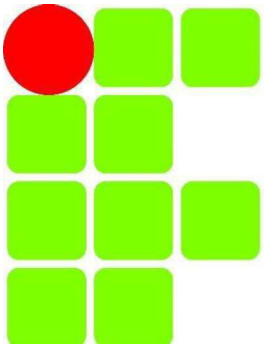

INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
SUL DE MINAS GERAIS
Campus Inconfidentes

MATEUS OLIVEIRA CABRAL

DATA WAREHOUSE: CONCEITOS E TECNOLOGIAS RELACIONADAS

INCONFIDENTES

2016

MATEUS OLIVEIRA CABRAL

DATA WAREHOUSE: CONCEITOS E TECNOLOGIAS RELACIONADAS

Trabalho de Conclusão de Curso apresentado como pré-requisito de conclusão do curso de Graduação Tecnológica em Redes de Computadores no Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Campus Inconfidentes, para obtenção do título de Tecnólogo em Redes de Computadores.

Orientador: Bruno Amarante Couto Rezende

INCONFIDENTES

2016

DEDICATORIA

Dedico este trabalho aos meus pais, por todo incentivo que me deram na minha jornada acadêmica e pelo suporte a alcançar todos os meus objetivos até hoje. Espero que com este trabalho possa retribuir de alguma forma o sacrifício que fizeram por mim.

AGRADECIMENTOS

Aos amigos, que sempre incentivaram meus sonhos e estiveram sempre ao meu lado mesmo com as grandes distâncias não somente física, mas como a do dia a dia.

Juntamente gostaria de agradecer aos meus colegas de classe pelo companheirismo e aos professores e servidores que estiveram me auxiliando nesta jornada também.

Ao estimado professor Bruno Amarante Couto Rezende que me acompanhou e auxiliou no desenvolvimento desse trabalho sempre muito paciente e muito atento ao meu projeto.

EPÍGRAFE

"Porque eu sou do tamanho do que vejo, e não do tamanho da minha altura. E o que vejo são os meus sonhos"

Fernando Pessoa

RESUMO

Neste trabalho discute-se os objetivos de um Data Warehouse (DW) e as diferenças entre outras tecnologias de armazenamento de dados operacionais, que podem possuir diferentes estruturas. Foram explorados os principais componentes e discutido o papel da modelagem dimensional e sua diferente abordagem, quando comparada com a modelagem Entidade-Relacionamento (ER). Para melhor apresentação do tema proposto é estabelecido um vocabulário preliminar sobre conceitos básicos do DW como, por exemplo, as diferentes arquiteturas, implementações, esquemas de modelagem e outros aspectos importantes. Brevemente são discutidos os principais processos e as melhores práticas de um projeto de DW até o esforço da sua implementação. Embora cada projeto seja diferente do outro, inevitavelmente, é necessário dar a devida atenção para cada uma das tarefas que garantem uma implementação de sucesso. Serão apresentados conceitos básicos, para tecnologias que podem trabalhar paralelamente com DW, tais como Big Data e mineração de dados. Big Data é uma combinação de tecnologias de gerenciamento de dados que têm evoluído ao longo do tempo, já a mineração de dados é uma tecnologia que trabalha com pequenas porções de dados, estabelecendo padrões que aconteceram no passado e criando hipóteses de acontecimentos futuros. Complementando os conceitos teóricos, este trabalho acompanha um estudo de caso focado na construção de um DW, partindo do processo de definição da estrutura e metas até a sua implementação física em um servidor de dados.

Palavras Chaves: Data Warehouse, Mineração de Dados, Big Data, Análise de Dados

ABSTRACT

In this paper, we discuss the goals of a Data Warehouse (DW) and the differences between others operational data storage technologies, which may have different data structures. It is explored the main components and it is discussed the role of dimensional modeling and its different approach compared with the Entity Relationship (ER) standard modeling. For the best presentation of the proposed theme is established a preliminary basic concepts' vocabulary of DW, such as the different architectures, implementations, modeling schemes and other aspects of the most importance. Briefly are discussed the main processes and best practices for a Data Warehouse project since implementation effort. Although each project is different from one another, inevitably will be need to give some attention for each tasks that ensure a successful implementation. Basic and brief concepts will be present about technologies that can work in parallel with DW, such as Big Data and data mining. Big Data is a combination of data management technologies that have evolved over time, as data mining is a technology that works with small amounts of data, setting standards that happened in the past and creating chances of events that may be happening in the future. Complementing the theoretical concepts, this work follows a focused case study in building a Data Warehouse in from the definition process of structure and scope until the physical implementation in the data server.

.

Key Words: Data Warehouse, Data Mining, Big Data, Data Analysis

LISTA DE ABREVIATURAS E SIGLAS

Business Intelligence (BI)

Data Marts (DM)

Data Warehouse (DW)

Database Administrator (DBA)

Entidade Relação (ER)

Extract,Transform and Load (ETL)

Hadoop Distributed File system (HDFS)

Instituto Brasileiro de Geografia e Estatística (IBGE).

