CURSO DE CIÊNCIA DA COMPUTAÇÃO

Rodrigo Frantz

DESENVOLVIMENTO DE UM AMBIENTE DE CONSULTAS EM BIG DATA CONSIDERANDO UM ESQUEMA ESTRELA DE DADOS

Santa Cruz do Sul

2016

Rodrigo Frantz

DESENVOLVIMENTO DE UM AMBIENTE DE CONSULTAS EM BIG DATA CONSIDERANDO UM ESQUEMA ESTRELA DE DADOS

Trabalho de Conclusão de Curso apresentado à disciplina de Trabalho de Conclusão II, do Curso de Ciência da Computação da Universidade de Santa Cruz do Sul - UNISC.

Orientador: Prof. Ms. Eduardo Kroth

Santa Cruz do Sul
2016

Rodrigo Frantz

DESENVOLVIMENTO DE UM AMBIENTE DE CONSULTAS EM BIG DATA CONSIDERANDO UM ESQUEMA ESTRELA DE DADOS

Esse trabalho foi submetido ao processo de avaliação por banca examinadora do Curso de Ciência da Computação da Universidade de Santa Cruz do Sul - UNISC, como requisito para obtenção do título de Bacharel em Ciência da Computação.

Prof. Ms. Eduardo Kroth
Professor Orientador – UNISC

Prof. Ms. Evandro Franzen
Professor Examinador – UNISC

Prof. Ms. Kurt Werner Molz
Professor Examinador – UNISC

Santa Cruz do Sul
2016

RESUMO

A geração e o armazenamento de informações em formato digital têm crescido consideravelmente nos últimos anos. Estes dados provêm das mais diversas origens e apresentam-se sobre várias formas. Este novo conceito denominado Big Data nos propõem novos desafios no que diz respeito ao armazenamento, a consulta e visualização dos dados. Deste modo formas para representar estas informações devem ser implementadas para auxiliar as organizações na análise destes dados, podendo ser assim um complemento aos sistemas de Data Warehouse que muitas empresas já possuem. O presente trabalho tem por objetivo desenvolver um ambiente de consultas em um Big Data considerando um esquema estrela de dados, esquema este que normalmente regula a organização dos fatos e das dimensões em um Data Warehouse. Portanto para captura dos dados e carga do Big Data será utilizado o Ontoclipping, uma ferramenta que utiliza ontologias para aprimorar e facilitar o trabalho de recuperação de informações na web. Sendo assim, para que seja possível integrar as consultas realizadas em um Data Warehouse a um Big Data, será realizado um mapeamento entre dimensões do esquema estrela com as ontologias usadas no Ontoclipping, assim possibilitando ao usuário a consulta e exibição dos resultados coletados tanto no Big Data como no Data Warehouse.

Palavras chaves: Big Data, Apache Cassandra, Apache Lucene, Ontologia, Data Warehouse e Modelo Estrela de dados.

ABSTRACT

The generation and storage of information in digital format has grown considerably in recent years. These data come from the most diverse origins and represent themselves in various forms. This new concept denominated as Big Data proposes to us new challenges when it comes of data storage, the query and visualization. Thus forms to represent this information should be implemented to help organizations in the analysis of these data and can be a complement to the systems of Data Warehouse that many companies already have. The objetive of the work is to developing a query environment in a Big Data considering the star scheme of Data, the star scheme that normally regulates the organization of facts and dimensions in a Data Warehouse. Therefore, in order to capture data and load the Big Data, the Ontoclipping, a tool that uses ontologies to improvement and faciliate the work of information retrieval on the web. To integrate Data Warehouse to a Big Data, a mapping between dimensions of the star scheme with the ontologies used in Ontoclipping as allowing the user to query and display the results collected both in Big Data as in the Data Warehouse.

Keywords: Big Data, Apache Cassandra, Apache Lucene, Ontology, Data Warehouse and Star Schema.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
COMEX	Comércio Exterior
CQL	*Cassandra Query Language*
DW	*Data Warehouse*
DFR	*Divergence From Randomness*
ERP	*Enterprise Resource Planning*
ETL	*Extract Transform Load*
HTML	*Hypertext Markup Language*
NoSQL	*Not Only SQL*
OLAP	*On-Line Analytic Processing*
OWL	*Web Ontology Language*
PPGC	Programa de Pós-Graduação em Computação
RDF	*Resource Description Framework*
SAD	Sistemas de Apoio à Decisão
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	*Structured Query Language*
UFRGS	Universidade Federal do Rio Grande do Sul
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A aplicação da computação nas mais diversas áreas do conhecimento, sempre foi relacionada ao processamento de dados. Tendo assim em vista essa imensa quantidade de dados e informações geradas atualmente, surge a necessidade de se buscar formas de se representar toda esta gama de informações que podem estar persistidas de forma estruturada, semiestruturada, ou sem nenhuma organização.

De acordo com Zikopoulos et al. (2012), o termo Big Data se aplica a todo este potencial de dados que não são passíveis de análise ou processamento através dos métodos e ferramentas tradicionais.

Durante muito tempo grande parte deste volume de informações era ignorado, pois não se tinha como armazenar estes dados a um custo benefício aceitável (BERNARDES, 2015). Porém com o avanço das tecnologias relativas à Big Data percebeu-se que a análise realizada sobre estas informações pode ser um fator determinante para a tomada de decisão em vários contextos.

As tecnologias de Big Data retratam uma nova geração de arquiteturas, pois precisam trabalhar com distribuição do processamento e permitir alta velocidade na captura ou análise dos dados, ou seja, precisam suportar aplicações com volumes de dados que crescem substancialmente em pouco tempo. Nesse sentido surge o Apache Cassandra um banco de dados NoSQL altamente escalável, indicado para gerenciar grandes quantidades de dados, sejam eles estruturados, semiestruturados ou não estruturados.

Buscando fazer a coleta dos dados referente a um domínio do conhecimento, surge o termo Ontologia. Aplicada a Ciência da Computação e Informática, a Ontologia é uma estrutura de dados utilizada para representar um conjunto de termos de uma determinada área do conhecimento ou domínio. A mesma pode ser utilizada para vários propósitos, como para melhorar as sugestões dadas em um sistema de recomendação de produtos ou notícias, ou na consulta, com a finalidade de aumentar a precisão de um sistema de recuperação da informação para posterior análise.

A realização de análises em sistemas de apoio a decisão, tem como ponto central de sua arquitetura de processamento um Data Warehouse, pois o mesmo possui sólida integração de dados corporativos e históricos.

Segundo Inmon (1997), Data Warehouse pode ser definido como uma coleção de dados orientados a assuntos, integrados, não voláteis e variáveis em relação ao tempo para apoio ao processo gerencial de tomada de decisão.

Na modelagem de dados de um DW, a configuração que regula a organização dos fatos e das dimensões para armazenamento corresponde geralmente a um esquema estrela (MACHADO, 2000).

Fatos de uma forma simples consistem em um assunto de negócio, enquanto que as dimensões são os elementos que participam do assunto. Como exemplo pode-se citar um fato vendas, onde suas dimensões seriam vendedor, cliente, produto, região e tempo.

Observando a necessidade de se ter um modelo que possibilite a integração dos dados do Big Data a arquitetura de um Data Warehouse para que posteriormente informações possam ser analisadas, foi desenvolvido um ambiente de consultas em um Big Data, fazendo-se valer da estrutura de fato e dimensões, utilizando a modelagem estrela de dados, empregada em um Data Warehouse.

1.1 Objetivo Principal

Este trabalho visa especificar e desenvolver um ambiente de consultas em um Big Data, considerando um esquema estrela de dados. Após os dados serem carregados no Big Data, bem como no Data Warehouse, o sistema irá possibilitar ao usuário realizar pesquisas no Big Data, através dos termos selecionados nas dimensões do esquema estrela de dados.

1.2 Objetivos Específicos

- Definir formas de mapeamento e identificação de objetos fato e objetos dimensão em um Big Data;
- Elaborar maneiras para consulta e transformação dos dados extraídos para visualização;
- Descrever formas de visualização das informações para posteriormente serem analisadas.

O próximo capítulo descreve a Fundamentação Teórica, onde são abordados os temas desse trabalho. No capítulo 3 são apresentados os Trabalhos Relacionados. Já no capítulo 4 é

mostrada a Solução Proposta, seguido do capítulo 5 que apresenta a validação do trabalho. Por fim, no último capítulo são apresentadas as Conclusões do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo objetiva abordar os fundamentos teóricos para o desenvolvimento do trabalho, visando apresentar conceitos literários das áreas e tecnologias que fazem parte do escopo desse trabalho. Apresenta também quais os conhecimentos necessários para o desenvolvimento da solução desenvolvida.

2.1 Big Data

A constante evolução da computação nos faz produzir uma quantidade imensa de informações. Este vasto volume acaba por nos trazer dificuldades em maximizar os proveitos que podem ser obtidos através da análise destes dados. (MARTINS, 2014).

Conforme Taurion (2013) estes dados provêm das mais diversas fontes, pois, além dos gerados pelos sistemas transacionais das empresas, se tem a imensidão de dados gerados pelos objetos na Internet das Coisas, como sensores e câmeras, e os gerados nas mídias sociais via PCs e dispositivos móveis.

Devido a essa grande variedade de informações surge o conceito de Big Data. De acordo com Pereira (2014), este conceito não possui uma definição formal única adotada por todos. De uma forma simples e pragmática o Big Data pode ser definido formalmente como um grande volume de dados, que são disponibilizados com diferentes graus de complexidade, sendo gerados a diferentes velocidades e graus de ambiguidade. O que resulta numa complexidade que está além da suportada pelas tecnologias, métodos de processamento e algoritmos tradicionais (PEREIRA, 2014).

2.1.1 Principais Características

Os dados Big Data possuem três características que quando em conjunto, os diferenciam de todos os outros tipos de dados. Estes aspectos são conhecidas como os três “Vs” dos dados Big Data sendo eles, volume, velocidade e variedade (SINGH, 2012). Encontram-se alguns autores que associam mais “Vs”, estes retratam que valor e veracidade dos dados também são duas características que distinguem os dados Big Data de todos os outros tipos de dados (DEMCHENKO et al., 2013; SATHI, 2012).

2.1.1.1 Volume

Apesar de atualmente se ter uma quantidade enorme de informações armazenadas, grande parte dela não está a ser analisada. No entanto, as organizações ainda continuam a armazenar dados na esperança de que eles venham a revelar importantes descobertas no futuro. Estas informações são provenientes de diversas fontes tais como, aplicações de gestão, redes sociais, sensores acoplados a dispositivos eletrônicos, dispositivos móveis entre outros (MARTINS, 2104).

Conforme Pereira (2014), a necessidade de fazer com que todos estes dados possam estar acessíveis ao mesmo tempo, em que se demostram de fácil pesquisa, processáveis e gerenciáveis, tornam o volume a característica mais importante e desafiadora no mundo Big Data.

2.1.1.2 Velocidade

Conseguir gerir todos os dados que chegam com extrema rapidez as organizações torna-se uma tarefa difícil e desafiante. Para se tirar proveito de todas as potencialidades dos dados Big Data, os dados muitas vezes carecem ser analisados ainda em fluxo e não quando se encontram em repouso (ZIKOPOULOS et al., 2012). O simples fato de os dados necessitarem ser analisados em breves instantes acaba se tornando um problema para algumas organizações, pois a velocidade com que os dados se apresentam, acabam ultrapassando em muito a capacidade disponível para o processamento desses dados (PEREIRA, 2014).

Muitas instituições necessitam de em pequenos instantes de tempo, verificar se os dados que se apresentam são relevantes e requerem ser melhor analisados, ou se esses dados precisam ser associados com outros dados para assim fornecerem um conhecimento relevante (PEREIRA, 2014).

Uma situação que demostra a necessidade de um dado ser analisado de forma veloz, pode ser vista em uma câmera de vigilância ao identificar um criminoso procurado, onde um evento precisa ser gerado instante depois de estas imagens terem sido obtidas. Outro cenário são as instituições que sustentam as redes sociais, estas têm a necessidade de criar publicidade personalizada instantes depois de uma pessoa ter demostrado interesse por um determinado produto em específico.

2.1.1.3 Variedade

Os dados do Big Data originam-se de diversas fontes internas e externas. Estes dados podem ser estruturados, semiestruturados e não estruturados (PEREIRA, 2014).

Os estruturados, conforme Martins (2014), são dados organizados em entidades sendo estas agrupadas através de relações e classes. Todos são armazenados em Sistemas de Gerenciamento de Banco de Dados (SGBD). Denominam-se estruturados porque possuem sua estrutura rígida, que foi previamente projetada através de um modelo de entidades e relacionamentos.

Nos semiestruturados não há a necessidade de se armazenar os dados em um Sistema de Gerenciamento de Banco de Dados (SGBD). Estes acompanham padrões heterogêneos, são mais difíceis de serem identificados, visto que podem seguir diversos padrões. Como exemplo temos, *Extensible Makup Language* (XML), o conteúdo de um e-mail, entre outros (MARTINS, 2014).

Já os não estruturados, são dados que não possuem necessariamente um formato. Segundo Martins (2014), estes dados são o foco de muita atenção na atualidade, em virtude da proliferação de dispositivos móveis responsáveis por uma grande variedade de dados. No entanto encontram-se outras fontes de dados como: sensores de máquinas, dispositivos inteligentes e as redes sociais. Portanto, estes dados não são dados relacionados mas sim variados. Alguns exemplos deste tipo de dados que podem ser citados são textos, vídeos e imagens.

Esta variedade coloca alguns problemas e desafios aos sistemas tradicionais, que não são adequados para executar as requeridas análises e extrair conhecimento de dados com tamanha complexidade. (ZIKOPOULOS et al., 2012).

2.1.1.4 Valor e Veracidade

A maior parte dos dados Big Data tem origem fora do controle da organização. Desse modo, a veracidade dos dados se torna um elemento fundamental, visto que decisões importantes serão tomadas com base no conhecimento obtido através destes dados. A veracidade no contexto Big Data representa a confiança nos dados apresentados, a

credibilidade da fonte desses dados e também a adaptação dos mesmos ao público desejado (SATHI, 2012).

Em um Big Data conforme Pereira (2014), a veracidade garante que os dados sejam confiáveis, legítimos e protegidos de acessos e alterações não concedidas. Para garantir a veracidade dos dados, precauções devem ser tomadas para que se assegure a integridade dos dados ao longo de sua existência. Estes dados devem possuir fontes confiáveis, bem como devem ser manipulados e armazenados em estruturas confiáveis.

O valor é uma característica muito importante dos dados. Este é estabelecido pelo conhecimento e informação que os dados trazem para o processo de análise. O valor a ser retirado de um conjunto de dados irá determinar qual o volume de dados deve ser excluído, assim como o período que estes devem ser armazenados. É provável que um grupo de dados se mostre importante apenas quando associado a outros dados, pertinentes a eventos que ainda não aconteceram, ou somente quanto um padrão equivalente surgir. Assim sendo, o valor dos dados está profundamente associado com o volume e a variedade dos mesmos (PEREIRA, 2014).

2.2 Apache Cassandra

Cassandra trata-se de um banco de dados de código aberto NoSQL (*Not Only SQL*), com o seu armazenamento orientado à colunas e escrito em Java. Foi criado pelo Facebook e posteriormente incorporado a Fundação Apache. Cassandra é utilizado por empresas como Netflix, Twitter, Ebay, Adobe e Cisco (MARROQUIM; RAMOS, 2012).

Conforme Leite (2014) Cassandra é um banco de dados altamente escalável e é indicado para gerenciar grandes quantidades de dados estruturados, semiestruturados ou não estruturados. O mesmo é focado em alta disponibilidade do serviço e durabilidade das informações gravadas. Sua arquitetura distribuída é descentralizada e, portanto, livre de pontos únicos de falha (MARROQUIM; RAMOS, 2012).

2.2.1 Modelo de Dados

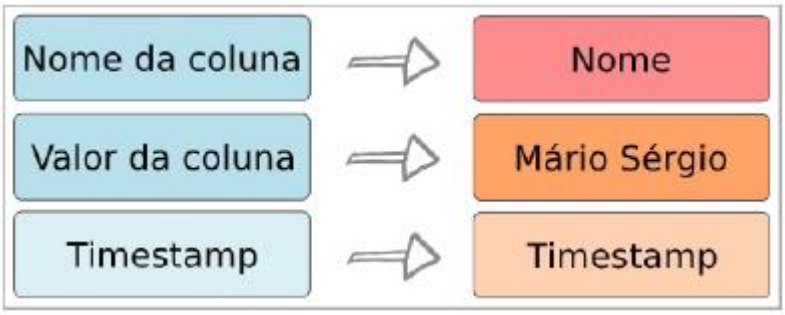
O Cassandra é baseado no modelo de dados orientado à coluna. Sua estrutura é formada por um *keyspace* contendo famílias de colunas, que por sua vez é um conjunto de

linhas englobando várias colunas (LEITE, 2014). A seguir será apresentado cada parte desta estrutura.

2.2.1.1 Coluna

Unidade básica de armazenamento a coluna, é formada pelo par nome e valor, sendo que o nome funciona como uma chave da coluna. Além destas informações cada coluna possui um valor *timestamp* que fornece informações acerca do valor temporal em que os dados foram inseridos ou atualizados permitindo entre outras coisas resolver conflitos em operações ou lidar com dados que estejam desatualizados (CUNHA, 2015).

Figura 1: Coluna como Elemento Básico



Fonte: Marroquim e Ramos (2012).

2.2.1.2 Super Coluna

Uma super coluna pode ser entendida como uma coluna com sub-colunas. Conforme Marroquim e Ramos (2012) da mesma forma que uma coluna, uma super coluna é formada por um par chave/valor ainda que neste caso o valor esteja associado a um mapa que contém várias colunas.

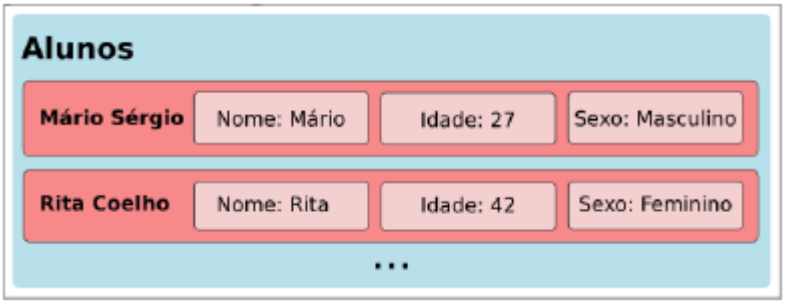
Figura 2: Super Coluna

Fonte: Marroquim e Ramos (2012).

2.2.1.3 Família de Colunas

A família de colunas é formada por um número infinito de linhas, sendo que cada linha contém um chave e um conjunto de colunas ordenadas pela sua chave (LEITE, 2014). De acordo com Sadalage e Fowler (2013) uma família de colunas pode ser comparada a um conjunto de linhas de uma tabela do modelo relacional em que cada linha possui uma chave e está associada a um conjunto de colunas. A grande diferença entre os dois modelos é o fato de que no modelo não relacional, não existe nenhum esquema pré-definido, sendo que as linhas não têm que ter exatamente as mesmas colunas.

Figura 3: Família de Colunas

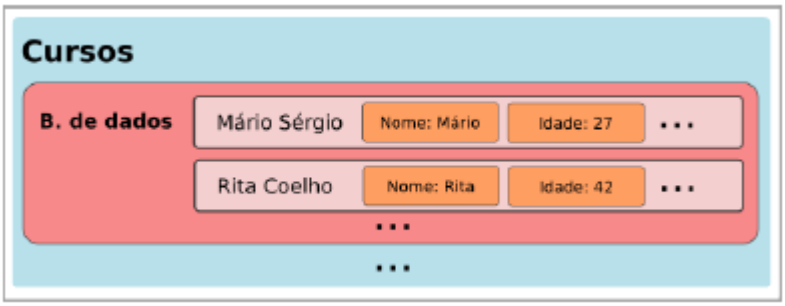


Fonte: Marroquim e Ramos (2012).

2.2.1.4 Super Família de Colunas

Semelhante a uma família de colunas, uma super família de colunas é formada por um conjunto de super colunas, sendo que são úteis para manter dados que estão relacionados juntos (SADALAGE; FOWLER, 2013).

