ajuda:

O conhecimento é tido como “novo”, no caso em que o respondente não tem noção sobre o fato.

O conhecimento é tido como útil, para casos em que é possível utilizá-lo no desenvolvimento da coleção.

Conhecimentos extraídos das postagens da rede social “*Twitter*” referentes ao desfile de moda denominado **New York Fashion Week 2017**.

Participou deste evento? ( ) sim ( ) não

Obs: \* terceira coluna – apenas responder caso tenha participado do evento.

Conhecimentos - 01		
Dentre as postagens no <i>Twitter</i> , tiveram maiores destaque a model-x, atriz-x e os designers: design-y; designer-x; e designer-w.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
Conhecimento - 02		
Dentre as postagens que citaram a atriz-x, foi identificado maior destaque do relacionamento da atriz com um vestido azul marinho da coleção de designer-y.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
Conhecimento - 03		
Comentários destacam que a atriz model-x também compareceu ao evento com vestido branco e uma capa emparelhada com sobre tudo creme,		

pertencente à coleção designer-x. No entanto, a coleção de designer-y obteve uma quantidade maior de comentários.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 04</b>		
Estilos de rua ( <i>street style</i> ) tiveram um significativo número de postagens		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 05</b>		
Metade das postagens que utilizaram a palavra “favorito” foram associadas às cores colegiais.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 06</b>		
Juntamente a palavra “estação” foi encontra a palavra “yeezi”.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 07</b>		
A palavra “estilo de rua” obteve maior destaque em relação ao designer-w.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 08</b>		
Metade das postagens citando o designer desinger-w menciona a model-y		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 09</b>		
Comentários destacam a coleção de outono de designer-w		

Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
--	--	--

\*responder caso tenha participado do evento.

Conhecimentos extraídos das postagens da rede social “*Twitter*” referentes ao desfile de moda denominado ***Milan Fashion Week 2017***.

Os dados extraído foram das palavras chaves: #MFW; #MFW17; E MILANFASHION.

Participou deste evento? ( ) sim ( ) não

Conhecimento - 01		
Dentre os dados extraídos, tiveram maiores freqüências: as modelos model-g, model-b e model-e; os designers-c, designer-h e designer-b; marcas brand-b e brand-c.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
Conhecimento - 02		
Assim como em NYFW em MFW as palavra “estilo de rua” foram significas na frequencia de vezes que foram citadas.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
Conhecimento - 03		
As palavras referentes a “inverno” e “coleção” estão associadas a um blog de moda alternativo denominado “ <i>theimonation</i> ”. Os proprietários desse blog afirmam que não possui vínculo com nenhuma marca.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
Conhecimento - 04		
As palavras “desfilar” e “novo” estão associadas à modelo sueca chamada model-e.		

Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 05</b>		
A palavra “tendência” está associada a palavras “arco Iris” e “ <i>moeetztal</i> ”, sendo essa última, referente a um aplicativo de celular cujo objetivo é seguir celebridades e blogueiros relacionados à moda.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 06</b>		
Na semana da moda sugere-se que a coleção da marca brand-b obteve destaque.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 07</b>		
Destaque para as modelos model-g e model-b que desfilaram a coleção do designer-c.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 08</b>		
Foram identificados padrões que destacaram novamente a coleção brand-b com o “look” de outono.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não
<b>Conhecimento - 09</b>		
Foram identificados padrões que agruparam a palavra “look” ao designer desingner-h.		
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não	Caso não tivesse participado do evento, este conhecimento seria novo? * ( ) Sim ( ) Não

\*responder caso tenha participado do evento.

Os próximos cinco conhecimentos são derivados das palavras chaves “**estilo de rua**”.