Knowledge Discovery in Databases (KDD)

On-Line Analytical Processing (OLAP)

Online Transaction Processing (OLTP)

Sistema de Gerenciamento de Banco de Dados (SGBD)

Sistemas de apoio à decisão (SAD)

Slowly Change Dimension (SCD)

Structured Query Language (SQL)

Support Vector Machine (SVM)

Tecnologia de Informação (TI)

Terceira Forma Normalizada (3NF)

LISTA DE FIGURAS

Figura 1. Exemplo da armazenagem dos dados por assuntos (tópicos).. 10
Figura 2. Diferenças de níveis de granularidade. 13
Figura 3. Exemplo de Data Warehouse de Arquitetura Global. 15
Figura 4. Exemplo de uma Arquitetura Independente de Data Marts. 15
Figura 5. Arquitetura Interconectada de um Data Mart..... 16
Figura 6. Implementação Top Down. 17
Figura 7. Implementação Bottom Up. 18
Figura 8. Modelo de Kimball para Componentes de um Data Warehouse. 19
Figura 9. Exemplo de Modelagem de Banco de Dados Relacionais. 24
Figura 10. Exemplo de Visualização Dimensional 26
Figura 11.Exemplo de Esquema Estrela..... 29
Figura 12.Exemplo de Esquema Floco de Neve.. 30
Figura 13. Esquema de nós Hadoop. 42
Figura 14 Exemplo de um Data Warehouse integrado ao Hadoop 45
Figura 15. Ciclo da Metodologia baseada em Herdem 46
Figura 16. Modelo Conceitual do Sistema 53
Figura 17. Esquema Estrela do Sistema 54
Figura 18. Pentaho Input passo..... 56
Figura 19. Transformação de Dimensões 56
Figura 20. Janela metric do programa Pentaho 58
Figura 21. Transformação da Tabela Fator 58
Figura 22. Modelo para SCD 2 e 1 59
Figura 23. Tela de configuração do Data Warehouse 60
Figura 24. Área de trabalho do Tableau 61
Figura 25. Total de Municípios 62
Figura 26. Cálculos Estatísticos por região. 63
Figura 27. Médicos e Professores por estado 64
Figura 28. Gestores da Educação e Saúde por gênero..... 65
Figura 29. Grau de Escolaridade 66
Figura 30. Formação dos gestores municipais 67
Figura 31. Inclusão Social 68
Figura 32. Quadro de funcionários municipais de Inconfidentes e Região..... 69
Figura 33. Perfil dos Gestores de Inconfidentes e Região..... 70

LISTA DE TABELAS

Tabela 1. Diferenças entre banco de dados operacionais e Data Warehouse. 8
Tabela 2 Diferenças entre Data Warehouse e Data Mart. 9
Tabela 3. OLTP vs OLAP 32
Tabela 4. Tipos de SCD 35

SUMÁRIO

1.INTRODUÇÃO 1
1.1 Relevância do Trabalho 2
1.2 Proposta 3
1.3 Trabalhos Relevantes 3
2.CONCEITOS BÁSICOS DO DATA WAREHOUSE 5
2.1 Históricos da Armazenagem de Dados e a Evolução 5
2.2 Diferenças entre Data Warehouse e Banco de Dados Relacional 7
2.3 Data Mart 9
2.4 Características do Data Warehouse 10
2.4.1 Orientado a Objeto (Organizado em assuntos) 10
2.4.2 Integrado 10
2.4.3 Variações de tempo 11
2.4.4 Não Volátil 11
2.4.5 Granularidades dos dados 12
2.5. Deveres de um Data Warehouse 13
2.6 Arquiteturas do Data Warehouse 14
2.6.1 Arquitetura Global de um Data Warehouse 14
2.6.2 Arquitetura Independente de um Data Warehouse (Data Mart) 15
2.6.3 Arquitetura Interconectada de um Data Mart 16
2.7 Implementações 16
2.7.1 Implementação Top Down 17
2.7.2 Implementação Bottom Up 17
2.7.3 Implementação Combinada (Top Down e Bottom Up) 18
2.8 Componentes de um Data Warehouse 19
2.8.1 Fontes de dados 20
2.8.2 Área de Trabalho (Data Staging Area) 20
2.8.3 Área de Apresentação 20
2.8.4 Ferramentas de acesso aos dados 21
3. TÉCNICAS DE MODELAGEM DE DADOS PARA DATA WAREHOUSE 22
3.1 Modelagem Tradicional 23
3.2 Modelagem Dimensional 25
3.2.1 Fatores, Dimensões e Medidas. 27
3.2.2 Esquemas de Estrutura de dados 28
3.3 Processamento OLAP e OLTP 31
3.4. Carregando o Data Warehouse (Processo ETL) 33
3.4.1 Captura dos dados 33
3.4.2 Transformações 34
3.4.3 Carregamento (Aplicação) 35
3.5 Slowly Change Dimension 35

