EDUARDO SOARES DE PAIVA

METODOLOGIA PARA EXTRAÇÃO DE CONHECIMENTO DE DADOS: OBTENDO VALOR A PARTIR DO VOLUME, VARIEDADE E VELOCIDADE DOS DADOS

Trabalho de Conclusão de Curso - Monografia apresentada ao Departamento de Estudos da Escola Superior de Guerra como requisito à obtenção do diploma do Curso de Altos Estudos de Política e Estratégia.

Orientador: Cel R1 Carlos Alberto Gonçalves de Araújo.

Rio de Janeiro
2017

C2017 ESG

_____
Assinatura do autor

Biblioteca General Cordeiro de Farias

Paiva, Eduardo Soares de.
Metodologia para extração de conhecimento de dados: obtendo valor a partir do volume, variedade e velocidade dos dados / Eduardo Soares de Paiva. - Rio de Janeiro: ESG, 2017.

49 f.: il.

Orientador: Cel R1 Carlos Alberto Gonçalves de Araújo.
Trabalho de Conclusão de Curso – Monografia apresentada ao Departamento de Estudos da Escola Superior de Guerra como requisito à obtenção do diploma do Curso de Altos Estudos de Política e Estratégia (CAEPE), 2017.

1. Análise de Dados. 2. Mineração de Dados. 3. Big Data. I.Título.

RESUMO

Atualmente, tem sido gerada uma grande quantidade de dados, que devidamente integrados e tratados, podem fornecer informações valiosas para as diversas áreas de conhecimento. Porém, a potencial quantidade de informações presentes nesses dados não costuma ser explorada. Essa subutilização se dá principalmente pela ausência de uma forma sistematizada de se manipular esse grande volume e variedade de dados. Sendo assim, o objetivo dessa monografia é definir uma metodologia capaz de sistematizar a sequência de atividades necessárias para se extrair conhecimentos implícitos nas diversas fontes de dados. Para isso, definiu-se a seguinte hipótese: se forem identificadas as atividades necessárias para a obtenção de conhecimento a partir de múltiplas fontes de dados e os relacionamentos existentes entre essas atividades, então, é possível se definir uma metodologia capaz de sistematizar um processo genérico de descoberta de conhecimento a partir dessas múltiplas fontes de dados. Para se avaliar a metodologia proposta e a validade da hipótese, foi desenvolvido um estudo de caso, com dados reais, para verificar se seguindo todos os passos elencados na metodologia seria possível chegar a conclusões a partir das bases de dados analisadas. Os resultados apresentados demonstraram que a metodologia desenvolvida foi capaz de extrair conhecimentos que anteriormente estavam implícitos, o que apontou a sua validade.

Palavras chave: Análise de Dados. Mineração de Dados. Big Data.

ABSTRACT

Currently, a large amount of data is being generated, which, when properly integrated and treated, can provide valuable information for the various knowledge areas. However, the potential amount of information present in such data is often not explored. This underutilization is mainly due to the absence of a systematized way of manipulating this large volume and variety of data. Thus, the objective of this monograph is to define a methodology capable of systematizing the activities sequence necessary to extract implicit knowledge in the various data sources. For this, the following hypothesis was defined: if the activities necessary to obtain knowledge from multiple data sources and the relationships among these activities are identified, then it is possible to define a methodology capable of systematizing a generic knowledge discovery process from these multiple sources of data. In order to evaluate the proposed methodology and the hypothesis validity, a study case with real data was developed to verify if following all listed steps in the methodology it would be possible to reach conclusions from the databases analyzed. The results showed that the developed methodology was capable extracting knowledge that was previously implicit, which pointed its validity.
Keywords*: Data analysis. Data Mining. Big Data.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CGU	Ministério da Transparência e Controladoria Geral da União
CRISP-DM	Cross Industry Standard Process for Data Mining
DW	Data Warehouse
KDD	knowledge-discovery in databases (ou Descoberta de conhecimento em bases de dados)
SIAFI	Sistema Integrado de Administração Financeira do Governo Federal
SIASG	Sistema Integrado de Administração de Serviços Gerais
SQL	Structured Query Language (ou Linguagem de Consulta Estruturada)
TCU	Tribunal de Contas da União

SUMÁRIO

1 INTRODUÇÃO

A evolução das tecnologias digitais traz como consequência o aumento da quantidade de dados gerados. Atualmente, o homem gera informações tanto nas suas atividades de trabalho quanto de lazer, através da criação de arquivos textos, planilhas, fotos, interação com sistemas informatizados, postagens em redes sociais, entre outros. Adicionalmente, também há os dados que são gerados sem a interação humana, como é o caso das câmeras de monitoramento de trânsito, radares, informações geradas pela interação entre dispositivos autônomos conectados à internet (internet das coisas) e etc. Estima-se que de todos os dados digitais existentes, 90% foram criados nos últimos dois anos (MARQUESONE, 2016).

Esses dados, devidamente integrados e tratados, podem fornecer informações valiosas para as diversas áreas de conhecimento. Por exemplo, na área de segurança pode-se prever a ocorrência de crimes através da análise de boletins de ocorrências anteriores. Já na área de Defesa, uma possível aplicação seria a análise de dados gerados em redes sociais para se identificar possibilidades de ataques terroristas. Com relação ao desenvolvimento nacional, uma possível aplicabilidade seria a utilização de dados para se identificar as prioridades de políticas públicas a serem implementadas no território nacional. Ou seja, existe uma infinidade de possibilidades de uso para a grande quantidade dados que é gerada.

Porém, a potencial quantidade de informações presentes nesses dados não costuma ser explorada. Segundo relatório publicado por Gantz; Reinsel (2012), de todos os dados produzidos até 2012 somente 3% foram utilizados, e apenas 0,5% desses dados foram devidamente analisados. Essa subutilização dos dados se dá por uma série de motivos, dentre os quais pode-se destacar a ausência de uma forma sistematizada de se manipular esse grande volume e variedade de informações, o que evidencia a necessidade de pesquisas nessa área.

Portanto, o problema a ser tratado nessa monografia pode ser descrito da seguinte forma: como tratar as fontes de dados existentes a fim de se extrair conhecimento desses dados?

O trabalho possui o objetivo de definir uma metodologia capaz de sistematizar a sequência de atividades necessárias para se extrair conhecimentos implícitos no grande volume de dados que são produzidos.

Para se atingir o objetivo final da pesquisa, foram identificados os seguintes objetivos intermediários:

- Identificar as atividades necessárias para se extrair conhecimentos implícitos no grande volume de dados que são produzidos; e
- Definir a sequência das atividades identificadas como necessárias para se extrair conhecimentos implícitos no grande volume de dados que são produzidos.

Segundo Pang-Ning; Steinbach; Kumar (2006), a descoberta de conhecimento em banco de dados é um processo de conversão de dados em informações úteis e consiste de um conjunto de atividades, tais como, pré-processamento de dados, mineração de dados[1] e pós-processamento de dados.

O pré-processamento de dados é uma área abrangente e consiste de um grande número de diferentes estratégias que são inter-relacionadas de forma complexa (BATISTA, 2003). Da mesma forma, as tarefas executadas na fase de mineração de dados também são complexas. O objetivo da fase de mineração é criar um modelo preditivo. Logo, decidir qual algoritmo é o mais adequado para o problema que está sendo tratado não é uma tarefa trivial (PANG-NING; STEINBACH; KUMAR, 2006).

Esse trabalho não pretende utilizar o processo de descoberta de conhecimento descrito em Pang-Ning; Steinbach; Kumar (2006), pois o objetivo da monografia como já definido é propor um modelo capaz de sistematizar as diversas atividades necessárias para a extração de conhecimento a partir de diferentes fontes de dados. No entanto, as atividades de pré-processamento e de mineração de dados são essências em qualquer processo de extração de conhecimento de dados. Sendo assim, levando-se em consideração a complexidade e peculiaridades das técnicas empregadas em cada uma dessas fases, essa pesquisa não irá especificar as técnicas e algoritmos a serem usadas em cada uma das atividades.

Dessa forma, esse trabalho pretende propor apenas as macroatividades envolvidas no processo de obtenção de conhecimento a partir das diferentes fontes

[1] Mineração de dados é o processo de exploração de grandes quantidades de dados com o objetivo de encontrar anomalias, padrões e correlações para suportar a tomada de decisões e proporcionar vantagens estratégicas (“Mineração de Dados | SAS”, 2017).

de dados, e os relacionamentos entre elas, sem abordar peculiaridades de cada uma dessas atividades.

A Constituição Federal (BRASIL, 1988) prevê em seu artigo 3º os Objetivos Nacionais Fundamentais. Alinhado a esses objetivos nacionais, foi criada a Política Nacional de Defesa, que é o documento de mais alto nível do país em questão de Defesa (MINISTÉRIO DA DEFESA, 2017c). Esse documento busca harmonizar as iniciativas de todas as expressões do Poder Nacional intervenientes com o tema, visando melhor aproveitar as potencialidades e as capacidades do país (MINISTÉRIO DA DEFESA, 2017c). Essa política faz uma análise do atual ambiente internacional e destaca um cenário propício para o desenvolvimento da denominada “guerra híbrida”, que combina distintos conceitos de guerra (MINISTÉRIO DA DEFESA, 2017c). Nesse contexto está inserido essa nova cultura da análise dos grandes volumes de dados. Esses dados devidamente tratados e integrados podem revelar hábitos populacionais, tendências econômicas e outros “insights” que são capazes de fornecer informações valiosas para que os tomadores de decisão possam considerar em um momento de crise.

Com o objetivo de orientar como se fazer o que está previsto na Política Nacional de Defesa, foi criada a Estratégia Nacional de Defesa (MINISTÉRIO DA DEFESA, 2017a). Segundo Ministério da Defesa (2017a), em face da análise dos atuais cenários, nacional e internacional, torna-se essencial adaptar a configuração das expressões do Poder Nacional às novas circunstâncias e, por conseguinte, buscar estruturar os meios de defesa em torno de capacidades. Dentre essas capacidades está a Capacidade de Gestão da Informação, que visa garantir a obtenção, a produção e a difusão dos conhecimentos necessários à coordenação e ao controle dos meios de que dispõe a Nação, proporcionando o acesso à Inteligência aos tomadores de decisão e aos responsáveis pelas áreas de Segurança Pública e de Defesa Nacional, em todos os escalões (MINISTÉRIO DA DEFESA, 2017a). Dessa forma, a essa pesquisa também recebe suporte da Estratégia Nacional de Defesa, uma vez que ela está diretamente ligada ao desenvolvimento da capacidade de Gestão da Informação.

Complementando o arcabouço literário da defensa nacional, o Livro Branco de Defesa Nacional (MINISTÉRIO DA DEFESA, 2017b) destaca os novos temas, que passaram a influir no ambiente internacional desse século, enfatizando a crescente transversalidade dos temas de segurança e defesa. Nesse sentido, essa pesquisa também traz contribuições, uma vez que, propicia a integração e o tratamento dos

dados gerados no enfrentamento desses novos temas, facilitando assim ações que precisem se valer do caráter transverso desses assuntos.

Levando-se em consideração a dificuldade em se iniciar um processo de extração de conhecimento a partir de múltiplas fontes de dados e a necessidade de se formalizar um processo capaz de orientar pesquisadores na busca de respostas através da análise de dados, formulou-se a seguinte hipótese:

• Se forem identificadas as atividades necessárias para a obtenção de conhecimento a partir de múltiplas fontes de dados e os relacionamentos existentes entre essas atividades, então, é possível se definir uma metodologia capaz de sistematizar um processo genérico de descoberta de conhecimento a partir dessas múltiplas fontes de dados.

Esse trabalho está organizado da seguinte forma: inicialmente é apresentado o referencial teórico da pesquisa fazendo-se um levantamento dos trabalhos relacionados ao tema tratado e identificado as lacunas ainda existentes na área. A seguir, a proposta de solução para o problema tratada, sendo que, a primeira parte do capítulo apresenta a teoria que serviu de base para o desenvolvimento da solução, enquanto que a segunda apresenta a proposta propriamente dita. Na sequência, é feita a avaliação da solução proposta. Finalmente, é apresentada a conclusão do trabalho.

2 REFERÊNCIAL TEÓRICO

A crescente velocidade com que os dados estão sendo gerados faz com que também se aumente o interesse pela análise desses grandes volumes de dados, a fim de se extrair informações úteis. Dessa forma, atualmente já existe uma série de trabalhos, nas mais diversas áreas de conhecimento, que se propõem a utilizar diferentes fontes de dados com o objetivo de obter novos conhecimentos a respeito de diversos assuntos.