Figura 4: Família de Super Colunas

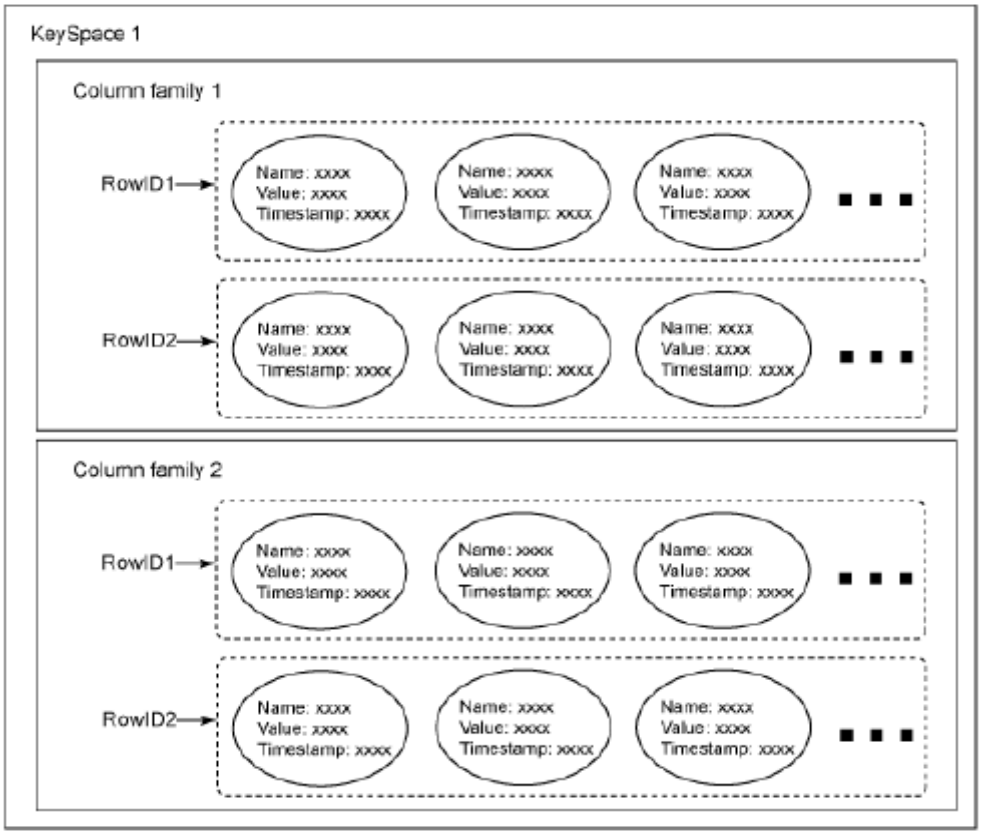


Fonte: Marroquim e Ramos (2012).

2.2.1.5 Keyspace

Keyspace é a estrutura mais externa para os dados do Cassandra onde se encontram todas as famílias de colunas relacionadas com a mesma aplicação. O *keyspace* quando da sua criação define a forma como os dados serão replicados ao longo do *cluster* (CUNHA, 2015).

Figura 5: Modelo de Dados Cassandra

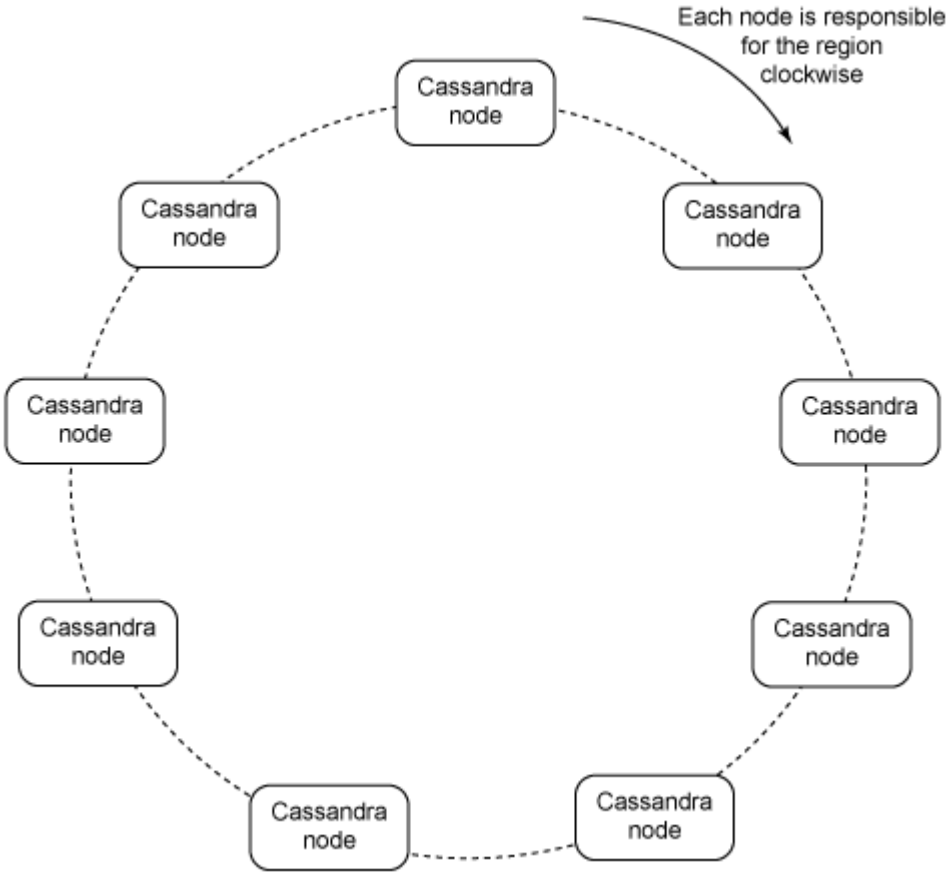


Fonte: Leite (2014).

2.2.1.6 Cluster

Cassandra é projetado especificamente para ser distribuído em várias máquinas operando em conjunto. Sua estrutura mais externa é o *cluster*, também chamado de anel, visto que os nós do *cluster* são organizados em forma de anel. Um nó mantém uma réplica para diferentes intervalos de dados, de modo que se um nó falhar, uma réplica em outro nó poderá responder às consultas (LEITE, 2014).

Figura 6: Cluster do Cassandra



Fonte: Perera (2012).

2.2.2 Modelo de Consulta

O Cassandra possui uma linguagem de consulta muito similar ao SQL denominada de CQL - *Cassandra Query Language*. Uma vez que a sua sintaxe é muito semelhante ao SQL torna-se mais fácil utiliza-la, verificando-se assim uma curva de aprendizagem reduzida (ABRAMOVA et al. 2014). No entanto, CQL não permite efetuar joins ou subconsultas dos dados (SADALAGE; FOWLER, 2013).

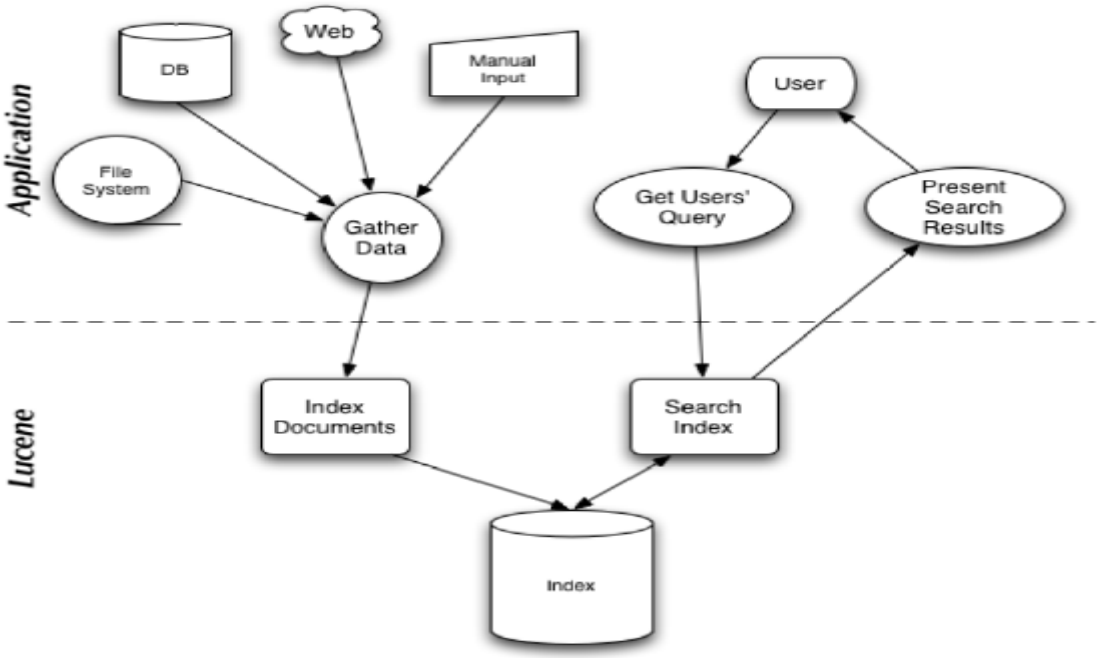
2.3 Apache Lucene

Lucene trata-se de uma biblioteca de software livre para indexação e recuperação de informações que originalmente é escrita em java. Criada por Doug Cuttingol no ano 2000,

posteriormente foi aprimorada e depois incorporada a Fundação Apache (MILHOMEM, 2013).

Conforme Andrade (2010) Lucene concede um bom nível de abstração para um conjunto poderoso de técnicas baseadas no modelo Vetorial e Booleano. A biblioteca Lucene é composta por duas etapas: indexação e pesquisa. Fundamentado em palavra-chave o algoritmo processa os dados gerando uma estrutura que possibilita a realização de consultas. A figura 7 demonstra uma típica aplicação que está integrada ao Lucene.

Figura 7: Arquitetura Apache Lucene



Fonte: Correia (2016).

As etapas ilustradas na figura 7 podem ser descritas da seguinte forma:

- *Gather Data*: Momento em que são recolhidos os conteúdos dos diferentes documentos.
- *Index Documents*: Etapa na qual onde é feita a análise do documento, sendo que após esta, tem início o processo de indexação.
- *Index*: Base de dados de índices.
- *Get Users' Query*: Uma vez feito o pedido de pesquisa do usuário, a aplicação constrói a query com base no texto de pesquisa, com a finalidade de utilizar a mesma na interrogação às bases de dados de índices.

- *Search Index*: Etapa onde se realiza a pesquisa por índices segundo a query construída a partir dos dados de pesquisa introduzidos pelo usuário.
- *Present Search Results*: Responsável por apresentar o resultado da pesquisa ao usuário.

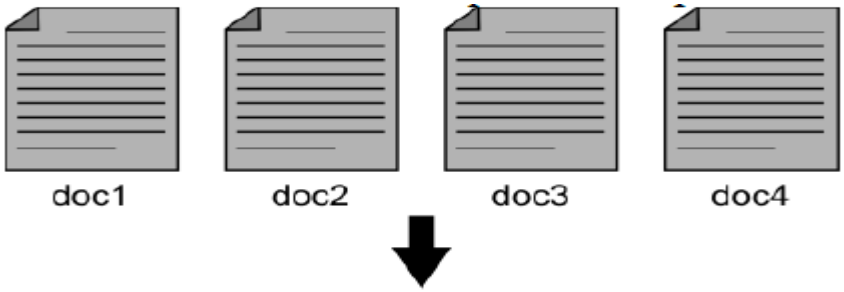
Lucene é usado para indexar e pesquisar dados em páginas de web, documentos armazenados em sistemas locais de arquivo, arquivos de texto simples, HTML ou qualquer outro formato a partir do qual é possível coletar informações textuais (ANDRADE, 2010).

No entanto para pesquisar grandes quantidades de texto de forma rápida, em um primeiro momento de acordo com Andrade (2010), Lucene indexa o texto e o converte em um formato que o permiti deixá-lo mais rápido na busca. Este processo é chamado de indexação, e sua saída é chamada de índice. Já a busca ou pesquisa se refere ao processo de procurar palavras em um índice para localizar documentos onde as mesmas aparecem. Logo há basicamente duas funcionalidades importantes: o processo de indexação, acessível através do comando *Indexer*, e o processo de busca, disponível por meio do comando *Searcher*.

2.3.1 Indexação

O Lucene conforme Machado (2013) utiliza em seu índice a estrutura de dados chamada de índice invertido, onde cada termo adicionado possui uma referência para o arquivo que o contém, conforme ilustrado na figura 8.

Figura 8: Representação da Indexação com Índice Invertido



Extração de termos relevantes
para identificação do documento

ID	Termo	Documento
1	termo1	doc1, doc4
2	termo2	doc2,doc3
3	termo3	doc1, doc2, doc3
4	termo4	doc1, doc2
5	termo5	doc3,doc4
6	termo6	doc3
7	termo7	doc2

Fonte: Prado (2012).

Usualmente, o processo de indexação na área de recuperação de informação segue alguns passos, sendo que primeiramente ocorre a aquisição dos termos necessários através de um coletor. Após os termos terem sido coletados é necessária à construção de uma abstração do conteúdo a ser indexado chamado documento, que é constituído por unidades denominadas campos, que são elementos nominais associados aos termos. Assim sendo, os documentos serão retornados de acordo com o pareamento dos termos requisitados durante a busca com os termos contidos no índice. No entanto antes da indexação propriamente dita pode ocorrer a etapa de análise do documento que tem como função o aprimoramento da fase de busca através da separação dos termos em uma série de elementos atômicos chamados *tokens*. Por fim o documento é adicionado ao índice seguindo um formato próprio do Lucene (PRADO, 2012).

2.3.2 Busca

No Lucene para cada documento presente no resultado de alguma busca é atribuído uma pontuação que representa a similaridade de tal documento com a consulta. O cálculo

dessa pontuação é feito baseando-se no modelo de recuperação de informação escolhido. Segundo Machado (2013) Lucene suporta os seguintes modelos:

- Modelo Booleano;
- Modelo Espaço Vetorial;
- Modelo Probabilístico, como Okapi BM25 e DFR;
- Modelo baseado em Linguagem Natural.

Sendo que por padrão a busca no Lucene ocorre através da combinação de duas técnicas de recuperação de informação: Modelo Espaço Vetorial e Modelo Booleano.

No entanto ao utilizar as funcionalidades do Lucene, o programador não tem a necessidade de implementar algoritmos de busca e classificação, pois a biblioteca possui mecanismos para calcular a pontuação de cada documento que corresponda a uma consulta e retornar documentos relevantes de acordo com essas pontuações. Também possui suporte a consultas como, *PhraseQuery*, *WildcardQuery*, *RangeQuery*, *FuzzyQuery*, *BooleanQuery*, além de permitir busca e indexação de forma simultânea (MILHOMEM, 2013).

Portanto para o cliente que faz a requisição de uma busca, o processo de pesquisa é efetuado apenas em uma base de índices indexada pelo Lucene, que faz uma lista de resultados relevantes ao contexto da busca (MILHOMEM, 2013).

2.4 Ontologia

Esta seção apresenta conceitos de ontologias, seus objetivos e propósito geral, algumas características, linguagens empregadas, bem como passos que devem ser observados na construção de uma ontologia.

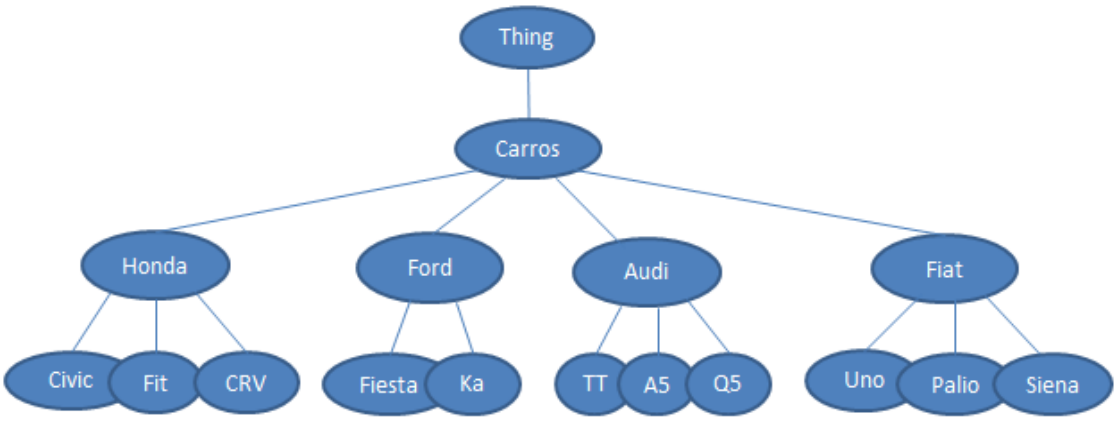
2.4.1 Conceitos e Objetivos

Segundo Almeida e Bax (2003), a palavra ontologia deriva do grego "*ontos*", ser, e "*logos*", discurso escrito ou falado. Esse termo vem do campo da filosofia que se atenta ao estudo do ser ou existência. Na filosofia, pode-se falar de uma ontologia como uma teoria sobre a natureza da existência.

Na literatura encontram-se várias definições para o termo ontologia em Ciência da Computação. Segundo Gruber (1992), a ontologia aplicada à Ciência da Computação e

Informática é um termo que expressa um objeto que é projetado para um propósito: permitir a modelagem de conhecimento sobre algum domínio, sendo ele real ou imaginário. Diferente do conceito filosóficos, na área de TI as ontologias são consideradas como conjuntos de definições e conceitos.

Figura 9: Exemplo de uma ontologia de carros e marcas



Fonte: Autor (2016).

Para Souza e Alvarenga (2004), o termo "ontologia" foi adaptado e, para os profissionais da ciência da computação, uma ontologia é um documento ou um arquivo que especifica formalmente as relações entre termos e conceitos, mantendo assim, semelhanças com os tesauros utilizados para a definição de vocabulários controlados.

Um tesauro, como explica Cavalcanti (1978), citado por Marques de Jesus (2002, p. 3), é uma lista estruturada de termos empregada por analistas de informação e indexadores, para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procura.

Conforme Souza e Alvarenga (2004), o objetivo da construção de uma ontologia é suprir a necessidade de um vocabulário compartilhado para a troca de informações entre os membros de uma comunidade, sejam eles seres humanos ou agentes inteligentes.

Segundo Moraes e Ambrósio (2007) ontologias são utilizadas em projetos de domínios como gestão do conhecimento, comércio eletrônico, processamento de linguagens naturais, recuperação da informação na Web, áreas educacionais, entre outros.

Quanto à recuperação da informação na Web, as ontologias podem ser usadas com o propósito de melhorar a exatidão de buscas. Desta forma, um programa de busca baseado em ontologias poderá ser capaz de recuperar somente as páginas relevantes para o usuário.

2.4.2 Características

A ontologia define as regras que regulam a combinação entre os termos e as relações em um domínio do conhecimento. Sendo que as relações entre os termos são criadas por especialistas. Os conceitos especificados pela ontologia podem assim ser consultados pelos usuários, buscando assim melhorias no processo de recuperação da informação (ALMEIDA; BAX, 2003).

Baroni (2011) diz que as ontologias podem ser categorizadas em Horizontal e Vertical. Onde na Horizontal se representa as relações lexicais entre os conceitos da linguagem, procurando representar todos os conceitos possíveis com uma descrição não muito detalhada. Já a Vertical se aplica aos conceitos de uma área específica, tendo ao contrário da Horizontal, em seu conceito, uma descrição completa relativa ao contexto em que está inserida.

Apesar de as ontologias terem elementos diferentes entre si, a maioria possuem elementos básicos semelhantes entre elas. Alguns desses elementos conforme Almeida e Bax (2003) são:

- Classes – Normalmente organizados em taxonomias. As classes representam alguma relação entre o domínio escolhido e a ontologia;
- Relações – Representam o tipo de interação entre os elementos de um domínio (classes);
- Axiomas – Usados para modelar sempre sentenças verdadeiras;
- Instâncias – Utilizadas para representar elementos específicos, ou seja, os próprios elementos da ontologia;
- Atributo – Uma junção de nome e valor e é usado para guardar informações relevantes sobre determinados objetos.