4.MINERACAO DE DADOS E BIG DATA 36
4.1.Mineracao de Dados 36
4.1.1 Estatísticas 37
4.1.2 Aprendizados de Máquina 37
4.1.3 Bancos de Dados 38
4.2. Big Data 39
4.2.1. Hadoop 41
4.2.2 Sistemas de arquivos distribuídos Hadoop 41
4.2.3 MapReduce 42
4.3 A Integração De Big Data Com Data Warehouse 43
5. METODOLOGIA A SER UTILIZADA 46
5.1 Definições do projeto 47
5.2 Levantamentos dos Requisitos 47
5.3 Modelagem Conceitual 47
5.4 Projeto Lógico 48
5.5 Projeto Físico 48
6. ESTUDO DE CASO: IMPLEMENTANDO UM AMBIEMTE WAREHOUSE PARA PESQUISAS DO IBGE. 49
6.1 Ambiente e Estabelecimento do Pré-projeto 49
6.1.1 Ponto de Partida 49
6.1.2 O Foco do Projeto 50
6.1.3 Responsabilidades de Gerência do DW 50
6.2 Levantamento de Pré-requisitos 51
6.2.1 Softwares Utilizados 51
6.2.1.1 MySQL Workbench 6.3 51
6.2.1.2.Pentaho 52
6.2.1.3 Tableau 52
6.3 Modelagem dos dados 52
6.4 Projeto Lógico 54
6.4.1 Processo ETL 55
6.5 Projeto Físico 59
6.6 Analise dos dados de forma dinâmica 60
6.6.1 Interface do Software 60
6.6.2 Gráficos 61
7.CONSIDERAÇÕES FINAIS 71
REFERENCIAS 73
APENDICE A: DIAGRAMA ESTRELA DO PROJETO 75

1.INTRODUÇÃO

Pode-se afirmar que Data Warehouse (DW) provê uma abordagem na transformação da grande quantidade de dados gerados diariamente nas organizações em informações úteis e confiáveis que podem solucionar questões levantadas durante alguns processos, além de auxiliar nas tomadas de decisões (KIMBALL, 2013). Um DW normalmente armazena os dados recolhidos a partir de múltiplas fontes de dados de uma organização, tais como bancos de dados transacionais

Para este processo, os dados são limpos e transformados em um formato padrão e consistente, antes de serem armazenados no DW. Pode-se também armazenar subconjuntos de dados em um DW, que podem ser extraídos e configurados como Data Mart (DM) para atender aos requisitos específicos de uma divisão organizacional. Ao contrário de bancos de dados transacionais onde os dados são constantemente atualizados, em um DW eles são atualizados apenas periodicamente.

O conceito que diferencia o DW de um sistema operacional de banco de dados comum é a sua estrutura dos dados. Isto é, a construção da estrutura dos dados e das aplicações em volta do banco são mais importantes que estruturar aplicações e trazer os dados até eles (KELLY, 1997).

Há outras ferramentas e tecnologias que podem trabalhar paralelamente ao DW. Um exemplo é a mineração de dados, um processo onde analisa-se padrões e relações em um determinado grupo de dados. Destaca-se que o armazenamento de dados é um processo que deve ser executado antes que qualquer mineração de dados ocorra. Em outras palavras, o armazenamento de dados é o processo de compilar e organizar dados em um banco de dados comum, e mineração de dados é o processo de extração de dados significativos a partir desse banco de dados (HAN, 2012). O processo de mineração de dados baseia-se nos dados compilados na fase de armazenamento de dados. A principal característica de mineração de

dados é a capacidade de entender eventos que ocorreram no passado e a capacidade de prever o que vai acontecer no futuro (CAMILO, 2009).

Este trabalho está organizado da seguinte forma, no capitulo 2 são apresentados os principais conceitos de Data Warehouse em níveis de hierarquia e funcionamento, no capitulo 3 é apresentado técnicas de modelagem de dados, no capitulo 4 são apresentados breves resumos sobre data mining, Big Data e seus papeis em relação a Data Warehouse, no capitulo 5 e 6 são apresentados a metodologia do estudo de caso e o estudo de caso em si e por último as considerações finais.

1.1 Relevância do Trabalho

Data Warehouse (DW) é uma solução não um produto (SHERMAN, 2014), o designer, a implementação do processo junto das ferramentas e os gerenciadores fazem a completa entrega da informação, que precisa ser entendida para uma tomada de decisão corporativa. Isto inclui todas as etapas e passos para que uma organização crie, gerencie e mantenha um DW em funcionamento.

As organizações têm vastas quantidades de dados armazenados, porém está mais difícil mantê-las organizadas e acessíveis para que se possa utiliza-las. A etapa de modelagem consiste em levantamento de requisitos, análise, validação e modelagem. Portanto o desenvolvimento do DW requer muito mais que uma aplicação bem pensada e técnicas de modelagem.

O papel e a proposta do DW nas organizações evoluíram consideravelmente desde do seu surgimento e continuam a evoluir. DW não deve ser identificado como um sistema de banco de dados, que suporta funções de consultas e relatórios para usuários finais, muito menos deve ser construído como *snapshots* de dados operacionais. Os bancos de dados do DW devem ser considerados como fontes novas de informações que podem ser consideradas "usáveis" por toda a organização, por setores menores de usuários e analistas de dados dentro das organizações. Por isso, Inmon (2012) afirma que uma simples reengenharia dos modelos da fonte de dados do padrão tradicional não irá satisfazer os requisitos para a construção e utilização efetiva de um DW. Hoje, a utilidade do DW não é apenas a criação de consultas ad hoc ou geração de relatórios. A real importância consegue-se quando se trabalha, analisa-se os dados na busca de extrair informações que fazem uma grande diferença nas atividades das organizações.

