UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

BACHARELADO EM ENGENHARIA DE SOFTWARE

ANDRÉ LUIZ OLIVEIRA MARTINS

ANÁLISE DE DADOS ABERTOS DE DESPESAS E RECEITAS DOS ESTADOS E MUNICÍPIOS BRASILEIROS UTILIZANDO DATA MART

QUIXADÁ

2016

ANDRÉ LUIZ OLIVEIRA MARTINS

ANÁLISE DE DADOS ABERTOS DE DESPESAS E RECEITAS DOS ESTADOS E MUNICÍPIOS BRASILEIROS UTILIZANDO DATA MART

Trabalho de Conclusão de Curso apresentada à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel. Área de concentração: Computação.

Orientador: Profª. Ticiana Linhares Coelho da Silva

QUIXADÁ

2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

M341a Martins, André Luiz Oliveira
Análise de dados abertos de despesas e receitas dos estados e municípios brasileiros utilizando Data Mart/ André Luiz Oliveira Martins. – 2016.
40 f.: il. color., enc.; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2016.
Orientação: Profª. Me. Ticiana Linhares Coelho da Silva
Área de concentração: Banco de dados

1. Transparência na administração pública. 2. Banco de dados. 3. Informações eletrônicas governamentais. 4. Visualização de informação. I. Universidade Federal do Ceará (Campus Quixadá). II. Título.

CDD 005.75

ANÁLISE DE DADOS ABERTOS DE DESPESAS E RECEITAS DOS ESTADOS E MUNICÍPIOS BRASILEIROS UTILIZANDO DATA MART

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharelado. Área de concentração: Computação.

Aprovada em: ____/____/_____.

BANCA EXAMINADORA

_____
Prof. Me. Ticiana Linhares Coelho da Silva
(Orientador) Universidade Federal do Ceará (UFC)

_____
Prof. Me. Regis Pires Magalhães
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Flávio Rubens de Carvalho Sousa
Universidade Federal do Ceará (UFC)

A Deus.

Aos meus pais, Amigos e Professores.

AGRADECIMENTOS

Aos meus pais, Márcio Greyck e Maria Dilma, que mesmo na simplicidade se esforçaram para me dar a melhor educação, sempre me motivando a trilhar bons caminhos.

Agradeço especialmente a minha orientadora Ticiana Linhares, pelo conhecimento e oportunidades que adquiri como aluno dela, engrandecendo minha vida acadêmica e me motivando a ser um grande profissional.

Aos professores participantes da banca examinadora Regis Pires Magalhães e Flávio Rubens de Carvalho Sousa pelo tempo, pelas valiosas colaborações e sugestões.

Aos colegas da turma de graduação, pelas reflexões, críticas e sugestões recebidas.

“Não desista nas primeiras tentativas, a persistência é amiga da conquista. Se você quer chegar aonde a maioria não chega, faça o que a maioria não faz. ”

(William Bill Gates)

RESUMO

A informação referente às receitas e despesas do governo é disponibilizada de forma ilegível para a maior parte da população, muitas dessas informações em formatos desconhecido pela grande maioria, o que acaba por dificultar o entendimento e interpretação da informação. Visando implantar uma forma de apresentação dos dados governamentais do Brasil de forma légivel para todos, este trabalho propõe a construção de um *Data Mart* através da ferramenta *Pentaho* para fornecer legibilidade aos dados dos recursos financeiros dos governos dos estados e municípios brasileiros, dessa forma consistindo em prover acesso, através de um website, à população que não possui o conhecimento necessário para visualizar, de forma intuitiva, as informações disponibilizadas pelo governo.

Palavras chave: Legibilidade, Análise, Dados Abertos.

ABSTRACT

Information related to government revenues and expenditures is available illegibly for most of the population, much of this information in formats unknown by most, which makes it difficult to understand and interpret the information. Aiming to implement a form of presentation of government data in Brazil legibly for all, this paper proposes the construction of a Data Mart through Pentaho tool to provide clarity to the data of the financial resources of state governments and municipalities, thus consisting of provide access, through a website, the population that does not have the knowledge to see, intuitively, the information provided by the government.

Keywords: Legibility. Analysis. Open Data.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras *(*OPENDATAHANDBOOK, 2012*).*

A informação referente às receitas e despesas do governo é disponibilizada de forma ilegível para a maior parte da população, muitas dessas informações em formato CSV, o que dificulta o entendimento para aqueles que não compreendem esse tipo de arquivo, somente aqueles com conhecimentos suficientes para interpretar os dados conseguem obter algum entendimento.

A população com interesse nas informações dos gastos públicos é aquela que exerce de seus direitos de voto em sua maioria, interessada assim em que direção o dinheiro de seus impostos pagos foram direcionados. Governadores do país também são grandes interessados, pois estes estão preocupados em disponibilizar os dados para a população a fim de manter a transparência entre governo e população.

Visando implantar uma forma de apresentação dos dados governamentais do Brasil, este trabalho propõe a construção de um *Data Mart* através da ferramenta *Pentaho*[1] para fornecer legibilidade aos dados dos recursos financeiros dos governos dos estados e municípios brasileiros. Foi selecionada a ferramenta *Pentaho* como principal meio de gerar informação legível devido ao fato dele ser um software de código aberto. *Pentaho* é um software de código aberto para B.I. (*Business Inteligence*) que engloba áreas de ETL *(Extraction, Transformation and Load)*, *Reporting*, OLAP(*Online Analytical Processing*) e mineração de dados (PENTAHO, 2005).

A importância deste trabalho se dá devido à necessidade da população em tomar conhecimento sobre as informações referentes às despesas e receitas do governo, seja ele estadual ou municipal, para que possam visualizar onde os recursos estão sendo investidos dentro do país, quais áreas estão sendo priorizadas, quais estados e municípios recebem mais ou menos receitas de acordo com o ano.

As maiores contribuições deste trabalho consistem em prover acesso, através de um website, à população que não possui o conhecimento necessário para visualizar, de forma intuitiva, as informações disponibilizadas pelo governo e apresentar ao longo dos anos quais áreas do âmbito governamental estavam sendo priorizadas.

---

[1] http://www.pentaho.com/

Na literatura, é possível encontrar diversos trabalhos que propõem *Data Mart* com dados abertos e inclusive dados privados. Em Sousa (2014), é proposto um modelo de *Data Mart*e para a realização de uma análise dos dados abertos do PROCON, utilizando algumas técnicas de extração de conhecimento. Em Bico et al.(2012), é desenvolvida uma aplicação web para apresentar os dados da Câmara Municipal de São Paulo no intuito de gerar visualizações gráficas, tornando assim mais democrática a iniciativa a abertura de dados. No caso de Akintola et al. (2011), é relatado o desenvolvimento de um *Data Mart* para gestão de um ambiente acadêmico e para isso utiliza-se de tecnicas de mineração de dados usando-se ferramentas como *Microsoft SQL Server Analysis Services*. Outro trabalho relacionado é o relato de experiência da empresa Wal-Mart em Ohlinger (2006), que relata a construção de um *Data Warehouse* para prover informações acerca de seus clientes, lojas e produtos ao redor do mundo sendo assim um caso de sucesso de aplicação de Inteligência Empresarial para fornecer informações com legibilidade que venha a oferecer suporte na tomada de decisões.

Além da introdução, as Seções que dividem este trabalho são apresentadas como segue: na Seção 2 são especificados os objetivos gerais e específicos; a Seção 3 apresenta a fundamentação teórica deste trabalho; na Seção 4 são citados os trabalhos relacionados; na Seção 5, são apresentados os procedimentos metodológicos que descrevem os passos para se alcançar a solução; na Seção 6 é apresentado os principais resultados; na Seção 7 apresenta as considerações finais; e finalmente, na Seção 8 é apresentado os trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

- Propor um *Data Mart* para análise dos dados abertos da base de dados governamentais referentes a receitas e despesas dos estados e municípios brasileiros.

2.2 Objetivos específicos

- Modelar um *Data Mart* para os dados referentes às despesas e receitas do governo federal.
- Popular o *Data Mart*.
- Analisar os dados abertos para gerar informação sobre as receitas e despesas dos estados e municípios brasileiros em uma ordem cronológica do tempo.
- Publicar a análise dos dados em um meio de comunicação de fácil acesso a população.