2.4.3 Vocabulário e Linguagens

O vocabulário concebido por predicados lógicos forma a rede conceitual das ontologias. A ontologia determina as regras que regulam a combinação entre os termos e as relações. Já as relações entre os termos são criadas por especialistas, sendo que os usuários formulam consultas usando os conceitos especificados. Uma ontologia representa assim uma

linguagem (conjunto de termos) que será utilizado para formular consultas (ALMEIDA; BAX, 2003).

A escolha de uma linguagem para a especificação de ontologias varia conforme o tipo a ser especificado. Portanto conforme Baroni (2011) para construir e representar uma ontologia pode-se utilizar linguagens como RDF (*Resource Description Framework*), DAML (DARPA *Agent Markup Language*), OIL (*Ontoly Interface Layer*) e o padrão hoje utilizado, OWL (*Web Ontology Language*).

A linguagem *Web Ontology Language* conhecida como OWL, trata-se de uma linguagem para ser utilizada, quando as informações presentes em documentos web precisam ser processadas por aplicações em situações em que seu conteúdo precisa mais do que ser apresentado apenas para humanos. Os elementos básicos para construção de uma ontologia OWL são as classes, as instâncias das classes e os relacionamentos entre estas instâncias (BULSING, 2013).

Segundo McGuinness e Harmelen (2004), para uso de diferentes usuários e comunidades de desenvolvedores, a linguagem OWL possui três sub-linguagens que são:

- OWL Lite – Desenvolvida para usuários que necessitam de uma simples hierarquia. Como principais pontos positivos, possui facilidade na hora de fornecimento de suporte, possibilita migração rápida para enciclopédias e outras taxonomias e possui uma complexidade formal menor que as outras duas sub-linguagens;
- OWL DL – Para usuário que buscam máxima expressividade, mantendo a integridade computacional e a decidibilidade, ou seja, há a garantida de serem computáveis todas as conclusões e todas as computações terminarão em tempo finito;
- OWL Full – Voltada para usuários que buscam a liberdade sintática do RDF sem garantias computacionais. O RDF é uma linguagem para intercâmbio de dados na web, sendo um de seus principais objetivos a criação de um modelo simples de dados que possa suportar o uso de XML.

Cada uma das sub-linguagens aqui mencionadas é uma extensão de seu antecessor, ou seja, toda a ontologia de um OWL Lite é legal para uma ontologia OWL DL, assim como toda a ontologia OWL DL é legal para uma ontologia OWL Full, mas o contrário disso é ilegal (MCGUINNESS; HARMELEN, 2004).

2.4.4 Construção da Ontologia

Ontologias são empregadas em diversas áreas do conhecimento, principalmente na Web Semântica, área essa que está crescendo à medida que a internet está migrando em direção a ela. Por ter várias utilidades, há, então, a necessidade de métodos e técnicas que apoiem o desenvolvimento de ontologias.

Baroni (2011) definiu alguns passos para a construção de ontologia em OWL, que segundo ele, são de suma importância:

- Enumerar os termos do domínio – Momento em que deve-se observar as repetições de termos, similaridade e possíveis relações entre eles. Ou seja, escolher quais nomes serão dados aos objetos da nossa ontologia, levando em consideração sua classificação. Por exemplo, em uma ontologia de automóveis, definir suas classes e subclasses, podendo ser tração, quatro rodas, duas rodas, ou tipo de combustível, gasolina, álcool.

- Definir as classes – Etapa em que deve-se tomar cuidado para que não haja confusão entre conceitos e classes no que se refere aos nomes. Ou seja, é nesse momento que se define que “tração” e “tipo de combustível” são classes, e “gasolina”, “álcool”, “duas rodas”, “quatro rodas” são instâncias.

- Definir a hierarquia das classes – Nesse passo é onde são criadas subclasses. É realizado no mesmo instante que o passo anterior, e é considerado o mais importante passo para o desenvolvimento. Aqui se deve observar para não serem criadas muitas subclasses e verificar se o uso de classes intermediárias não seria o mais adequado a ser feito. Ou seja, é nesse momento que dizemos que os objetos “duas rodas” e “quatro rodas” pertencem ao objeto “tração”.

- Definir os atributos e facetas de cada classe – Exceto em classes terminológicas, novos atributos são responsáveis pela definição de uma classe, consequentemente, esse passo deve interagir com os dois anteriores. Por exemplo, podemos criar atributos para um automóvel como número de portas, tipo de motor, etc.

- Criar as instâncias – Instâncias são os conceitos de menor granularidade de uma ontologia.

- Definir convenções de nomes – Devem ser evitadas abreviações para não gerar confusão aos usuários. Para que a ontologia seja legível a todos que querem utilizá-la, devem ser colocados nomes compreensíveis, de fácil entendimento, adotando convenções diferentes para classes, instâncias e atributos. Por exemplo, não adotar abreviações que podem ser difíceis de compreender ou abreviações que podem gerar mais de um significado.

2.5 Data Warehouse

O Data Warehouse conforme Inmon e Hackathorn (1997) é o ponto central da arquitetura de processamento de informações para sistemas de apoio á decisão (SAD), visto que suportam o processamento informacional através de um alicerce sólido de integração de dados corporativos e históricos para a realização de análises gerenciais.

Segundo Inmon (1997), Data Warehouse pode ser definido como uma coleção de dados orientados a assuntos, sendo eles integrados, não voláteis e variáveis em relação ao tempo para apoio ao processo gerencial de tomada de decisão.

Esta definição acadêmica trás características importantes de um DW, que são melhor analisadas nas próximas sessões.

2.5.1 Características

Nas seguintes sessões são apresentadas características de suma importância para um Data Warehouse.

2.5.1.1 Orientação por Assunto

A primeira característica notável do DW é a orientada por assunto, pois conforme Machado (2000), ela significa que o DW armazena as informações agrupadas por assuntos de interesse da empresa que são de maior importância.

Também cabe salientar que segundo Machado (2000), os projetistas de DW devem ter seu foco na modelagem dos dados e no projeto de banco de dados. Sendo que em um DW somente importam os dados que sejam importantes para a tomada decisão.

2.5.1.2 Variação Tempo

Conforme Inmon e Hackathorn (1997), todos os dados no DW são precisos em algum instante no tempo. Esta característica básica dos dados do DW é muito diferente das encontradas em um ambiente operacional, pois nela quando se acessa uma unidade de dados, é esperado que esta reflita valores corretos no momento do acesso.

Em virtude de os dados no DW serem corretos como em algum momento no tempo, conforme Inmon e Hackathorn (1997) são ditos que estes variam com o tempo.

A variação do tempo dos dados do DW segundo Inmon e Hackathorn (1997) apresentam-se de diversas maneiras. Conforme Machado (2000), a primeira e a mais simples é aquela que os dados representam informações sobre espaços de tempos de cinco a dez anos. Já a segunda maneira são as estruturas básicas, onde cada estrutura contém um elemento tempo. E por fim a terceira maneira em que apresenta-se são os dados do DW, que uma vez armazenados corretamente, não podem ser atualizados (INMON; HACKATHORN, 1997).

2.5.1.3 Não Volatibilidade

Conforme Inmon e Hackathorn (1997), outra característica definidora do DW trata-se do mesmo não ser volátil. A manipulação básica dos dados em um DW é muito mais simples do que em ambientes operacionais, pois de acordo com Machado (2000) existem apenas dois tipos de operações que podem ocorrer em um DW, a carga inicial do dados e o acesso em modo de leitura.

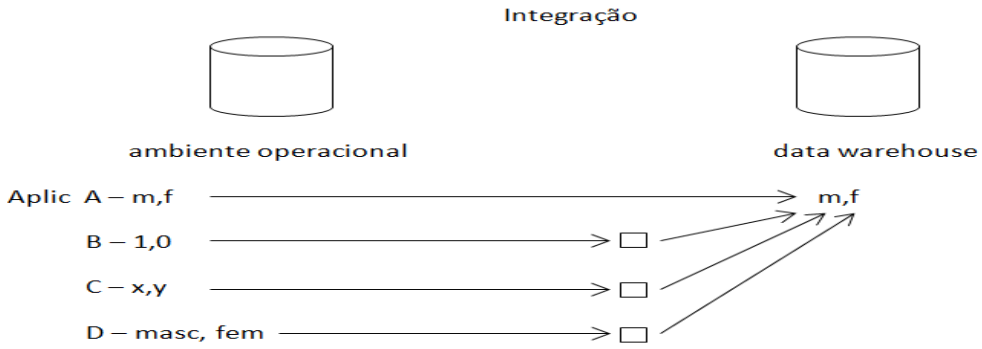
De acordo com Machado (2000), os dados ficam no DW até o momento que seja decido que os mesmos não fazem mais parte dele, ou que se tonaram irrelevantes para a análise de tomada de decisão.

Segundo Inmon e Hackathorn (1997), após os DW ter sido carregado, ele somente possui operações de consulta, e sem necessidade de qualquer tipo de bloqueio por concorrência de usuários no acesso.

2.5.1.4 Integração

Segundo Machado (2000), esta é uma das características de suma importância em um DW, pois todos os seus dados possuem um alto nível de integração.

Figura 10: Integração de Dados



Fonte: Adaptado de Inmon e Hackathorn (1997).

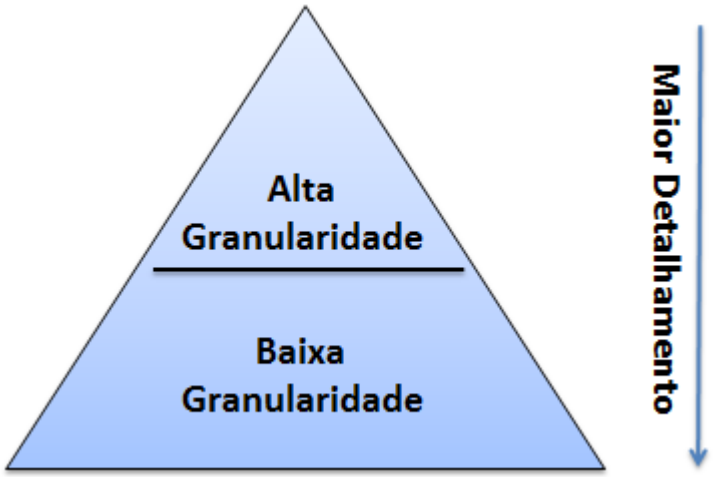
Conforme Inmon e Hackathorn (1997), os dados ao serem movidos de um ambiente operacional orientado a aplicações para o DW, são integrados antes de serem incluídos no DW. Estes mesmos autores afirmam também que os dados precisam ser armazenados no DW de uma forma única, mesmo quando as aplicações armazenam os dados de modo diferente.

2.5.2 Granularidade de Dados

Conforme Inmon (1997), o aspecto mais importante do projeto de um Data Warehouse faz referência a questão da granularidade, pois diz respeito ao nível de detalhe ou de resumo

contido nas unidades de dados existentes no DW. Segundo Machado (2000), quanto mais detalhe, mais baixo o nível de granularidade, quanto menos detalhe, mais alto o nível de granularidade.

Figura 11: Nível de Granularidade / Detalhamento



Fonte: Autor (2016).

O motivo no qual torna a granularidade a principal questão do projeto, de acordo com Inmon (1997) está no fato de que ela afeta profundamente o volume de dados que encontra-se no DW e, ao mesmo tempo, afeta o tipo de consulta que pode ser atendida.

Segundo Inmon (1997), o nível de granularidade exerce um profundo efeito tanto sobre as perguntas que podem ser respondidas, bem como sobre os recursos necessários para responder a uma pergunta.

A escolha do nível ou níveis de granularidade a serem utilizados em um projeto é indispensável para o sucesso. O método mais indicado para definir os níveis de granularidade conforme Machado (2000) está na utilização do bom senso e da análise detalhada das necessidades de informações levantadas para o projeto.

2.5.3 Modelagem Multidimensional

Esta sessão tem como objetivo apresentar conceitos e terminologias empregadas no processo de modelagem e na sequência deste trabalho.

2.5.3.1 Fatos

Conforme Machado (2000), um fato trata-se de uma coleção de itens de dados, composta de dados de medidas e de contexto. Um fato consiste em um item de negócio, uma transação de negócio ou um evento de negócio. O mesmo é utilizado para verificar o processo de negócio de uma empresa.

De acordo com Kimball (1998), tudo aquilo que reflete a evolução dos negócios do dia-a-dia de uma instituição, é um fato.

2.5.3.2 Dimensões

Quando trata-se de dimensão, está se referindo aos elementos que participam de um fato, assunto de negócio. As dimensões determinam o contexto de um assunto de negócios (MACHADO, 2000).

As tabelas dimensionais conforme Kimball (1998), normalmente não possuem atributos numéricos, uma vez que são somente textuais e classificatórias dos elementos que participam de um fato.

Uma dimensão de acordo com Machado (2000) pode conter membros e hierarquias. Os membros são uma classificação de dados dentro de uma dimensão. Estes membros de uma dimensão podem ser arranjados em uma ou mais hierarquias, que por sua vez podem conter vários níveis hierárquicos.

2.5.3.3 Variáveis

As variáveis ou medidas são os atributos numéricos de um fato. Elas representam o desempenho de um indicador de negócios referente às dimensões que participam desse fato. Uma medida é estabelecida pela combinação das dimensões que pertencem a um fato (MACHADO, 2000).

2.5.3.4 Operações Básicas

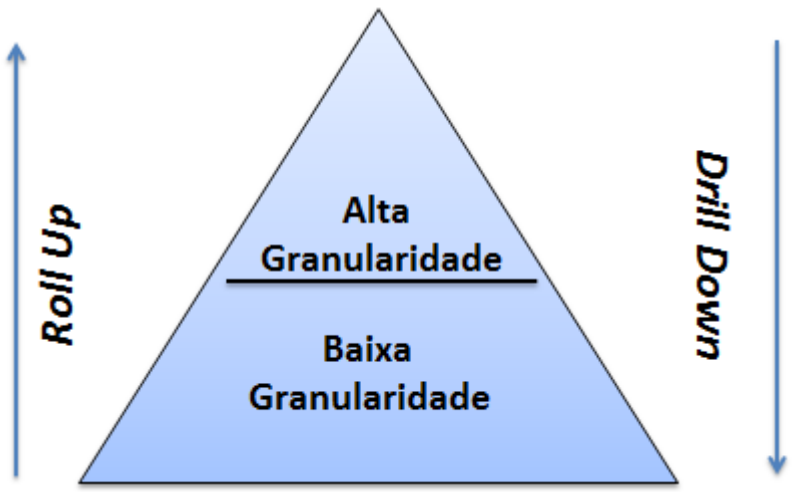
Em um modelo de dados multidimensional se possuí operações básicas de OLAP (*On-Line Analytic Processing*).

Conforme Kimball (1998), OLAP é um termo inventado para descrever uma abordagem dimensional para o suporte à decisão.

Estas operações são usadas para analisar dados, sendo duas delas "*drill down*" e "*roll up*". Para poder-se utilizar estas operações devesse fazer valer da granularidade (MACHADO, 2000).

Com a capacidade do "*drill down*" se esta diminuindo o nível da granularidade, aumentando assim o nível de detalhes. De maneira oposta a isso, o "*roll up*" aumenta o nível da granularidade, diminuindo desta maneira, o nível de detalhes das informações.

Figura 12: Drill Down – Roll Up



Fonte: Autor (2016).

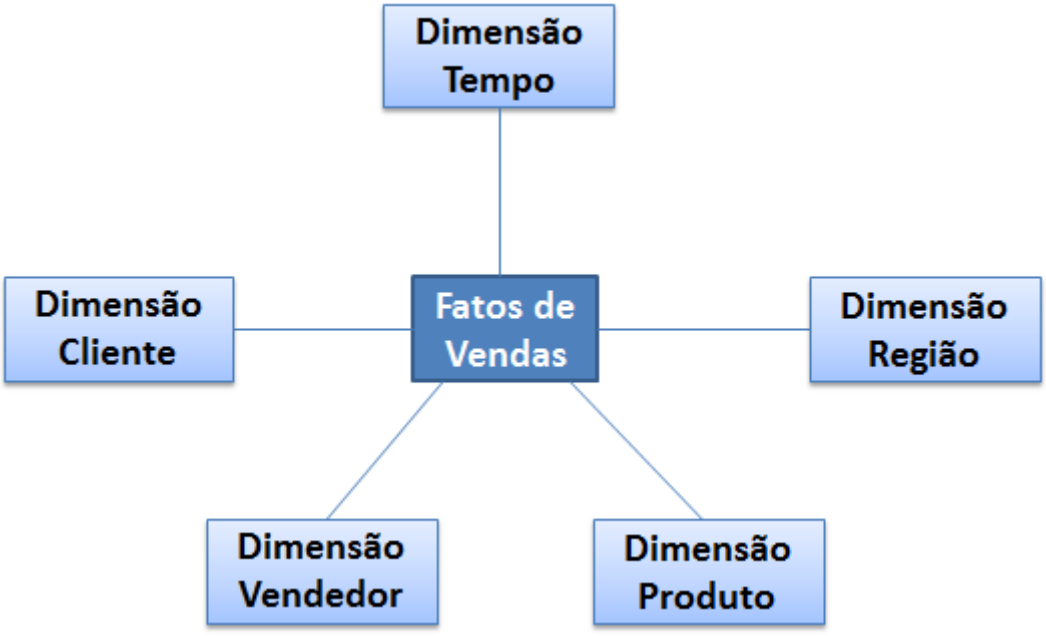
Portanto conforme Machado (2000), estas operações permitem movimentar nossa visão dos dados ao longo dos níveis hierárquicos de uma dimensão.

2.5.3.5 Modelo Estrela

Em um modelo de dados multidimensional, a configuração que regula a organização dos fatos e das dimensões para armazenamento corresponde geralmente, a um esquema em estrela (MACHADO, 2000).

Este é composto por uma grande entidade central denominada tabela de fatos, e um conjunto de entidades menores denominadas tabelas de dimensões, que por sua vez estão organizadas ao redor da entidade central, formando assim uma estrela, conforme mostra a figura 13.

Figura 13: Modelo Estrela



Fonte: Adaptado de Machado (2000).

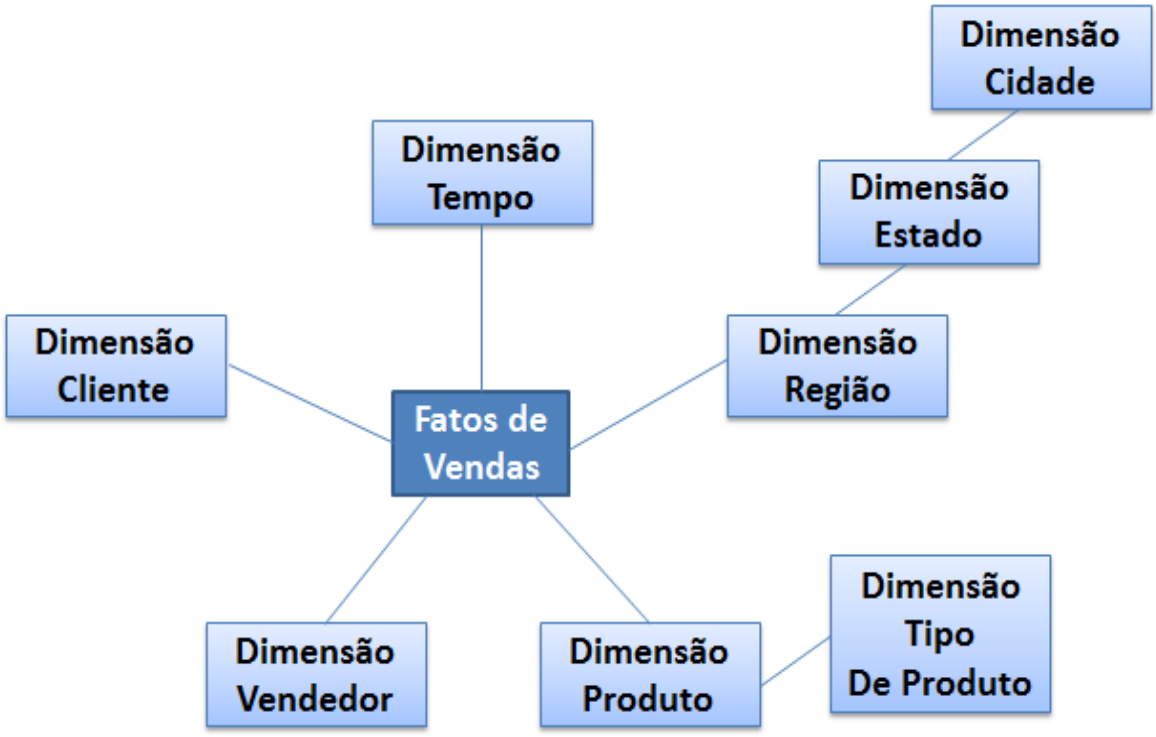
Neste modelo o relacionamento entre a entidade fato e as dimensões é uma simples ligação entre duas entidades, em um relacionamento de um para muitos no sentido da dimensão para o fato.