Na área de Administração Pública, Paiva e Revoredo (2016) propõem uma metodologia de tratamento de informação para a obtenção de visões mais elucidativas sobre os gastos públicos, que são apresentados no Portal da Transparência do Governo Federal. O objetivo desse trabalho é utilizar a base de dados de Itens de Empenho do Poder Executivo Federal, que possui uma taxa de atualização de aproximadamente 25.000 registros por dia, para fazer a identificação de um conjunto pré-determinado de produtos comprados pela Administração Pública, através de uma técnica de mineração de texto, e propor uma forma de consolidação dessas informações de maneira que permita a fácil visualização de disparidades encontradas no grande volume de dados apresentados. Esse mecanismo propicia uma melhor avaliação dos gastos públicos, permitindo com que o cidadão possa entender melhor como o governo está empregando seu dinheiro.

Já na área de planejamento urbano, José (2015) desenvolve um método para identificar perfis de mobilidade humana em grandes eventos usando dados de telefonia celular, sendo que, para a avaliação da metodologia utilizou-se os dados de localização gerados pelos celulares das pessoas que participaram de 2 eventos distintos na cidade do Rio de Janeiro no ano de 2014: o Réveillon na Praia de Copacabana e o bloco de Carnaval do Cordão da Bola Preta.

Dessa forma, o trabalho permite identificar o trajeto e o horário de deslocamento do público dos eventos, o que é útil para o planejamento urbano, facilitando ao governo prever as vias mais utilizadas e os meios de transportes mais apropriados para eventos futuros.

Considerando o grande volume de dados de pacientes, gerados pelos diversos sistemas utilizados pela área de saúde, Sahoo; Mohapatra; Wu (2016) propuseram um mecanismo capaz de fazer previsões das futuras condições de saúde dos pacientes a partir dos parâmetros gerado pelos sistemas ligados à área de saúde.

Dessa forma, o artigo se propõe a desenvolver um mecanismo que faz a coleta de dados e analisa as correlações desses dados. Posteriormente, um modelo de previsão estocástico[2] é projetado para prever o estado de saúde dos pacientes cujos dados foram analisados.

Estes são apenas alguns exemplos dentre muitos outros em que grandes volumes de dados gerados, aliados à variedade de diferentes fontes e à velocidade de geração de informações, estão sendo aplicados. No entanto, uma limitação comum a esses trabalhos é o fato deles se proporem a estudar casos específicos, cujas soluções não são passíveis de generalizações e replicações em outras situações.

Algumas iniciativas já se propõem a sistematizar o processo de obtenção de conhecimento a partir da análise de dados. Nesse sentido, Shearer (2000) propõem o CRISP-DM ("Cross Industry Standard Process for Data Mining"), esse modelo está dividido em 6 fases (entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação) e tem como finalidade definir os passos a serem seguidos em um projeto de mineração de dados. Porém, essa metodologia parte da premissa de que os dados a serem utilizados já estejam disponíveis para o uso, o que nem sempre é verdade. Outro limitante dessa abordagem é a não previsão de múltiplas fontes de dados[3] no processo de descoberta de conhecimento.

Por outro lado, Marquesone (2016) prevê uma sequência de 5 passos que devem ser seguidos para a obtenção de conhecimento a partir de múltiplas fontes de dados. No entanto, esse trabalho, diferentemente do que é proposto em Shearer (2000), não prevê a possibilidade de iteração desse processo, nem possíveis reexecuções de atividade devido à identificação de novas necessidades levantadas em uma fase subsequente.

Dessa forma, o trabalho ora proposto pretende combinar as metodologias introduzidas em Shearer (2000) e Marquesone (2016) a fim de gerar um método para a descoberta de conhecimento utilizando diferentes fontes de dados. Sendo assim, a principal contribuição dessa monografia será propor uma metodologia capaz de

[2] Estocástico: Cujos resultados ou conclusões são determinados pelas variáveis aleatórias, segundo as leis da probabilidade; aleatório (MICHAELIS, 2010).

[3] O fato da metodologia não prever a utilização de múltiplas fontes de dados não quer dizer que não seja possível utilizá-la com diferentes bases de dados, porém, como não há uma previsão formal para essa situação, todas as atividades relacionadas à integração e uniformização dos dados provenientes de diferentes bases terão que ser executadas por atividades não previstas na metodologia.

sistematizar a sequência de atividades necessárias para se extrair conhecimentos implícitos no grande volume de dados que são produzidos diariamente.

3 PROPOSTA

O objetivo desse capítulo é formular uma solução capaz de responder ao problema tratado nessa pesquisa. Dessa forma, a primeira parte do capítulo apresenta a teoria que serviu de base para o desenvolvimento da solução, enquanto que a segunda parte apresenta a proposta propriamente dita.

A atividade de extração de conhecimento a partir de diferentes bases de dados, apesar de demonstrar uma excelente forma de obter vantagem competitiva, apresenta uma série de dificuldades para a equipe envolvida com essa tarefa. Os responsáveis pela análise desses dados devem dominar uma série de ferramentas e técnicas computacionais, precisam estudar com profundidade os dados relacionados ao assunto e, principalmente, devem dominar as regras de negócio da área que está sendo tratada. A carência de metodologias capazes de sistematizar tal atividade aumenta o grau de incerteza dos projetos de análise de grandes volumes de dados.

Essa pesquisa conseguiu identificar duas metodologias, propostas por Shearer (2000) e Marquesone (2016), que tentam guiar os analistas nos projetos de análise de dados. Apesar de tais metodologias não atenderem aos objetivos dessa monografia, elas serviram de base para o desenvolvimento da solução proposta.

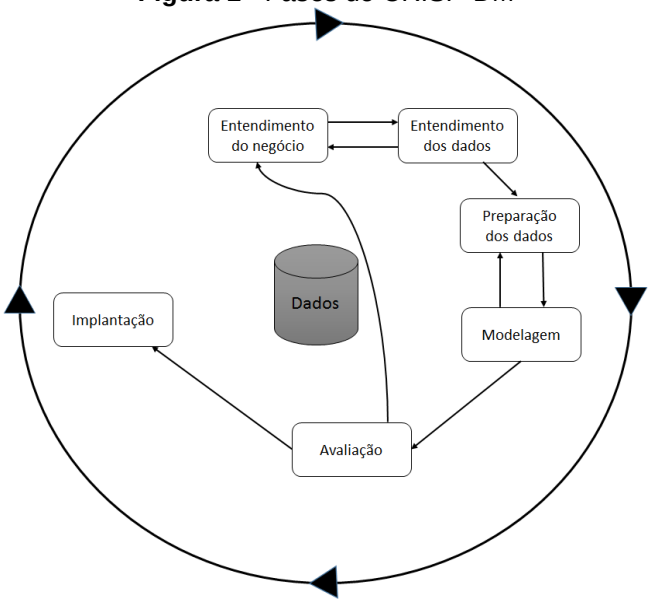
3.1 CRISP-DM

O CRISP-DM tem como objetivo fornecer orientações básicas para a condução do processo de KDD[4] (GOLDSCHMIDT; BEZERRA, 2015). Esse modelo está dividido em 6 fases e tem como finalidade definir os passos a serem seguidos em um projeto de mineração de dados.

Na Figura 1 são indicadas as fases do modelo, sendo que, as setas internas representam as dependências mais importantes e frequentes entre essas fases e a seta externa indica que a metodologia é iterativa, ou seja, pode ser necessário se passar mais de uma vez por essas fases até se atingir o objetivo final.

[4] KDD: knowledge-discovery in databases (ou Descoberta de conhecimento em bases de dados)

Figura 1 - Fases do CRISP-DM



Fonte: o autor, adaptado de SHEARER, 2000.

O entendimento do negócio é a fase inicial do processo e se concentra em entender os objetivos e requisitos do projeto a partir da perspectiva do negócio. Essa fase é responsável por converter esse conhecimento em uma definição do problema de mineração e em um plano preliminar projetado para atingir os objetivos (AZEVEDO; SANTOS, 2008).

A fase de entendimento dos dados geralmente ocorre ao mesmo tempo do entendimento do negócio (GOLDSCHMIDT; BEZERRA, 2015). Essa etapa é responsável por realizar um estudo minucioso dos dados disponíveis, a fim de propiciar maior familiaridade com estes dados. Ela envolve atividades como: identificação de problemas de qualidade de dados, descrição do formato dos dados, quantidade de registros, atributos e relacionamento entre atributos.

A atividade de preparação é responsável pelo pré-processamento dos dados envolvidos no processo de extração de conhecimento. Nessa fase, os dados são selecionados, limpos, integrados, formatados e adequados para as atividades subsequentes. O objetivo principal dessa fase é deixar os dados prontos para serem utilizados na etapa de modelagem. Na fase de modelagem, são definidos e aplicados os algoritmos de mineração e, também é feita a escolha dos parâmetros, utilizados pelos algoritmos, que melhor se adaptam ao conjunto de dados em questão.

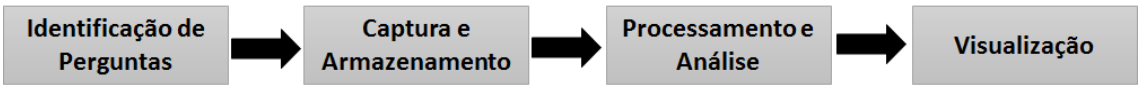
Durante a avaliação, os resultados são analisados, a fim de se verificar a qualidade dos modelos gerados. Normalmente essa análise é baseada em indicadores e métricas comparativas, cujo objetivo é confrontar os resultados obtidos com os esperados.

Finalmente, a implantação consiste no planejamento e acompanhamento das ações a serem realizadas pelo modelo de conhecimento gerado. Nessa fase, os modelos gerados são colocados à disposição do usuário para serem usados.

3.2 TÉCNICA BIG DATA[5]

Marquesone (2016) descreve uma sequência de 4 passos para o desenvolvimento de projetos de Big Data. Segundo essa metodologia, um projeto de big data deve ser composto pelas seguintes fases: Identificação de perguntas, Captura e Armazenamento de Dados, Processamento e Análise de Dados e Visualização dos resultados, conforme apresentado na Figura 2.

Figura 2 - Fases do Projeto de Big Data



Fonte: o autor, criado com base em MARQUESONE, 2016

Segundo Marquesone (2016), o primeiro passo é identificar quais perguntas se deseja responder com os dados. Nessa fase deve-se determinar quais informações pretende-se extrair de um conjunto de dados.

Já o segundo passo diz respeito a captura e armazenamento dos dados. Nessa fase deve-se se identificar quais fontes serão utilizadas, e como os dados serão capturados. Além disso, deve-se também identificar a ordem em que os dados serão usados.

O terceiro passo definido por Marquesone (2016), é o processamento e análise dos dados. Nessa fase emprega-se métodos estatísticos e técnicas de aprendizado de máquina. Por fim, Marquesone (2016)define a última fase como sendo a de visualização de dados, situação em que se criam gráficos dinâmicos e interativos para a apresentação dos resultados obtidos. Segundo a autora, essa fase também

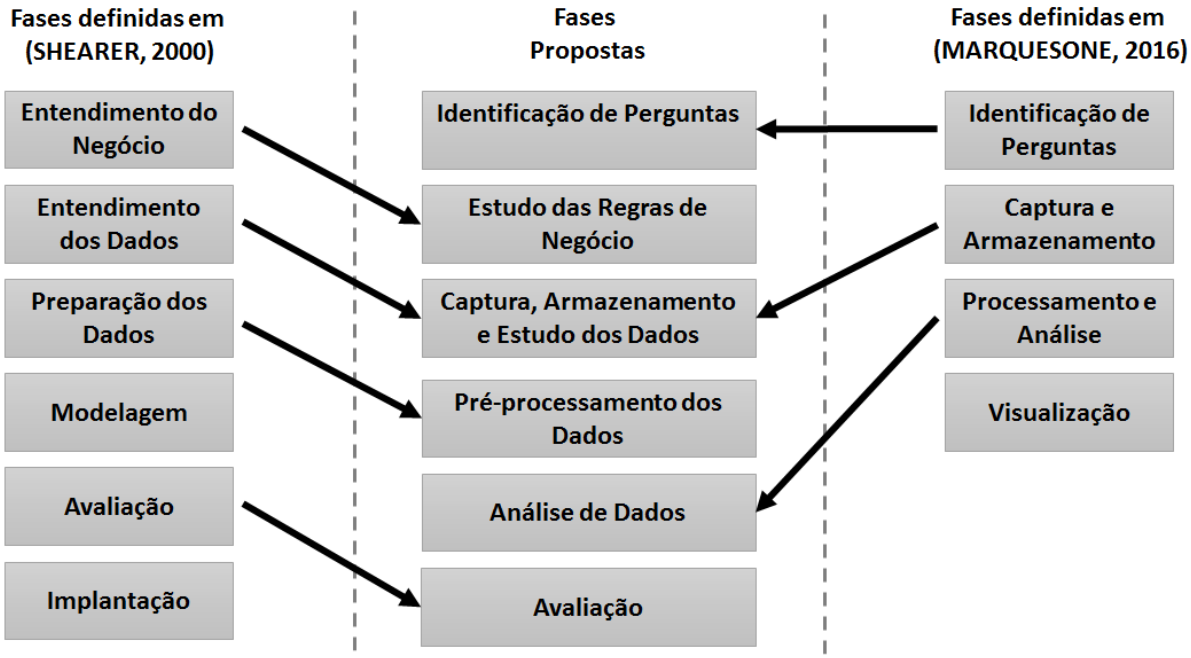
[5] Big Data é o termo que descreve o imenso volume de dados – estruturados e não estruturados – que impactam os negócios no dia a dia (“O que é Big Data? | SAS”, 2017).

pode ser utilizada em conjunto com a fase de análise de dados, para facilitar o processo de descoberta de dados.