3 FUNDAMENTAÇÃO TEÓRICA

Na fundamentação teórica do projeto são abordados os conceitos utilizados em seu desenvolvimento. Na primeira Subseção será definido o que é Dados Abertos e o que é Dados Abertos Governamentais. Na segunda, é definido a Legibilidade de Dados e qual sua relevância para este trabalho. Por fim, na última Subseção é abordado os conceitos relacionados com a análise de dados para atender o propósito deste projeto.

3.1 Dados Abertos

Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras *(*OPENDATAHANDBOOK, 2012*)*.

Na Califórnia, Estados Unidos da América, foram definidos por um grupo[2] os oitos princípios de Dados Abertos Governamentais:

> 1 - **Completos**. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.
> 2 - **Primários**. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.
> 3 - **Atuais**. Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.
> 4 - **Acessíveis**. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.
> 5 - **Processáveis por máquina**. Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.
> 6 - **Acesso não discriminatório**. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.
> 7 - **Formatos não proprietários**. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.
> 8 - **Livres de licenças**. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Os dados abertos são, no Brasil, pautados por três leis e oito princípios. As três leis foram propostas por David Eaves (2009), para contemplar o significado de Dados Abertos Governamentais:

> 1 - Se o dado não pode ser encontrado e indexado na Web, ele não existe;
> 2 - Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado;
> 3 - Se algum dispositivo legal não permitir sua replicação, ele não é útil.

[2] https://opengovdata.org/

Este trabalho utiliza-se de dados abertos disponibilizados pelo Portal da Transparência[3] que é uma iniciativa da Controladoria-Geral da União (CGU) para assegurar a boa e correta aplicação dos recursos públicos e pPALOelo Portal Brasileiro de Dados Abertos[4] que é uma ferramenta disponibilizada pelo governo para que todos possam encontrar e utilizar os dados e as informações públicas. As informações geradas através deste trabalho estão limitadas aos dados fornecidos por esses dois portais citados acima.

3.2 Legibilidade dos Dados

Legibilidade dos dados pode ser compreendida, segundo Sousa (2014), como "[...] o aspecto de tornar facilmente compreensível os dados disponibilizados.". Neste caso, os dados referentes às despesas e receitas dos estados e municípios brasileiros encontram-se disponibilizados no formato CSV e no formato XML, em uma estrutura de lista, que para cidadãos leigos torna-se de difícil visualização e compreensão.

Um trabalho relacionado à legibilidade dos dados foi Legibilidade em Dados Abertos: uma Experiência com os Dados da Câmara Municipal de São Paulo (BICO et al., 2012). Neste trabalho o objetivo foi fornecer uma aplicação capaz de fornecer os dados, que até então só existiam no formato XML, para a população de São Paulo que não possuía os conhecimentos necessários para interpretar as informações, tornando assim mais democrática a iniciativa de abertura de dados.

Desta forma, este trabalho visa à implantação de uma aplicação, onde o público alvo possa obter toda a informação necessária e com a devida legibilidade dos dados para que assim possa obter-se um entendimento das informações disponibilizadas.

3.3 Análise de Dados

A seguir serão discutidos os conceitos: Data Warehouse e Descoberta de conhecimento em Banco de Dados.

3.3.1 *Data Warehouse*

Um depósito de dados (*data warehouse*) é um repositório (ou arquivo) de informações colhidas de várias origens, armazenadas sob um esquema unificado, em um único

[3]http://www.portaltransparencia.gov.br/
[4]http://dados.gov.br/

local. Uma vez reunidos, os dados são armazenados por muito tempo, permitindo acesso a dados históricos (Silberschatz et al. 2012).

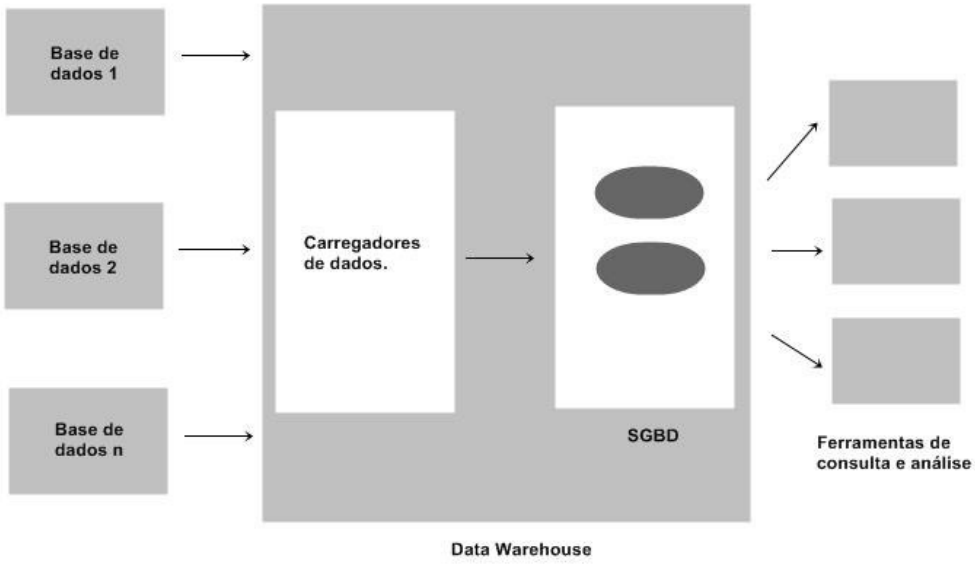
Os *data warehouses* oferecem acesso a dados para análise complexa, descoberta de conhecimento e tomada de decisão. Eles dão suporte à demanda de alto desempenho sobre os dados e informações de uma organização (Elmasri, 2011).

Uma das aplicações que podem estar vinculadas ao *data warehouse* é a aplicação OLAP que segundo Elmasri, (2011) é "[...] um termo usado para descrever a análise de dados complexos do *data warehouse.* Nas mãos de trabalhadores de conhecimento habilidosos, as ferramentas OLAP utilizam capacidades de computação distribuída para análises que exigem mais armazenamento e poder de processamento [...]".

Assim, existem dois tipos de sistemas, os sistemas transacionais (OLTP) e analíticos (OLAP). Os sistemas OLTP (*Online Transaction Processing*) se diferenciam dos sistemas OLAP pela sua finalidade dos dados, consultas, velocidade de processamento, requisitos de espaço e entre outras características. Optou-se neste trabalho por um sistema OLAP, devido ao suporte a consultas complexas que envolvem agregações.

Na Figura 1 é apresentada a arquitetura de um depósito de dados típico.

Figura 1 – Arquitetura de um depósito de dados



Fonte: Elaborado pelo autor

Em Codd E. F., (1993) foi definido 12 regras que as aplicações OLAP devem atender:

1. Conceito de visão multidimensional;
2. Transparência;
3. Acessibilidade;

4. Performance consistente de relatório;
5. Arquitetura cliente/servidor;
6. Dimensionamento genérico;
7. Tratamento dinâmico de matrizes esparsas;
8. Suporte a multiusuários;
9. Operações de cruzamento dimensional irrestritas;
10. Manipulação de dados intuitiva;
11. Relatórios flexíveis;
12. Níveis de dimensões e agregações ilimitados.