2.5.3.6 Modelo Floco de Neve

O modelo floco de neve da mesma forma que o modelo estrela possui uma entidade central denominada de fatos e um conjunto de entidades dimensão ao seu redor formando uma estrela.

No entanto o modelo floco de neve de acordo com Machado (2000), é o resultado da decomposição de uma ou mais dimensões que possuem hierarquias entre seus membros. Conforme ilustrado na figura 14.

Figura 14: Modelo Floco de Neve



Fonte: Adaptado de Machado (2000).

Segundo Machado (2000), este modelo é o resultado da aplicação da terceira forma normal sobre entidades dimensão.

Na construção de DWs, desenvolvedores frequentemente preferem a utilização deste modelo pelo fato de conservar a utilização de meios de armazenamento. Pois como é um modelo normalizado, evita a redundância de valores textuais em tabelas.

No entanto, conforme Machado (2000), um DW não possui inclusão de dados por meio de digitação, não necessitando assim garantir que os valores textuais sejam únicos, e nem tão pouco se preocupar com a economia de espaço, mas sim garantir o preceito de informação rápida.

O modelo floco de neve é esteticamente melhor para visualização de hierarquias, no entanto para se realizar consultas neste modelo são necessários mais joins, resultando assim em um gasto maior de tempo.

2.5.4 Fatos e Dimensões

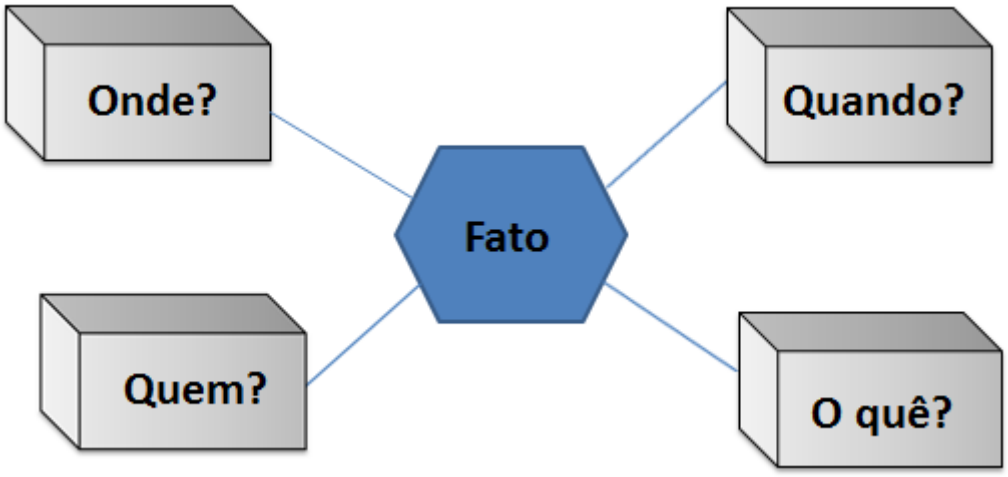
Um fato é considerado tudo aquilo que pode ser representado por um valor aditivo, ou seja, por meio de valores numéricos. Este conjunto de valores é denominado de métrica (MACHADO, 2000).

Outra característica relevante que serve para identificar um fato segundo Kimball (1998) é que este é evolutivo, suas medidas variam ao logo do tempo, podendo ser sempre questionado sobre esta evolução ao longo de um espaço de tempo.

Porém, para a modelagem e correta identificação de um fato segundo Machado (2000) é necessário descobrir quais elementos ou objetos participam de um fato.

Se um fato qualquer acontece, esse fato tem participantes, indicação do tempo em que acontece, onde acontece, quem está no fato, e o que está nele.

Figura 15: Esquema Fato Dimensões



Fonte: Adaptado de Machado (2000).

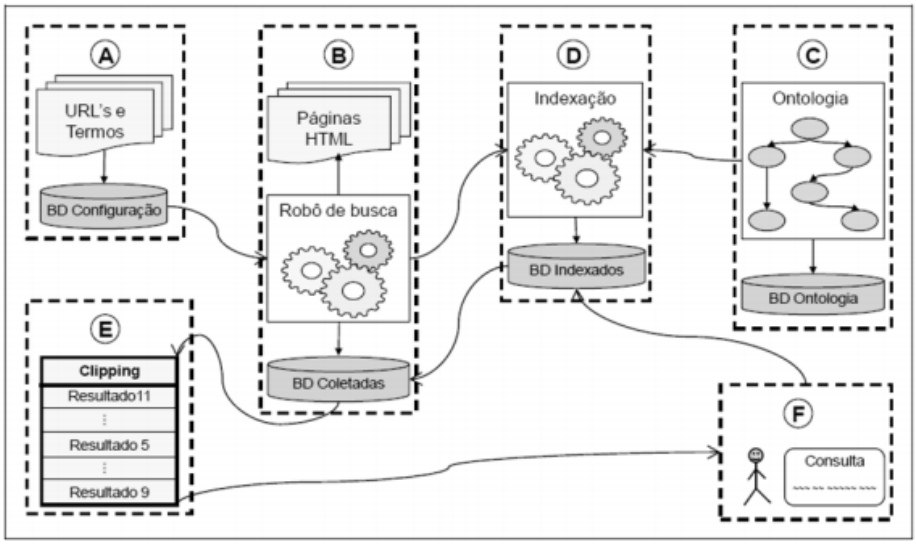
Estes elementos são referências que são utilizadas para facilitar a identificação de um fato, mas não obrigatoriamente todos os elementos devem estar presentes, somente um deles é obrigatório, o elemento de tempo. Tais elementos são conceituados como dimensões de um fato (MACHADO, 2000).

3 TRABALHOS RELACIONADOS

Esse capítulo tem como objetivo, apresentar alguns trabalhos que abordam assuntos já apresentados no capítulo 2 Fundamentação Teórica. Sendo que o foco principal dos trabalhos relacionados é: Modelagem dimensional, captura e indexação de textos para clipagem on-line com base em Ontologias , mapeamento entre Data Warehouse e Big Data, e modelagem de cubo OLAP através de ontologias.

Resultado do trabalho de conclusão de curso de Roberto Schuster Filho (2013), o Ontoclipping2, trata-se de uma evolução do Ontoclipping, desenvolvido por Claudio Omar Correa Carvalho Junior, que é uma ferramenta de clipagem on-line que utiliza uma ontologia como forma de representação do conhecimento, em conjunto com técnicas de recuperação da informação aliadas a um motor de busca de páginas web (SCHUSTER, 2013).

Figura 16: Fluxo do Ontoclipping



Fonte: Schuster Filho (2013).

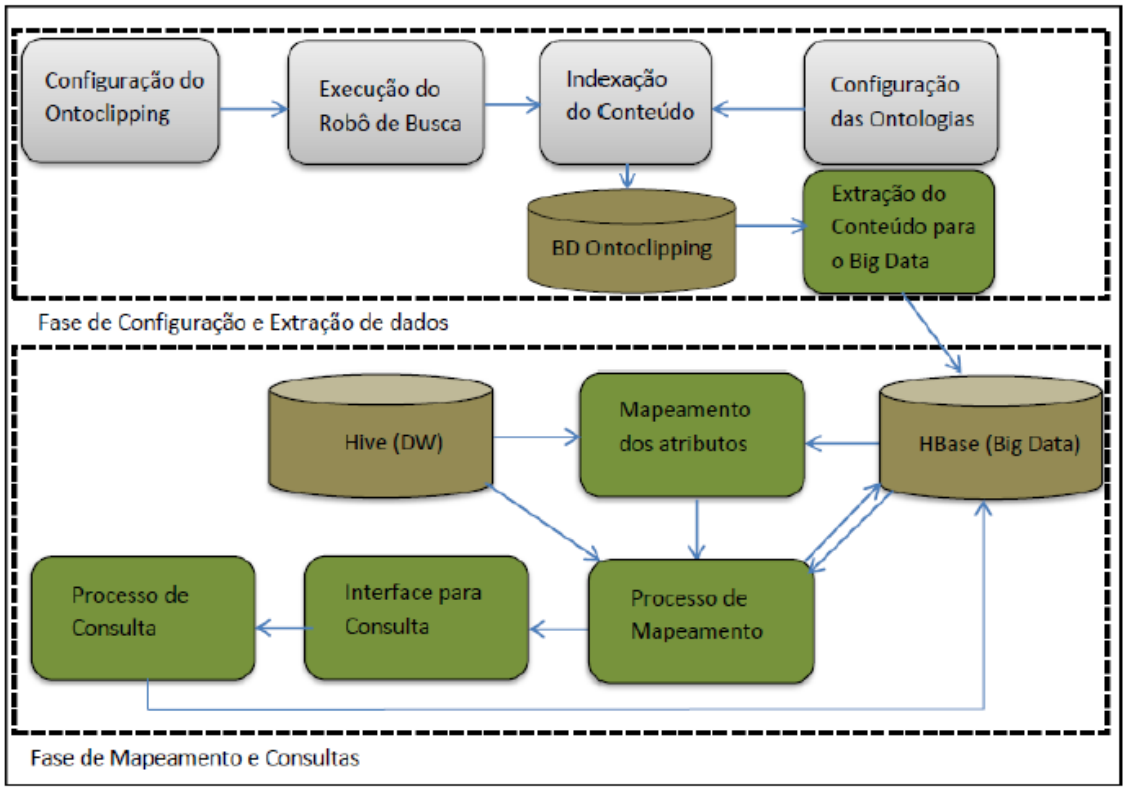
A figura 15 ilustra o fluxo do Ontoclipping, que possui as seguintes etapas: (A) cadastros de sites relevantes para o processo de *clipping*, denominados sites sementes ou sites raízes; (B) processo de *clipping*, ou seja, o robô de busca recupera informações a partir dos sites raízes, baseadas nas informações da etapa A; (C) criação de uma ontologia a fim de representar o conhecimento conforme as necessidades reais; (D) indexação do conteúdo coletado pelo robô de busca; (E) busca do conteúdo indexado, feito por um algoritmo próprio,

utilizando o conteúdo armazenado pelas etapas anteriores B e C e pela ontologia da etapa D; (F) consulta do usuário em uma interface gráfica.

Logo o trabalho realizado por Kuplich (2013) tem o objetivo de criar uma solução para auxiliar na análise da produção acadêmica do Programa de Pós-Graduação em Computação (PPGC) do Instituto de Informática da UFRGS. A solução toma como base o conceito de modelagem dimensional no desenvolvimento de um data warehouse para armazenar os dados de produção acadêmica e utiliza uma ferramenta OLAP para análise dos dados. Esses dados tem origem no Aplicativo de Coleta de Dados CAPES, que tem como objetivo coletar informações dos programas de pós-graduação no Brasil. A solução desenvolvida neste trabalho permite visualizar, de forma flexível e dinâmica, o desempenho da produção do PPGC ao longo do tempo, auxiliando nas decisões gerenciais relativas às pesquisas acadêmicas realizadas no programa.

Já o trabalho de conclusão de curso de Ricardo Schroeder (2013), nomeado de HiveHBase Integrator, consiste em buscar informações do HBase (Big Data), com base em informações mapeadas do Hive (Data Warehouse). Seu desenvolvimento foi direcionado para a solução de Big Data da IBM, o IBM BigInsights na versão 2.1. Seu funcionamento se baseia em duas fases: Fase de Configuração e Extração de dados e Fase de Mapeamento e Consulta (SCHROEDER, 2013). Conforme ilustrado na figura 17.

Figura 17: Fases do HiveHBase Integrator



Fonte: Schroeder (2013).

A **fase de configuração e extração dos dados** tem como objetivo carregar o banco de dados do Big Data, HBase, com dados coletados da internet. Para a realização desta fase foi utilizado o programa Ontoclipping, desenvolvido no trabalho de conclusão de Claudio Omar Correa Carvalho Junior. Sendo que no Ontoclipping são realizadas as configurações inicias para coleta das informações, execução do robô de busca, configuração das ontologias e indexação do conteúdo coletado (SCHROEDER, 2013).

Depois de realizadas as etapas citadas anteriormente, deve ser realizada a etapa de extração das informações coletadas através do Ontoclipping. Este processo de extração deve ser feito manualmente, analisando-se caso a caso, pois deve ser programado um job MapReduce para que sejam criados os arquivos HFile, os quais são utilizados para realizar a carga do HBase. Depois da extração e carga das informações no HBase, pode ser realizado o mapeamento dos atributos (SCHROEDER, 2013).

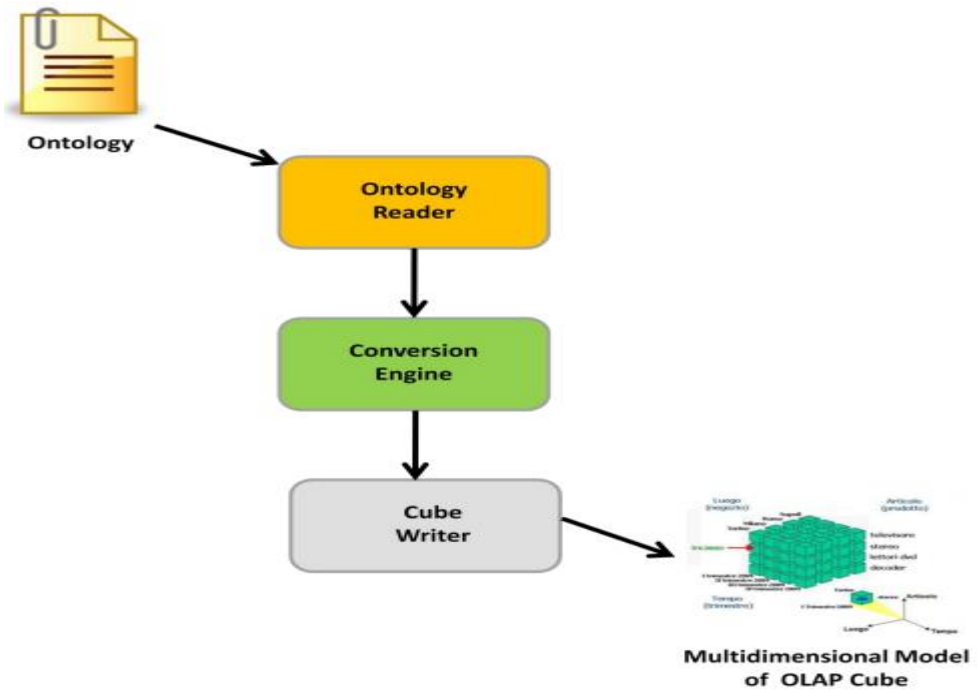
Na **fase de mapeamento e consultas** é realizado o mapeamento de atributos do banco de dados do Data Warehouse, Hive, com atributos do HBase. Para selecionar as colunas do Hive que serão mapeadas, deve-se escolher uma base de dados e após a tabela na qual a coluna existe. No entanto, para selecionar as colunas do HBase, além de selecionar a tabela, é

necessário escolher a família da coluna. Quando o processo de mapeamento tiver terminado, seleciona-se uma das colunas mapeadas e informa-se um termo de busca, para então mostrar os resultados (SCHROEDER, 2013).

Já a ferramenta desenvolvida por Flávius Anderson Félix Pereira (2011) nomeada Ontolap, tem como principal objetivo sugerir um modelo multidimensional de um cubo OLAP, a partir da ontologia que representa um domínio sobre o qual deseja-se realizar análises dos dados. Possui como principais características ter uma ontologia como entrada e a partir dela gerar o cubo OLAP e também não necessitar a existência de repositórios de dados como um Data Warehouse, para que o modelo possa ser gerado.

A arquitetura da ferramenta Ontolap é composta por três módulos, conforme ilustrado na figura 18.

Figura 18: Arquitetura da Ferramenta Ontolap



Fonte: Pereira (2011).

Os módulos conforme Pereira (2011) são responsáveis pelas seguintes atividades:

- *Ontology Reader* – Módulo encarregado por realizar a leitura da ontologia. Identifica os principais componentes como classes e propriedades da ontologia. Pode realizar a leitura de uma ontologia em qualquer uma das linguagens descritas em RDF, RDF Schema e OWL.

- *Conversion Engine* – Módulo responsável por realizar a geração do cubo OLAP, a partir da ontologia de entrada. Pode ser considerado como parte fundamental da ferramenta, visto que nele fica o algoritmo de conversão para a geração do cubo.
- *Cube Writer* – Módulo responsável por escrever a saída do modelo do cubo. Realiza a comunicação com a ferramenta de cube designer.

3.1 Comentários do Autor

Os trabalhos citados e descritos anteriormente contribuíram para a elaboração deste trabalho, fazendo com que fosse possível se ter um melhor entendimento de como os conceitos descritos no capítulo 2 Fundamentação Teórica podem ser utilizados na prática.

Já o trabalho desenvolvido por Schuster (2013), além de possibilitar um melhor entendimento dos conceitos empregados neste trabalho, é utilizado na solução desenvolvida para coletar os dados da *web* e realizar a carga do Big Data. A ferramenta também é utilizada na integração da estrutura do Big Data com o Data Warehouse, isto porque o termo da ontologia configurada no Ontoclipping é mapeado a coluna da dimensão do Data Warehouse.

4 SOLUÇÃO DESENVOLVIDA

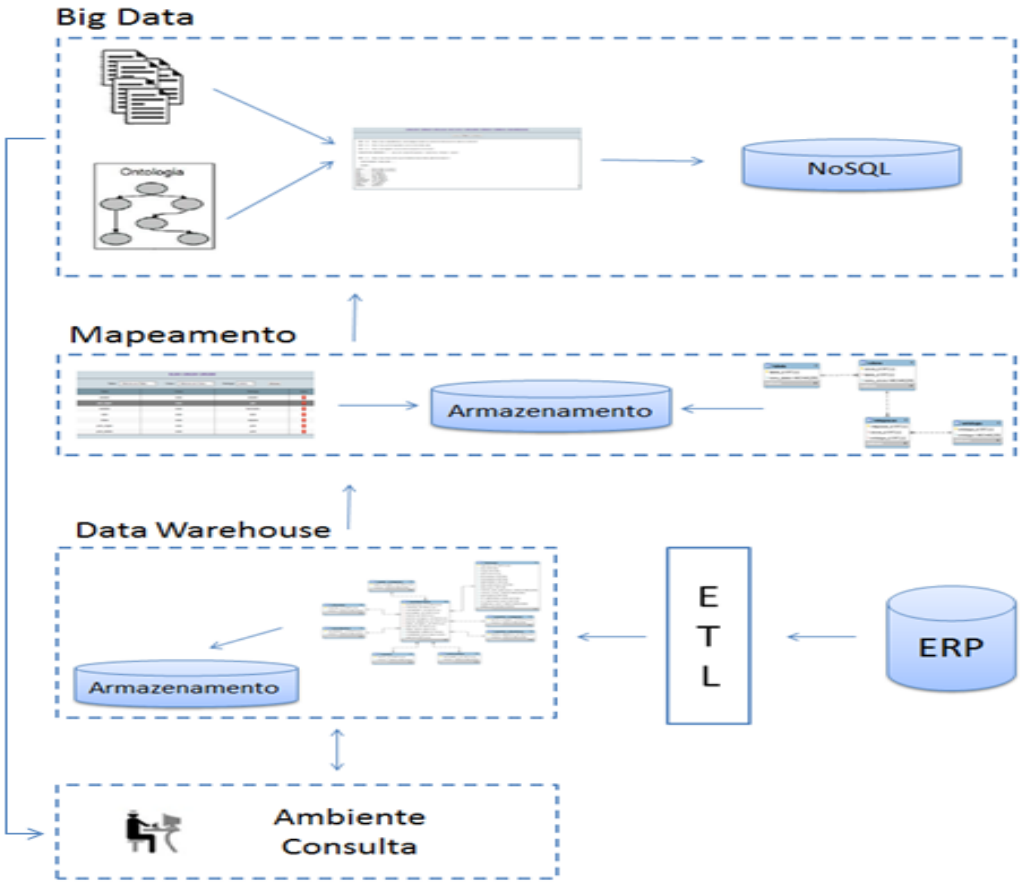
Após apresentar nos capítulos anteriores, os principais conceitos relacionados aos objetivos deste trabalho, este capítulo descreve a aplicação dos conceitos ao apresentar a solução desenvolvida neste projeto.