No mercado atual muitas empresas estão adotando o modelo de DW, seguindo grandes grupos que já utilizam esse tipo de implementação, como a Coca Cola, Itaú, Vivo, Walmart entre outras grandes empresas (TURBAN, 2009).

1.2 Proposta

No mercado atual há uma enorme comparação entre as tecnologias de DW e Big Data. Ambas tecnologias possuem quase as mesmas funções, o que praticamente separa uma da outra são os tipos de dados aceitos nas plataformas de armazenamento. DW possui dados estruturados, ou seja, dados provenientes de banco de dados em sua maioria ou se não, os dados são filtrados e transformados para entrar nos padrões do DW. Os dados armazenados do Big Data podem ser dados estruturados e não estruturados (e-mails, fotos, vídeos, mídias sociais), além disso o processo ETL (*Extract, Transform and Load* – Extrair, Transformar e Carregar) é utilizado apenas no DW.

Neste trabalho esclarece-se os conceitos básicos, técnicas, ferramentas para se utilizar na implementação do DW, definições básicas das tecnologias Big Data e mineração de dados. Com tantas tecnologias (DW, mineração dados e Big Data) há uma dificuldade de usuários compreenderem as funções básicas de cada uma e a influência direta que cada tecnologia exerce sobre a outra.

Para evidenciar a utilização do DW, foi realizado um estudo de caso baseado em pesquisas sobre o munícipio brasileiros realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Neste caso a metodologia utilizada para a implementação foi a metodologia de Herdem (discutida no capitulo 5) e DW construído utilizando o MySQL.

1.3 Trabalhos Relevantes

Nessa seção são apresentados os principais trabalhos na área de DW e análise de dados.

Os estudos de Sherman (2014) trazem conceitos de alto nível de *Business Intelligence* (BI) e instruções para utilizar as ferramentas que são essenciais para a implementação de uma arquitetura, desenho e processo. Seu estudo apresenta temas muitas vezes negligenciados, mostrando de forma sucinta o conhecimento necessário para que usuários possam projetar processos de BI utilizando integrações solidas dos dados em um ambiente de DW ou Big

Data. Sherman (2014) tira a necessidade de adivinhação sobre os problemas, transformando-os em sistemas que são rentáveis, reutilizáveis e essenciais para transformar dados brutos em informações valiosas para decisões empresariais.

Os primeiros estudos de Kimball (1996) introduziu às indústrias, as suas próprias técnicas de modelagem dimensional na primeira edição do seu livro "*The Data Warehouse Toolkit*". Desde então, a modelagem dimensional tornou-se a abordagem mais aceita para apresentação de informações no DW e sistemas de BI. O Data Warehouse Toolkit é reconhecido como a fonte definitiva para as técnicas de modelagem dimensional, padrões e melhores práticas.

Outro autor de estudos de DW é Inmon (2012). A sua perspectiva é uma visão mais *top down* (de cima para baixo) olhando para a arquitetura geral e, em seguida, investigando as questões subjacentes dos componentes do DW. Com um olhar mais futurístico, ele tenta determinar quais novas tecnologias podem ser compradas, como planejar as extensões para o DW, o que pode ser recuperado a partir do sistema atual e como justificar os gastos com a implementação.

2.CONCEITOS BÁSICOS DO DATA WAREHOUSE

O DW é a base para o processamento de informação de qualquer corporação (KIMBALL, 2013). Ele contém dados históricos granulares e integrados. A integração dos dados permite a uma corporação ter uma verdadeira visão empresarial de todos os seus dados. Ao invés de olhar para apenas uma parte deles, o analista poderá ver toda a informação de diversas fontes de dados como se estivesse vindo de apenas uma única fonte definida.

Usar o DW para visualizar fatos ou acontecimentos cotidianos através de dados da corporação é a sua primeira grande vantagem (BALLARD et al, 1998). Utilizando a granularidade (o mais fino nível de detalhe) permite que os dados sejam mais flexíveis e possam ser examinados de uma maneira por um grupo e de outra maneira por um outro grupo de usuários. Outra grande característica e a capacidade de armazenar os dados em um lugar por vários anos, permitindo que consultas possam traçar uma linha temporal das mudanças ocorridas ao longo da gestão do DW (KIMBALL, 2013).

Mesmo com tantas características positivas, a implementação apresenta algum grau de dificuldade, tais como o processo de integração dos dados, volume de dados e as diferentes abordagens desenvolvidas.

2.1 Históricos da Armazenagem de Dados e a Evolução

Bem no começo dos sistemas computacionais, os usuários conseguiam informações brutas dos computadores, e ainda havia a existência de muitos obstáculos como, por exemplo, os simples mecanismos para o armazenamento de dados, que era algo bem limitado e caro. Nessa mesma época, os usuários usavam cartões perfurados e fitas de papeis para armazenar

os dados, e o modo de leitura era feito através de vários arquivos hexadecimais em uma minúscula partícula de informação escondida atrás de mil páginas de código criptografado (BALLARD et al, 1998).