<b>Conhecimento - 01</b>	
A palavra “estilo de rua” faz referência às palavras “ <i>fashionblogger</i> ” e “ <i>blogger</i> ”. Outra palavra que chama atenção é “ <i>doweaststyle</i> ”, que representa uma loja nos Estados Unidos, e mantém um blog para discutir tendências da moda.	
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não
<b>Conhecimento - 02</b>	
Nas postagens relacionadas a “ <i>fashionblogger</i> ”, destacam-se as cores amarelo e branco.	
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não
<b>Conhecimento - 03</b>	
As palavras relacionadas “ <i>doweaststyle</i> ” destacam as palavras “ <i>polka</i> ” (bolinhas), “ <i>plaid</i> ” (xadrez) e “ <i>sweater</i> ” (suéter).	
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não
<b>Conhecimento - 04</b>	
A palavra “ <i>outfit</i> ” (roupa) destacam as palavras “ <i>polka</i> ” (bolinhas); “ <i>plaid</i> ” (xadrez); “ <i>hoodie</i> ” (moletom com capuz).	
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não
<b>Conhecimento - 05</b>	
A palavra “ <i>mystyle</i> ” (meu estilo) destacam as palavras “ <i>hoodie</i> ” (moletom com capuz) e “ <i>styleinspo</i> ”; essa última faz menção a um portal on-line de moda.	
Este conhecimento é novo? ( ) Sim ( ) Não	Este conhecimento é útil? ( ) Sim ( ) Não