Existem três variações de banco de dados OLAP, cada um nomeado com base no formato de armazenamento que é usado. BOUMAN (2009, p. 122), descreve as três variações:

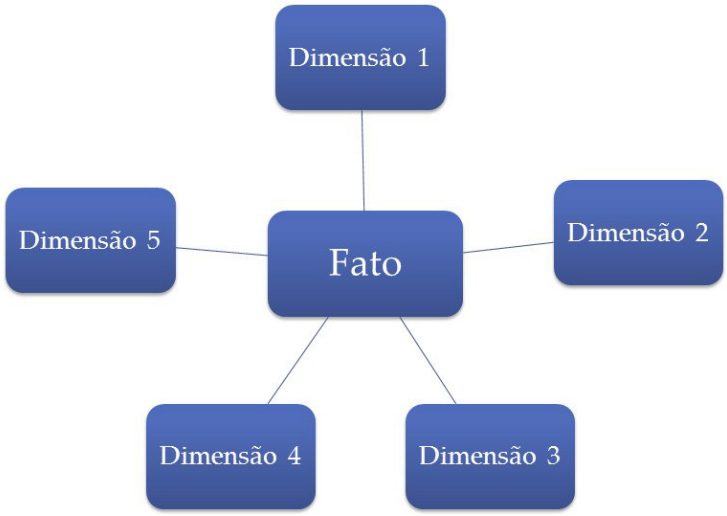
> MOLAP (OLAP Multidimensional)-O formato original em OLAP que os dados são armazenados em um formato proprietário multidimensional. Todos dados detalhados e agregados são armazenadas no arquivo de cubo. Um bom exemplo de uma fonte aberta de banco de dados MOLAP é PALO, desenvolvido pela empresa alemã Jedox.
> ROLAP (OLAP Relacional) - Neste caso, os dados e todos os agregados são armazenadas em um banco de dados relacional padrão. O motor traduz ROLAP consultas multidimensionais em SQL otimizado e, geralmente, acrescenta o cache capacidades, bem como para acelerar as consultas analíticas subsequentes. Pentaho Mondrian é um exemplo perfeito de um motor de ROLAP.
> HOLAP (OLAP Híbrido) - Em HOLAP, agregados e dados de navegação são armazenados em uma estrutura MOLAP, mas dados detalhados são mantidos no banco de dados relacional. Até o momento, não há solução open source HOLAP disponíveis, mas algumas das vantagens foram incorporadas no Mondrian com além de quadros gerados automaticamente agregados para agilizar consultas.

Neste trabalho será utilizada a variação ROLAP, devido a ela oferecer melhor suporte para bancos relacionais. Os dados contidos em um D*ata Warehouse* são multidimensionais e possuem atributos de dimensão e atributos de medidas. Os dados estão contidos em tabelas que são denominadas de **tabelas de fatos** e estas podem estar ligadas a diversas **tabelas de dimensão.** A tabela de fatos armazena principalmente informações quantitativas, como por exemplo, valor do produto, quantidade de itens vendidos e etc., já uma tabela de dimensão armazena principalmente informações descritivas aos dados armazenados na tabela fato, como por exemplo, nome do produto, tipo de produto que esse produto pertence e etc.

Um depósito de dados pode ser projetado em dois esquemas diferentes sendo eles: esquema estrela e esquema flocos de neve. O esquema estrela apresenta uma tabela de fato e várias tabelas de dimensão. As chaves estrangeiras na tabela de fatos estão relacionadas a atributos de alguma tabela de dimensão. A Figura 2 representa o esquema estrela. Enquanto que no esquema flocos de neves, as tabelas dimensionais relacionam-se com a tabela de fatos, mas

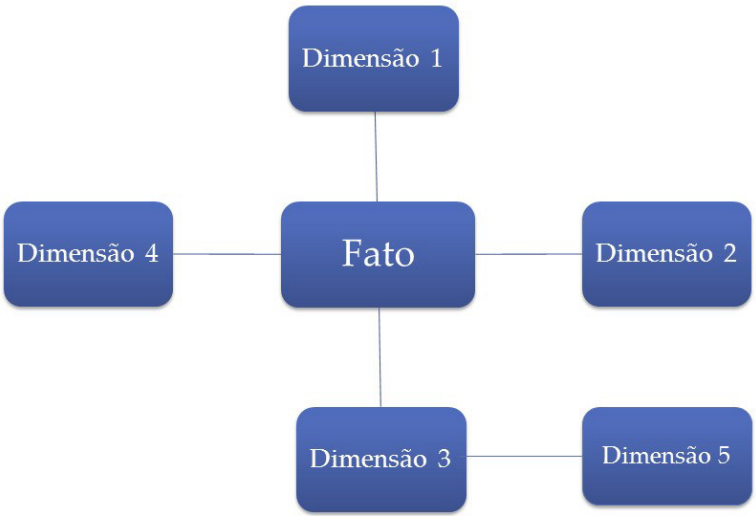
algumas dimensões relacionam-se apenas entre si. A Figura 3 representa o esquema flocos de neve. Pode ainda ocorrer o caso em que exista uma constelação de fatos, que consiste em várias tabelas fatos compartilhando uma ou várias tabelas de dimensão. A Figura 4 representa uma constelação de fatos.

Figura 2 – Representação do Esquema Estrela.



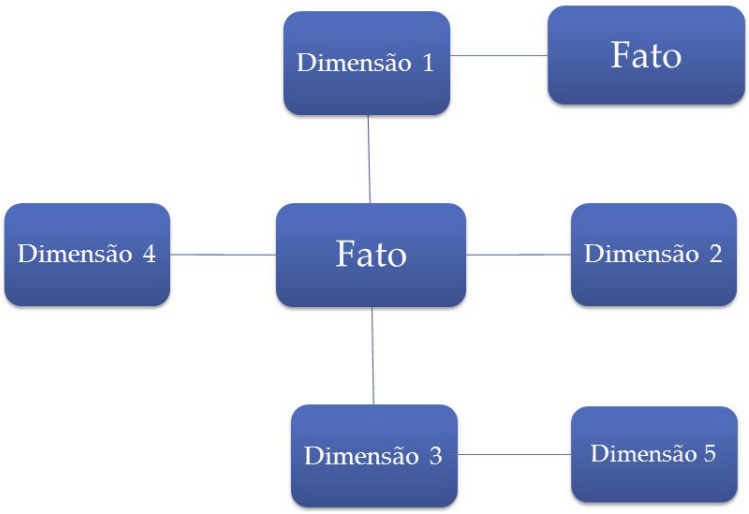
Fonte: Elaborado pelo autor.

Figura 3 – Representação do Esquema Flocos de Neve.



Fonte: Elaborado pelo autor.

Figura 4 – Representação uma Constelação de fatos.



Fonte: Elaborado pelo autor.

Devido à grande quantidade de tabelas e o relacionamento entre elas, neste trabalho optou-se pelo modelo de flocos de neve, pois este se adapta melhor ao contexto do trabalho, além de que o modelo proposto contém uma constelação de fatos, sendo as tabelas de fato de receita e despesas e estas compartilhando tabelas de dimensão como estado, município, entre outras.

Um *Data Mart* é uma parte de um *Data Warehouse* e segundo Bouman et al. (2009), contém informações relativas para uma função específica do negócio, como vendas ou quadro de pessoal. Estas informações podem ser visualizadas a partir de perspectivas diferentes, chamadas dimensões.

Um *Data Mart*, após construído, pode ser alimentado de duas formas diferentes:

- Por meio de um *Data Warehouse* já existente.
- Por sistemas transacionais ou arquivos de dados em geral.

A justificativa de se construir um Data Mart ao invés de um Data Warehouse para muitas organizações, se dá ao elevado custo de manutenção e construção do Data Warehouse. Bem como, é necessário mais tempo para ser concluído sua implementação e obter resultados. Basicamente as principais diferenças entre um Data Mart e um Data Warehouse são acerca do tamanho e do escopo a ser resolvido. Um Data Mart por ser utilizado principalmente para escopos menores, pode conter informações de uma área específica da organização, como por exemplo, dados relativos ao departamento de Recursos Humanos. Já o Data Warehouse, em

geral, apresenta um tamanho maior e sua construção pode abranger todos os departamentos da organização.

Este trabalho propõe um *Data Mart*, devido ao seu contexto ser apenas no setor financeiro da administração do governo brasileiro (apenas despesas e receitas dos estados e municípios brasileiros), não incluindo assim dados referentes a outros setores do Governo que também são disponibilizados pelo Portal da Transparência.

Segundo Anzanello (2006), um Cubo Olap "[...] é uma estrutura que armazena os dados de negócio em formato multidimensional, tornando-os mais fácil de analisar.". O cubo deve ser construído para que seja possível fazer a análise sobre os dados, é a partir dele que são obtidas as agregações entre as dimensões e a tabela fato. Neste trabalho foram construídos dois cubos, um sendo para receitas e outro para despesas acerca do setor financeiro do governo federal.