4.1 Visão Geral

A solução desenvolvida para este trabalho consiste em um ambiente para realizar consultas em um Big Data, com base nos termos selecionados nas dimensões de um esquema estrela de dados, estrutura esta que normalmente regula a organização dos fatos e dimensões de um Data Warehouse. Seu objetivo é o de buscar e acrescentar a consulta realizada no esquema estrela, informações referentes à área do conhecimento para o qual o Data Warehouse foi modelado.

O cenário para a realização da solução desenvolvida neste trabalho de conclusão pode ser acompanhado na figura 19, segundo as etapas especificadas.

Figura 19: Cenário da Solução Desenvolvida



Fonte: Autor (2016).

A etapa do **Big Data** consiste na captura e carga dos dados em um banco de dados NoSQL, o Cassandra. Para realizar esta fase utiliza-se o programa Ontoclipping, desenvolvido no trabalho de conclusão de Roberto Schuster Filho, trata-se de uma ferramenta de clipagem on-line que utiliza uma ontologia como forma de representação do conhecimento, em conjunto com técnicas de recuperação da informação aliadas a um motor de busca de páginas web.

Já a etapa do **Data Warehouse** pode ser dividida em duas fases: modelagem e carga dos dados. Na fase da modelagem do Data Warehouse utiliza-se o banco de dados MySQL, sendo que o mesmo é organizado em um esquema estrela de dados possuindo uma entidade central denominada tabela fato, e um conjunto de entidades menores denominadas de tabelas dimensão. Já a fase da carga dos dados é realizada utilizando uma base de dados de um sistema ERP, sendo que antes de ser realizada a carga, os dados passam por um processo de ETL.

Na etapa de **Mapeamento** entre a estrutura do Big Data com a do Data Warehouse, é realizada uma modelagem em um banco de dados MySQL, para que seja possível a integração entre estas duas estruturas. O modelo criado para realizar a integração possui as seguintes tabelas: Tabela, Coluna, Ontologia e Integração.

As tabelas “Tabela” e “Coluna”, fazem referência as tabelas do modelo estrela, estrutura esta utilizada para organizar os fatos e dimensões do Data Warehouse. Já a tabela “Ontologia”, contém os termos da ontologia utilizada no Ontoclipping para realizar a captura dos dados na web e consultas no Big Data. Logo, a tabela “Integração” possui o mapeamento realizado entre as estruturas.

Por fim, no **Ambiente de Consultas** desenvolveu-se uma interface que possibilita realizar consultas no Data Warehouse modelado. As consultas realizadas neste ambiente além de retornar resultados do Data Warehouse, retornam resultados do Big Data. Esta integração das consultas é possível devido ao mapeamento descrito na etapa anterior, visto que são submetidos ao processo de Busca do Apache Lucene, os termos selecionados nas dimensões do Data Warehouse bem como as ontologias que estão associadas a coluna da dimensão.

4.2 Funcionalidades

Nesta seção, as funcionalidades citadas na visão geral são descritas detalhadamente para se ter uma melhor compreensão da aplicabilidade delas no sistema desenvolvido.

4.2.1 Configuração e Extração dos dados para o Big Data

Tendo como objetivo a coleta de dados da internet e o carregamento do Big Data, utiliza-se o programa Ontoclipping, desenvolvido no trabalho de conclusão de Roberto Schuster Filho (2013). Abaixo segue a descrição das etapas para configuração do Ontoclipping, sendo que não cabe a este trabalho explicar detalhadamente cada etapa:

- Configuração do Ontoclipping: Esta etapa consiste em cadastrar e atribuir um índice de reputação aos sites relevantes ao robô de buscas, visando aperfeiçoar o processo de busca e recuperação, evitando assim a coleta de informações

desnecessárias. Também é possível cadastrar gêneros e formatos jornalísticos, atribuindo um índice de prioridade para cada um deles.

- Execução do robô de busca: Consiste em recuperar o conteúdo dos sites cadastrados.

- Configuração das ontologias: Consiste na construção da base de conhecimento, representando uma específica área para análise. A ontologia tem papel fundamental nos resultados apresentados pela ferramenta, sendo que a maneira como a mesma é modelada influencia diretamente na qualidade do processo de busca do robô.

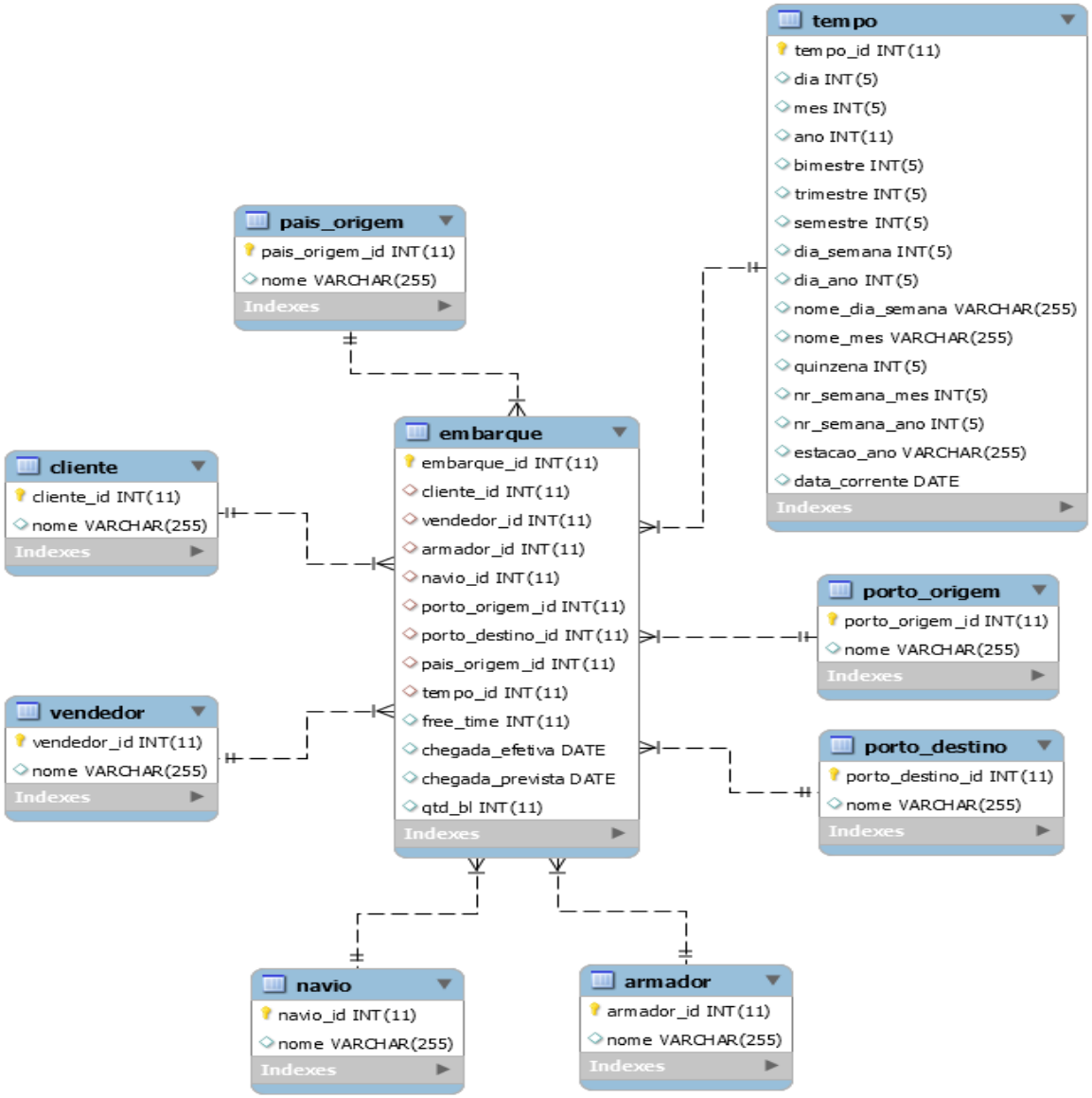
- Indexação do conteúdo recuperado dos sites cadastrados: Nesta etapa a ferramenta Apache Lucene é responsável por controlar o processo de indexação. Para realizar a indexação, esta ferramenta baseia-se nos dados coletados pelo robô de buscas e na configuração das ontologias.

O banco de dados utilizado pelo Ontoclipping para armazenar os dados extraídos da internet, bem como toda sua configuração é o Cassandra, um banco de dados NoSQL com seu armazenamento orientado à colunas.

4.2.2 Modelagem e Carga do Data Warehouse

O Data Warehouse desenvolvido para esta solução é modelado no esquema estrela. Este modelo é composto por uma entidade central nomeada de tabela fato, e um conjunto de entidades menores nomeadas de tabelas dimensão, que estão organizadas ao redor da entidade central, formando assim um esquema em estrela. A figura 20 representa a estrutura modelada para o trabalho desenvolvido.

Figura 20: Modelagem do Data Warehouse



Fonte: Autor (2016).

O banco de dados utilizado para a modelagem do Data Warehouse é o MySQL, sendo que para carga dos dados utiliza-se a base de dados de um sistema ERP. No entanto antes de serem carregados no Data Warehouse, os dados contidos no ERP passam por um processo de ETL.

O processo de ETL consiste basicamente de três etapas que são: a extração dos dados de origem, a transformação dos dados e por último a carga dos dados.

A **extração dos dados** pode dar-se de diversas origens como: SGBD, planilhas, arquivos texto entre outros. Neste trabalho a extração é realizada em um SGBD que comporta uma ferramenta de ERP.

A etapa de **transformação dos dados** trata-se do ponto onde concentra-se o maior esforço de análise. Alguns dos trabalhos de transformação abordados são a limpeza de dados desnecessários, a conversão dos tipos de campos e o tratamento da integridade dos dados, isso para que a falta de padronização do ambiente transacional não venha a comprometer a qualidade das informações no ambiente do Data Warehouse.

Já para a fase de **carga dos dados**, são definidas as estratégias de alimentação das tabelas para o ambiente do Data Warehouse de acordo com o esquema dimensional que é adotado.

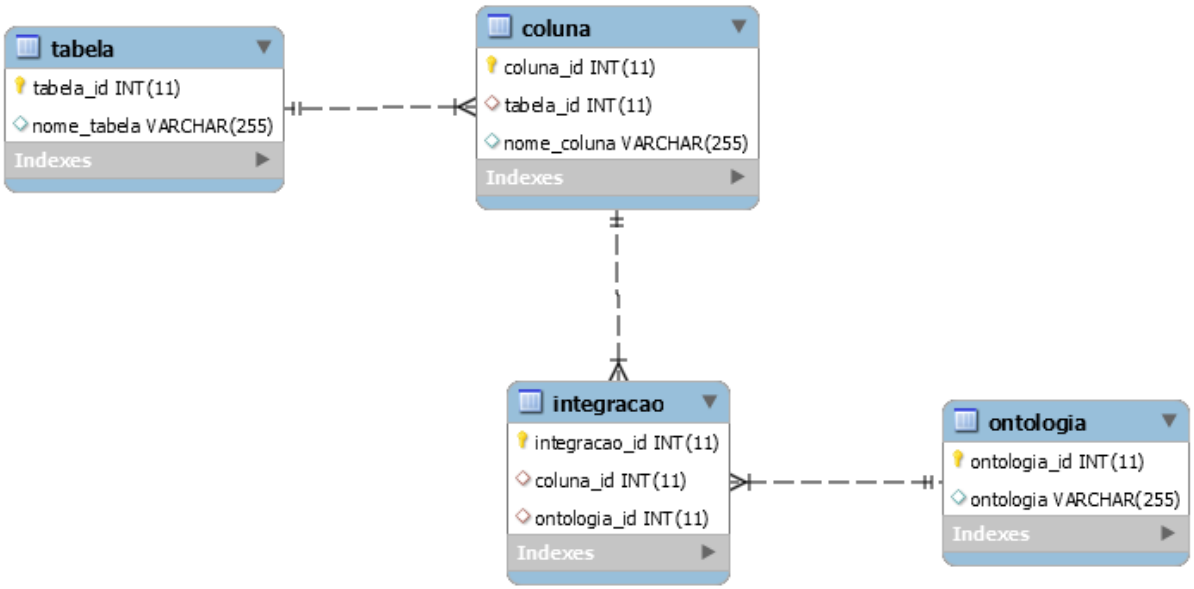
Todo este processo de extração, transformação e carga dos dados do ERP, para o Data Warehouse, é realizado com o apoio do Talend Open Studio, uma ferramenta para ETL e integração de dados.

4.2.3 Consultas e Mapeamento entre estrutura relacional e NoSQL

Para que seja possível integrar as consultas do Data Warehouse, que possui uma estrutura relacional com o Big Data, que possui uma estrutura não relacional, é necessário um mapeamento entre estas estruturas distintas, fazendo assim com que seja possível extrair resultados do Big Data através das consultas realizadas no Data Warehouse.

O trabalho desenvolvido possibilita a integração destas consultas, através do mapeamento entre as dimensões do esquema estrela, modelado para o Data Warehouse, e a ontologia utilizada pelo Ontoclipping. Este mapeamento é armazenado no MySQL, sendo que a estrutura que torna possível o mapeamento é composta por quatro tabelas, que são: Tabela, Coluna, Ontologia e Integração, conforme figura 21.

Figura 21: Modelagem do Mapeamento Dimensão Ontologia

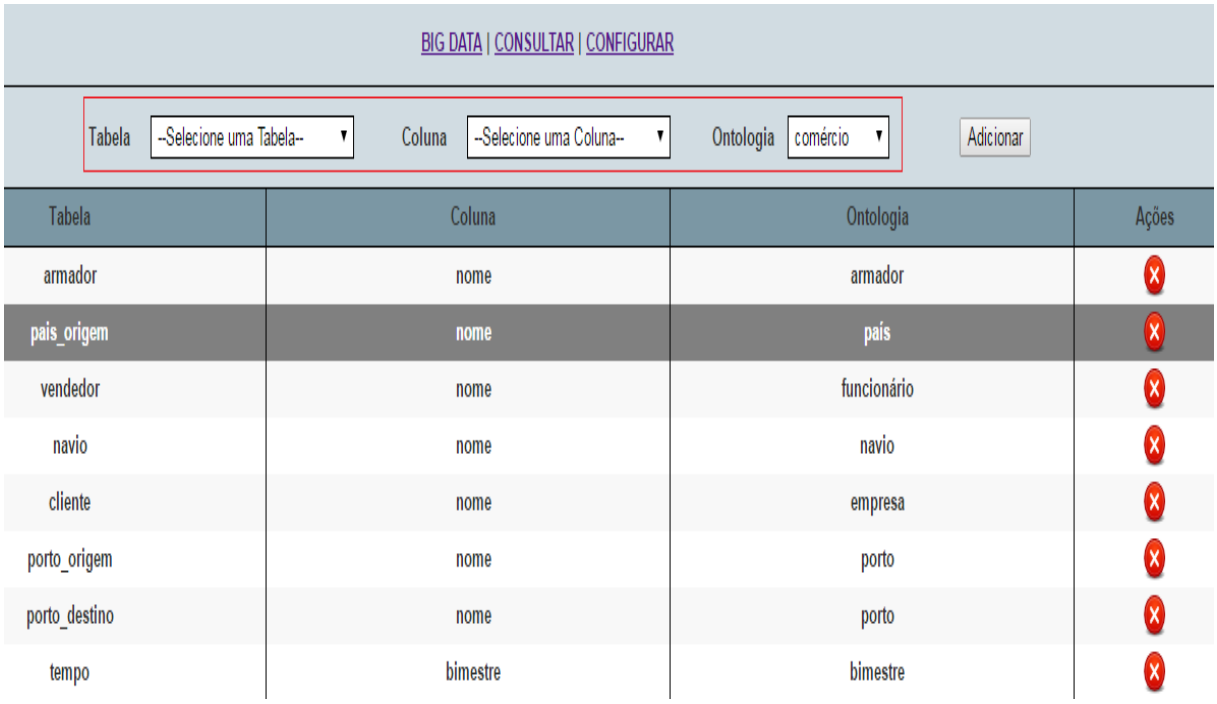


Fonte: Autor (2016).

As tabelas denominadas “tabela” e “coluna” possuem os nomes das tabelas e os nomes das colunas do Data Warehouse respectivamente. Já a tabela designada “ontologia”, possui os termos utilizados pelo Ontoclipping para realizar a coleta de informações na web e busca no Big Data. Logo, a tabela “integracao”, contém o mapeamento realizado entre a coluna da dimensão com o termo da Ontologia.

Para realizar o mapeamento, a solução apresenta uma interface onde deve-se informar a tabela, a coluna e a ontologia que se deseja associar. Conforme ilustrado na figura 22.

Figura 22: Interface de Mapeamento



Fonte: Autor (2016).

Conforme figura 22, o campo “Tabela” lista as tabelas modeladas no Data Warehouse. De acordo com a tabela selecionada, é listado no campo “Coluna”, as colunas existentes na tabela. Além da listagem das tabelas com suas respectivas colunas, são listados no campo “Ontologia”, os termos da ontologia que poderão ser associados à coluna da dimensão.

Este mapeamento torna possível a busca por informações no Big Data, que não somente tenham termos idênticos aos selecionados na consulta do Data Warehouse, mas também termos que tenham alguma relação com o que está se buscando.

Desta maneira ao se realizar uma consulta dois processos ocorrem, sendo o primeiro a busca pelos termos da ontologia que foram mapeados para a coluna da dimensão, onde conforme demonstrado na figura 23, se consultou no Data Warehouse os termos “China” e “Santos”, e a partir do mapeamento realizado foi adicionado a consulta os termos “país” e “porto”, sendo estes termos da ontologia.

Figura 23: Termos Data Warehouse e Ontologia

Fonte: Autor (2016).

Já o segundo processo que ocorre ao realizar uma consulta é a adição dos sucessores e antecessores dos termos da ontologia mapeada. Para que esta adição ocorra é realizada uma busca na ontologia configurada no Ontoclipping, com os termos da ontologia mapeada as colunas das dimensões envolvidas na consulta, a fim de encontrar termos que tenham ligação imediata aos termos utilizados nesta consulta, conforme mostra figura 24, onde o termo da ontologia mapeada “país” possui ligação imediata aos termos “política”, “clima”, “governo”, “sindicato” e “importação”, já o termo da ontologia mapeada “porto”, possui ligação com os termos “importação” e “marítimo”.

Figura 24: Antecessores e Sucessores da Ontologia

Fonte: Autor (2016).