3.3 DESENVOLVIMENTO DA PROPOSTA

Após a análise das metodologias propostas em Shearer (2000) e Marquesone (2016), foi possível definir-se um conjunto de atividades necessárias para projetos que tenham o objetivo de extrair informações relevantes a partir de várias fontes de dados. Dessa forma, combinou-se os dois modelos estudados a fim de se propor um modelo mais completo, capaz de combinar os pontos fortes dos dois modelos que foram utilizados como base, conforme é apresentado na Figura 3.

Figura 3 – Composição das fases do modelo proposto



Fonte: **o autor**

3.3.1 Especificação das Fases

A seguir são apresentadas as principais atividades a serem executadas durante cada uma das fases do modelo proposto.

- **Identificação de Perguntas**

A primeira atividade a ser feita, antes mesmo de se iniciar o trabalho é definir-se os objetivos do projeto de extração de conhecimento. Normalmente, esses projetos são motivados por alguma necessidade de negócio, algum conhecimento que precisa ser adquirido a fim de se subsidiar alguma decisão. Sendo assim, a melhor forma para guiar um projeto desse tipo é definindo as perguntas que se pretende responder com a análise de dados.

- **Estudo das Regras de Negócio**

O principal objetivo dessa fase é conhecer o contexto em que o projeto está inserido. O conhecimento das rotinas peculiares da área em que o projeto está inserido é tão importante quanto os dados em si, pois uma falha no entendimento de algum processo de negócio específico pode acarretar em conclusões erradas.

Dessa forma, essa fase destina-se ao estudo das regras de negócio da área que está sendo pesquisada. Esse estudo pode incluir a identificação de pessoas conhecedoras dos processos que possam ajudar no entendimento das regras, a leitura da doutrina e da legislação relacionada, entre outras.

Por exemplo, caso esteja sendo desenvolvido um projeto que tente encontrar respostas para perguntas ligadas à área de licitação públicas, boas fontes de consultas a serem utilizadas nessa fase são as leis 8666/93[6] (BRASIL, 1993) e 10520/02[7] (BRASIL, 2002), livros relacionados à área de licitações e contratos, acórdãos e jurisprudência do TCU[8], tais como, pessoas que trabalham com licitações.

- **Captura, Armazenamento e Estudo dos Dados**

Após o entendimento do negócio, o passo seguinte é a captura, o armazenamento e estudo dos dados. Essa fase começa com a identificação das possíveis fontes de dados que podem ser usadas para responder as questões levantadas. Sendo que, um bom entendimento do negócio irá facilitar bastante a identificação das fontes de dados a serem utilizadas.

---

[6] Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências.

[7] Institui, no âmbito da União, Estados, Distrito Federal e Municípios, nos termos do art. 37, inciso XXI, da Constituição Federal, modalidade de licitação denominada pregão, para aquisição de bens e serviços comuns, e dá outras providências.

[8] O Tribunal de Contas da União (TCU) é instituição brasileira prevista na Constituição Federal para exercer a fiscalização contábil, financeira, orçamentária, operacional e patrimonial da União e das entidades da administração direta e administração indireta, quanto à legalidade, à legitimidade e à economicidade e a fiscalização da aplicação das subvenções e da renúncia de receitas.

Essas bases de dados podem ser internas ou externas. As fontes internas são aquelas que pertencem a própria organização que está desenvolvendo o projeto em questão. Já as fontes externas são aquelas que estão fora da organização, e que por essa razão o esforço de obtenção tende a ser maior. Uma fonte de dados externa pode ser um conjunto de dados abertos disponibilizado pelo governo, pode ser um conjunto de posts em alguma rede social, algum site público, pode ser uma base de dados comercializada por empresas específicas, entre outros.

Essas possíveis fontes de dados também podem se dividir em dados estruturados e não estruturados, sendo que, os dados estruturados são aqueles que são mantidos em tabelas em bancos de dados relacionais[9] e possuem um formato específico e rígido. Já os dados não estruturados não possuem uma estrutura definia e normalmente são caracterizados por documentos textos, imagens, vídeos e etc. Cabe ressaltar que a maioria dos dados produzidos atualmente são não estruturados e por isso são mais difíceis de serem tratados.

Nessa fase, deve-se dar especial atenção a qualidade dos dados que serão colhidos, pois os resultados obtidos no projeto de descoberta de conhecimento estão diretamente relacionados com a qualidade dos dados utilizados.

Conforme ressaltado por Nehmy; Paim (1998), o termo "qualidade da informação", tal como é abordada na literatura pertinente ao assunto, revela que se trata de uma noção vaga, imprecisa, situando-se muito próxima ao entendimento do senso comum. Dessa forma, para o contexto desse trabalho, considera-se que um dado de qualidade é aquele capaz de preservar os atributos de precisão, atualização e consistência das informações disponíveis em uma base de dados.

O estudo desses dados serve para identificar como eles se relacionam. Tal estudo é subsidiado pelas informações colhidas durantes o estudo das regras de negócio, e serve para alimentar a fase seguinte, o pré-processamento dos dados.

- **Pré-processamento dos dados**

O pré-processamento dos dados é uma das fases mais difíceis do processo, onde se emprega grande parte do esforço. Essa fase consiste das atividades de limpeza e integração dos dados oriundos das diferentes bases de dados, bem como de outros tipos de tratamentos necessários à análise dos dados.

[9] O banco de dados relacional é um banco de dados que modela os dados em formato de tabelas (com colunas e linhas), que podem se relacionar entre si.

A atividade de limpeza de dados tem a função de minimizar a probabilidade de se obter um resultado incorreto devido a sujeiras existentes nos dados de entrada. O processo de limpeza requer uma inspeção minuciosa dos dados, bem como a realização de operações de correção e remoção de registros, conforme a necessidade (MARQUESONE, 2016). Ou seja, durante a limpeza dos dados verifica-se a existência de dados duplicados, dados incompletos, de dados inconsistentes ou faltantes.

A atividade de integração consiste da junção dos dados oriundos de diferentes bases de dados. Segundo Lenzerini (2002), a integração de dados resulta na combinação de diferentes fontes para prover ao usuário uma visão unificada desses dados.

No entanto, para se integrar bases de dados distintas, é necessário se unificar os esquemas[10] dessas bases. Os esquemas são as formas de representação dos dados utilizadas pelos bancos de dados. As técnicas de integração de esquemas são consideradas como uma das operações básicas requeridas pelo processo de integração de dados (BERNSTEIN e MELNIL 2004).

O objetivo da integração de esquemas é produzir um conjunto de correspondência entre diferentes esquemas (DOAN et al. 2012). As técnicas de integração de esquema oferecem a possibilidade para unificar a representação de vários esquemas em um esquema global (ARFAOUI e AKAICHI 2015). No entanto, durante essa integração podem ser detectados conflitos entre os diferentes esquemas (DOAN et al. 2012).

Um exemplo desses tipos de conflito são os conflitos de nome. Batini (2001) divide os conflitos de nome em sinonímia e homonímia. A Homonímia ocorre quando um mesmo nome referencia dois ou mais conceitos em diferentes bancos de dados (Bellström 2006). Por exemplo, dois bancos de dados podem ter cada um deles uma tabela chamada veículo, porém, em um banco ela só recebe valores “SIM” e “NÃO”, enquanto que no outro banco a tabela veículo é povoada com os modelos. Ou seja, apesar das referidas tabelas terem o mesmo nome, elas armazenam informações diferentes e por isso não podem ser representadas por uma mesma tabela no esquema integrador. Já a sinonímia ocorre quando dois ou mais nomes se referem a um mesmo conceito (BELLSTROM, 2006). Por exemplo, caso dois bancos distintos possuam tabelas para armazenar modelos de carros, porém, em um banco essa

[10] Esquema de banco de dados é a representação da estrutura lógica desse banco de dados.

tabela se chama veículo, enquanto que no outro ela se chama automóvel. Ou seja, as duas tabelas armazenam o mesmo tipo de informação, apesar de possuírem nomes diferentes, logo, no esquema integrador elas devem ser representadas por uma única tabela.

Dessa forma, a principal dificuldade dessa atividade de integração é encontrar uma forma única de representar dados oriundos de diferentes sistemas, que foram concebidos em momentos distintos e para propósitos diversos.

Além disso, o pré-processamento pode exigir outros tipos de tratamento de dados, que podem variar de acordo com os propósitos do projeto. Dessa forma, ao final da fase de pré-processamento os dados devem estar prontos para serem utilizados na próxima fase, que é a fase de análise.

- **Análise**

É na fase de análise que as perguntas formuladas são respondidas. Nessa fase deve-se utilizar os dados pré-processados a fim de se extrair conhecimentos implícitos.

As técnicas de análise de dados a serem utilizadas variam de acordo com o tipo de problema que está sendo tratado, bem como com os tipos de dados disponíveis. No entanto, as principais técnicas de análise de dados que podem ser utilizadas são: análises estatísticas, buscas diretas em bancos de dados, data warehouse[11] (DW), analise de regressão, aprendizado de máquina e mineração de dados, dentre outras.

Análise estatística é um método que busca descrever e resumir o dado, bem como fazer comparações de grandezas estatísticas das variáveis de interesse. As buscas diretas em bancos de dados são consultas feitas diretamente nos bancos de dados através do uso da linguagem de consulta em banco de dados SQL (Structured Query Language, ou Linguagem de Consulta Estruturada). Essa técnica geralmente é mais indicada para a respostas de problemas mais simples e diretos. Data warehouse (DW) é uma forma de armazenamento dos dados que facilita a recuperação de dados históricos, assim como o cruzamento de informações que possam auxiliar no processo de tomada de decisão. A análise de regressão é uma técnica que permite explorar e inferir a dependência de uma variável em relação a outra. O aprendizado de máquina

[11] data warehouse é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais (INMON, 1997).

e a mineração de dados são métodos de análise de dados que automatizam o desenvolvimento de modelos analíticos, que permitem a descoberta de padrões ou tendências implícitas nos dados.

No entanto, não existe um método melhor ou pior, e sim aquele que melhor se adapta ao problema que está sendo tratado, bem como ao tipo de dados utilizados na análise. Cabe ressaltar que, podem-se utilizar essas técnicas conjuntamente na solução de um mesmo problema. Ou seja, independente da técnica utilizada, o principal objetivo dessa fase é encontrar as repostas para as perguntas definidas como alvo do projeto.

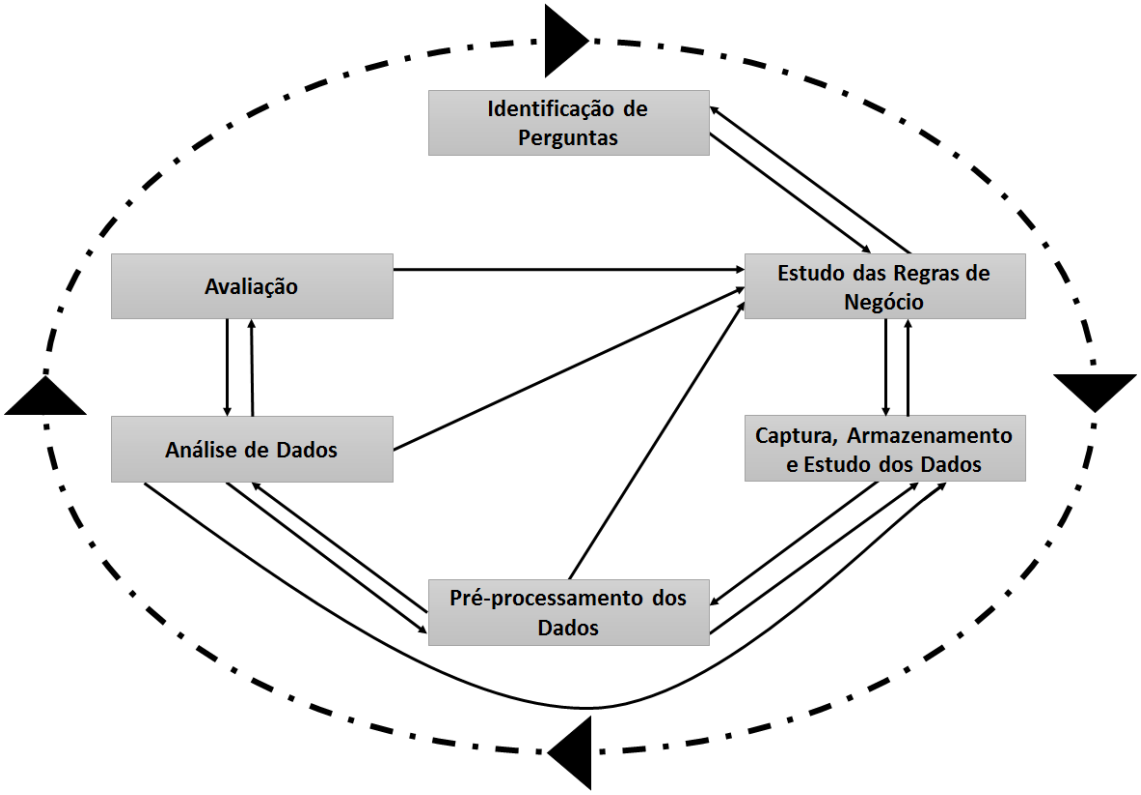
- **Avaliação**

A sexta fase é a de avaliação das respostas obtidas na etapa anterior. O principal objetivo dessa fase é verificar se os resultados são consistentes e atendem aos objetivos estabelecidos. A participação de especialistas das áreas de negócio é de extrema importância nessa fase, pois, estes têm melhores condições de avaliar a consistência dos resultados gerados.