3.3.2 Descoberta de Conhecimento em Banco de Dados *(Knowledge Discovery in Database)*

Para Prass (2012):

> É sabido que conhecer o perfil do cliente traz uma série de benefícios para a empresa, sendo o principal deles, a capacidade de melhorar a qualidade de seus serviços prestados. Conhecendo o público alvo é possível montar uma melhor estratégia de marketing e com isto obter resultados mais significativos com a venda de produtos e/ou serviços.
> O problema é que estes registros, muitas vezes, representam apenas dados e não conhecimento. Visando transformar estes dados em conhecimento, surge o processo chamado de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD).

Descoberta de conhecimento em bancos de dados é o processo não trivial de identificar em dados padrões que sejam válidos, novos (previamente desconhecidos), potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão (FAYYAD et al., 1996). Tal descoberta de dados, chamada de KDD (*Knowledge Discovery in Databases*) é constituída das seguintes etapas (Silva, 2014):

1. Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar.

2. Criação de um conjunto de dados alvo (Selection): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada.

3. Limpeza de dados e pré-processamento (Preprocessing): operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração.

4. Redução de dados e projeção (Transformation): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.

5. Mineração de dados (Data Mining): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.

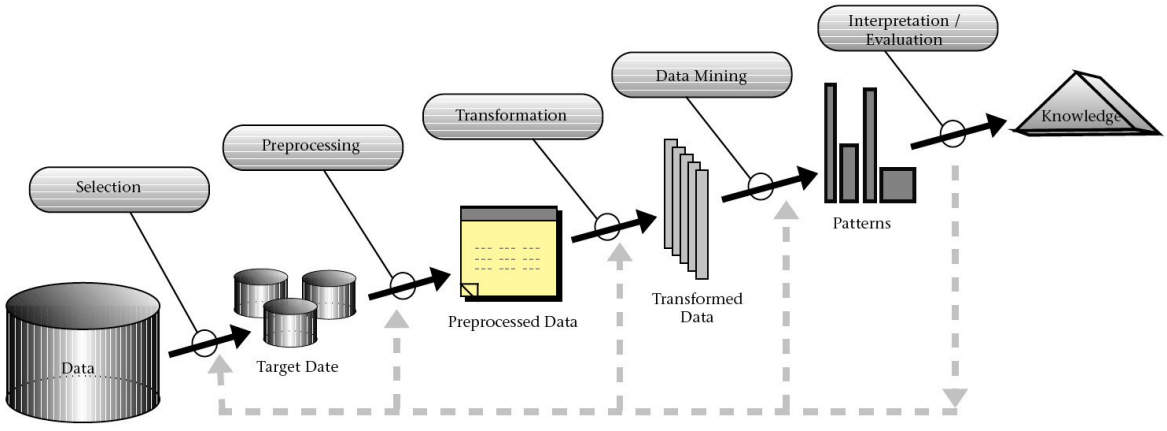
6. Interpretação dos padrões minerados (Interpretation/Evaluation), com um possível retorno aos passos 1-6 para posterior iteração.

7. Implantação do conhecimento descoberto (Knowledge): incorporar este conhecimento à performance do sistema, ou documentá-lo e reportá-lo às partes interessadas.

Na Figura 5 são demonstrados esses processos mostrando assim as relações entre cada um deles.

O processo de KDD será utilizado nas bases de dados utilizadas por esse trabalho para garantir que as demais atividades sejam concluídas e assim atingir o objetivo final.

Figura 5 – Etapas do processo de KDD.



Fonte: Fayyad et al. (1996)

4 TRABALHOS RELACIONADOS

Segundo Bico (2014):

> Com o movimento em prol da divulgação de dados públicos e com a efetiva disponibilização de dados, começaram a surgir iniciativas que tem como objetivos analisar os dados, processa-los e disponibilizar aplicações para visualização de informação ou para prestação de serviços a partir das informações geradas.

Diante deste movimento, surgiram diversos trabalhos com esses propósitos como o Sousa (2014), que é proposto um modelo de *Data Mart* e foi realizada uma análise dos dados abertos do PROCON utilizando-se da ferramenta *Pentaho*. Os dados do PROCON foram disponibilizados pelo Governo Federal no Portal Brasileiro de Dados Abertos[5]. Sousa (2014) utilizou algumas técnicas de extração de conhecimento como mineração de dados e consultas OLAP. O trabalho aqui proposto utilizou uma base de dados diferente, que corresponde a receitas e despesas dos estados e municípios brasileiros, além disso o esquema utilizado também é diferente devido à relação dos dados coletados. Em Sousa (2014), não foi proposto uma aplicação para apresentar ao público alvo as informações geradas pelas análises, algo que este trabalho já propõe.

Em Bico et al. (2012), é desenvolvida uma aplicação web para apresentar os dados da Câmara Municipal de São Paulo no intuito de gerar visualizações gráficas. Os dados encontravam-se no formato XML e foram providos pela própria prefeitura do município. Foi utilizado o modelo de dados dimensional e hierárquico flocos de neve, assemelhando-se assim com o trabalho aqui descrito. Foram utilizadas consultas SQL (*Structured Query Language*) para a obtenção de informações válidas diferenciando-se assim do trabalho aqui descrito já que este utiliza consultas OLAP.

No caso de Akintola et al. (2011), é relatado o desenvolvimento de um *Data Mart* para gestão de um ambiente acadêmico e para isso utiliza técnicas de mineração de dados usando a ferramenta *Microsoft SQL Server Analysis Services,* além disso faz uso do modelo de dados dimensional estrela, diferenciando-se assim deste trabalho que se utiliza do *Pentaho Business Analytics.* Este trabalho propõe ainda um modelo de dados dimensional e hierárquico flocos de neve. Uma semelhança entre o trabalho de Akintola e o deste projeto é que ambos buscam prover a informação para aqueles que não possuem conhecimentos necessários para interpretar as informações de seus devidos contextos.

[5]http://dados.gov.br/aplicativos/

5 PROCEDIMENTOS METODOLÓGICOS

Esta Seção apresenta os procedimentos metodológicos adotados neste trabalho. As ferramentas utilizadas foram o brModelo[6], para modelagem do *Data Mart*; SGBD PostgresSQL[7], para armazenar e gerenciar a base de dados; e o framework *Pentaho* para criar, gerenciar e realizar as operações de ETL e OLAP, sendo o processo de ETL realizado através do PDI (*Pentaho Data Integration*). Para criação dos cubos OLAP a ferramenta *Schema Workbench*. Os passos para a execução estão descritos de acordo com a Figura 6.

Os passos apresentados envolvem a exploração dos resultados esperados do processo de análise dos dados das despesas e receitas do governo do Brasil ao longo dos anos. A seguir esses passos são detalhados.

Figura 6 – Passos que serão realizados durante a execução deste trabalho.

Fonte: Elaborado pelo autor.

5.1 Processo 1 KDD: Definição do tipo de conhecimento a descobrir

Nesta etapa foi definido em qual contexto o projeto iria ser desenvolvido, abrangendo assim a área de finanças do governo brasileiro. Foi decidido um estudo sobre as

[6]http://www.sis4.com/brModelo/
[7]http://www.postgresql.org/

despesas e receitas dos estados brasileiros a fim de gerar informação para o público alvo deste trabalho.

5.2 Processo 2 *KDD*: Seleção

Nesta etapa do projeto, foram identificados os principais dados relacionados a receitas e despesas dos estados e municípios brasileiros, a fim de obter-se uma maior quantidade de informações para povoamento do *Data Mart* e assim gerar mais informação. Para identificar quais dados estavam disponíveis foi realizada uma busca no Portal da Transparência e no Portal Brasileiro de Dados Abertos usando como critérios os arquivos que possuíam estados e municípios como chave, a data e um valor correspondente à despesa total ou receita total.

5.3 Processo 3 *KDD*: Pré-processamento

Foi realizado um pré-processamento em algumas das bases de dados selecionadas, devido a algumas bases não estarem no formato padrão necessário para o PDI realizar seus objetivos. Muitos dos dados CSV continham, em um determinado ponto, uma quebra de linha não esperada, fazendo com que o arquivo fugisse do padrão desejado. Os arquivos que necessitavam de um pré-processamento foram os arquivos relacionados as informações do programa Garantia Safra.