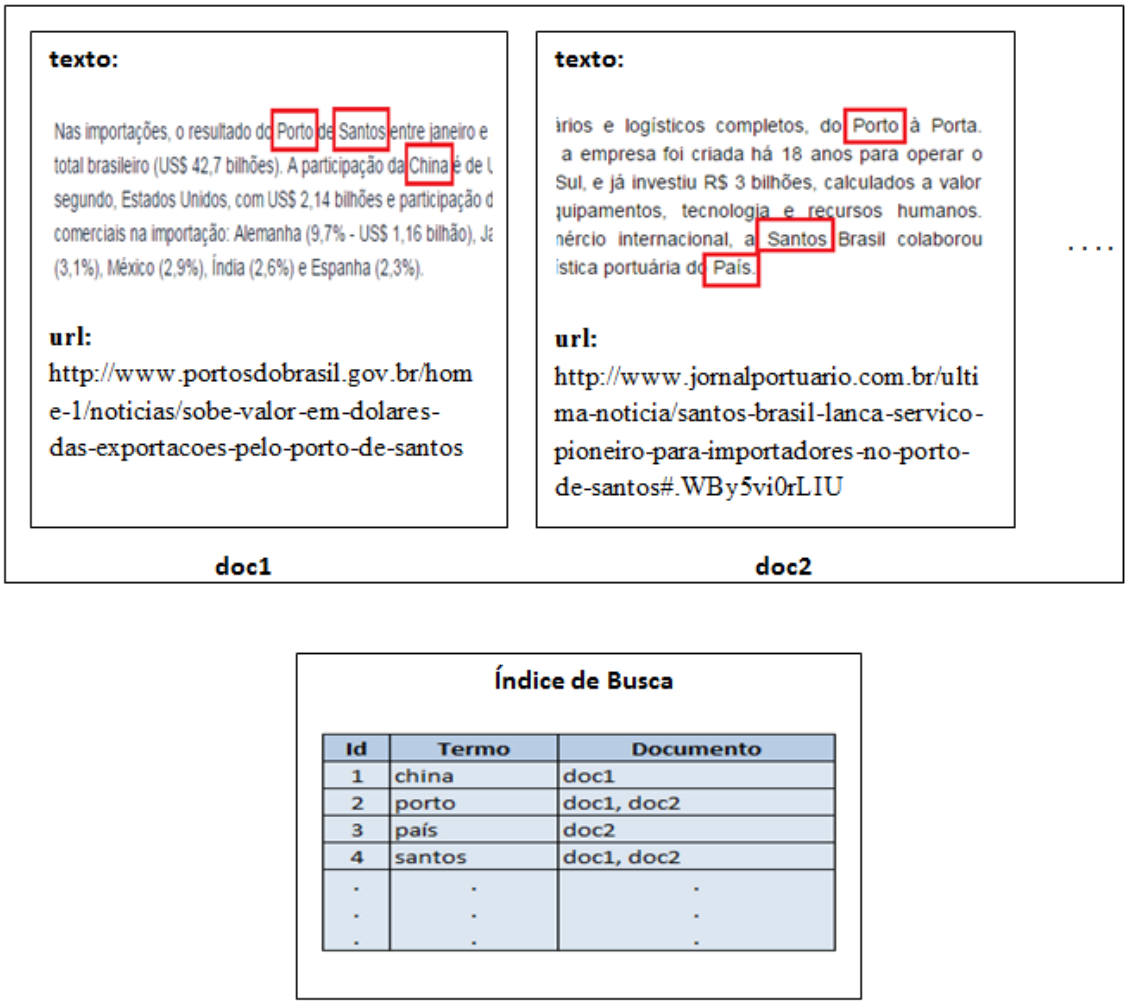
Desta forma termos selecionados nas dimensões do Data Warehouse, bem como termos da ontologia que foram mapeados para a coluna da dimensão, juntamente aos sucessores e antecessores dos termos da ontologia mapeada, formam um conjunto de termos.

Este conjunto de termos obtido através da consulta realizada no Data Warehouse, é submetido ao processo de busca do Apache Lucene. No entanto antes de submeter os termos para a procura, o Lucene cria uma estrutura de índices com documentos associados para

tornar a pesquisa mais rápida. Este índice é criado a partir do texto que está armazenando no Cassandra e que foi coletado pelo Ontoclipping.

Neste índice cada termo adicionado possui uma referência para o documento que o contém, conforme figura 25.

Figura 25: Índice de Buscas do Lucene



Índice de Busca

Id	Termo	Documento
1	china	doc1
2	porto	doc1, doc2
3	país	doc2
4	santos	doc1, doc2
.	.	.
.	.	.
.	.	.

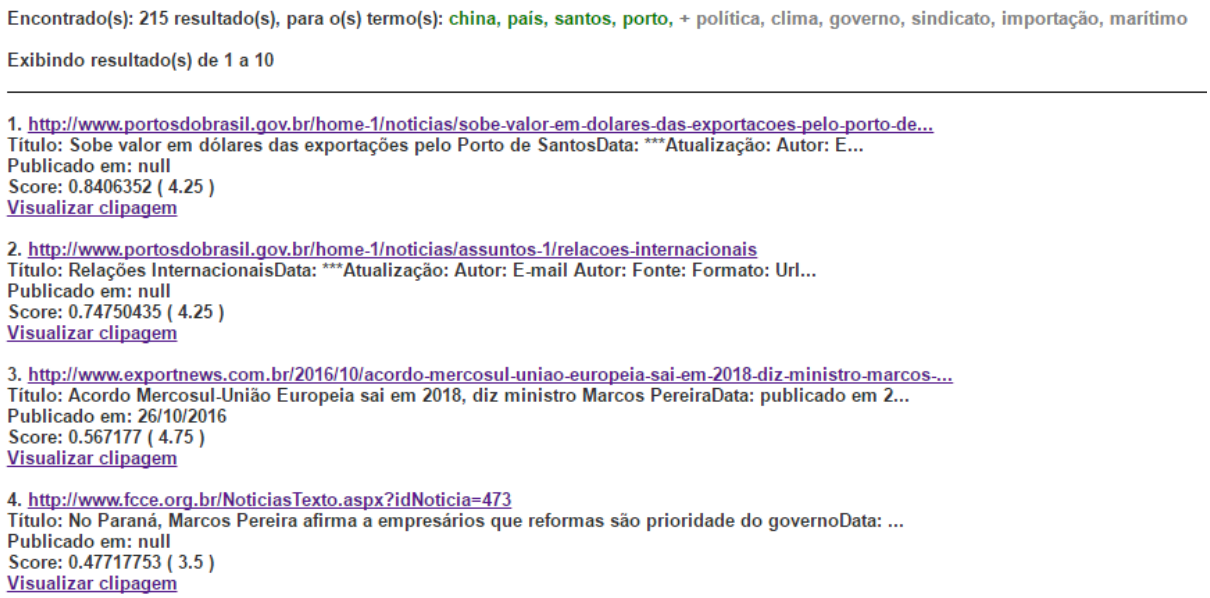
Fonte: Autor (2016).

Tendo o processo de indexação do Lucene terminado, os termos da consulta podem ser submetidos ao processo de busca do Lucene. Neste processo é atribuído para cada resultado obtido, uma pontuação que representa a similaridade do documento com a consulta.

Portanto para cada requisição de busca, o processo de pesquisa é efetuado apenas em uma base de índices indexada pelo Lucene, que traz uma lista de resultados relevantes ao

contexto da busca, sendo que os mesmos são ordenados pelo índice de similaridade da consulta com o documento.

Figura 26: Resultados Apache Lucene



Fonte: Autor (2016).

4.2.4 Casos de uso

Nesta seção a solução desenvolvida é detalhada na forma de casos de uso, a fim de oferecer uma percepção mais prática e detalhada ao leitor. Os casos de uso relacionados às etapas desenvolvidas no trabalho de conclusão de Roberto Schuster Filho (2013) não foram expostos.

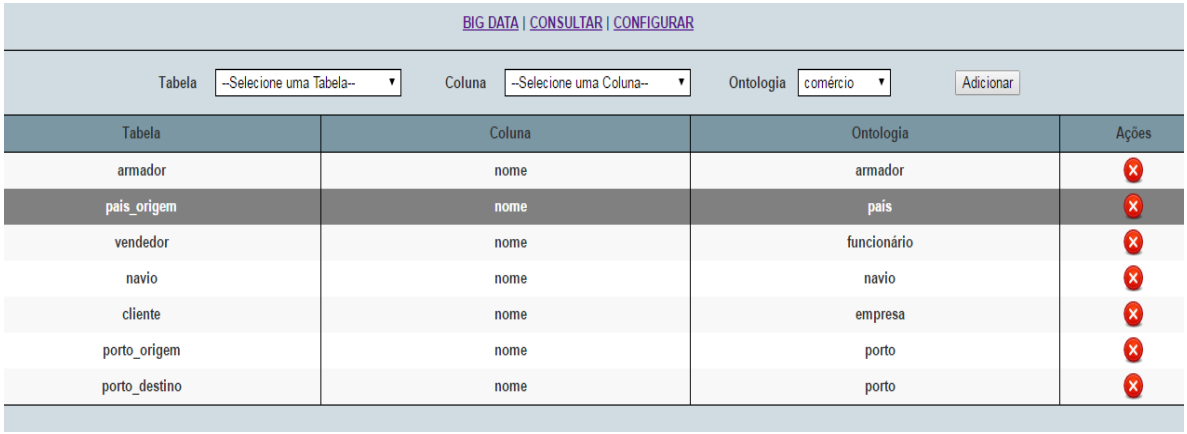
Tabela 1 - Caso de uso: Configurar Mapeamento

Descrição	Este caso de uso tem como objetivo realizar a configuração do mapeamento entre as colunas das dimensões do esquema estrela e a ontologia modelada no Ontoclipping.
Atores	Administrador
Pré-condições	Todas as etapas do Ontoclipping devem ter sido realizadas.

Fluxo Principal	P1 – Deve ser selecionada a dimensão que se deseja realizar o mapeamento; P2 – Selecionar a coluna da dimensão que se deseja mapear; P3 – Informar o termo da ontologia que se deseja associar a coluna da dimensão; P4 – Clicar em Adicionar o mapeamento; P5 – O sistema irá armazenar está informação na base de dados.
Pós-condições	O processo de mapeamento ocorre com sucesso e a coluna da dimensão com sua ontologia mapeada é exibida para o ator.

Fonte: Autor (2016).

Figura 27: Etapa de Configuração do Mapeamento



Fonte: Autor (2016).

Tabela 2 - Caso de uso: Montar Consulta

Descrição	Este caso de uso tem como objetivo montar a consulta no esquema modelado para o Data Warehouse.
Atores	Usuário
Pré-condições	Todas as etapas do Ontoclipping devem ter sido realizadas.
Fluxo Principal	P1 – Deve-se selecionar o fato; P2 – Após ter sido selecionado o fato, deve-se selecionar uma métrica do fato; P3 – Clicar em Incluir o fato; P4 – Selecionar uma dimensão que se deseja consultar;

	P5 – Selecionar a coluna da dimensão que se pretende verificar; P6 – Selecionar o conteúdo da coluna que se deseja consultar; Se for preciso adicionar mais alguma dimensão ou restrição P7 - Selecionar operador; P8 - Clicar em Incluir restrição; P9 - Repetir P4, P5 e P6. Senão P7 - Clicar em Incluir restrição.
Pós-condições	A consulta está pronta para ser submetida ao processo de pesquisa no Data Warehouse, como também no Big Data.

Fonte: Autor (2016).

Figura 28: Etapa de Montagem da Consulta

Fonte: Autor (2016).

Tabela 3 - Caso de uso: Consulta Ontologia Mapeada

Descrição	Este caso de uso tem por objetivo consultar os termos da ontologia, mapeados a coluna da dimensão.
Atores	Sistema
Pré-condições	Realização das etapas do Ontoclipping. Etapa de configuração do mapeamento.
Fluxo Principal	Para cada inclusão de restrição ou dimensão, o sistema faz a leitura da dimensão, coluna e conteúdo selecionado; Busca uma lista com os termos da ontologia mapeados a coluna da dimensão; Se existir algum elemento na lista

	Para cada elemento da lista Adiciona ou concatena o elemento a uma string; O sistema adiciona a string e o conteúdo selecionada a uma string geral. Senão O sistema adiciona o conteúdo selecionado a uma string geral.
Pós-condições	O sistema está pronto para enviar os termos para o processo de pesquisa do Big Data.

Fonte: Autor (2016).

Tabela 4 - Caso de uso: Montagem Consulta DataWarehouse

Descrição	Este caso de uso tem por objetivo montar a consulta que será realizada no Data Warehouse.
Atores	Sistema
Pré-condições	
Fluxo Principal	O sistema coloca o fato em uma string *fato* e as métricas selecionadas em uma string *select*; Para cada nova dimensão incluída Se dimensão não havia sido selecionada Inclui a dimensão em uma string para realizar inner join dela; Adiciona dimensão a lista de dimensões selecionadas; Para cada nova restrição de coluna incluída Se coluna da dimensão já havia sido selecionada Inclui conteúdo da coluna selecionada a uma string de restrições; Senão Inclui conteúdo da coluna selecionada a uma string de restrições; Inclui coluna selecionada a string *select*; Adiciona coluna a lista de colunas selecionadas; O sistema monta uma string geral para se realizar a consulta.
Pós-condições	O sistema está pronto para realizar a consulta na base de dados do Data Warehouse.

Fonte: Autor (2016).

Na montagem da consulta a ser realizada no Data Warehouse, conforme descrito pelo fluxo principal da tabela 4, a tabela selecionada como fato é atribuída a uma variável “fato”. Tendo sido o fato selecionado é possível atribuir a variável “select” as métricas pertencentes ao fato selecionado. Após terem sido inicializadas as variáveis “fato” e “select” é possível selecionar as dimensões que deseja-se consultar.

Para cada nova dimensão inserida na consulta, o sistema verifica se a dimensão já não havia sido selecionada. Esta verificação ocorre para que não realize-se *inner join* da mesma dimensão mais de uma vez na consulta, pois isso resultaria em um erro de SQL.

Após ter sido selecionada a dimensão, as colunas da mesma são listadas para que seja possível incluir as restrições da consulta. Sendo assim, para cada nova restrição incluída o sistema verifica se a coluna da dimensão já não havia sido selecionada. Esta verificação é realizada para que não se inclua na variável “select” o nome da coluna mais de uma vez, pois isto faria com que a coluna aparece-se mais de uma vez na listagem.

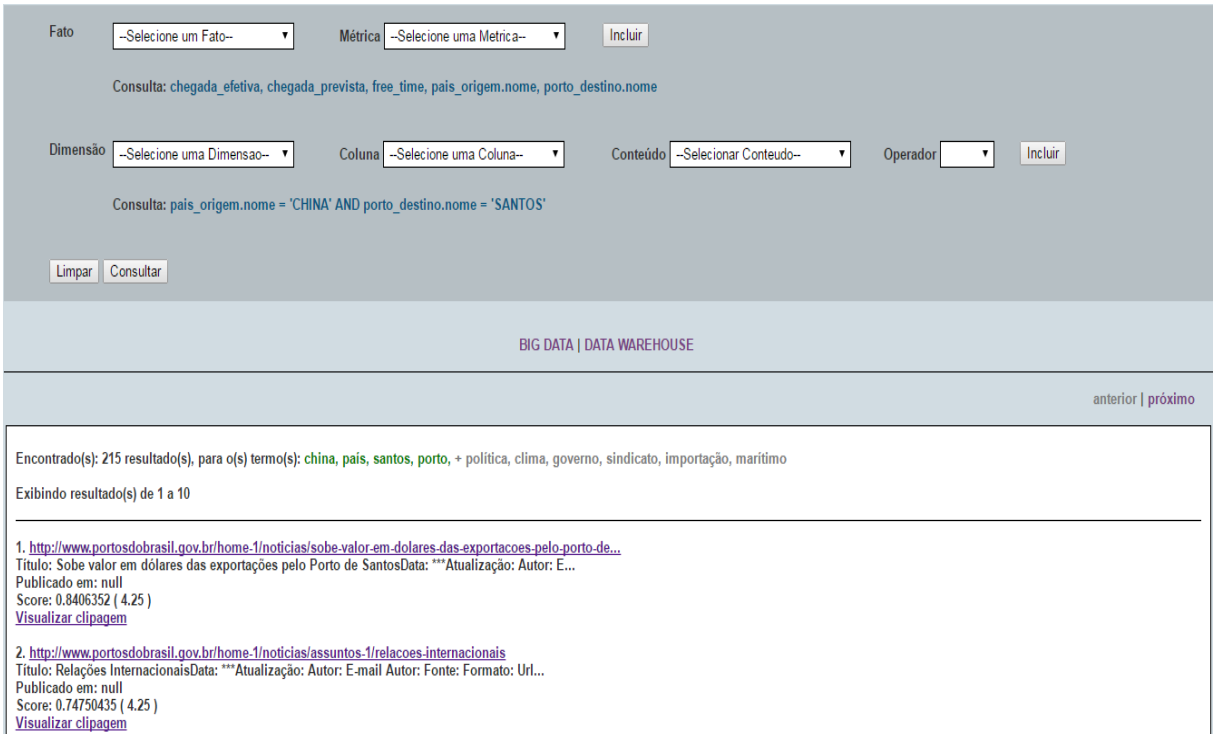
Tabela 5 - Caso de uso: Processo de Pesquisa Big Data

Descrição	Este caso de uso tem o objetivo de mostrar como é feito, e como ocorre o processo de pesquisa das informações no Big Data.
Atores	Usuário
Pré-condições	Realização das etapas do Ontoclipping. Etapa de configuração do mapeamento. Etapa de montagem da consulta. Etapa de consulta das ontologias mapeadas.
Fluxo Principal	P1 – Após ter sido realizada a etapa de Montagem da Consulta (descrita na tabela 2), clicar em Consultar; P2 – É realizada uma busca na ontologia configurada no Ontoclipping, com os termos retornados da Consulta da Ontologia Mapeada (descrita na tabela 3), a fim de encontrar termos que tenham ligação imediata aos termos utilizados na consulta; P3 – Os termos de consulta, juntamente com os termos da ontologia, são submetidos aos procedimentos de busca da ferramenta Apache Lucene, que realiza a pesquisa no material indexado. O Índice de Qualificação interfere na geração dos resultados, pois foi associado aos documentos coletados no momento da indexação; P4 – Os resultados são apresentados contendo o link da página coletada, a data de publicação (se houver), bem como uma prévia do conteúdo coletado; P5 – Clicar em Visualizar clipagem para acessar o conteúdo.

Pós-condições	Dados estão disponíveis para análise.

Fonte: Autor (2016).

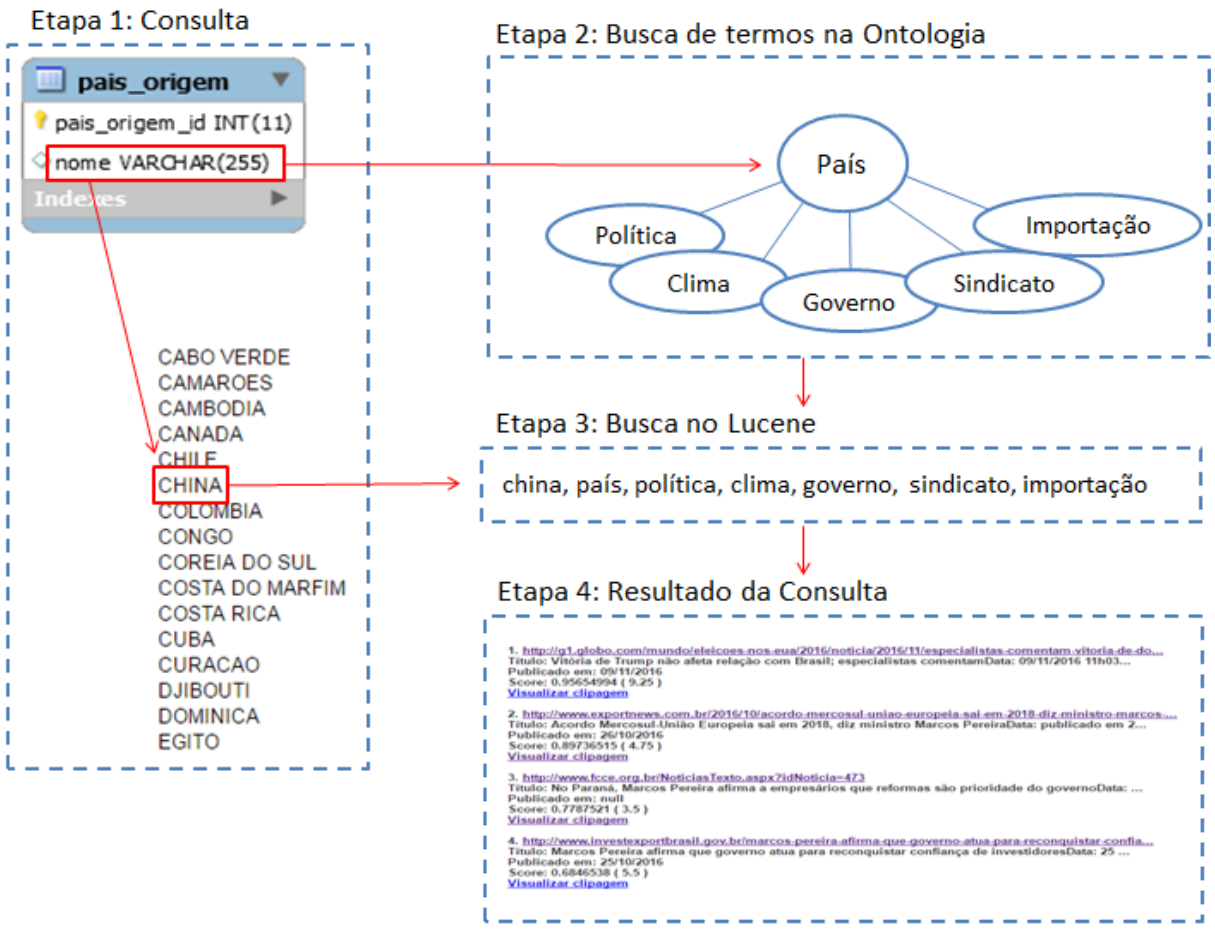
Figura 29: Etapa do Processo de Pesquisa no Big Data



Fonte: Autor (2016).