3.3.2 Relacionamento entre as Fases

Apesar de já ser apresentada uma ordem sequencial das fases propostas na figura 3, essas não necessariamente devem seguir uma sequência fixa de execução, pois alguma deficiência encontrada em uma determinada fase pode exigir a reexecução de uma ou mais fase anterior. Dessa forma, na figura 4 é apresentado o modelo proposto já com os possíveis fluxos sequenciais existentes entre as diversas fases, sendo que, as setas internas e contínuas indicam possibilidades de sequencias entre as fases, enquanto que a linha externa tracejada indica a possibilidades de reexecuções do ciclo todo, uma vez que, ao final de cada interação podem ser identificadas novas perguntas a serem respondidas.

Figura 4 – Modelo Proposto



Fonte: o autor

Dessa forma, iniciando-se pela fase de identificação de perguntas, a única fase possível a ser seguida é o Estudo das Regras de Negócio. A seguir, pode-se identificar que as perguntas não foram bem formuladas e por isso deve-se voltar para a atividade anterior ou pode-se prosseguir para a fase de Captura, Armazenamento e Estudo dos Dados. Após a execução da fase 3, verifica-se que há necessidade de novos estudos sobre as regras de negócio, retornando-se assim para a fase anterior, ou pode-se seguir para a fase seguinte. Como citado anteriormente, o pré-processamento é a fase mais trabalhosa, onde identifica-se a necessidade de novos estudos sobre as regras de negócio, ou a necessidade de captura de novas fontes de dados. Caso tudo corra bem, passa-se para a fase de análise de dados. Durante a análise de dados, pode-se voltar para qualquer uma das três fases imediatamente anteriores, ou ir para a fase de avaliação. Finalmente, após a avaliação, existe a possibilidade de voltar para a análise de dados ou para a fase de entendimento das regras de negócio, ou finalizar o ciclo.

4 VALIDAÇÃO

Esse capítulo tem o objetivo de avaliar a veracidade da hipótese formulada e a eficácia da metodologia desenvolvida. Para isso, foi proposto um estudo de caso, com dados reais, para verificar se seguindo todos os passos elencados na metodologia seria possível chegar a conclusões a partir das bases de dados analisadas.

Dessa forma, o estudo de caso aplica a metodologia para a solução de uma questão ligada à área da Administração Pública. Logo, esse estudo de caso se propõe a percorrer os seis passos sugeridos na metodologia, a fim de analisar as compras realizadas pelo Poder Executivo Federal, para que, ao final do processo se possa avaliar a eficácia dessa metodologia.

4.1 INFRAESTRUTURA PARA O ESTUDO DE CASO

O estudo de caso foi desenvolvido utilizando-se uma máquina com processador Intel Core i7 com 8 GB de memória RAM e sistema Operacional Windows 7 Home Premium de 64 Bits. As bases de dados intermediárias foram armazenadas no Sistema de Gerenciamento de Banco de Dados Microsoft SQL Server, versão 2012.

Especificamente para as fases de preparação e análise de dados, que manipulam grandes volumes de dados e necessitam de maior poder de processamento, foi utilizado o ambiente de computação nas nuvens disponibilizado pela empresa Amazon Web Services[12]. Esse processamento dos dados foi feito com a utilização de um cluster[13]. O cluster utilizado era composto por 4 CPUs com Processadores Intel Xeon E5-2670 v2 (Ivy Bridge) de alta frequência, sendo que, essas máquinas possuíam 15 GB de memória, 80 GB de armazenamento de instância local baseado em SSD (solid-state drive) e plataforma de 64 bit.

4.2 ESTUDO DE CASO

[12] https://aws.amazon.com
[13] Clusters de computadores são máquinas interligadas que trabalham de forma conjunta a fim de executar uma mesma tarefa de processamento.

O estudo de caso proposto está ligado à área da Administração Pública pelo fato dos dados governamentais serem mais fáceis de serem obtidos, no entanto, a metodologia poderia ser testada para qualquer outra área do conhecimento.

Sendo assim, as próximas subseções descrevem as atividades realizadas em cada uma das etapas enunciadas pela metodologia proposta, a fim de se resolver uma questão ligada às compras realizadas pela Administração Pública Federal.

4.2.1 Identificação da Pergunta

A sensação geral que a população tem a respeito das compras realizadas pelo poder público é que essas compras, em geral são superfaturadas. Essa sensação ocorre principalmente pelas notícias e escândalos de corrupção que são veiculados nos telejornais, nos jornais impressos e nos grandes portais de notícias.

No entanto, para se fazer qualquer afirmação a respeito da economicidade das compras realizadas pela Administração Pública, é necessário um estudo mais aprofundado, analisando-se uma grande variedade de situações, e não apenas alguns casos extremos que possam tornar as conclusões tendenciosas.

Sendo assim, formulou-se a seguinte pergunta a ser respondida: Quanto o governo paga nas compras que faz? O governo paga mais ou menos que o preço praticado no mercado?

4.2.2 Estudo das Regras de Negócio

A partir das perguntas levantadas na etapa anterior, o passo seguinte é estudar as regras de negócio do contexto em que tal pergunta está inserida. Logo, como a pesquisa em questão trata de assuntos associados a compras públicas, faz-se necessária uma análise dos mecanismos ligados aos processos de compras públicas, para se evitar que uma má interpretação dos dados possa levar a conclusões erradas.

A Constituição Federal (BRASIL, 1988) prevê em seu artigo 37 os cinco princípios fundamentais que devem nortear as atividades da Administração Pública: legalidade, impessoalidade, moralidade, publicidade e eficiência. A fim de atender a esses princípios, no que diz respeito às compras e contratações públicas, foi editada

a lei 8666/93 (BRASIL, 1993), que instituiu as normas para a execução de licitações e contratos no âmbito da Administração Pública.

Licitação é um conjunto de procedimentos administrativos para compras ou serviços contratados por órgãos públicos (governos Federal, Estadual ou Municipal) e todos os entes federativos, onde são convocadas empresas interessadas em apresentar propostas para oferecimento de bens e serviços (DA SILVA, 2016). Segundo Pires (2002), o objetivo da licitação é proporcionar à Administração, meios para, ao instaurar a competição entre licitantes, assegurar a seus administradores a possibilidade de disputarem a participação nos negócios do Governo e receberem o mesmo tratamento jurídico, sem discriminação, obedecendo somente os preceitos do edital.

Dessa forma, as compras na Administração Pública devem ser precedidas de um processo licitatório. A lei 8666/93 (BRASIL, 1993) também prevê o estabelecimento de um contrato administrativo entre a Administração Pública e a pessoa (física ou jurídica) vencedora da licitação, para que o fornecimento do produto, ou a execução do serviço licitado possa ocorrer.

Meirelles (2001) conceitua contratos administrativos como o ajuste que a Administração Pública, agindo nessa qualidade, firma com particular ou outra entidade administrativa para a consecução de objetivos de interesse público, nas condições estabelecidas pela própria Administração.

Além da necessidade do processo licitatório e do contrato administrativo, todas as compras executadas pela Administração Pública devem seguir as regras do processo de execução de uma despesa pública. Essas regras são disciplinadas pela Constituição Federal (BRASIL, 1988) e pela lei 4320 (BRASIL, 1964).

A despesa pública possui uma sequência de três estágios que deve ser respeitada: Empenho, Liquidação e Pagamento.

O empenho, primeiro estágio da despesa, é uma garantia ao fornecedor de que a repartição pública tem autorização legal para realizar o gasto. Ele oferece um documento denominado nota de empenho como suporte para essa despesa.

O artigo 60 da lei 4320 (BRASIL, 1964) veda a realização de despesa sem o empenho prévio. Quando um empenho é emitido, a Administração faz uma reserva do seu orçamento no valor desse empenho a fim de evitar gastos acima da sua capacidade de pagamento. É importante ressaltar que o empenho ainda não cria nenhuma obrigação de pagamento para a Administração Pública, ele apenas reserva

uma parte do orçamento para um provável pagamento futuro, porém, esse pagamento pode não vir a se concretizar e a reserva pode ser cancelada.

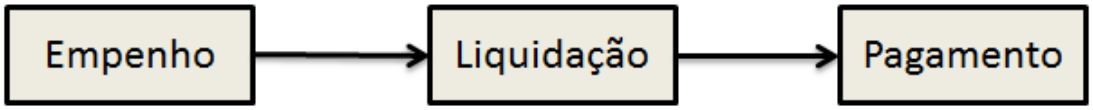
O empenho é a fase da execução da despesa em que o gasto é descrito de forma mais detalhada. A Nota de Empenho especifica o que será comprado, bem como a quantidade. Todas as demais fases da execução da despesa se baseiam no empenho.

A segunda fase da despesa é a liquidação. Nesse momento, o Estado reconhece o compromisso de pagar aos seus fornecedores. A liquidação pode ser entendida como o ateste do recebimento do material ou serviço e consequentemente o reconhecimento por parte da Administração da sua obrigação de pagar.

O Pagamento é o último estágio da execução da despesa. É nessa fase que a Administração faz o desembolso dos valores necessários para o efetivo pagamento dos fornecedores.

Todas essas fases são sequenciais, ou seja, para que um determinado estágio ocorra, é necessário que o seu estágio anterior já tenha ocorrido. Logo, não se pode liquidar despesa que ainda não tenha sido previamente empenhada e nem tão pouco é possível realizar o pagamento sem a anterior liquidação. A Figura 5 ilustra a sequência dos estágios da despesa pública.

Figura 5 – Estágios da Execução da Despesa



Fonte: o autor

Dessa forma, as compras executadas pela Administração Pública devem passar, necessariamente, pelos três estágios da execução da despesa. Quanto à licitação, existem algumas situações em que a lei dispensa ou declara a inexigibilidade do processo licitatório. No entanto, o contrato administrativo sempre será necessário. Sendo assim, esse estudo poderia utilizar os dados oriundos do contrato administrativo, ou de uma das três fases da execução da despesa: empenho, liquidação ou pagamento. Porém, como as informações apresentadas na fase de

empenho são as que possuem um maior grau de detalhamento, optou-se pela utilização dos dados oriundos dessa fase na pesquisa que está sendo realizada.

4.2.3 Captura, Armazenamento e Estudo dos Dados

Uma vez entendida as regras de negócio do assunto pesquisado, o passo seguinte é a busca pelos dados que possam ser utilizados para responder à pergunta levantada. Como definido anteriormente, a principal fonte de informações a ser utilizada são os dados da fase de empenho, da execução da despesa pública. Porém, faz-se necessário definir-se a forma de obtenção dessas informações, bem como outras fontes de dados complementares utilizados na pesquisa.

4.2.3.1 Fonte de Dados 1: Portal da Transparência

As informações relativas aos empenhos do Poder Executivo Federal são oriundas do sistema SIAFI (assim como as demais informações relativas à execução da despesa), muito embora, o acesso a tal base é restrito. No entanto, apesar de tal restrição, a lei complementar 131 (BRASIL, 2009) passou a exigir que todos os entes federados disponibilizassem, em tempo real, todas as informações relativas à execução orçamentária.

Sendo assim, atualmente a União disponibiliza na Internet, com atualização diária, não só as informações de empenho, como também das fases de liquidação e pagamento, bem como os dados relativos à arrecadação de receita. Logo, uma das fontes de dados utilizadas nesse trabalho são as informações de empenho do governo federal, obtidas a partir do Portal da Transparência do Governo Federal[14].