5.4 Processo 4 *KDD*: Transformação

A etapa de transformação consistiu em adicionar, em algum dos arquivos selecionados, um valor que representaria o ano referente a cada item do arquivo, isso foi necessário devido a alguns arquivos não possuírem o ano dentro do arquivo, mas sim somente no título do arquivo. Esse problema influenciaria no momento de consultas no *Data Mart* impossibilitando assim ter uma visão sobre a dimensão tempo.

5.5 Proposta do modelo ROLAP para a geração do Data Mart

Após a coleta dos dados e a realização dos primeiros passos do processo de KDD, foi realizada a modelagem do *Data Mart* que consiste em inicialmente conhecer os modelos já existentes. Após a análise de qual modelo utilizar, o esquema do *Data Mart* baseado no modelo ROLAP foi proposto para os dados utilizados neste trabalho.

5.6 Processo de ETL

Para o povoamento e implementação do *Data Mart* foi utilizado a ferramenta *Spoon,* que é uma ferramenta do PDI que tem como função transferir os dados dos arquivos obtidos na etapa de definição e coleta das bases de dados do Portal da Transparência e do Portal Brasileiro de Dados Abertos a fim de transferi-los ao *Data Mart*.

5.7 Proposta de questões

Foi realizado um estudo de quais questões devem ser respondidas e quais informações seriam providas para o público alvo, para que assim o trabalho possua uma justificativa e seja útil para a população.

A seguir são listadas as questões as quais este trabalho responde:

- Qual o valor da despesa de cada estado e município ao longo do tempo?
- Qual o valor da receita de cada estado e município ao longo do tempo?
- Qual a receita e despesa de cada região do Brasil ao longo do tempo?
- Qual a renda de cada área prioritária dos estados e municípios ao longo do tempo?
- Quantas pessoas saíram e se mantiveram no programa Pescador Artesanal?
- Quanto se investiu, em média, em cada pessoa no programa Pescador Artesanal ao longo do tempo?
- Quanto se investiu, em média, em cada pessoa nos programas sociais do governo ao longo do tempo?
- Quanto foi investido, em média, nos programas do governo ao longo do tempo em cada estado e em cada município brasileiro?
- Quais estados e municípios recebem mais e menos verbas da União?
- Quais estados e municípios contem mais e menos despesas?

5.8 Processo 5 *KDD*: Análise via OLAP

A análise dos dados se deu por meio de consultas OLAP que, usa um formato de armazenamento otimizado para análise de dados em um formato multidimensional para

oferecer a flexibilidade ao usuário e acesso rápido. Foi utilizada a ferramenta *Schema Workbench* para criação dos cubos OLAP.

5.9 Processo 6 *KDD:* Interpretação/ Avaliação

Esta etapa consistiu na análise dos dados a fim de se verificar a veracidade e consistência das informações gerada por todo o processo e a averiguação da consistência dos dados.

5.10 Definição da apresentação do conhecimento

Foi utilizado o framework *Pentaho Business Analytics* para gerar gráficos que auxiliem na interpretação dos resultados e tornando assim a informação mais legível para a população. Foram analisados meios por onde esses gráficos sejam disponibilizados para o público alvo, para que com isso o acesso à informação esteja mais acessível e facilmente interpretado a todos.

Nesta etapa ocorreu todo o planejamento da aplicação que contém as informações geradas pelas consultas OLAP.

5.11 Processo 7 *KDD: (*Conhecimento*)* Publicação das informações.

Após a definição da forma de publicar os resultados deste projeto foi implementada a aplicação que contém os resultados, construindo assim um meio de apresentação dos resultados obtidos.

6 RESULTADOS

Esta seção apresenta os resultados obtidos neste trabalho sendo obtidos na execução dos procedimentos metodológicos.

6.1 Processo 1 KDD: Definição do tipo de conhecimento a descobrir

Nesta etapa foi definido qual o contexto do projeto, abrangendo assim a área de finanças do governo brasileiro. As informações tratadas neste trabalho abordam a receita e despesas dos estados, municípios e regiões do Brasil havendo ainda a classificação por ano. Quanto as despesas, foram contempladas 6 (seis) tipos, são elas despesas referentes à: Bolsa Família, Erradicação do Trabalho Infantil, Garantia Safra, Pescador Artesanal, Transferências (Pagamentos) e Transferências (Outros Programas Sociais).

6.2 Processo 2 *KDD*: Seleção

No Quadro 1 é apresentado o tipo, o conjunto e a descrição dos dados obtidos nesta etapa.

Quadro 1 – Conjunto de dados obtidos na fase de seleção.

Tipo	Conjunto	Descrição
Receita	Total das Transferências para Estados Constitucionais, legais e voluntárias.	Valor das transferências voluntárias da União para Estados obtidos pela agregação das diversas contas do SIAFI, agrupadas em constitucionais, legais e voluntárias.
Receita	Total das Transferências para Municípios Constitucionais, Legais e Voluntárias.	Total das transferências da União para Municípios. Dados obtidos no SIAFI. Resulta da soma das transferências constitucionais, legais e voluntárias.
Receita	Transferência de Receitas de Estado e Municípios.	Transferência da União diretamente para estado e municípios brasileiros.
Despesa	Bolsa Família	Transferências de recursos federais diretamente repassados a cidadãos, referentes ao pagamento do Bolsa Família.
Despesa	Erradicação do Trabalho Infantil	Concessão de Bolsa a Crianças e Adolescentes em Situação de Trabalho
Despesa	Garantia Safra	Contribuição ao Fundo Garantia-Safra
Despesa	Pescador Artesanal	Recursos destinados ao Seguro Defeso - Pescador Artesanal.
Despesa	Transferências Outros Programas Sociais	Transferências de recursos federais diretamente repassados a cidadãos, referentes a diversos programas sociais.
Despesa	Transferências Pagamentos	Transferências Constitucionais e as Decorrentes de Legislação Específica

Fonte: Elaborado pelo autor.

6.3 Processo 3 *KDD*: Pré-processamento

Nesta etapa, o arquivo era submetido a processamento no PDI, caso o arquivo estivesse no padrão correto nenhum pré-processamento se fazia necessário, caso contrário, o PDI informava a linha do arquivo em que havia uma quebra de linha inesperada e então bastava remover a quebra de linha indesejada e submeter novamente o arquivo para processamento.