No processo de pesquisa que o sistema realiza no Big Data, conforme descrito pelo fluxo principal da tabela 5, verifica-se a coluna que foi selecionada na consulta, e se a ela tem mapeada um termo da ontologia. Caso a coluna tenha algum mapeamento, o sistema busca na ontologia configurada no Ontoclipping, termos que tenham ligação imediata ao termo mapeado à coluna. Desta forma, o termo da ontologia mapeada, bem como os termos que possuem ligação imediata, são submetidos juntamente com o conteúdo da coluna selecionada, ao processo de busca do Lucene. O Lucene por sua vez, realiza a busca no material indexado e retorna os resultados de acordo com o índice de qualificação que a ferramenta atribui ao resultado. Estas etapas são ilustradas na figura 30.

Figura 30: Processo de Pesquisa no Big Data



Fonte: Autor (2016).

Tabela 6 - Caso de uso: Processo de Pesquisa no Data Warehouse

Descrição	Este caso de uso tem o objetivo de mostrar como é feito, e como ocorre o processo de pesquisa das informações no Data Warehouse.
Atores	Usuário
Pré-condições	
Fluxo Principal	P1 – Após ter sido realizada a etapa de Montagem da Consulta (descrita na tabela 2), clicar em consultar; P2 – É realizada uma busca na base de dados do Data Warehouse com a consulta modelada na fase de Montagem da Consulta do Data Warehouse (descrita na tabela 4).
Pós-condições	Dados estão disponíveis para análise.

Fonte: Autor (2016).

O sistema disponibiliza uma interface para realizar consultas no Data Warehouse. Nesta interface é possível selecionar fato, métricas e dimensões do esquema modelado, conforme os passos descritos na tabela 2. No entanto para que o sistema consiga buscar resultados na base de dados do Data Warehouse, ele monta uma *string* geral conforme descrito na tabela 4 e representado no exemplo da figura 31.

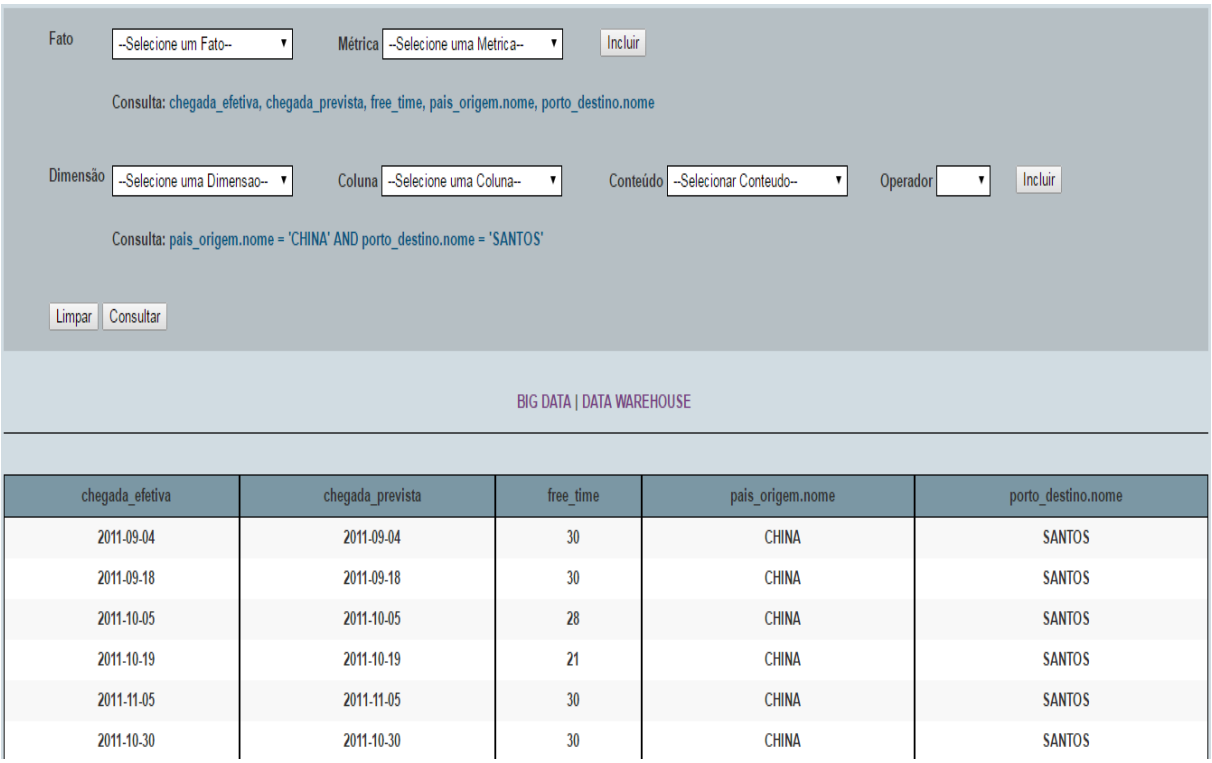
Figura 31: String Geral de Consulta para o Data Warehouse

```
SELECT chegada_efetiva, chegada_prevista, free_time, pais_origem.nome, porto_destino.nome FROM embarque
INNER JOIN pais_origem ON pais_origem.pais_origem_id=embarque.pais_origem_id
INNER JOIN porto_destino ON porto_destino.porto_destino_id=embarque.porto_destino_id
WHERE pais_origem.nome = 'CHINA' AND porto_destino.nome = 'SANTOS'  ;
```

Fonte: Autor (2016).

Após a consulta ter sido montada, o sistema busca as informações na base de dados do Data Warehouse e lista os resultados conforme demonstrado na figura 32.

Figura 32: Etapa do Processo de Pesquisa no Data Warehouse



chegada_efetiva	chegada_prevista	free_time	pais_origem.nome	porto_destino.nome
2011-09-04	2011-09-04	30	CHINA	SANTOS
2011-09-18	2011-09-18	30	CHINA	SANTOS
2011-10-05	2011-10-05	28	CHINA	SANTOS
2011-10-19	2011-10-19	21	CHINA	SANTOS
2011-11-05	2011-11-05	30	CHINA	SANTOS
2011-10-30	2011-10-30	30	CHINA	SANTOS

Fonte: Autor (2016).

5 VALIDAÇÃO

Neste capítulo é abordada a validação da proposta implementada neste trabalho, de forma a verificar a eficiência do sistema em si.

5.1 Método de Validação

Com o objetivo de avaliar e validar de forma prática a solução desenvolvida realizou-se um estudo de caso. O estudo busca verificar se a consulta realizada no Data Warehouse, consiga através do mapeamento realizado, trazer resultados do Big Data que tenham alguma relação com os elementos que estão participando da transação do negócio, ou seja com os termos selecionados nas dimensões do esquema estrela.

Sendo assim, as configurações do Ontoclipping bem como a modelagem do Data Warehouse foram realizadas para uma empresa de comércio exterior. Dentre as atividades que a empresa exerce estão:

- A realização de toda a documentação para o embarque da mercadoria do cliente;
- A verificação da disponibilidade de espaço em navios para o embarque dos containers com as mercadorias;
- O rastreamento do navio, onde são verificadas as datas de saída e chegada do navio ao porto, bem como alguma alteração que possa ocorrer durante o percurso do navio;
- A venda de serviços de embarque.

Desta forma a principal transação de negócio da empresa é o embarque de cargas, sendo que os elementos envolvidos nesta transação podem ser porto de origem e destino, navio, armador, vendedor, cliente, país de origem, entre outros.

Desta maneira, primeiramente no Ontoclipping foram cadastrados sites raízes relacionados à área de COMEX e também alguns jornais com abrangência nacional e internacional, conforme relação da tabela 7.

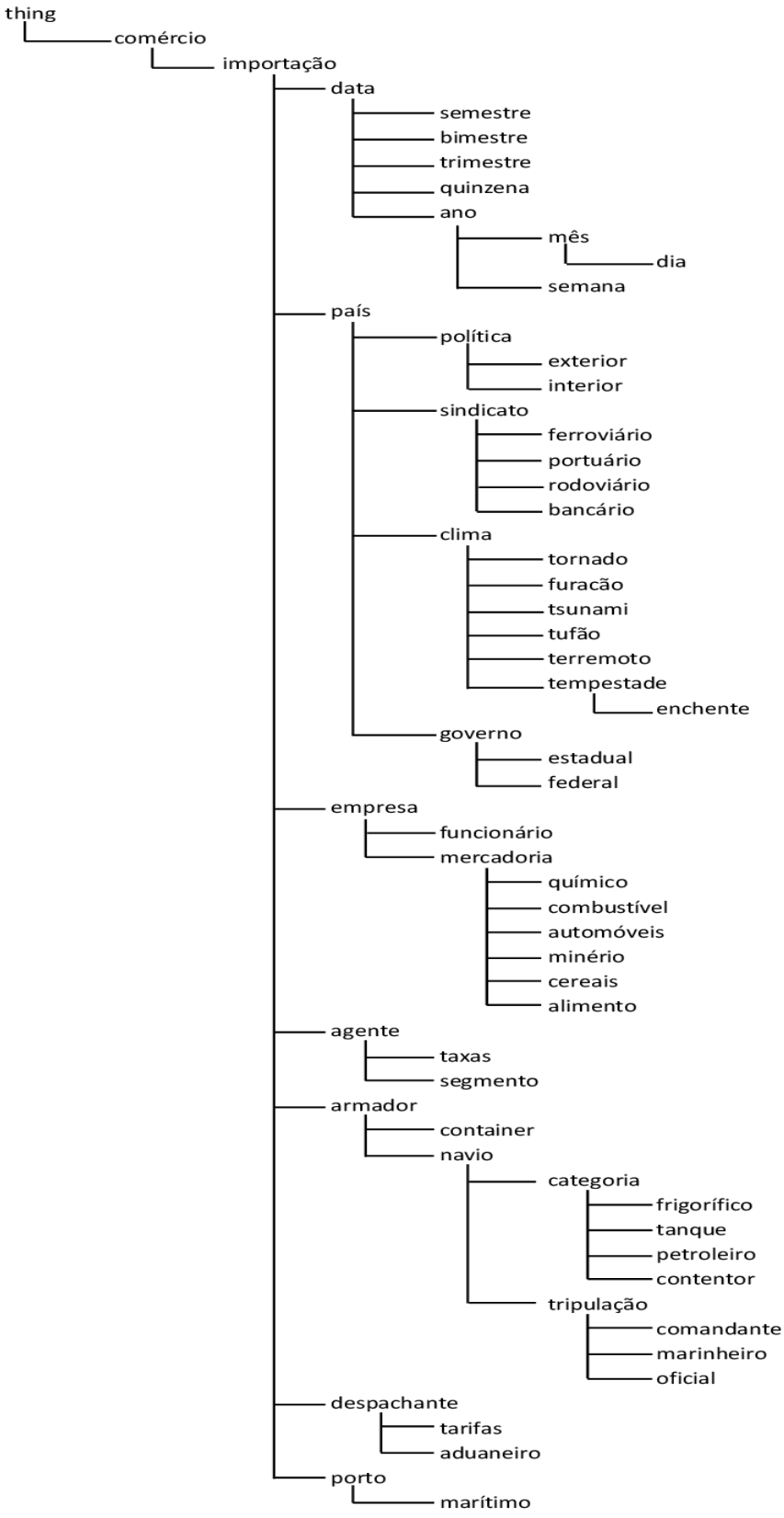
Tabela 7 - Relação de Sites Raízes

Endereço
http://www.portosdobrasil.gov.br/home-1/noticias
http://www.jornalportuario.com.br/
http://www.comissaoportos.com.br/noticias.php
http://www.aeb.org.br
http://www.investexportbrasil.gov.br/
http://www.fnportuarios.org.br/
http://www.comexdobrasil.com/
http://portuariosrio.org.br/category/noticias/
http://g1.globo.com
http://www.exportnews.com.br
http://www.aduaneiras.com.br
http://diariocatarinense.clicrbs.com.br/sc/
http://www.sindaport.com.br/noticias.php
http://diariogaucho.clicrbs.com.br/rs/
http://exame.abril.com.br/topicos/portos
http://www.revistaportuaria.com.br/novo/
http://www.portoriogrande.com.br/site/index.php
http://www.gaz.com.br/gazetadosul
http://www.correiodopovo.com.br
http://noticias.uol.com.br/politica
http://www.fcce.org.br
http://exame.abril.com.br/topicos/comercio-exterior
http://www.portoriogrande.com.br/site/index.php
http://portogente.com.br/noticias/portos-do-brasil
http://www.fenccovib.org.br/modules/news/index.php?storytopic=1

Fonte: Autor (2016).

Cada site raiz teve um Índice de Reputação conforme a importância que se deu ao site. Após, foi modelado uma ontologia que representa o conhecimento sobre o domínio de Comércio Exterior, mais especificamente sobre importações marítimas, conforme figura 33.

Figura 33: Ontologia Configurada no Ontoclipping



Fonte: Autor (2016).

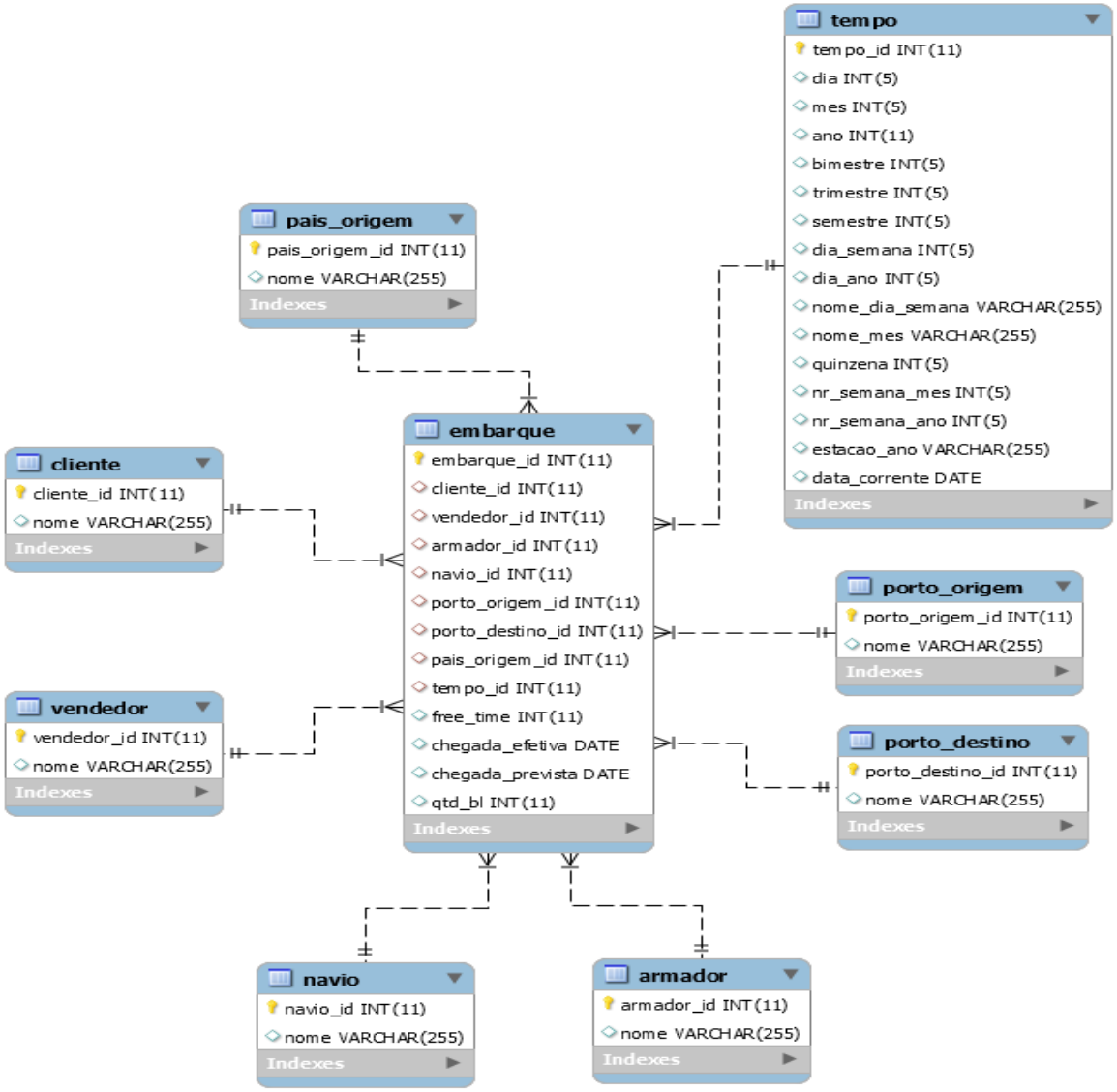
Posteriormente foram inseridos gêneros e formatos jornalísticos, atribuindo um Índice de Prioridade para cada um deles. Então, foram definidos os pesos para os critérios de qualificação:

- Peso Reputação Sites Raízes: 1.5;
- Peso Completude: 2.5;
- Peso Prioridade Gêneros: 0.5;
- Peso Prioridade Formatos: 0.5.

Depois da configuração da ferramenta, a coleta de dados foi realizada pelo *crawler*, obtendo uma base com 7318 páginas que receberam um Índice de Qualificação e foram indexadas.

Após a configuração e coleta dos dados realizada no Ontoclipping, foram realizadas consultas no esquema estrela, esquema este usado para organizar os fatos e dimensões do Data Warehouse. O esquema foi modelado para analisar a transação de embarque de importação marítima, onde os elementos envolvidos nesta transação são: porto de origem, porto de destino, armador, navio, vendedor, cliente, país de origem e tempo em que ocorreu a transação. O esquema para esta transação foi modelado conforme figura 34.

Figura 34: Representação do Modelo do Data Warehouse



Fonte: Autor (2016).

No entanto com o intuito de verificar se as consultas realizadas no Data Warehouse, consigam através do mapeamento realizado, trazer resultados do Big Data que tenham alguma relação com os elementos que estão participando da transação do negócio, foram criadas duas validações de consulta. Os próximos itens detalham os procedimentos destas validações.

5.2 Validação 1

No primeiro teste da validação foi realizada uma consulta no Data Warehouse, buscando saber a data da chegada efetiva dos embarques com país de origem China, e porto de destino, Santos. Neste teste o mapeamento entre ontologia e dimensão não havia sido realizado, sendo assim foi submetido ao processo de busca somente os termos das dimensões selecionadas na consulta, ou seja, os termos “china” e “santos”. Nesta consulta foram encontrados 84 resultados para os termos china e santos, conforme apresentado na figura 35.

Figura 35: Resultados sem o mapeamento na validação 1

Fonte: Autor (2016).