Uma evolução se iniciou com a introdução da fita magnética, na qual era possível gravar um volume maior de dados. Mas o grande avanço foi o desenvolvimento dos discos de armazenagem de dados representando outro grande passo para o futuro da armazenagem de dados. Com eles os dados podiam ser escritos e reescritos como também poderiam ser acessados diretamente e em massa (BALLARD et al, 1998).

Pouco tempo depois do desenvolvimento dos discos, sistemas de gerenciamento de banco de dados começaram a ser criados. O Sistema de Gerenciamento de Banco de Dados (SGBD) é um software computacional que gerencia os bancos de dados e pode usar diversos modelos de estrutura para um banco de dados. Foram desenvolvidos com o propósito de executar atividades como: identificar a localização exata dos dados; resolver conflitos quando duas ou mais unidades de dado são mapeados na mesma localização física; permitir que os dados sejam deletados, explorar e otimizar uma localização física quando um registro de dado não encaixaria num espaço físico limitado.

Com avanço do armazenamento de dados uma nova preocupação iniciou-se: como interpretar as informações armazenadas. Para um suporte maior das informações armazenadas alguns formulários foram criados, porém por conta da interface pouco desenvolvida, outras soluções surgiram, e uma delas foram os relatórios.

Outro passo significativo se deu com a disponibilização instantânea (online) das informações. Possibilitando que os usuários integrassem as informações corporativas em apenas um meio (fonte de dados) de forma que todas as informações fossem mantidas em um único histórico de dados.

Tal progresso iniciou-se simultaneamente com o desenvolvimento de arquiteturas e tecnologias da primeira geração de DW, que facilitou a localização, interpretação de dados e geração de relatórios. Sem o desenvolvimento de um DW, os usuários seriam deixados apenas com uma fração de informação do que eles realmente precisavam para tomadas de decisões nas empresas ou organizações. A necessidade dos usuários por informações corporativas que suportassem melhor as decisões da empresa foi o motivo pelo qual se deu a evolução do DW. Essa constante busca para o melhoramento não somente da armazenagem de dados, mas também pela necessidade de usar os dados armazenados para interpretar e corrigir algumas condutas da corporação com as informações armazenadas (BALLARD et al, 1998).

O DW representou uma grande mudança no pensamento dos profissionais de Tecnologia de Informação (TI). Contudo, antes disso, o pensamento padrão da época era que o banco de dados deveria ser algo que serviria todas as necessidades, porém com o avanço das tecnologias e com a entrada do Big Data, a necessidade de melhora na parte da administração e análise de dados, serviu como propulsor para o desenvolvimento de diversos projetos na área de DW.

2.2 Diferenças entre Data Warehouse e Banco de Dados Relacional

Os bancos de dados executam transações para prover respostas às requisições do sistema, enquanto os DW são baseados em relatório de estrutura ad hoc, geralmente para propósitos de administração gerencial. Os principais focos dos bancos de dados são a segurança e a coerências, que tornam as consultas lentas, especialmente os relatórios ad-hoc.

Mesmo que a maioria dos bancos de dados operacionais e DW sejam construídos numa tecnologia relacional, seus desenhos são substancialmente diferentes, como seus propósitos também. Banco de dados são desenhados para processamento de transações online e seu principal objetivo refere-se à eficiência de armazenar um grande volume de dados transacionais (BALLARD et al, 1998). Eles incluem informações atuais de atividades do dia-a-dia e o processo é orientado a transações. Como um resultado, o dado é dinâmico e bem volátil. As tarefas de tais sistemas são contraídas e repetitivas, e fazem transações concorrentes, curtas e isoladas, incluindo dados detalhados. Essas transações leem e atualizam alguns registros, principalmente acessados baseando-se nas suas chaves primárias. Produzem milhões de megabytes para gigabytes de dados, e sua consistência é essencial.

Como oposto a esse tipo de banco de dados, DWs são desenhados para serem o suporte de sistemas de tomada de decisão, ou seja, eles são desenhados para facilitar a análise de dados, e não a armazenagem. Nele se encontra as operações já executadas que foram carregadas e os acessos de dados históricos. A sumarização (resumo) e conciliação dos dados é mais importante que os dados detalhados, eles incluem dados consolidados de diversos sistemas operacionais, com diferentes intervalos de tempo e tem uma visão integrada e de evolução. Segundo Ballard (1998), os dados são estáticos e não voláteis, o tamanho do DW pode alcançar com o tempo milhões de gigabytes, terabytes ou até mesmo petabytes. Muitos relatórios ad-hoc podem ser feitos e milhões de registros podem ser acessados, como muitas junções e agregações podem ser executadas. A informação é orientada a objeto (assunto) e

os DWs fornecem uma visão multidimensional dos dados, baseado em um modelo intuitivo, desenhado para encontrar os requisitos da análise de dados e dos gestores (responsáveis pela toma de decisão) (SHERMAN, 2014).

Outra diferença é o estado do dado mostrado. DW's mostram o estado do dado em diferentes momentos no tempo e ainda provê uma leitura estendida do histórico do mesmo. Em um banco de dados, o estado dos dados que podem ser mostrados são somente os atuais na hora do acesso.