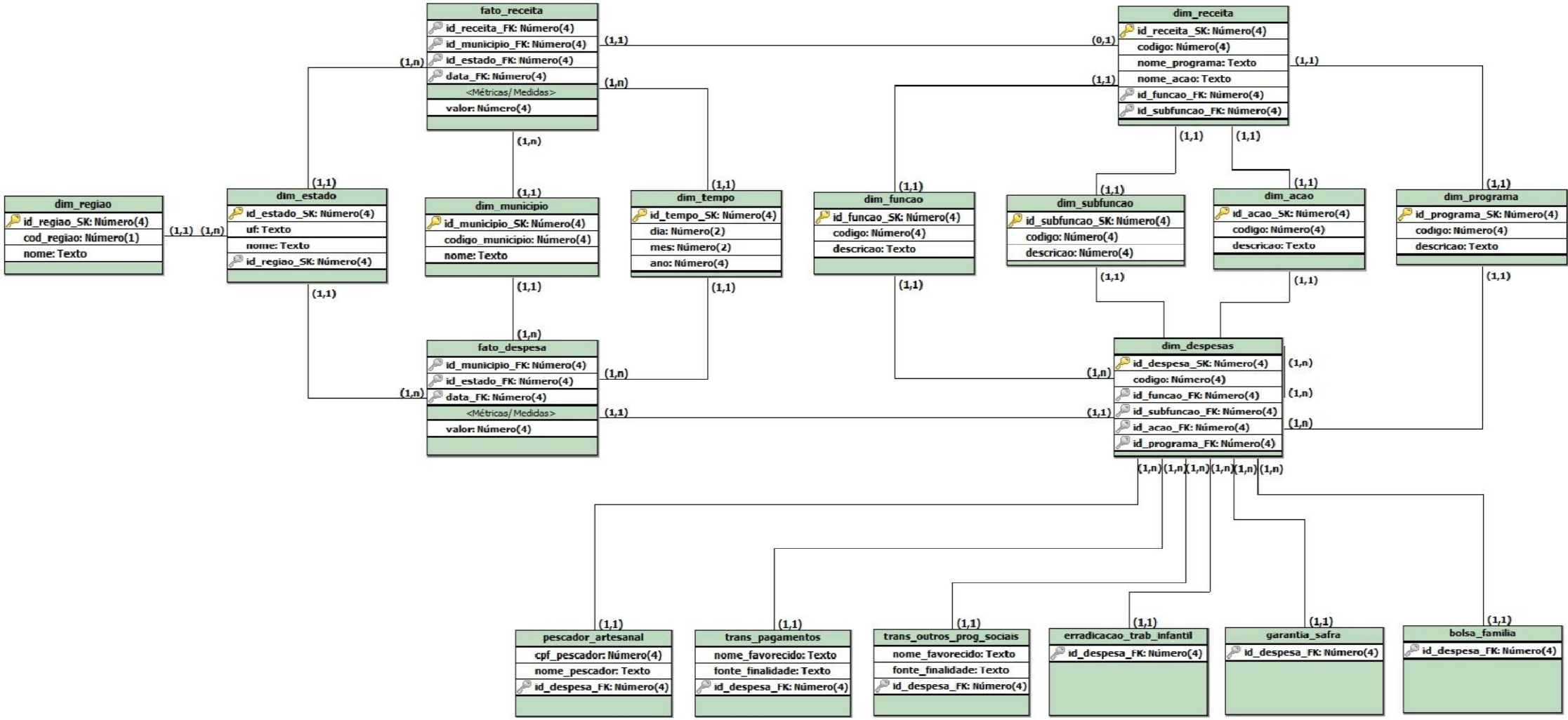
6.4 Processo 4 *KDD*: Transformação

Para auxiliar neste processo foi utilizada um elemento do PDI chamado de *Add Constants* que adiciona novos campos no decorrer do processo de transformação. Foram adicionados valores como os de data, que em muito dos arquivos a informação estava contida apenas no título do arquivo e não necessariamente dentro dele onde realmente havia a informação.

6.5 Proposta do modelo ROLAP para a geração do Data Mart

A Figura 7 apresenta o modelo proposto por este trabalho para a modelagem do *Data Mart* utilizado neste trabalho.

Figura 7- Esquema ROLAP utilizado neste trabalho

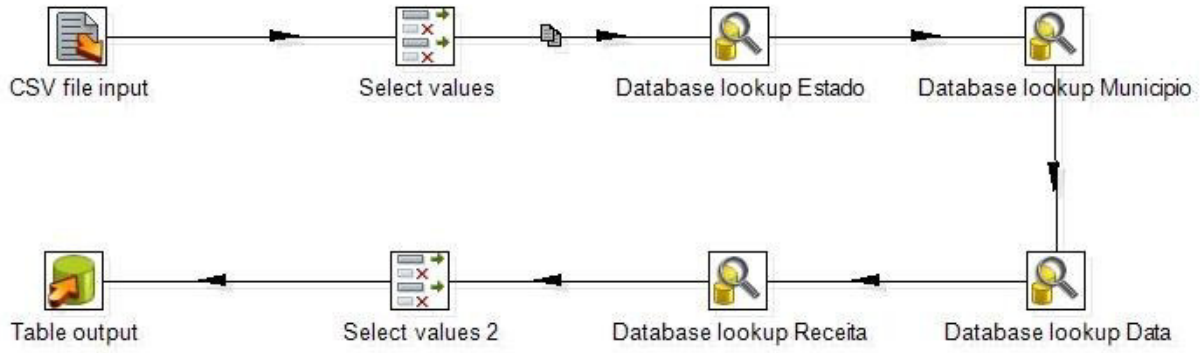


Fonte: Elaborada pelo autor.

6.6 Processo de ETL

Na Figura 8 é apresentado um dos passos realizado na ferramenta *Spoon* utilizada para o processo de ETL.

Figura 8 – Ferramenta Spoon.



Fonte: Elaborado pelo autor.

Na Figura 8 é demonstrado o passo para construção da tabela fato receita, sendo ela constituída de chaves estrangeiras para as dimensões estado, município, tempo e receita, bem como da medida valor. Além da tabela fato receita, também foi realizado o processo de ETL das dimensões estado, município, receita, tempo, região, função e subfunção.

Inicialmente, um arquivo do tipo .csv é dado de entrada no processo, através do item CSV file input demonstrado, este arquivo contém informações de receita a serem persistidas no banco de dados. No item Select values houve a filtragem de quais campos seriam inseridos no banco de dados. Nos itens Database lookup Estado, Database lookup Município, Database lookup Data e Database lookup Receita foram utilizados para retornarem a chave primaria das tabelas Estado, Município, Data e Receita respctivamente. No item Select values 2 foram selecionados somente as chaves primarias de cada dimensão mencionada e o campo referente ao valor da receita. Finalmente, no Table output, os dados foram persistidos na tabela fato de Receita.

6.7 Proposta de questões

Para responder a todas as questões levantadas, foi disponibilizada a aplicação, que se encontra online no seguinte endereço: http://despesasereceitaspublicas.com.br/.

Na aplicação é possível obter gráficos para as questões levantadas neste trabalho. Para cada questão existe um conjunto de gráficos que as representam. O trabalho não se limitou

apenas a essas questões apresentadas no texto, mas também há outras que foram respondidas em gráficos adicionais.

6.8 Processo 5 *KDD*: Análise via OLAP

A estrutura final dos cubos construídos a partir do *Schema Workbench* é apresentada na Figura 9, nela contém os cubos de Despesa e Receita, além das dimensões Estado, Município e Tempo, estas 3 últimas dimensões não foram adicionadas dentro de cada cubo por motivos de reutilização. Dentro dos cubos foi adicionado o que é chamado de *Dimension Usage* que nada mais é do que um link para as dimensões externas.

Figura 9 – Cubos na ferramenta Schema Workbench.

Fonte: Elaborado pelo autor.

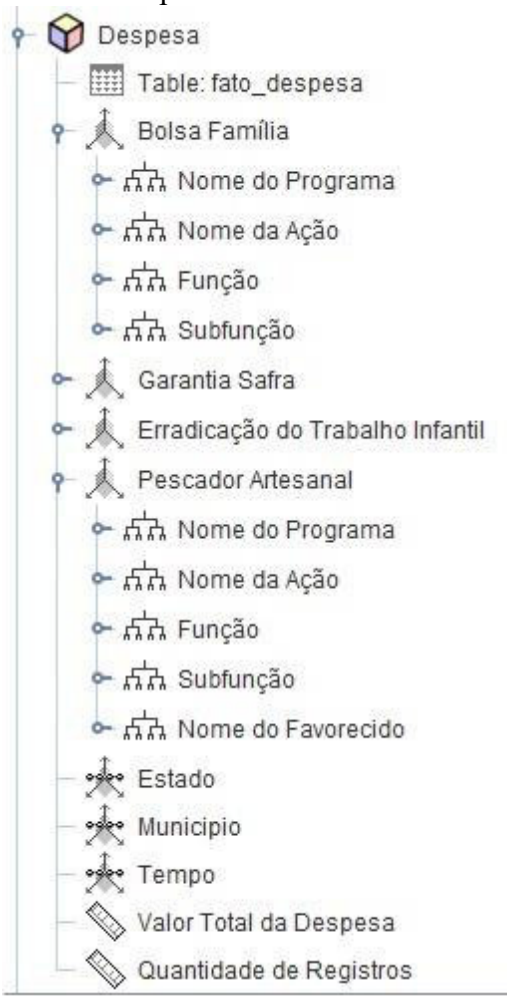
Na Figura 10 é apresentada a construção do cubo de Receita, sendo este formado pelas dimensões Receita, Estado, Município e Tempo além da medida Valor Total da Receita.

Figura 10 – Representação do cubo Receita.