Sendo assim, identificaram-se duas tabelas, oriundas do Portal da Transparência e utilizadas como base para o trabalho em questão: tabela de Empenhos e tabela de Itens de Empenho. Essas tabelas estão relacionadas, sendo que, uma nota de empenho corresponde a um registro na tabela de Empenhos e a um ou mais registros na tabela de Itens de Empenho. Ou seja, um empenho pode ter vários itens de empenho. Cabe ressaltar que todas as despesas possuem um empenho específico, porém, nem todas as despesas correspondem a compra de

[14] http://www.transparencia.gov.br/

materiais; visto que, existem outros tipos de gastos que não dizem respeito à compra de produtos, como por exemplo, contratação de serviços e pagamento de pessoal; e por isso não estão dentro do escopo dessa pesquisa. Dessa forma, existem empenhos que devem ser desconsiderados no momento do desenvolvimento da proposta.

A tabela de Empenhos traz informações gerais da Nota de Empenho, como por exemplo: quem está executando aquele gasto, quem será o favorecido da despesa, a data de emissão do empenho, a modalidade de licitação utilizada para selecionar o fornecedor, entre outros. Já a tabela de Itens de Empenho traz informações relativas a cada um dos itens que estão sendo comprados. Dessa forma, caso um determinado órgão esteja comprando, de um mesmo fornecedor cinco itens distintos, cada um desses itens corresponderá a uma linha na tabela de itens de empenho. Na Figura 6 é apresentada uma tela de empenho do Portal da Transparência do Governo Federal, com a indicação das tabelas fontes de cada uma das informações apresentadas.

Figura 6 – Tela de Empenho do Portal da Transparência do Governo Federal

Data:	19/09/2016
Tipo de Empenho:	ORDINARIO **Espécie de Empenho:** Original
Órgão Superior:	39000 - MINISTERIO DOS TRANSPORTES
Órgão / Entidade Vinculada:	39252 - DEPTO.NAC.DE INFRA+ESTRUT.DE TRANSPORTES-DNIT
Unidade Gestora Emitente:	393020 - SUPERINTENDENCIA REG. NO ESTADO MT - DNIT
Gestão:	39252 - DEPTO. NAC. DE INFRA+ESTRUTURA DE TRANSPORTES
Favorecido:	05.383.313/0001-90 - NOGUEIRA NOBRE COMERCIO E SERVICOS LTDA - ME
Valor:	R$ 6.549,90

DADOS DETALHADOS

Observação do Documento:	EMPENHO QUE SE EFETIVA PARA COBRIR DESPESAS COM MATERIAL DE EXPEDIENTE PARA ATENDER AS NECESSIDADES DA SR/MT. PROC ORIGEM: 2016PR00280
Esfera:	1 - ORÇAMENTO FISCAL **Tipo de Crédito:** A - INICIAL (LOA)
Grupo da Fonte de Recursos:	1 - RECURSOS DO TESOURO – EXERCÍCIO CORRENTE
Fonte de Recursos:	00 - RECURSOS ORDINARIOS
Unidade Orçamentária:	39252 - DEPTO.NAC.DE INFRA+ESTRUT.DE TRANSPORTES-DNIT
Funcional Programática	
Função:	26 - TRANSPORTE
Subfunção:	122 - ADMINISTRACAO GERAL
Programa:	2126 - PROGRAMA DE GESTAO E MANUTENCAO DO MINISTERIO DOS TRANSPORTE
Ação:	2000 - ADMINISTRACAO DA UNIDADE **Linguagem Cidadã:** Administração de unidade
Subtítulo (localizador):	0001 - ADMINISTRACAO DA UNIDADE - NACIONAL
Plano Orçamentário - PO:	0000 - ADMINISTRACAO DA UNIDADE **Autor da Emenda:** SEM EMENDA
Categoria de Despesa:	3 - Despesas Correntes **Grupo de Despesa:** 3 - Outras Despesas Correntes
Modalidade de Aplicação:	90 - Aplic. Diretas (Gastos Diretos do Governo Federal)
Elemento de Despesa:	30 - MATERIAL DE CONSUMO
Processo Nº:	50611003665201621
Modalidade de Licitação:	PREGAO **Inciso:** **Amparo:**
Referência da Dispensa ou Inexigibilidade:	
Nº Convênio / Contrato de Repasse / Termo de Parceria / Outros:	

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
16 - MATERIAL DE EXPEDIENTE	50	2,80	140,00	50,00000 UNIDADE ALMOFADA CARIMBO, MATERIAL CAIXA PLÁSTICO/METAL, TAMANHO MÉDIO, COR AZUL, TIPOENTINTAMENTO PERMANENTE, COMPRIMENTO 12 CM, LARGURA 9 CM MARCA: RADEX ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000251047
16 - MATERIAL DE EXPEDIENTE	10	13,70	137,00	10,00000 UNIDADE COLA, COR BRANCA, APLICAÇÃO PAPEL, CARACTERÍSTICAS ADICIONAIS ATÓXICA, TIPO BASTÃO MARCA: LEONORA ITEM DO PROCESSO: 00002 ITEM DE MATERIAL: 000292447
16 - MATERIAL DE EXPEDIENTE	10	15,40	154,00	10,00000 UNIDADE COLA, COMPOSIÇÃO POLIVINIL ACETATO- PVA, COR BRANCA, APLICAÇÃO PAPEL, TIPO PASTOSA MARCA: FRAMA ITEM DO PROCESSO: 00003 ITEM DE MATERIAL: 000282967
16 - MATERIAL DE EXPEDIENTE	5	7,80	39,00	5,00000 UNIDADE COLA, COMPOSIÇÃO CIANIACRILATO, COR INCOLOR, APLICAÇÃO VIDRO,BORRACHA,PLÁSTICO,PVC,METAL,ACRÍLICO,NÁILON, CARACTERÍSTICAS ADICIONAIS GEL, TIPO INSTANTÂNEA MARCA: TEKBOND ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000281629

Fonte:
Tabela de Empenho

Fonte:
Tabela de Itens de Empenho

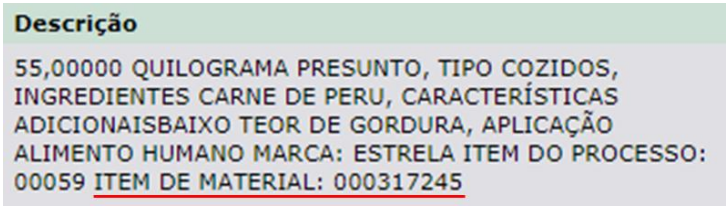
Fonte: o autor, adaptado do Portal da Transparência do Governo Federal

Para se realizar o estudo, fez-se necessária a definição de um limite temporal dos dados a serem analisados. Dessa forma, optou-se por utilizar-se os dados das notas de empenhos referentes aos meses de janeiro a maio de 2017, o que totalizou 1.012.109 registros na tabela de Empenhos e 1.562.223 registros na tabela de Itens de Empenho.

4.2.3.2 Fonte de Dados 2: Portal de Dados Abertos

Analisando de forma mais detalhada a tabela de Itens de Empenho, percebe-se que, apesar do campo descrição do item de empenho apresentar informações textuais, algumas informações estruturadas[15] podem ser extraídas desse campo. Dentre essas informações estruturadas está a informação de Item de Material, conforme pode ser observado na Figura 7.

Figura 7 – Descrição de um Item de Empenho



Fonte: o autor, adaptado do Portal da Transparência do Governo Federal

O Item de Material é uma informação que é incorporada nesse campo ainda na fase de licitação[16]. Essa informação corresponde ao código do material ou serviço a ser comprado. Apesar desse código não ser muito informativo, existe uma tabela no sistema SIASG, denominada tabela de Materiais e Serviços, que traz a descrição de cada um desses códigos. Dessa forma esse código será extraído (durante a fase de pré-processamento) de todas os itens de empenho analisados, a fim de auxiliar na identificação de cada um dos itens que estão sendo especificados no campo descrição da tabela de Itens de Empenho.

---

[15] Informação estruturada é a informação codificada, sistematizada, dentro de uma estrutura pré-estabelecida. Este tipo de informação geralmente é apresentada em combinação com outras variáveis a fim de permitir a leitura de uma determinada situação (FALSARELLA; JANNUZZI; BERAQUET, 2012).

[16] Apesar da Nota de Empenho ser emitida no sistema SIAFI, as compras realizadas pela Administração Pública Federal devem passar anteriormente pelo sistema SIASG, que trata das licitações e contratos administrativos. Dessa forma, ainda no sistema SIASG são feitos pré-empenhos que posteriormente são passados para o SIAFI, a fim de se emitir a Nota de Empenho. Durante a edição do pré-empenho essa informação do ITEM DE MATERIAL é inserida no texto da descrição do item de empenho.

Logo, também será necessária a utilização da tabela de Materiais e Serviços do Sistema SIASG. O banco de dados desse sistema está disponível no Portal de Dados Abertos do Governo Federal[17]. Sendo assim, essa será outra fonte de dados a ser utilizada na solução do problema proposto.

4.2.3.3 Fonte de dados 3: Aplicativos e Sites de Comparação de Preços

As informações apresentadas nas fontes de dados anteriores (Fontes de dados 1 e 2: Portal da Transparência do Governo Federal e Sistema SIASG) possibilitam a análise das compras que são realizadas pelo Poder Executivo Federal, porém, não fornecem nenhum tipo de informação que permita a comparação dessas compras com o preço praticado pelo mercado, a fim de se verificar se os preços pagos pela Administração Pública são melhores ou piores que os valores cobrados pelo mercado.

Sendo assim, tentou-se encontrar uma forma de se obter uma relação de produtos com seus preços associados, a fim de se ter algum tipo de subsídio para se comparar com os preços pagos pela Administração Pública.

No entanto, verificou-se que não existe um índice oficial de preços disponível para ser utilizado na pesquisa em questão. Dessa forma, decidiu-se buscar por aplicativos utilizados em telefones celulares e sites na internet que se propunham a auxiliar os consumidores a verificar se os produtos que eles pretendiam comprar estavam com preços bons ou não.

Durante a busca, foram analisadas algumas soluções disponíveis nas lojas virtuais dos dois sistemas operacionais para smartphones mais utilizados no momento (Android, desenvolvido pelo Google, e IOS, desenvolvido pela Apple) e alguns sites disponíveis na internet que tinham o objetivo de fazer comparações de preços. Dentre os aplicativos analisados, selecionou-se o aplicativo “Quanto Custa”, disponível para ambas as plataformas pesquisadas (Android e IOS). Selecionou-se também o site de comparação de preços Buscapé[18] e um site de cotação de produtos agrícolas[19].

Sendo assim, esse trabalho fez uso das informações disponibilizadas por meio desse aplicativo para smartphones e dos sites citados a fim de se obter os preços

[17] http://dados.gov.br
[18] http://www.buscape.com.br
[19] https://www.noticiasagricolas.com.br/cotacoes

cobrados pelos principais estabelecimentos comerciais para formular uma base de preço a ser utilizada como referência na comparação com os preços pagos pela Administração Pública. Cabe ressaltar que os dados coletados se referiam ao mês de julho de 2017.

4.2.3.4 Integrando as Bases de Dados

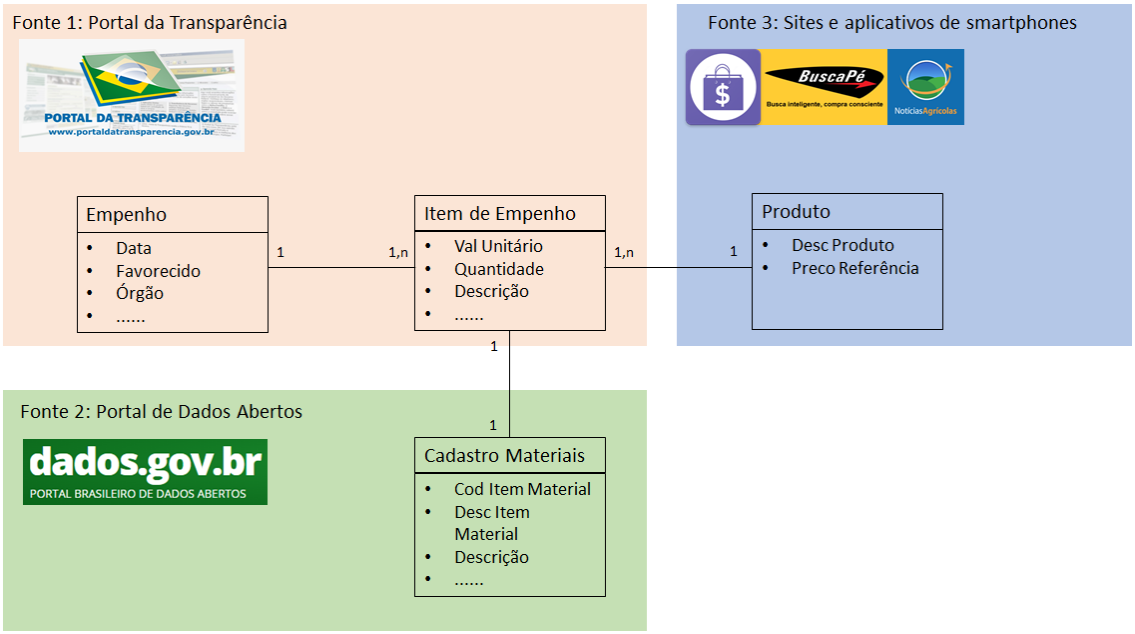
Todas as informações capturadas, das diversas fontes, devem ser armazenadas em um banco de dados central. Apesar das informações serem oriundas de diferentes origens, esses dados devem ser tratados de maneira harmônica, a fim de se poder extrair o máximo de conhecimento. Dessa forma, na Figura 8 é ilustrado um modelo de dados resumido com as principais tabelas utilizadas, bem como as origens de tais dados e os relacionamentos existentes entre eles.