Enquanto banco de dados operacionais são desenhados para fornecer a otimização do processamento de dados, segurança e escrita, o DW otimiza a análise e a leitura. Para esse propósito, o modelo multidimensional é usado para o desenho do DW para fazer consultas para a análise e sumarização de grandes volumes de dados mais eficiente. A estrutura do DW é simples, intuitiva e de fácil entendimento para uma pessoa não especializada, ao contrário da estrutura de um banco de dados, onde o desenho é baseado na estrutura do modelo entidade-relacionamento e é composto por técnicas específicas.

Na Tabela 1 é apresentado as principais diferenças entre um banco de dados comum e um DW.

Característica	**Banco de dados**	**Data Warehouse**
Usuários	Milhares	Poucos usuários
Carga de Trabalho	Transações do Presente	Consultas específicas de análise
Acesso	Há milhares de registros, modos de leitura e escrita	Há milhões de registros, basicamente fica em modo somente de leitura
Objetivo	Depende das aplicações	Suporte a tomada de decisões
Dados	Detalhados, ambos numéricos e alfanuméricos	Sumarizado, maioria das vezes numérico
Integração de dados	Baseado na aplicação	Orientado a assunto
Qualidade	Em termos de integridade	Em termos de consistência
Tempo de Cobertura	Apenas dados atuais	Dados atuais e históricos
Atualizações	Continuas	Periódicas
Modelos	Normalizado	Desnormalizado e multidimensional
Otimizações	Acesso para OLTP para parte do banco de dados	Acesso para OLAP para grande parte do banco de dados

Tabela 1. Diferenças entre banco de dados operacionais e Data Warehouse (Kelly, 1997).

2.3 Data Mart

Os Data Mart (DM) são, como DW, sistema de suporte a tomada de decisão destinados a um grupo ou setor, que foca em uma função específica ou atividade da empresa. Sua promessa é de simplicidade e foco numa implementação do sistema de suporte a tomada de decisão, com um rápido retorno na demanda do investimento pelo ritmo dos negócios. O DM pode inclusive se transformar um DW através de uma expansão de dados ou entre o compartilhamento de dados em diferentes DM

O seu conceito evolui-se do DW, seu escopo é altamente focado e concentrado em somente um objeto (assunto), ao invés de tratar o sistema inteiro da organização. De fato, um total controle do escopo dessa maneira faz com que o tempo e o dinheiro investimento sejam reduzidos drasticamente (KIMBALL, 2013).

Um DM de sucesso estratégico pode abrandar riscos, limites de expansão e reduzir o tempo requerido para distribuir a funcionalidade do DW. Por causa da escalabilidade (quando é corretamente implementado por alguém experiente no assunto) o DM pode trabalhar muito bem para organizações de qualquer tamanho e nível de complexidade. Desse modo é fácil entender o motivo pelo qual o DM é visto como um dos mecanismos mais efetivos para entrega rápida e capacidade de suporte de decisão confiável.

Uma vez que DM diminui os riscos associados com a construção do sistema de suporte de decisão, não requer grande habilidade e conhecimento para propriamente implementar um sistema como esse. Um DM pode ser transformado em um DW com o tempo devido a incorporação de mais dados ou mais tabelas.

Na Tabela 2 a, segue a comparação entre um DW e um DM.

Característica	Data Warehouse	Data Mart
Dados	Todas áreas de assuntos	Apenas um assunto
Escopo	Escopo bem amplo	Escopo focado
Duração da implementação	Anos	Meses
Investimentos	Milhões	Milhares

Tabela 2 Diferenças entre Data Warehouse e Data Mart. (fonte: O autor)

2.4 Características do Data Warehouse

Um DW possui características chaves destacando-se: a orientação à objeto (ou assunto), a sua integração, a variação do tempo, a não volatilidade e sua granularidade.

2.4.1 Orientado a Objeto (Organizado em assuntos)

É orientado ao assunto porque provê informações relativos a um tópico especifico e não das diversas operações constantes em andamento dentro das organizações (KIMBALL, 2013). Esses tópicos podem ser dos mais diversos grupos, mas normalmente se caracterizam por ser um grupo específico da empresa. Em outras palavras, os dados são armazenados em categorias nas quais um departamento ou um determinado grupo pode trabalhar, economizando tempo no processamento (INMON,2002).

A figura 1, traz um exemplo que destaca traz um exemplo que destaca a questão da orientação ao objeto que será dividida em três partes: produtos, clientes e vendas. E as aplicações desse sistema podem ser compra, estoque, data da venda, entre outros. Um grupo de usuários no departamento de análise cuida da parte de vendas ao qual apenas os dados ligados ao DM venda serão usados na análise e não todo o sistema de DW (Vendas, Produtos e Clientes). Todas as tabelas com o mesmo tópico irão ser agregadas e sumarizadas para ajudar melhor no processamento das consultas.