Fonte: Elaborado pelo autor.

Na Figura 11 é apresentada a construção do cubo de Despesa, sendo este formado pelas dimensões Bolsa Família, Erradicação do Trabalho Infantil, Garantia Safra, Pescador Artesanal, Estado, Município e Tempo além da medida Valor Total da Despesa e Quantidade de Registros. As dimensões Erradicação do Trabalho Infantil e Garantia Safra possuem a mesma estrutura que a dimensão Bolsa Família.

Figura 11 – Representação do cubo Despesa.



Fonte: Elaborado pelo autor.

6.9 Processo 6 *KDD:* Interpretação/ Avaliação

A etapa de avaliação consistiu em averiguar se as informações exibidas nos gráficos condiziam com as informações apresentadas no *Data Mart*, para isso foram executadas instruções em SQL e posteriormente comparado seus resultados com os resultados gerados pelos cubos.

6.10 Definição da apresentação do conhecimento

Para que o público alvo deste trabalho possa ter acesso às informações com mais facilidade, foi construído um *website*, que contém as informações gráficas acerca da área de finanças do governo, onde foram utilizadas tecnologias como: PHP, HTML, Bootstrap, JavaScript e API Highcharts.

6.11 Processo 7 *KDD: (*Conhecimento*)* Publicação das informações.

Para a publicação das informações o *website* foi hospedado (http://despesasereceitaspublicas.com.br/) e encontra-se online para que o objetivo do trabalho aqui descrito seja cumprido, e dessa forma disponível para seu público alvo.

A Figura 12 apresenta a interface da aplicação já disponível para o público alvo deste trabalho.

A Figura 13 apresenta as informações referentes ao programa Bolsa Família categorizado por estado e por ano.

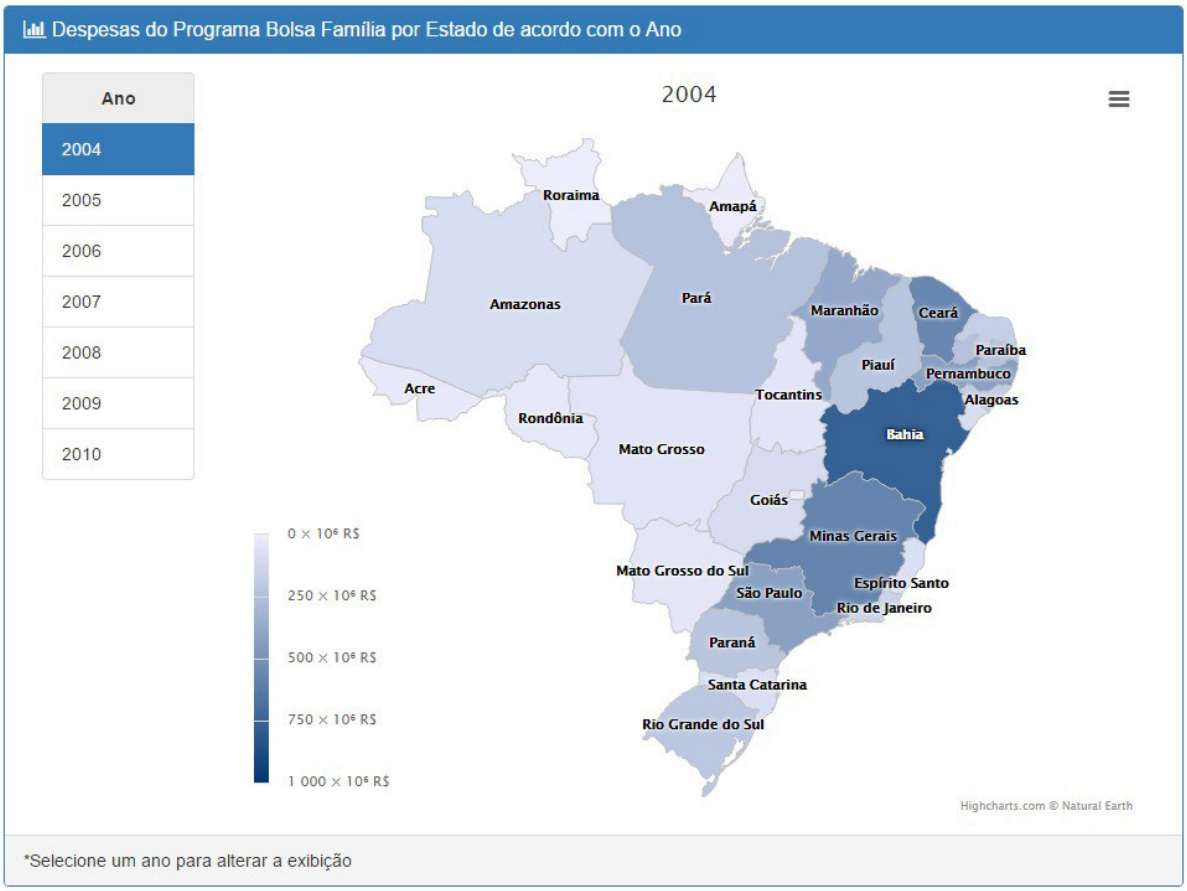
A Figura 14 apresenta as informações referentes ao programa Erradicação do Trabalho Infantil, fazendo uma comparação entre os valores repassados a cada região do Brasil ao longo do tempo.

A Figura 15 apresenta os valores repassados para o estado do Ceará ao longo do tempo.

Figura 12 - Interface da aplicação.

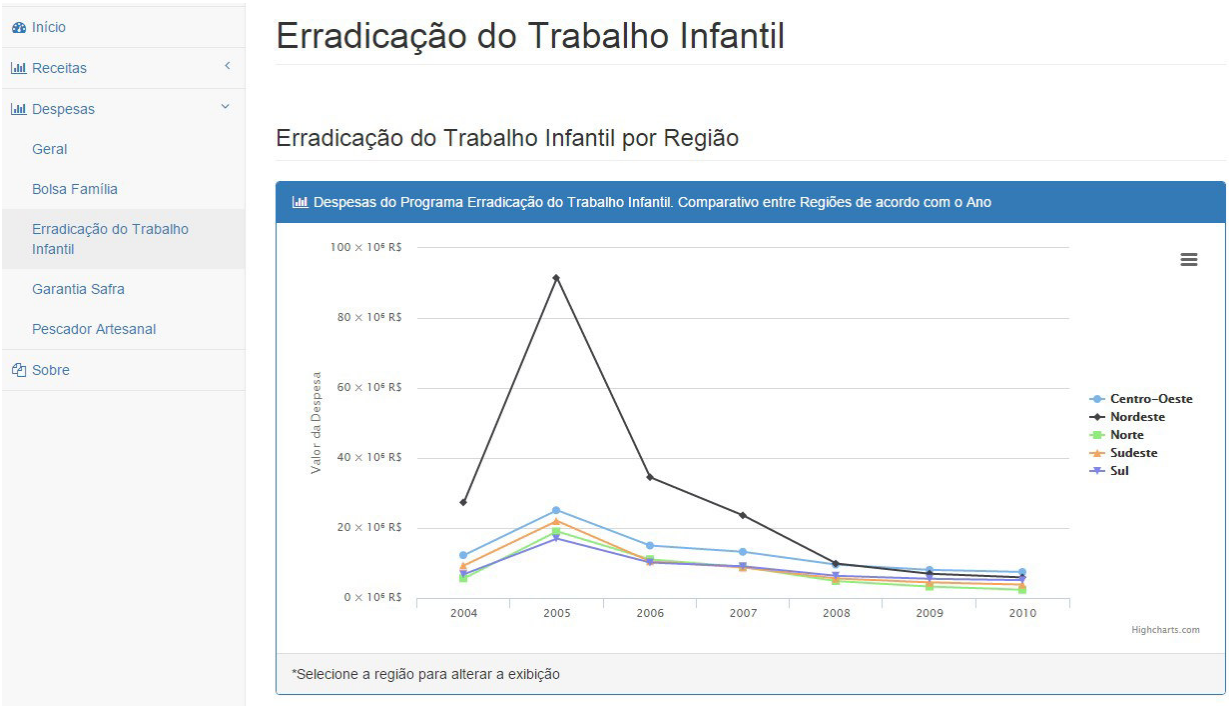
Fonte: Elaborado pelo autor.