## ANEXO A – TABELA PERIÓDICA PARA TÉCNICAS DE VISUALIZAÇÃO

<div><div>&gt;☀&lt;</div><div>☀</div><div>&lt;☀&gt;</div></div> <div>continuum</div>	<div><div>☀</div> Visualização de Dados</div> <div><div>☀</div> Visualizações Estratégicas</div>										<div><div>☀</div> G</div> <div>graphic facilitation</div>						
<div><div>&gt;☀&lt;</div><div>Tb</div><div>&lt;☀&gt;</div></div> <div>table</div>	<div><div>&gt;☀&lt;</div><div>Ca</div><div>&lt;☀&gt;</div></div> <div>cartesian coordinates</div>	<div><div>☀</div> Visualização de Informações</div> <div><div>☀</div> Visualizações Metafóricas</div>										<div><div>&gt;☀&lt;</div><div>Cs</div><div>&lt;☀&gt;</div></div> <div>concept skeleton</div>	<div><div>&gt;☀&lt;</div><div>Mm</div><div>&lt;☀&gt;</div></div> <div>metro map</div>	<div><div>☀</div><div>Tm</div></div> <div>temple</div>	<div><div>&lt;☀&gt;</div><div>St</div><div>&gt;☀&lt;</div></div> <div>story template</div>	<div><div>&gt;☀&lt;</div><div>Tr</div><div>&lt;☀&gt;</div></div> <div>tree</div>	<div><div>☀</div><div>Ct</div></div> <div>cartoon</div>
<div><div>&gt;☀&lt;</div><div>Pi</div><div>&lt;☀&gt;</div></div> <div>pie chart</div>	<div><div>&gt;☀&lt;</div><div>L</div><div>&lt;☀&gt;</div></div> <div>line chart</div>	<div><div>☀</div> Visualização Conceitual</div> <div><div>☀</div> Visualizações Compostas</div>										<div><div>&gt;☀&lt;</div><div>Me</div><div>&lt;☀&gt;</div></div> <div>meeting trace</div>	<div><div>&gt;☀&lt;</div><div>Fp</div><div>&lt;☀&gt;</div></div> <div>flight plan</div>	<div><div>&lt;☀&gt;</div><div>Cf</div><div>&gt;☀&lt;</div></div> <div>concept fan</div>	<div><div>☀</div><div>Br</div></div> <div>bridge</div>	<div><div>&gt;☀&lt;</div><div>Fu</div><div>&lt;☀&gt;</div></div> <div>funnel</div>	<div><div>☀</div><div>Ri</div></div> <div>rich picture</div>
<div><div>&gt;☀&lt;</div><div>B</div><div>&lt;☀&gt;</div></div> <div>bar chart</div>	<div><div>&gt;☀&lt;</div><div>Hi</div><div>&lt;☀&gt;</div></div> <div>histogram</div>	<div><div>&gt;☀&lt;</div><div>T</div><div>&lt;☀&gt;</div></div> <div>timeline</div>	<div><div>&gt;☀&lt;</div><div>Pa</div><div>&lt;☀&gt;</div></div> <div>parallel coordinates</div>	<div><div>&gt;☀&lt;</div><div>Hy</div><div>&lt;☀&gt;</div></div> <div>hyperbolic tree</div>	<div><div>&gt;☀&lt;</div><div>Cy</div><div>&lt;☀&gt;</div></div> <div>cycle diagram</div>	<div><div>&gt;☀&lt;</div><div>Sa</div><div>&lt;☀&gt;</div></div> <div>sankey diagram</div>	<div><div>&gt;☀&lt;</div><div>Ve</div><div>&lt;☀&gt;</div></div> <div>venn/euler diagram</div>	<div><div>&lt;☀&gt;</div><div>Mi</div><div>&gt;☀&lt;</div></div> <div>mindmap</div>	<div><div>&gt;☀&lt;</div><div>Sq</div><div>&lt;☀&gt;</div></div> <div>square of oppositions</div>	<div><div>&gt;☀&lt;</div><div>Co</div><div>&lt;☀&gt;</div></div> <div>concentric circles</div>	<div><div>&gt;☀&lt;</div><div>Ar</div><div>&lt;☀&gt;</div></div> <div>argument slide</div>	<div><div>&gt;☀&lt;</div><div>Co</div><div>&lt;☀&gt;</div></div> <div>communication diagram</div>	<div><div>&gt;☀&lt;</div><div>Gc</div><div>&lt;☀&gt;</div></div> <div>gant chart</div>	<div><div>&lt;☀&gt;</div><div>Pe</div><div>&gt;☀&lt;</div></div> <div>perspectives diagram</div>	<div><div>&gt;☀&lt;</div><div>D</div><div>&lt;☀&gt;</div></div> <div>dilemma diagram</div>	<div><div>&lt;☀&gt;</div><div>Pr</div><div>&gt;☀&lt;</div></div> <div>parameter ruler</div>	<div><div>☀</div><div>Kn</div></div> <div>knowledge map</div>