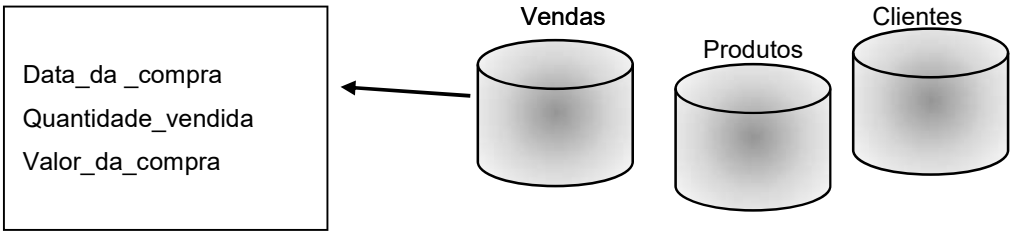

Figura 1. Exemplo da armazenagem dos dados por assuntos (tópicos). (Fonte: O autor).

2.4.2 Integrado

Pode-se enfatizar que o DW é consistente e integrado por conta das suas múltiplas fontes de dados, como por exemplo os dados extraídos da produção, os bancos de dados da empresa, ou ainda também dados de terceiros (sistema de logs de programas ou servidores)

podem ser armazenados em um único sistema (INMON, 2012). De maneira geral, a criação de um DW não requer que nova informação seja adicionada, mas sim que a informação já existente seja rearranjada, bastando um sistema de informação disponível para a sua criação (DW).

A passagem das fontes dos dados para o DW é conduzida de forma que as inconsistências sejam todas desfeitas, e que, se alguma codificação seja necessária (como por exemplo representações de letras atribuídas, ou valores para verdadeiro ou falso for convertidos para valores numéricos) será feita de forma consistente e independente da aplicação de origem (KIMBALL, 2013). Este processo de transformação será melhor discutido no capítulo 3.

2.4.3 Variações de tempo

Uma operação de banco de dados convencional geralmente abrange apenas curtos períodos de tempo porque a maioria das transações envolvem o dado mais recente. Com implementações em DW são fornecidas análises que, ao invés de cobrir um curto espaço de tempo, podem abranger longos períodos e por esta razão, os DW são atualizados com os dados operacionais e continuam a ser populados cada vez mais.

Pode ser usado como analogia para melhor entendimento dessa característica, uma situação na qual os dados do DW fossem visualmente representados como uma fotografia dos dados operacionais. Essa fotografia seria feita em intervalos regulares de tempo, as sequências dessas fotografias seriam armazenadas em um DW e o resultado seria mostrado em um filme que revelaria o progresso da empresa desde sua fundação até o período atual (INMON, 2012).

2.4.4 Não Volátil

Basicamente, os dados nunca são deletados do DW e atualizações normalmente ocorrem quando os DW estão desconectados (off-line). Em outras palavras, isso mostra que eles podem ser considerados como banco de dados de somente leitura e isto cumpre com a necessidade do usuário ter apenas um pequeno intervalo de tempo de respostas para suas análises e/ou consultas, Sempre trabalhando com todas as operações.

2.4.5 Granularidades dos dados

Segundo Inmon (2012) e Kimball (2013), a mais importante característica do DW é a granularidade, que algumas vezes pode tornar-se um problema. De fato, esta questão permeia toda a arquitetura do DW. Granularidade se refere ao nível de detalhe ou resumo (sumarização) das unidades de dados. Quanto maior o detalhamento, menor o nível de granularidade e quanto menor o detalhamento, maior o nível de granularidade.

Por exemplo, uma simples transação estaria no nível mais baixo de granularidade, e um resumo de todas as transações por mês estaria num nível mais alto de granularidade. Quando dados detalhados estão sendo atualizados é quase certo que o dado será armazenado no nível inferior de granularidade.

Esta característica pode afetar profundamente o volume de dados que está alocado no DW e o tipo de consulta a ser respondida. Em quase todos os casos, dados vão para o DW em um nível bem alto de granularidade (INMON, 2012). O que significa que o desenvolvedor deve ter muitos gastos com recursos para reconstruir os dados com uma maior segmentação.

A granularidade encontrada nos dados de um DW é a chave de tudo pois eles poderão ser reutilizados, sendo assim usados por muitas pessoas de diferentes formas (SHERMAN, 2014). Por exemplo, ao se colocar uma organização, com os mesmos dados que serão usados para satisfazer as necessidades dos departamentos de Marketing, Vendas e Contabilidade, todos esses departamentos analisarão basicamente o mesmo dado. O setor de Marketing pode querer consultar as vendas em um mês em uma determinada localização geográfica, as vendas podem querer visualizar as vendas de um determinado vendedor em uma semana e o setor de finanças talvez queira ver a renda de uma linha de produtos. Todas essas informações estão proximamente relacionadas, ainda que sejam um pouco diferentes. Com um DW, diferentes organizações são capazes de visualizar os dados como eles desejam ver. Pela Figura 2, pode-se afirmar que a quantidade de detalhes é proporcional ao espaço requerido para armazenamento, ou seja, quanto mais detalhes mais espaços necessários, quanto menos detalhes menos espaços necessários.