Sendo assim, no modelo de dados apresentado na Figura 8, cada tabela é representada por um retângulo, sendo que, na parte superior do retângulo coloca-se o nome da tabela e na parte inferior desses retângulos estão as principais colunas[20] dessas tabelas. Já o relacionamento existente entre duas tabelas é representado por uma linha contínua, que interliga essas duas tabelas. Por exemplo, as tabelas de Empenho e de Itens de Empenho estão relacionadas, e por essa razão existe uma linha interligando-as. Os símbolos (números e letras) que acompanham a representação do relacionamento indicam a cardinalidade desse relacionamento. Logo, no caso da ligação existente entre as tabelas citadas (Empenho e Itens de Empenho), a tabela de Empenho possui cardinalidade 1 e a tabela de Itens de Empenho possui cardinalidade (1,n), ou seja, um registro na tabela de Empenho pode possuir um ou mais registros correspondentes na tabela de Itens de Empenho.

Cabe ressaltar que, o relacionamento existente entre tabelas de fontes de dados diferentes não é explícito, sendo que, tais relacionamentos começam a ser identificados na fase de estudo das regras de negócio e completa-se nessa fase de captura, armazenamento e estudo das bases. Nas fases subsequentes da metodologia deve-se fazer os processamentos necessários para tornar tais relacionamentos viáveis.

[20] Optou-se pela representação apenas das colunas principais (e não de todas) para facilitar a apresentação de forma concisa dessas informações.

Figura 8 – Modelo de Dados Integrado, com as diferentes Fontes Utilizadas



Fonte: o autor

4.2.4 Pré-processamento dos Dados

O fato da maioria dos dados serem oriundos de banco de dados coorporativos que apresentam dados com boa qualidade (ou seja, sem valores faltantes ou preenchidos com dados incorretos) facilita a atividade de pré-processamento.

Dessa forma, as atividades dessa fase ficaram concentradas no campo descrição do item de empenho, que conforme citado anteriormente, armazena a descrição textual da especificação do item que está sendo adquirido. Sendo assim, a etapa de pré-processamento retira informações que estão presentes no campo de descrição da compra, mas que não fazem parte da especificação textual do produto.

Uma dessas informações presentes é o código do item de material. Essa informação é oriunda de outro sistema (Sistema SIASG, usado para o controle das licitações e contratos) que auxilia na identificação do produto que está sendo comprado.

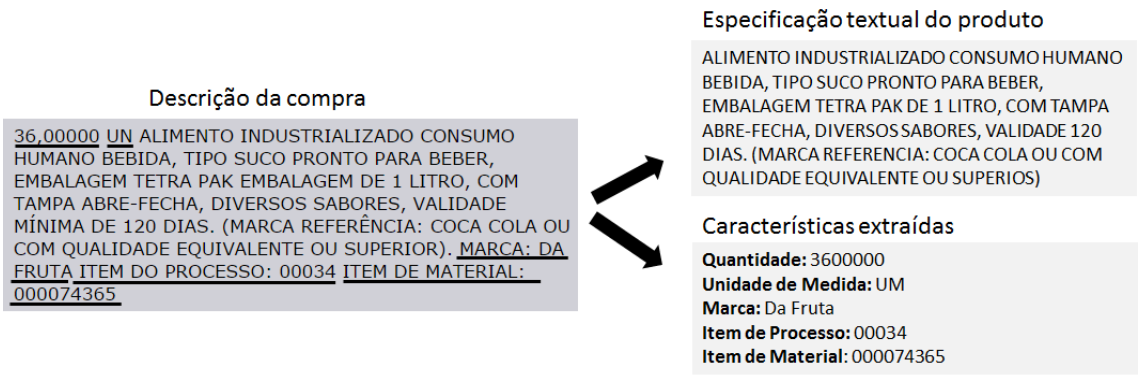
Logo, durante o pré-processamento extrai-se essa informação do campo textual, para que tal dado (código do item de material) possa ser cruzado com a tabela de Cadastro de Material e Serviço do sistema SIASG a fim de se recuperar a descrição do referido código. Para essa extração, formulou-se um processo que percorre todo o

campo de descrição, da tabela de Itens de Empenho buscando-se a expressão “ITEM DE MATERIAL”. Uma vez encontrada essa expressão, extrai-se os 9 dígitos seguintes (que indica o código do material, cuja descrição encontra-se no sistema SIASG).

Além disso, utilizou-se essa fase para extraírem-se outras características das compras que também estavam descritas nesse campo textual e que apesar de não serem utilizadas para a integração com outras fontes de dados, servem para agregar mais conhecimento a respeito de tais compras. Essas informações são a quantidade do que está sendo comprado, a unidade de medida utilizada para a quantificação do produto, a marca e o Item de processo.

Na Figura 9 é apresentado o resultado do pré-processamento de uma descrição de compra. Dessa forma, o texto que realmente especifica o produto é reduzido, enquanto que, algumas características específicas de cada compra são identificadas e retiradas do campo textual, para que, ao final de todo o processo, elas possam ser incorporadas aos produtos identificados, enriquecendo assim o conhecimento relativo às compras governamentais.

Figura 9 – Exemplo da Atividade de Pré-processamento



Fonte: o autor

4.2.5 Análise dos Dados

Após o pré-processamento dos dados, o passo seguinte é o de análise de dados. Nessa fase os dados são manipulados de forma a se extrair conhecimentos implícitos no conjunto de dados disponíveis.

4.2.5.1 Identificação dos Produtos e Preços Pagos pela Administração Pública

O objetivo do trabalho é analisar as compras que são feitas pelo Poder Executivo Federal, sendo que, para isso, optou-se pela utilização das informações da fase de empenho, apresentadas no Portal da Transparência do Governo Federal, que são armazenadas nas tabelas de Empenho e de Itens de Empenho. Para atingir esse objetivo, a primeira tarefa a ser executada é descobrir a que produto cada um dos itens de empenho analisados se refere, visto que, cada item de empenho deve corresponder a um produto específico.

Para isso, utilizaram-se as informações presentes no campo descrição da tabela de Itens de Empenho. Nessa tarefa, faz-se necessária o emprego de uma técnica de mineração de texto, sendo que, nesse estudo de caso, optou-se por uma técnica introduzida por Paiva e Revoredo (2016), que propõe a utilização de combinações de palavras chaves e o emprego do paradigma de programação Mapreduce , com a utilização da ferramenta Hadoop .

Logo, durante essa quinta fase da metodologia proposta são analisadas as descrições dos itens de empenho. O objetivo dessa atividade é fazer a identificação dos produtos que estão sendo descritos nas especificações das compras. Essa identificação se dá através da aplicação de regras de identificação, estabelecidas por especialistas. Tais regras combinam dados presentes nos campos estruturados e informações contidas no campo textual (descrição do item).

Além da combinação de palavras chaves, utilizou-se também, para auxiliar na identificação dos produtos que estão sendo especificados nas descrições textuais, os códigos de material e serviço, assim como as suas respectivas descrições (oriundas do sistema SIASG).

Ao final do processo obtém-se, uma relação com todas as compras cujos produtos adquiridos foram identificados, sendo que, para cada umas dessas compras computou-se as seguintes informações:

- Produto – nome do produto comprado.
- Valor - valor unitário pago pelo produto. Esse campo será utilizado para o cálculo dos valores estatísticos dos produtos.
- Nº do Empenho – identificação do número da nota de empenho do produto comprado.

- Nº Sequencial – ordem sequencial, dentro da nota de empenho, do item de empenho que contém o produto identificado.

Na Tabela 1 são apresentados alguns exemplos de saídas dessa primeira etapa da análise de dados.

Tabela 1 – Exemplo do Resultado da 1ª Parte do Processo de Análise de Dados

produto	Valor(R$)	Nº do Empenho	seq
Grapo p Grampeador - cx 5000 un	3,69	160171000012017NE800290	4
Grapo p Grampeador - cx 5000 un	8,65	160400000012017NE800289	1
Grapo p Grampeador - cx 5000 un	9,90	160296000012017NE800223	1
Laranja - Kg	0,84	153166152402017NE800566	4
Laranja - Kg	0,95	160088000012017NE800610	1
Laranja - Kg	0,95	167064000012017NE800027	1

Fonte: autor

Para efeitos desse estudo, aplicou-se a técnica de mineração textual citada para se fazer a identificação de um conjunto de 11 produtos distintos. No entanto, novos produtos podem ser acrescentados a essa lista. Dessa forma, para cada um dos itens de empenho que se referiam a um desses 11 produtos designados, calcularam-se algumas métricas a partir do número de compras consideradas (do referido produto) e do valor unitário pago em cada uma das compras identificadas.

Na Tabela 2 são apresentados os produtos selecionados, bem como algumas métricas relacionadas aos preços pagos por tais produtos.

Tabela 2 – Principais Métricas dos Produtos Identificados

Produtos	Qtd de compras	Val Max (R$)	Média (R$)	Mediana (preço de referência – R$)
Agua Mineral - 20l	300	37,05	5,57	7,77
Agua Mineral - 500 ml	38	1,93	0,82	0,60
Batata – Kg	260	8,45	3,08	2,93
Caneta Esferografia - cx 50 um	38	24,75	17,48	18,00
Cebola – Kg	225	19,67	3,20	2,64
Cenoura – Kg	283	13,85	2,86	2,32
Grapo p Grampeador - cx 1000 um	43	10,00	3,48	3,50
Grapo p Grampeador - cx 5000 um	76	22,83	6,42	5,33
Laranja – Kg	381	8,00	2,41	1,99
Papel A4 Branco – resma	154	74,00	22,87	17,22
Presunto – Kg	376	39,60	16,27	15,09

Fonte: autor

4.2.5.2 Identificação dos Preços de Mercado dos Produtos Analisados

Para o cálculo dos preços de mercado dos produtos estudados utilizou-se a base de dados do aplicativo para smartphones chamado Quanto Custa e dos sites Buscapé e de Cotação de produtos agrícolas (conforme citado na subseção 3.3). Para cada um dos produtos pesquisados, levantaram-se todas as ofertas desses produtos disponíveis nas bases de dados utilizadas, e seus respectivos preços.

A partir dessas informações, o passo seguinte é o cálculo de um preço de referência para cada um dos produtos considerados. A exemplo da estratégia utilizada para o cálculo do preço de referência das compras feitas pela Administração Pública, adotou-se como o preço de referência praticado pelo mercado a mediana dos preços dos produtos considerados, tomando-se por base as ofertas apresentadas nos aplicativos e sites utilizados para esse fim. Dessa forma, na tabela 3, são apresentados os preços de referência computados para os 11 produtos analisados.

Tabela 3 – Preços de Referência Praticados pelo Mercado

Produtos	**Preço** (R$)
Agua Mineral - 20l	8,50
Agua Mineral - 500 ml	1,06
Batata – Kg	2,99
Caneta Esferografia - cx 50 um	28,95
Cebola – Kg	3,24
Cenoura - Kg	2,29
Grapo p Grampeador - cx 1000 un	6,42
Grapo p Grampeador - cx 5000 un	8,92
Laranja - Kg	2,78
Papel A4 Branco - resma	18,70
Presunto - Kg	20,16

Fonte: o autor

4.2.6 Avaliação

Após a descoberta de novos conhecimentos obtidos a partir da análise dos conjuntos de dados estudados, o último passo da metodologia é a avaliação dos resultados obtidos.

Dessa forma, juntaram-se os conhecimentos extraídos, a fim de se avaliar esses resultados. Sendo assim, na Tabela 4 são apresentados os preços praticados pela Administração Pública Federal e os preços praticados pelo mercado para o conjunto de produtos analisados, lembrando que, em ambos os casos considerou-se como preço praticado a mediana dos preços dos produtos considerados.