Mesmo sem a realização do mapeamento entre a coluna da dimensão e o termo da ontologia, pode-se observar que através da consulta realizada no Data Warehouse foi possível buscar informações no Big Data. No entanto os resultados encontrados como válidos para a consulta fazem referência especificamente aos termos selecionados.

Já no segundo teste da validação foi realizada no Data Warehouse, a mesma consulta utilizada no primeiro teste. No entanto, neste teste o mapeamento entre ontologia e dimensão foi realizado, sendo que a coluna “nome” da dimensão do porto de destino foi mapeada para o termo da ontologia “porto” e a coluna “nome” da dimensão país de origem foi mapeada para o termo da ontologia “país”, conforme demonstrado na figura 36.

Figura 36: Mapeamento para validação 1

Tabela	Coluna	Ontologia
porto_destino	nome	porto
pais_origem	nome	país

Fonte: Autor (2016).

Sendo assim foi submetido ao processo de busca os termos das dimensões selecionadas na consulta, ou seja, os termos “china” e “santos”, os termos da ontologia mapeada “porto” e “país”, e os termos “política”, “clima”, “governo”, “sindicato”, “importação” e “marítimo”, adicionados a busca devido a ontologia . Nesta consulta foram encontrados 281 resultados para os termos china, país, santos, porto, política, clima, governo, sindicato, importação e marítimo, conforme apresentado na figura 37.

Figura 37: Resultados com o mapeamento na validação 1

Fonte: Autor (2016).

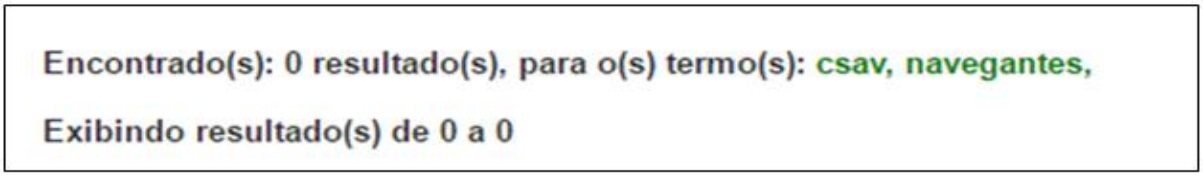
Com a realização do mapeamento entre a coluna da dimensão e o termo da ontologia, pode-se constatar que através da consulta realizada no Data Warehouse foi possível buscar informações no Big Data que não somente faziam referência especificamente aos termos selecionados nas dimensões, mas também informações que tinha alguma relação com os elementos pesquisados.

5.3 Validação 2

No primeiro teste da segunda validação foi realizado uma consulta no Data Warehouse, onde buscou-se saber a data da chegada prevista e a data da chegada efetiva dos embarques realizados pelo armador CSAV para o porto de destino Navegantes. Neste teste o mapeamento entre ontologia e dimensão não havia sido realizado, sendo assim submetido ao processo de busca somente os termos das dimensões selecionadas na consulta, ou seja, os

termos “csav” e “navegantes”. Nesta consulta não foram encontrados resultados, conforme apresentado na figura 38.

Figura 38: Resultados sem o mapeamento na validação 2



Fonte: Autor (2016).

Já no segundo teste da validação foi realizada no Data Warehouse, a mesma consulta utilizada no primeiro teste. No entanto, neste teste o mapeamento entre ontologia e dimensão foi realizado, sendo que a coluna “nome” da dimensão do armador foi mapeada para o termo da ontologia “armador”, conforme demonstrado na figura 39.

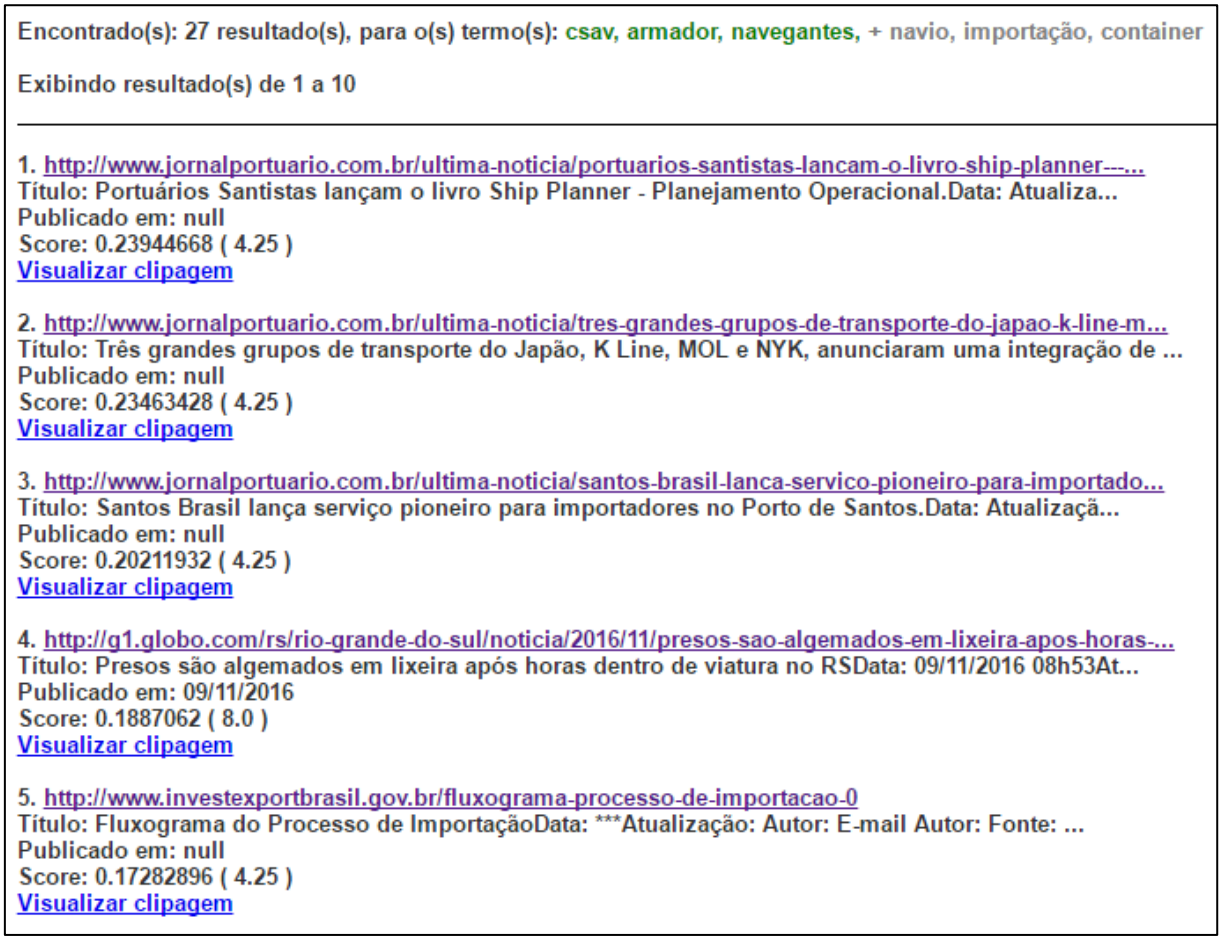
Figura 39: Mapeamento para validação 2

Tabela	Coluna	Ontologia
armador	nome	armador

Fonte: Autor (2016).

Desta maneira foi submetido ao processo de busca os termos das dimensões selecionadas na consulta, ou seja, os termos “csav” e “navegantes”, o termo da ontologia mapeada “armador”, e os termos “navio”, “importação” e “container, adicionados a busca devido a ontologia . Nesta consulta foram encontrados 27 resultados para os termos csav, armador, navegantes, navio, importação, container, conforme figura 40.

Figura 40: Resultados com o mapeamento na validação 2



Fonte: Autor (2016).

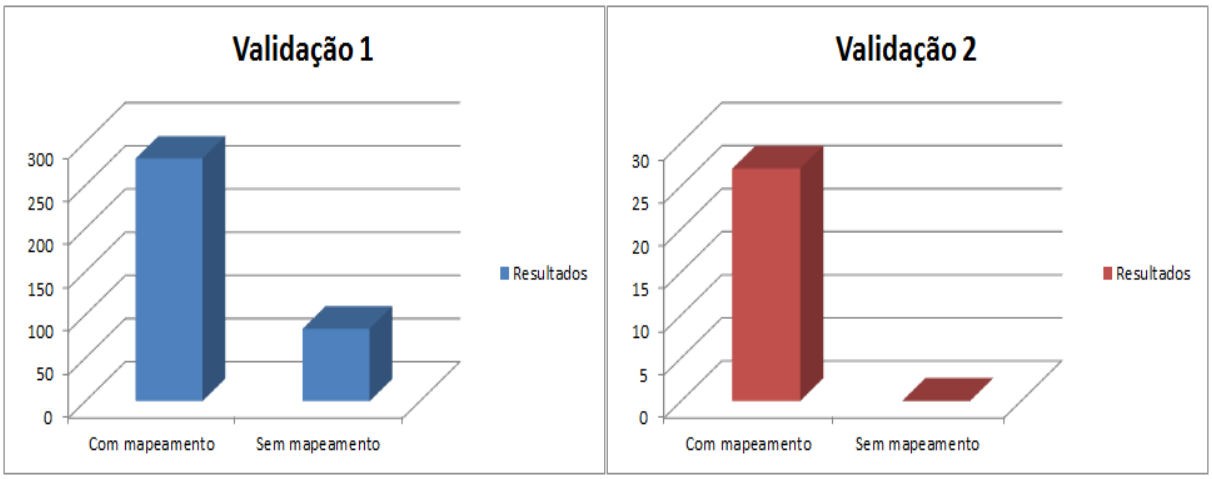
Com a realização do mapeamento entre a coluna da dimensão e o termo da ontologia, pode-se observar que através da consulta realizada no Data Warehouse foi possível buscar informações no Big Data que tenham alguma relação com os elementos pesquisados.

5.4 Comparação entre Validação 1 e Validação 2

Na primeira validação, bem como na segunda foram realizadas consultas no Data Warehouse com o mapeamento entre ontologia e dimensão, e sem o mapeamento. Em ambas as validações os resultados com o mapeamento foi superior, sendo que na validação 1, quando não foi utilizado o mapeamento se obteve 84 resultados e quando foi realizado o mapeamento se obteve 281 resultados. Já na validação 2, sem o mapeamento não se teve resultados e com o mapeamento se teve 27 resultados.

A seguir é apresentado um gráfico para uma visualização mais clara do comparativo realizado acima.

Figura 41: Comparativo entre validações



Fonte: Autor (2016).

Considerando o número de resultados obtidos nas duas validações pode-se observar que mesmo sem o mapeamento entre dimensão e ontologia, em alguns casos é possível coletar informações no Big Data. No entanto, com o mapeamento entre dimensão e ontologia foi possível ampliar os resultados com informações referentes ao contexto da busca, mesmo que os termos das dimensões consultadas não estejam contidos no Big Data.

6 CONCLUSÕES

Com base na necessidade de ter-se um modelo que possibilite a integração das consultas de um Data Warehouse a um Big Data, o trabalho aqui desenvolvido propôs a criação de um ambiente de consultas em um Big Data considerando um esquema estrela de dados. O ambiente possibilita ao usuário a busca por informações no Big Data através dos termos selecionados nas dimensões do esquema. Sendo também possível integrar mais resultados a busca, por meio do mapeamento entre as colunas da dimensão e os termos da ontologia configurada.

A principal função que pode influenciar diretamente no resultado final está na realização do mapeamento entre a coluna da dimensão e o termo da ontologia. Isto porque ao se realizar o mapeamento, não somente os termos selecionados nas dimensões do esquema estrela serão enviados ao processo de busca, mas também o termo da ontologia mapeada, que irá possibilitar a agregação de novos termos contidos na ontologia configurada. Diferentemente de quando não se realiza o mapeamento, onde somente os termos selecionados nas dimensões serão enviados ao processo de busca.

Os termos selecionados nas dimensões do esquema bem como os termos da ontologia, são submetidos ao processo de busca que fica a cargo da ferramenta Apache Lucene, responsável por buscar os documentos relacionados a consulta e classificar os resultados de acordo com o índice de similaridade.

A fim de validar o sistema foram realizadas consultas no ambiente, tendo o Big Data assim como o Data Warehouse terem sido carregados com informações referentes à área de comércio exterior. Estas consultas foram efetuadas com a realização do mapeamento e sem o mapeamento. Sendo que seus resultados foram comparados com a intenção de demonstrar os ganhos na busca, quando as consultas utilizam o mapeamento entre dimensão e ontologia.

No entanto mesmo que tenha sido possível realizar a integração das consultas entre as estruturas, não foi possível buscar informações no Big Data que condiziam com o período em que o fato ocorreu. Para que fosse possível a busca destas informações, seria necessário que a coleta de dados para o Big Data tivesse sido realizada no mesmo período em que se deu o fato.

Para a realização do trabalho desenvolvido foram necessários estudos que possibilitaram vislumbrar trabalhos futuros que possam vir a ser desenvolvidos através deste como: extração de dados na web através dos termos das dimensões, podendo levar em

consideração no momento da clipagem dos dados da web, não somente a ontologia configurada, mas também os termos da coluna da dimensão mapeada ao termo da ontologia; avaliação qualitativa dos resultados, onde seria verificado a importância dos resultados para a consulta realizada; e melhorias no próprio ambiente de consultas.

Assim, conclui-se que o trabalho teve relativo êxito no seu propósito ao apresentar resultados razoáveis através de uma validação feita por testes distintos, de forma a demonstrar a relevância das funcionalidades desenvolvidas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, M.; BAX, M. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção, 2003. Disponível em: <www.scielo.br/pdf/ci/v32n3/19019>. Acessado em: outubro de 2016.

ALVARENGA, L.; SOUZA, R. R. A Web Semântica e suas contribuições para a ciência da informação, 2004. Disponível em: http://www.scielo.br/pdf/ci/v33n1/v33n1a16.pdf. Acessado em: outubro de 2016.

ANDRADE, L. C.; SOUZA, M. C.; MAFORT, R. F. LMR Sistema gerenciador de documentos. Trabalho de conclusão de curso. Universidade Gama Filho, 2010.

BARONI, Rodrigo; Aplicação de ontologias para recuperação de informação em fóruns do ambiente virtual de aprendizagem. 2011.

BERNARDES, Guilherme de Lima. Desenvolvimento de software no contexto big data. Trabalho de conclusão de curso. Universidade de Brasília, 2014.

BULSING, Gabriel Merten. Ferramenta para Extração de Dados Semi-Estruturados para Carga de um Big Data. Trabalho de conclusão de curso. Universidade de Santa Cruz do Sul, 2013.

CUNHA, J. P.; PEREIRA, J. L. Column-based databases: Estudo exploratório no âmbito das bases de dados NoSQL, 2015. Disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/39180/1/CAPSI2015_JPC_JLP.pdf. Acessado em: outubro de 2016.

CORREIA, João de Sá Balão Calisto Correia. Indexação de documentos clínicos. Dissertação de mestrado. Faculdade de Engenharia da Universidade do Porto, 2016.

DEMCHENKO, Y.; GROSSO, P.; LAAT, C. D.; MEMBREY, P. Addressing Big Data Issues in Scientific Data Infrastructure. Collaboration Technologies and Systems (CTS), 2013 International Conference on, 20-24 Maio 2013 San Diego, California, EUA, p.48-55. 2013.

GRUBER, Tom. What is na Ontology?, 1992. Disponível em: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>. Acessado em: outubro de 2016.

INMON, W. H. Como Construir o Data Warehouse. Rio de Janeiro: Campus, 1997. 388 p.

INMON, W. H; HACKATHORN, R. D. Como Usar o Data Warehouse. Rio de Janeiro: Livraria e Editora Infobook, 1997. 277 p.

KIMBALL, Ralph. Data Warehouse Toolkit. São Paulo: Makron Books, 1998. 388 p.

KUPLICH, Cassiano Rocha. Desenvolvimento de Data Warehouse e Ferramenta OLAP para Análise da Produção Acadêmica de Pesquisadores: Estudo de Caso no PPGC. Trabalho de conclusão de curso. Universidade Federal do Rio Grane do Sul, 2013.

LEITE, Lizane Alvares. Banco de dados cassandra: Um estudo de caso para análise dos dados dos serviços públicos federais. Trabalho de conclusão. Universidade de Brasília, 2014.

MACHADO, Felipe Nery Rodrigues. Projeto de Data Warehouse uma Visão Multidimencional. São Paulo: Érica. 2000. 248 p.

MACHADO, Marco André Santos. Uma abordagem para indexação e busca full-text baseadas em conteúdos em sistemas de armazenamento em nuvem. Dissertação de mestrado. Universidade Federal de Pernambuco, 2013.

MARQUES DE JESUS, Jerocir Botelho. Tesauro: um instrumento de representação do conhecimento em sistemas de recuperação da informação, 2002. Disponível em: http://www.ndc.uff.br/OLD2013/repositorio/Tesauros.pdf. Acessado em: outubro de 2016.

MARROQUIM, M. S.C.; RAMOS, R. M. Distribuição de dados em escala global com cassandra, 2012. Disponível em: http://mariomarroquim.github.io/research/artigo-mariomarroquim-cassandra.pdf. Acessado em: outubro de 2016.

MARTINS, César Augusto da Silva. Arquitetura de um sistema de análise de dados big data no modelo cloud computing. Dissertação de mestrado. Universidade do minho escola de engenharia, 2014.

MCGUINNESS, D.; HARMELEN, F. OWL Web Ontology Language Overview. 2004. Disponível em: <http://www.w3.org/TR/2004/REC-owl-features-20040210/>. Acessado em: outubro de 2016.

MILHOMEM, Wisley Cristiano de Souza. Indexação de termos para um sistema de recuperação da informação utilizando computação distribuída. Trabalho de conclusão de curso. Centro Universitário Luterano de Palmas, 2013.

MORAIS, E.; AMBRÓSIO, A. P. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens, 2007. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-07.pdf>. Acessado em: outubro de 2016.

PEREIRA, Daniel José Pinto. Armazéns de dados em bases de dados nosql. Dissertação de mestrado. Instituto superior de engenharia do porto, 2014.

PEREIRA, Flávius Anderson Félix. Geração semi-automatizada de modelo multidimensional em um cubo OLAP a partir de ontologias. Requisito parcial para obtenção do grau de mestre em ciência da computação. Universidade Federal de Pernambuco, 2011.

PERERA, Srinath. Considerações sobre o banco de dados apache Cassandra, 2012. Disponível em: <http://www.ibm.com/developerworks/br/library/os-apache-cassandra/>. Acessado em: outubro de 2016.

PRADO, Thiago Coelho. Otimização da persistência de dados em pacs empregando modelos de dados hierárquicos indexados. Dissertação de mestrado. Universidade Federal de Santa Catarina, 2012.

SADALAGE, P. J.; FOWLER, M. NoSQL distilled: A brief guide to the emerging world of polyglot persistence. Estados Unidos da América: Addison-Wesley Professional, 2013. 192 p.
ABRAMOVA, V.; BERNARDINO, J.; FURTADO, P. Experimental evalution of nosql databases. International Journal of Database Management Systems. Vol. 6, No. 3, p.1-16. 2104.

SATHI, A. Big Data Analytics: Disruptive Technologies for Changing the Game: 1ª Edição. Canadá: MC Press Online, 2012. 97p.

SCHROEDER, Ricardo. Consulta integrada entre big data e data warehouse. Trabalho de conclusão de curso. Universidade de Santa Cruz do Sul, 2013.

SCHUSTER FILHO, Roberto Antonio. Adaptação temporal e qualitativa sobre mecanismos de clipagem eletrônica. Trabalho de conclusão de curso. Universidade de Santa Cruz do Sul, 2013.

SINGH, S.; SINGH, N. Big Data Analytics. Communication, Information & Computing Technology (ICCICT), 2012 International Conference on, 19-20 Outubro 2012 Mumbai, India, p.1-4. 2012.

TAURION, Cezar. Big data. Rio de Janeiro: Brasport Livros e Multimídia, 2013. 184 p.

ZIKOPOULOS, P. C.; EATON, C.; DEUTSCH, D. D. T.; LAPIS, G. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streming Data: 1ª Edição. Estados Unidos da América: McGraw-Hill, 2012. 142p.