Figura 13 - Bolsa Família por Estado



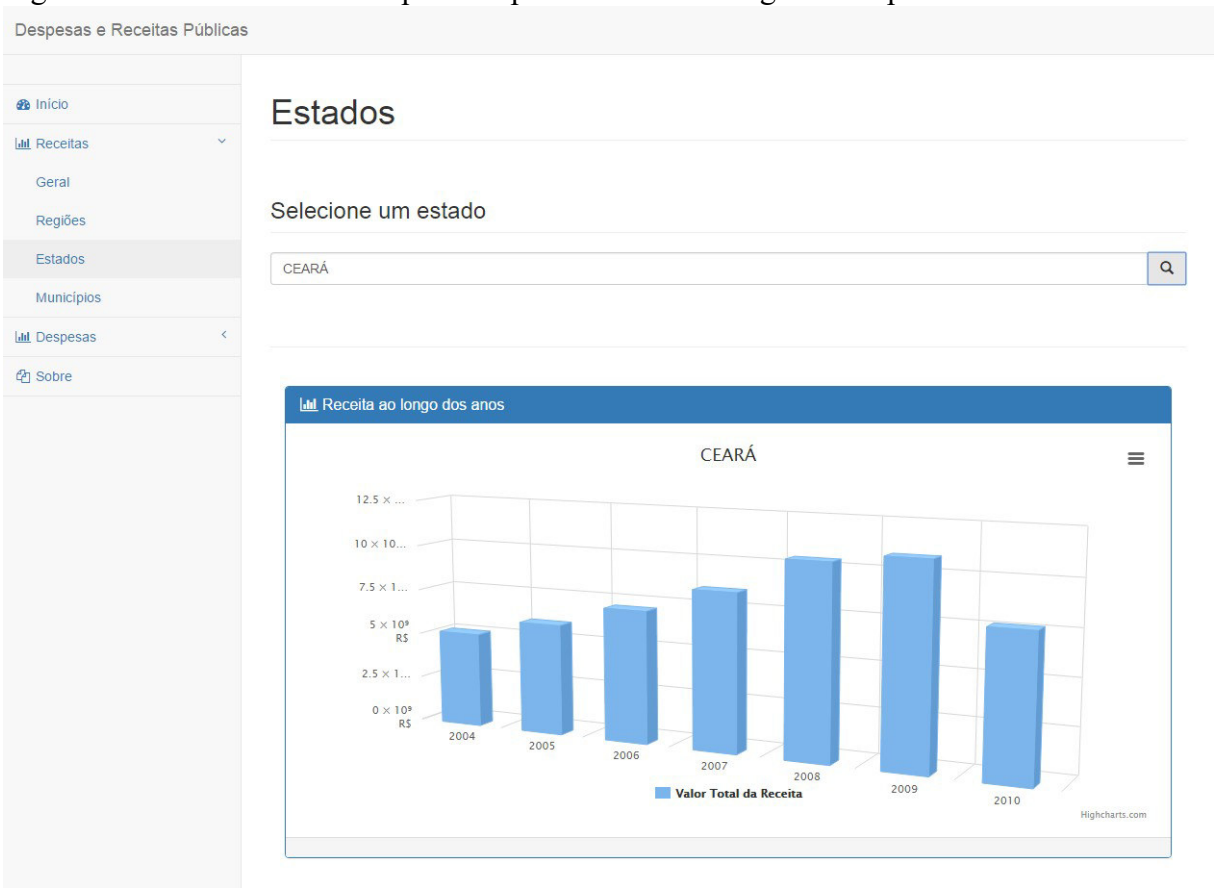
Fonte: Elaborado pelo autor.

Figura 14 - Erradicação do Trabalho Infantil



Fonte: Elaborado pelo autor.

Figura 15 - Valor da Receita repassado para o Ceará ao longo do tempo



Fonte: Elaborado pelo autor.

7 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi de extrair informações importantes dos dados acerca das despesas e receitas dos estados e municípios brasileiros disponibilizados pelo Portal da Transparência. Tendo objetivo realizar análises nestes dados, utilizando técnicas de extração de conhecimento, além de permitir a visualização das informações de maneira compreensível a qualquer cidadão, por meio de gráficos.

O objetivo foi alcançado construindo-se um *Data Mart* para armazenar os dados a serem analisados. As informações são exibidas através de gráficos e foram geradas a partir da API Highcharts, e para o povoamento do *Data Mart* foi utilizado o *Pentaho Data Integration.*

8 TRABALHOS FUTUROS

Diante do trabalho de análise realizado, como trabalhos futuros será planejada uma forma de manter o *Data Mart*, com incremento de novos dados e novas tabelas no modelo aqui proposto.

No que diz respeito a novos dados que serão incrementados ao *Data Mart*, esses dados poderão ser outros tipos de despesas, havendo ainda a possibilidade de dividir as receitas também em categorias.

Outro trabalho futuro será a implementação de uma ferramenta de *Business Intelligence* a fim de propiciar as pessoas, que possuem um conhecimento envolvido na área, realizarem consultas livres e gerarem gráficos a partir delas, diferente das consultas e gráficos deste trabalho que são estáticos.

Por fim, verificar a possibilidade da criação de um aplicativo para dispositivo móvel para que o conhecimento seja disseminado com maior facilidade.

REFERÊNCIAS

AKINTOLA K.G., ADETUNMBI A.O. ADEOLA O.S. **Building Data Warehousing and Data Mining from Course Management Systems: A Case Study of FUTA Course Management Information Systems.** International Journal of Database Theory and Application. Vol. 4, No. 3, September, 2011.

BICO, Fernanda C. et al. Legibilidade em Dados Abertos: uma experiência com os dados da Câmara Municipal de São Paulo. In: Simpósio Brasileiro de Sistemas de Informação, 8., 2012, São Paulo. **Anais...** São Paulo, 2012. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2012/0068.pdf>. Acesso em: 15abr. 2015.

BOUMAN, Roland; DONGEN, Jon van. **Business Intelligence e DataArmazenamento com Pentaho e MySQL**. 1.ed. Indianapolis, Indiana: Atlas, 2009. 602p.

Codd E.F., Codd S.B., and Salley C.T. "**Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate**". Codd & Date, Inc 1993

DATE, C. J. **Introdução a Sistemas de Bancos de Dados**. 8 ed. Rio de Janeiro: Campus, 2004.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistema de Banco de Dados**. 6.ed. São Paulo: Atlas, 2011. 788p.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, EUA: AAAI Press, 1996. 611 p.

OPENDATAHANDBOOK. **O que são dados abertos**. , 2012. Disponível em: <http://opendatahandbook.org/pt_BR/what-is-open-data/index.html>. Acesso em: 14abr. 2015.

OHLINGER, Patrick. Wal-Mart's Data Warehouse. In: **SCODAWA**., 2006. Vienna University of Technology: June 19, 2006.

PENTAHO. **Pentaho**, 2005. Disponível em: <http://www.pentaho.com/>. Acesso em: 14 abr. 2015.

PRASS, Fernando Sarturi. **Um visão geral sobre as fases do Knowledge Discovery in Databases (KDD)**. , 2012. Disponível em: <http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd/>. Acesso em: 22 abr. 2015.

SILVA, Marcelino P. Santos. **Mineração de Dados-Conceitos, Aplicações e Experimentos com Weka**. In: Artigo. Instituto Nacional de Pesquisas Espaciais (INEP),2004. São José dos Campos-SP.

ANZANELLO, Cynthia Aurora. **OLAP Conceitos e Utilização.** In: Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS), 2006.Porto Alegre – RS.

SILBERSCHATZ, Abrahan; KORTH, Henry F.; SUDARSHAN, S.**Sistema de Banco de Dados**. 6.ed. Rio de Janeiro: Atlas, 2012. 861p.

SOUSA, Pedro José Rodrigues. **Uma proposta para análise de dados abertos do PROCON utilizando Data Mart** / Pedro José Rodrigues de Sousa. Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2014.