Tabela 4 – Comparação entre os Preços de Mercado e os Valores Pagos pela Administração Púbica

Produtos	**Preços Praticados (Em R$)**		**Diferençade preços praticados pelo Mercado e pela Adm. Pública**	
	Pelo Mercado	**Pela Adm. Pública**	**Em (R$)**	**Em (%)**
Agua Mineral - 20l	8,50	7,77	0,73	8,59
Agua Mineral - 500 ml	1,06	0,60	0,46	43,40
Batata – Kg	2,99	2,93	0,06	2,01
Caneta Esferografia - cx 50 um	28,95	18,00	10,95	37,82
Cebola – Kg	3,24	2,64	0,6	18,52
Cenoura - Kg	2,29	2,32	-0,03	-1,31
Grapo p Grampeador - cx 1000 un	6,42	3,50	2,92	45,48
Grapo p Grampeador - cx 5000 un	8,92	5,33	1,43	26,83
Laranja - Kg	2,78	1,99	0,79	28,42
Papel A4 Branco - resma	18,70	17,22	1,48	7,91
Presunto - Kg	20,16	15,09	5,07	25,15

Fonte: o autor

Analisando-se a Tabela 4, percebe-se que os preços praticados pela Administração Pública tendem a ser mais baixos que os valores cobrados pelo mercado em geral, sendo que, a única situação em que o governo estava praticando um preço maior do que o levantado no mercado foi na compra da cenoura, situação essa em que o governo pagou 3 centavos a mais no preço do quilo da cenoura, o que equivale a um valor 1,31% mais caro.

Para todos os outros valores constatou-se que o governo comprou os produtos a um preço médio abaixo do preço praticado pelo mercado, chegando a uma situação extrema em que foi pago um valor 45,48% mais baixo do que ao preço de mercado na caixa de caneta esferográfica com 50 unidades.

No entanto, é importante se ter em mente que essas comparações foram feitas apenas para se responder à pergunta inicialmente formulada, que tinha um objetivo de apenas traçar um diagnóstico geral das compras realizadas pelo poder público quando comparadas com os preços de mercado. Pois tanto os valores praticados pela Administração Pública quanto pelo mercado, careceriam de técnicas mais apuradas e especializadas para os seus cálculos, levando-se em conta outros

atributos das compras como por exemplo, região do país, marca e outras especificidades dos produtos, o que fugiria dos propósitos desse estudo.

Uma razão que pode justificar a diferença de preços para menor nas compras feitas pela Administração Pública em relação aos preços de mercado é o fato das compras públicas geralmente serem precedidas de procedimentos licitatórios, o que estimula a concorrência entre os fornecedores e acaba diminuindo o preço a ser pago pelos produtos. No entanto, o objetivo desse estudo de caso é apena responder as perguntas: “Quanto o governo paga nas compras que faz? Será que o governo paga mais ou menos que o preço praticado no mercado?”, sem ser preocupar com as condicionantes que possam ter motivado a constatação levantada.

Sendo assim, após a análise dos dados tratados constatou-se que a Administração Pública, via de regra, faz suas compras pagando valores abaixo ou igual ao preço de mercado, desconstruindo-se assim a idéia de que, em geral, o governo costuma realizar suas compras de forma superfaturada.

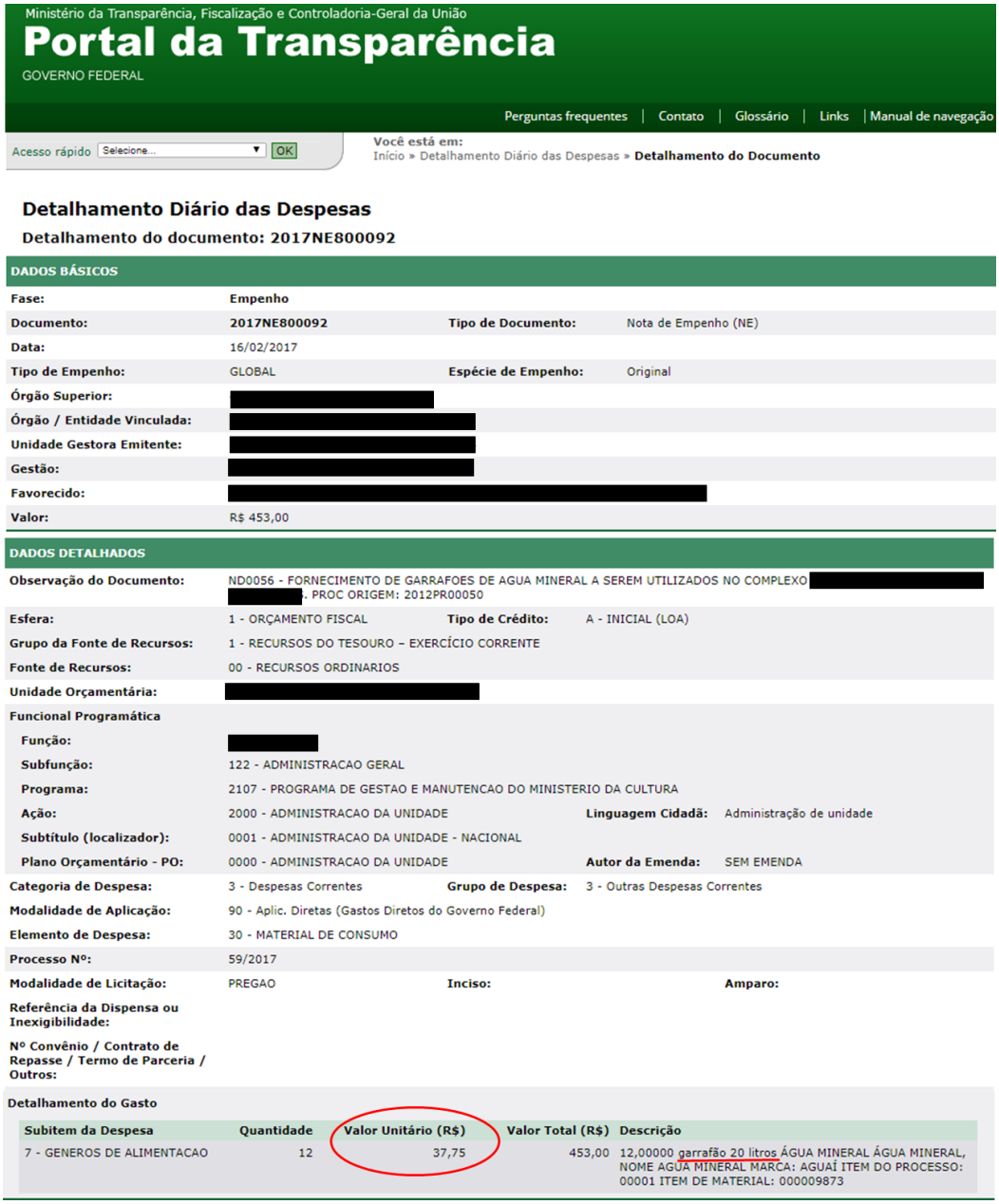
No entanto, é importante ressaltar que essa conclusão foi obtida a partir de valores médios, ou seja, a grande maioria das Compras feitas pela Administração Pública atende ao princípio da economicidade, porém, não se pode esquecer que existem casos extremos que carecem de análises mais aprofundadas.

Verificando-se a coluna que apresenta os valores máximos pagos nas compras realizadas pelo Governo Federal, na Tabela 2, podem-se constatar valores extremamente altos, que evidenciam disparidades que carecem de análises mais detalhadas. A título de exemplo, analisou-se o caso do Valor máximo pago na compra de galão de 20 litros de Água Mineral, apresentado na primeira linha da Tabela 2. Nesse caso, constatou-se que foi pago um valor de R$ 37,75 na compra na compra do galão de água, enquanto que o preço praticado pela Administração Pública, de acordo com a metodologia adotada, era de R$ 7,77. Sendo assim, verificou-se o empenho dessa compra e constatou-se que realmente a referida compra realmente era de galão de 20 litros de água mineral e que o valor unitário pago por esse produto realmente foi R$ 37,75, evidenciando que a identificação do produto comprado ocorreu de maneira correta, e que o preço pago nessa compra realmente está muito acima do valor esperado.

Na Figura 10, é apresentada a Nota de Empenho dessa compra citada, sendo que, todas as informações que pudessem identificar a unidade que executou esse gasto, bem como a empresa que realizou essa venda foram tarjadas para que não

houvesse exposição das partes envolvidas, visto que os propósitos desse trabalho são meramente acadêmicos, e que qualquer afirmação sobre o processo de compra em questão careceria de estudos mais aprofundadas que iriam além da simples análise de dados. No entanto esse exemplo foi apresentado para ilustrar a possibilidade de se identificar compras que fujam de um determinado padrão de normalidade, e que apesar das compras governamentais em geral serem realizadas a preços satisfatórios para a Administração Pública, existem situações extremas que merecem ser investigadas.

Figura 10 – Tela do Portal da Transparência com um Exemplo de Compra com Preço Discrepante



Ministério da Transparência, Fiscalização e Controladoria-Geral da União

Portal da Transparência

GOVERNO FEDERAL

Perguntas frequentes | Contato | Glossário | Links | Manual de navegação

Acesso rápido Selecione... OK

Você está em: Início » Detalhamento Diário das Despesas » Detalhamento do Documento

Detalhamento Diário das Despesas

Detalhamento do documento: 2017NE800092

DADOS BÁSICOS

Fase:	**Empenho**		
Documento:	**2017NE800092**	**Tipo de Documento:**	Nota de Empenho (NE)
Data:	16/02/2017		
Tipo de Empenho:	GLOBAL	**Espécie de Empenho:**	Original
Órgão Superior:	[redacted]		
Órgão / Entidade Vinculada:	[redacted]		
Unidade Gestora Emitente:	[redacted]		
Gestão:	[redacted]		
Favorecido:	[redacted]		
Valor:	R$ 453,00		

DADOS DETALHADOS

Observação do Documento:	ND0056 - FORNECIMENTO DE GARRAFOES DE AGUA MINERAL A SEREM UTILIZADOS NO COMPLEXO [redacted]. PROC ORIGEM: 2012PR00050		
Esfera:	1 - ORÇAMENTO FISCAL	**Tipo de Crédito:**	A - INICIAL (LOA)
Grupo da Fonte de Recursos:	1 - RECURSOS DO TESOURO – EXERCÍCIO CORRENTE		
Fonte de Recursos:	00 - RECURSOS ORDINARIOS		
Unidade Orçamentária:	[redacted]		
Funcional Programática			
Função:	[redacted]		
Subfunção:	122 - ADMINISTRACAO GERAL		
Programa:	2107 - PROGRAMA DE GESTAO E MANUTENCAO DO MINISTERIO DA CULTURA		
Ação:	2000 - ADMINISTRACAO DA UNIDADE	**Linguagem Cidadã:**	Administração de unidade
Subtítulo (localizador):	0001 - ADMINISTRACAO DA UNIDADE - NACIONAL		
Plano Orçamentário - PO:	0000 - ADMINISTRACAO DA UNIDADE	**Autor da Emenda:**	SEM EMENDA
Categoria de Despesa:	3 - Despesas Correntes	**Grupo de Despesa:**	3 - Outras Despesas Correntes
Modalidade de Aplicação:	90 - Aplic. Diretas (Gastos Diretos do Governo Federal)		
Elemento de Despesa:	30 - MATERIAL DE CONSUMO		
Processo Nº:	59/2017		
Modalidade de Licitação:	PREGAO	**Inciso:**	**Amparo:**
Referência da Dispensa ou Inexigibilidade:			
Nº Convênio / Contrato de Repasse / Termo de Parceria / Outros:			

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	12	37,75	453,00	12,00000 garrafão 20 litros ÁGUA MINERAL ÁGUA MINERAL, NOME AGUA MINERAL MARCA: AGUAÍ ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000009873

Fonte: Portal da Transparência do Governo Federal

4.3 CONSIDERAÇÕES SOBRE A VALIDAÇÃO

O objetivo desse capítulo foi verificar se a hipótese formulada: “Se forem identificadas as atividades necessária para a obtenção de conhecimento a partir de múltiplas fontes de dados e os relacionamentos existentes entre essas atividades, então, é possível se definir uma metodologia capaz de sistematizar um processo genérico de descoberta de conhecimento a partir dessas múltiplas fontes de dados” era verdadeira.

Sendo assim, de acordo com a hipótese citada, foram identificadas as principais atividades que deveriam fazer parte de um processo de obtenção de conhecimento a partir de várias bases de dados e seus respectivos relacionamentos, o que possibilitou a proposta da metodologia apresentada no capítulo 3 dessa monografia.

Por fim, com o intuito de verificar a aplicabilidade e a eficácia da metodologia proposta, foi desenvolvido um estudo de caso que tinha o objetivo de extrair conhecimento sobre uma atividade da Administração Pública, a partir da aplicação da metodologia proposta.