<div><div>&gt;☀&lt;</div><div>Ar</div><div>&lt;☀&gt;</div></div> <div>area chart</div>	<div><div>&gt;☀&lt;</div><div>Sc</div><div>&lt;☀&gt;</div></div> <div>scatterplot</div>	<div><div>&gt;☀&lt;</div><div>R</div><div>&lt;☀&gt;</div></div> <div>radar chart cobweb</div>	<div><div>&gt;☀&lt;</div><div>Ch</div><div>&lt;☀&gt;</div></div> <div>chernoff faces</div>	<div><div>&gt;☀&lt;</div><div>E</div><div>&lt;☀&gt;</div></div> <div>entity relationship diagram</div>	<div><div>&gt;☀&lt;</div><div>Fb</div><div>&lt;☀&gt;</div></div> <div>feedback cycle diagram</div>	<div><div>&gt;☀&lt;</div><div>Pa</div><div>&lt;☀&gt;</div></div> <div>pareto chart</div>	<div><div>&lt;☀&gt;</div><div>Cl</div><div>&gt;☀&lt;</div></div> <div>clustering</div>	<div><div>&gt;☀&lt;</div><div>L</div><div>&lt;☀&gt;</div></div> <div>layer chart</div>	<div><div>&lt;☀&gt;</div><div>Py</div><div>&gt;☀&lt;</div></div> <div>minto pyramid technique</div>	<div><div>&gt;☀&lt;</div><div>Ca</div><div>&lt;☀&gt;</div></div> <div>cause-effect chains</div>	<div><div>&gt;☀&lt;</div><div>Tl</div><div>&lt;☀&gt;</div></div> <div>toulmin map</div>	<div><div>&lt;☀&gt;</div><div>Dt</div><div>&gt;☀&lt;</div></div> <div>decision tree</div>	<div><div>&gt;☀&lt;</div><div>Gp</div><div>&lt;☀&gt;</div></div> <div>cpm critical path method</div>	<div><div>&lt;☀&gt;</div><div>Ev</div><div>&gt;☀&lt;</div></div> <div>evocative knowledge maps</div>	<div><div>&gt;☀&lt;</div><div>Co</div><div>&lt;☀&gt;</div></div> <div>concept map</div>	<div><div>☀</div><div>Ic</div></div> <div>iceberg</div>	<div><div>☀</div><div>Cm</div></div> <div>cognitive mapping</div>
<div><div>&gt;☀&lt;</div><div>Tk</div><div>&lt;☀&gt;</div></div> <div>tukey box plot</div>	<div><div>&gt;☀&lt;</div><div>Sp</div><div>&lt;☀&gt;</div></div> <div>spectrogram</div>	<div><div>&gt;☀&lt;</div><div>Te</div><div>&lt;☀&gt;</div></div> <div>tensor diagram</div>	<div><div>&gt;☀&lt;</div><div>Tr</div><div>&lt;☀&gt;</div></div> <div>treemaps</div>	<div><div>&gt;☀&lt;</div><div>N</div><div>&lt;☀&gt;</div></div> <div>nassi shneiderman diagram</div>	<div><div>&gt;☀&lt;</div><div>Se</div><div>&lt;☀&gt;</div></div> <div>semantic network</div>	<div><div>&gt;☀&lt;</div><div>Fl</div><div>&lt;☀&gt;</div></div> <div>flow chart</div>	<div><div>&gt;☀&lt;</div><div>Sy</div><div>&lt;☀&gt;</div></div> <div>system dyn./loop diagrams</div>	<div><div>&gt;☀&lt;</div><div>So</div><div>&lt;☀&gt;</div></div> <div>soft system modeling</div>	<div><div>&gt;☀&lt;</div><div>Sm</div><div>&lt;☀&gt;</div></div> <div>synergy map</div>	<div><div>&gt;☀&lt;</div><div>Fo</div><div>&lt;☀&gt;</div></div> <div>force field diagram</div>	<div><div>&gt;☀&lt;</div><div>Ib</div><div>&lt;☀&gt;</div></div> <div>ibis argumentation map</div>	<div><div>&gt;☀&lt;</div><div>Pr</div><div>&lt;☀&gt;</div></div> <div>process event chains</div>	<div><div>&gt;☀&lt;</div><div>Pe</div><div>&lt;☀&gt;</div></div> <div>pert chart</div>	<div><div>&gt;☀&lt;</div><div>Sw</div><div>&lt;☀&gt;</div></div> <div>swim lane diagram</div>	<div><div>&gt;☀&lt;</div><div>V</div><div>&lt;☀&gt;</div></div> <div>vee diagram</div>	<div><div>&lt;☀&gt;</div><div>Hh</div><div>&gt;☀&lt;</div></div> <div>heaven 'n' hell chart</div>	<div><div>☀</div><div>I</div></div> <div>infomural</div>