Os resultados apresentados demonstraram que a metodologia desenvolvida foi capaz de extrair conhecimentos que anteriormente estavam implícitos nas diversas bases de dados analisados, o que leva-nos a concluir que a hipótese formulada é verdadeira.

5 CONCLUSÃO

A principal contribuição deste trabalho é a proposta de uma metodologia para extração de conhecimentos a partir de diferentes bases de dados. No entanto, a pesquisa realizada também apresentou uma segunda contribuição, que foi a revisão bibliográfica, exposta no referencial teórico, que apresenta outros trabalhos que fazem uso de diferentes bases de dados a fim de resolver problemas relevantes da atualidade. Apesar desse levantamento bibliográfico não ser exaustivo, uma vez que, as pesquisas nessa área estão crescendo muito rapidamente, tornando inviável um levantamento completo a respeito de um tema tão genérico, ele serve de suporte para embasar novos estudos relacionados a esse assunto, e evidencia as diversas opções de emprego de técnicas de extração de conhecimento a partir de múltiplas bases de dados.

Esse levantamento também serviu para a percepção da inexistência de um guia que pudesse conduzir o analista de dados nas tarefas necessárias para a atividade de extração de conhecimento em bases de dados. Essa limitação serviu de motivação para os estudos que embasaram contribuição principal desse trabalho. Dessa forma, a principal contribuição dessa monografia é a proposta de uma metodologia capaz de guiar os analistas de dados na atividade de descoberta de conhecimento a partir da utilização das diferentes bases de dados que estão disponíveis para seu uso.

O desenvolvimento dos trabalhos para a formulação da metodologia sugerida começou pelo estudo de algumas soluções que também se propõem a sistematizar algum tipo de procedimento capaz de auxiliar no processo de obtenção de conhecimento em bancos de dados.

Dessa forma, foram identificados dois trabalhos (SHEARER, 2000) e MARQUESONE (2016), cujos objetivos eram guiar os analistas nos projetos de análise de dados. Shearer (2000) propõem o CRISP-DM (“Cross Industry Standard Process for Data Mining”), cujo objetivo é fornecer orientações básicas para a condução do processo de descoberta de conhecimentos em bancos de dados. Esse modelo está dividido em 6 fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. Essas fases definem os passos a serem seguidos em um projeto de mineração de dados. No entanto, uma limitação dessa metodologia é que ela parte da premissa de que os dados a serem

utilizados já estejam disponíveis para o uso, o que nem sempre é verdade. Outro limitante dessa abordagem é a não previsão de múltiplas fontes de dados no processo de descoberta de conhecimento.

No segundo trabalho, que também se propunha a guiar projetos de análise de dados, MARQUESONE (2016) enumera 4 passos para o desenvolvimento de projetos de Big Data: Identificação de perguntas, Captura e Armazenamento de Dados, Processamento e Análise de Dados e Visualização dos resultados. Um limitante da metodologia prevista por Marquesone (2016) é que ela não prevê a possibilidade de iteração entre suas fases, além de também não considerar possíveis reexecuções de atividades devido à identificação de novas necessidades levantadas em uma fase subsequente.

Sendo assim, esse trabalho combinou algumas das atividades previstas em cada um dos estudos analisados, a fim de propor uma nova metodologia que fosse capaz de suprir as carências levantadas, mantendo os aspectos positivos das pesquisas utilizadas como base.

Dessa forma, foi definido um conjunto de atividades necessárias para projetos que tenham o objetivo de extrair informações relevantes a partir de várias fontes de dados. Essas fases são Identificação de Perguntas; Estudo das Regras de Negócio; Captura, Armazenamento e Estudo dos Dados; Pré-processamento dos dados; Análise de Dados; e Avaliação.

Na fase de Identificação de Perguntas, são definidas as perguntas a serem respondidas no processo de análise de dados. Já na fase de Estudo das Regras de Negócio busca-se conhecer o contexto em que o projeto está inserido. Nessa fase são realizados os estudos das regras de negócio da área que está sendo pesquisada. A etapa de Captura, Armazenamento e Estudo dos Dados tem início com a identificação das possíveis fontes de dados que podem ser usadas para responder as questões levantadas. Ela também é responsável pelos estudos dos relacionamentos, explícitos e implícitos, existentes entre as diversas fontes de dados. Durante o Pré-processamento dos dados são realizadas as atividades de limpeza e integração dos dados oriundos das diferentes bases de dados, bem como outros tipos de tratamentos necessários à análise dos dados. Na fase de Análise de Dados são formuladas as respostas para as perguntas levantadas. Portanto, deve-se utilizar os dados pré-processados a fim de se extrair conhecimentos implícitos. Por fim, a Avaliação serve

para verificar se os resultados encontrados são consistentes e atendem aos objetivos estabelecidos.

A solução proposta e a hipótese formulada foram validadas pela execução de um estudo de caso que tem como objetivo analisar as compras executadas pela Administração Pública Federal. Dessa forma, o estudo percorreu os seis passos da metodologia, a fim de analisar as compras realizadas pelo Poder Executivo Federal. Sendo assim, ao final desse processo de validação verificou-se que a metodologia desenvolvida foi capaz de extrair conhecimentos que anteriormente estavam implícitos nas diversas bases de dados analisados, o que possibilitou a conclusão de que a metodologia desenvolvida era eficaz, e que hipótese formulada era verdadeira.

A principal limitação encontrada durante o desenvolvimento dessa pesquisa ocorreu durante o processo de validação da solução proposta. A metodologia foi validada a partir de um estudo de caso que buscava resolver uma questão relacionada à análise das compras que são realizadas pelo Poder Executivo Federal. No entanto, a inexistência de um resultado esperado, ou seja, de um gabarito que evidenciasse a que resultado o processo de descoberta de conhecimento deveria chegar, a fim de utilizá-lo como parâmetro de comparação para as respostas obtidas pelo processo de validação, dificultou a avaliação da solução proposta de uma forma mais objetiva.

Porém, como foi possível percorrer-se todas as fases da metodologia e como os resultados obtidos demonstram-se consistentes e coerentes, pode-se dizer que durante a validação não foram encontrados erros que pudessem refutar a validade da metodologia e da hipótese formulada.

Essa pesquisa permite o desenvolvimento de trabalhos futuros tanto relacionados a melhoria da metodologia proposta quanto trabalhos relacionados a aplicação de tal metodologia em outros conjuntos de dados e áreas de conhecimento.

Com relação à aplicação da metodologia desenvolvida, uma possível melhoria seria o desenvolvimento de outras metodologias específicas para a condução de cada uma das seis fases propostas, a fim de direcionar o analista de dados, não apenas na condução geral do processo de obtenção de conhecimento, mas também guiá-lo nas especificidades peculiares a cada uma das fases que compõem esse processo de descoberta de conhecimento.

Com relação às possibilidades de aplicação da metodologia em questão, existe uma grande variedade de conjuntos de dados que estão disponíveis, tanto na internet, quanto em outros repositórios de dados, públicos e privados, que podem ser

utilizados para atender demandas do mundo atual. Pode-se citar, por exemplo, a utilização dos dados de redes sociais para se identificar redes de relacionamentos existentes entre indivíduo. Esse tipo de conhecimento pode ser de extrema importância para a defesa nacional, principalmente se tal conhecimento for enriquecido com informações oriundas de outras fontes de dados.

A principal motivação para essa pesquisa foi a possibilidade de se dar outros usos a grande quantidade de dados que está sendo gerada atualmente, visto que, tais dados podem armazenar uma gama de conhecimentos implícitos, que poderiam ser utilizados para resolver problemas ligados a diversas áreas de conhecimento.

Apesar de já existirem uma série de aplicações que se propõem a explorar o potencial implícito nos grandes volumes de dados, a falta de uma metodologia que sistematize as atividades a serem executadas durante os procedimentos de descoberta de conhecimento acabavam dificultando o processo como um todo.

Sendo assim, essa pesquisa propôs uma metodologia capaz de guiar os analistas de dados nas principais tarefas que eles devem executar durante as análises e processamentos das diferentes bases de dados utilizadas como fontes para a obtenção de novos conhecimentos, tornando assim esses procedimentos mais efetivos.

REFERÊNCIAS

ARFAOUI, Nouha; AKAICHI, Jalel. Automating schema integration technique case study: generating data warehouse schema from data mart schemas. **Beyond Databases, Architectures and Structures**. Springer, 2015. p. 200–209. 3-319-18421-0.

BATINI, Carlo; LENZERINI, Maurizio; NAVATHE, Shamkant B. A comparative analysis of methodologies for database schema integration. **ACM computing surveys (CSUR)** v. 18, n. 4, p. 323–364, 1986.

BATISTA, Gustavo Enrique de Almeida Prado. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Universidade de São Paulo, 2003.

BELLSTRÖM, Peter. Bridging the gap between comparison and conforming the views in view integration. 2006, [S.l: s.n.], 2006. p.184–199.

BERNSTEIN, Philip A.; MELNIK, Sergey. Meta data management. 2004, [S.l.]: IEEE, 2004. p.875. 0-7695-2065-0.

BRASIL. Constituição Federal, de 5 de outubro de 1988. CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 5 out. 1988.Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>.Acesso em: 24 abr. 2017.

_____. Lei no 10.520, de 17 de julho de 2002. . Lei 10.520/2002. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 17 jul. 2002., 2002 . Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/2002/l10520.htm>. Acesso em: 16 jun. 2017.

_____. Lei no 8.666, de 21 de junho de 1993. . Normas para licitações e contratos da Administração Pública. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 21 jun. 1993., 1993 . Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L8666cons.htm>. Acesso em: 16 jun. 2017.

_____. Ministério da Defesa. Estratégia Nacional de Defesa. Brasília, DF, 2016a, p. 15-46. Versão sob apreciação do Congresso Nacional. Lei Complementar n. 97/1999, art. 9°, § 3°. Disponível em: <http://www.defesa.gov.br/arquivos/2017/mes03/pnd_end.pdf>. Acesso em: 30 mar. 2017a.

_____. Ministério da Defesa. **Livro Branco de Defesa Nacional**. Brasília, DF, 2016b. Versão sob apreciação do Congresso Nacional. Lei Complementar n. 97/1999, art. 9°, § 3°. Disponível em: <http://www.defesa.gov.br/arquivos/2017/mes03/livro_branco_de_defesa_nacional_minuta.pdf>. Acesso em: 30 mar. 2017b.

_____. Ministério da Defesa. **Política de Defesa Nacional**. Brasília, DF, 2016c, p. 03-14.. Versão sob apreciação do Congresso Nacional. Lei Complementar n. 97/1999, art. 9°, § 3°. Disponível em:

<http://www.defesa.gov.br/arquivos/2017/mes03/pnd_end.pdf>. Acesso em: 30 mar. 2017c.

DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: simplified data processing on large clusters. **Communications of the ACM** v. 51, n. 1, p. 107–113, 2008.

DOAN, AnHai; HALEVY, Alon; IVES, Zachary. **Principles of data integration**. Elsevier, 2012. .0-12-391479-5.

FALSARELLA, Orandi Mina; JANNUZZI, Celeste Aida Sirotheau Corrêa; BERAQUET, Vera Sílvia Marão. Informação empresarial: dos sistemas tradicionais à latência zero. **Transinformação-ISSNe 2318-0889** v. 15, n. 3 , 2012.

GANTZ, John; REINSEL, David. 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, n. 2012, p. 1–16, 2012.

INMON, William H. **Como Construir o Data Warehouse**. 2a. Edição. Rio de Janeiro: Campus, 1997.

JOSÉ, Fernanda Franco Dias. **Análise de grandes eventos na cidade do Rio de Janeiro usando dados de telefonia celular**. Rio de Janeiro - RJ, 2015.

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados**. 1. ed. São Paulo - SP: Casa do Código, 2016. .978-85-5519-231-9.

MICHAELIS, DICIONÁRIO. Disponível em <http://michaelis.uol.com.br/moderno/portugues/index.php>. Acesso em v. 24/04/2017.

Mineração de Dados | SAS. Disponível em: <https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html>. Acesso em: 24 abr. 2017.

NEHMY, Rosa Maria Quadros; PAIM, Isis. A desconstrução do conceito de" qualidade da informação". Ciência da Informação v. 27, n. 1, p. 0–0 , 1998.

O que é Big Data? | *SAS*. Disponível em: <https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html>. Acesso em: 24 abr. 2017.

PAIVA, Eduardo; REVOREDO, Kate. Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos. XII Simpósio Brasileiro de Sistemas de Informação, Florianópolis-SC, 2016 2016.

SAHOO, Prazam Kumar; MOHAPATRA, Suvendu Kumar; WU, Shih-Lin. Analyzing Healthcare Big Data With Prediction for Future Health Condition. **IEEE Access** v. 4, p. 9786–9799, 2016.

SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing** v. 5, n. 4, p. 13–22, 2000.

TAN, Pang-Ning. **Introduction to data mining**. Pearson Education India, 2006. .81-317-1472-1.