**Cy** Visualização de processo

**Hy** Visualização de estrutura de informação

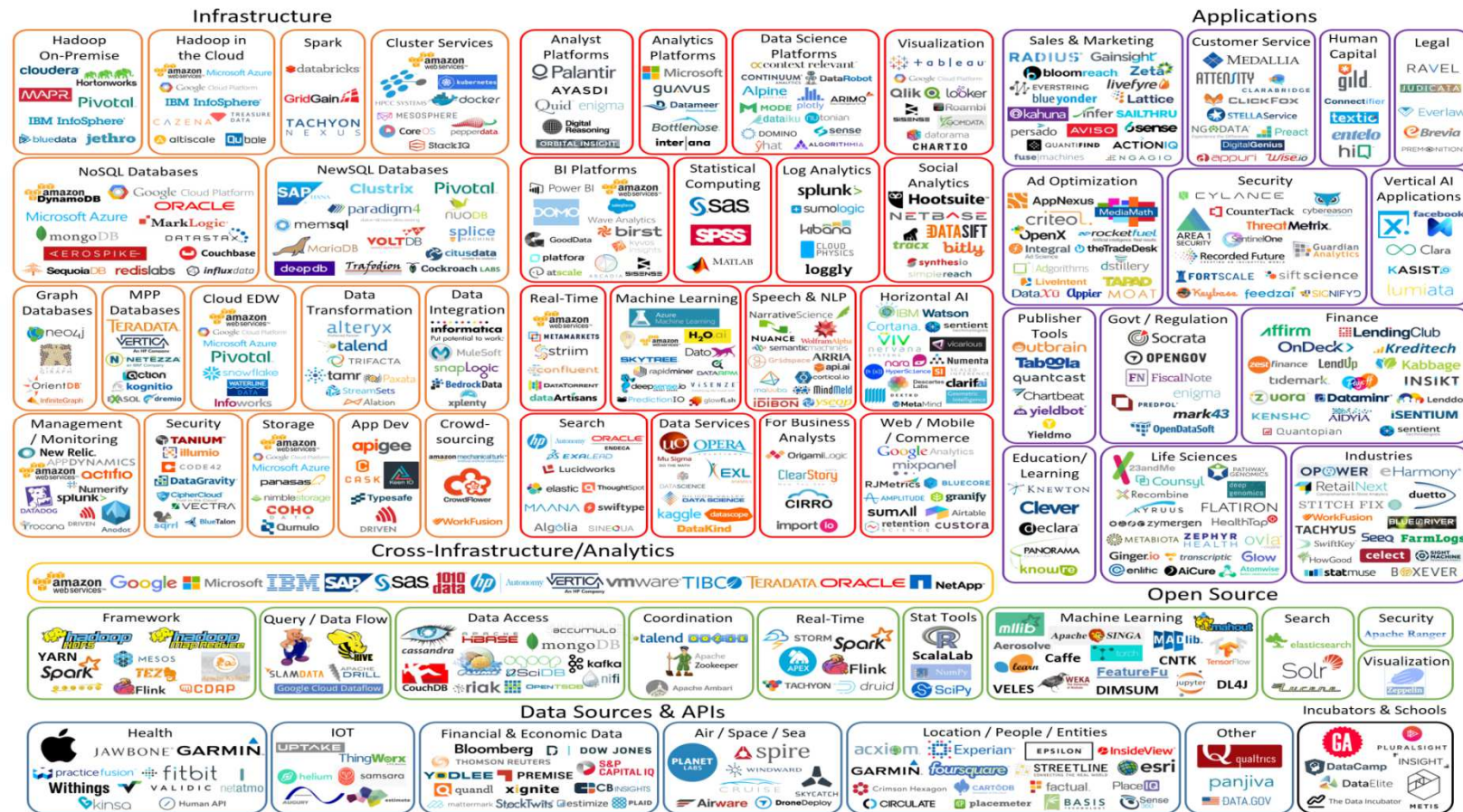
- ☀ Visão
- ☀ Detalhe e visão
- ☀ Detalhe
- < > Divergente
- > < Convergente

>☀< <b>Sd</b> supply demand chain	>☀< <b>Pr</b> performance charting	>☀< <b>St</b> strategy map	>☀< <b>Oc</b> organisation chart	<☀> <b>Ho</b> house of quality	>☀< <b>Fd</b> feedback diagram	☀ <b>Ft</b> failure tree	>☀< <b>Mq</b> magic quadrant	>☀< <b>Sr</b> stakeholder rating map	>☀< <b>Po</b> porter's five forces	<☀> <b>S</b> s-cycle	>☀< <b>Sm</b> stakeholder map	>☀< <b>Ld</b> life-cycle diagram	☀ <b>Tc</b> technology roadmap
☀ <b>Ed</b> edgeworth box	>☀< <b>Pf</b> portfolio diagram	☀ <b>Sg</b> strategic game board	>☀< <b>Mz</b> mintzberg's organigraph	<☀> <b>Z</b> zwick's morphological box	<☀> <b>Ad</b> affinity diagram	☀ <b>De</b> decision discovery diagram	>☀< <b>Bm</b> bcg matrix	>☀< <b>Stc</b> strategy canvas	>☀< <b>Vc</b> value chain	<☀> <b>Hy</b> hype-cycle	☀ <b>Is</b> ishikawa diagram	>☀< <b>Ta</b> taps	<☀> <b>Sd</b> spray diagram

FONTE: LENGLER E EPPLER (2007)



## ANEXO B – AMBIENTE PARA O BIG DATA



Last Updated 3/23/2016

FONTE: TURCK E HAO (2016)