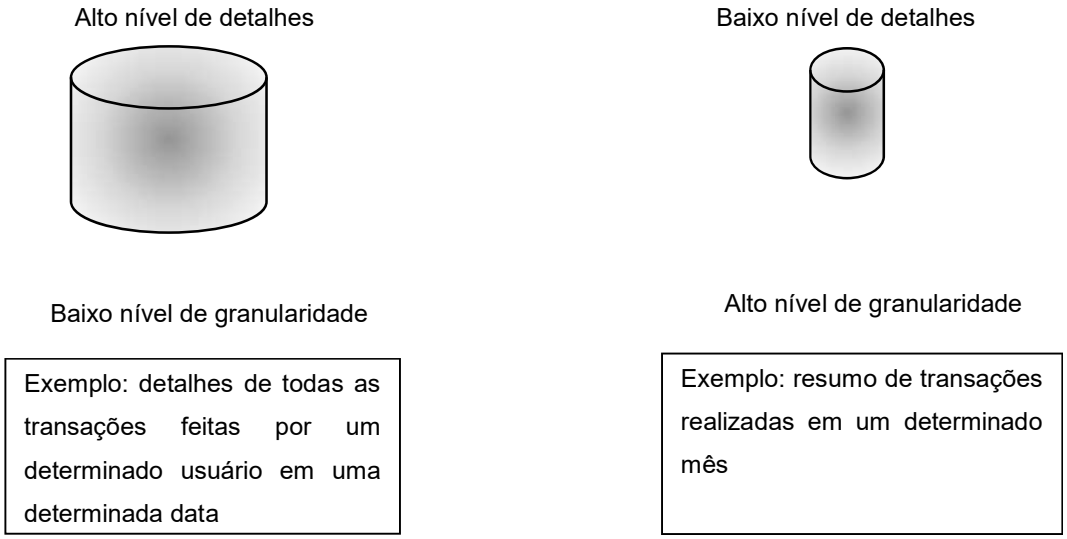

Figura 2. Diferenças de níveis de granularidade (Fonte: O autor).

2.5. Deveres de um Data Warehouse

O DW deve fazer com que a informação gerencial da organização seja de fácil acesso, os conteúdos legíveis e os dados intuitivos e óbvios para os gestores da organização. A capacidade de compreender um DW implica que seu conteúdo esteja devidamente rotulado. Usuários que querem separar e unir os dados no DW em infinitas possibilidades, utilizam processos comumente referidos como operação *slicing* e *dicing.* Ambas operações envolve a utilização do DW como um cubo, a diferença é que a operação *slicing* agrega apenas uma dimensão do cubo, enquanto a operação do *dicing* oferece múltiplas dimensões. As ferramentas que acessam o DW devem ser simples e fáceis de usar. Elas devem também retornar os resultados da consulta para o usuário com um intervalo mínimo de espera.

A consistência nos dados deve estar presente, pois tanto a estrutura como os dados devem ser confiáveis. Os dados devem ser cuidadosamente agregados de várias fontes ao redor da organização, eles devem ser limpos com uma qualidade assegurada e que seja liberado apenas quando se encaixar nos padrões de consumo do usuário.

Duas medidas de desempenho com o mesmo nome, devem significar a mesma coisa. Convencionalmente, se as duas medidas não significam a mesma coisa, então eles devem ser nomeados diferentemente. Com a apresentação de informações consistentes, há uma alta qualidade na informação, e isto significa que todos os dados são contabilizados, íntegros e disponíveis para o usuário (BALLARD et al, 1998).

O DW deve ser adaptável e resistente a mudanças. Necessidade dos usuários, condições da organização, dados e tecnologia são todos quesitos fundamentais a serem observados. Por conta disso, o desenho do DW deve ser planejado para que seja adaptável a mudanças, não invalidando dados já existentes ou aplicações, mesmo quando novos dados são adicionados ao DW.

A proteção da informação é outro dever importante do DW. Uma organização deve conter informações sigilosas como, por exemplo, o que estão vendendo, para quem estão vendendo e a que preço. Isto pode se tornar uma informação perigosa em mãos de concorrentes. O DW deve efetivamente controlar o acesso para informações confidenciais da organização (BALLARD et al, 1998).

2.6 Arquiteturas do Data Warehouse

A seleção de uma arquitetura irá determinar ou ser determinada pela localização do DW ou DM, ou aonde o controle central estiver alocado (INMON, 2012). Por exemplo o banco pode ser alojado e administrado numa localização central, ou o banco pode ser alojado distributivamente em vários locais remotos e ser administrado centralmente ou independentemente.

Os tipos de arquitetura existentes são: global, independente, interconectada ou alguma combinação das três. As implantações podem ser: *top down, bottom up* ou uma combinação das duas (BALLARD et al, 1998).

2.6.1 Arquitetura Global de um Data Warehouse

De maneira resumida pode se dizer que é considerada uma arquitetura que suportará toda ou uma grande parte da corporação. Tem como requisito um DW totalmente integrado com um grau de acesso e uso sobre todos departamentos. Seu desenho e construção, demonstrado na figura 3, se baseia nas necessidades da empresa como um todo. Tem como característica um repositório comum para a tomada de decisão e o suporte ao banco que estão disponíveis através de toda a organização, ou num grande subconjunto.

O termo global é usado para refletir o escopo do acesso ao banco de dados e o uso, não a sua estrutura física. Os dados para o DW são tipicamente extraídos dos sistemas de operação e possivelmente também proveniente de dados externos às organizações como, por

exemplo, um processamento de dados batch durante períodos de operação com baixa utilização. Os dados são filtrados, é eliminado qualquer dado não requisitado, e os demais são transformados para encontrar a qualidade e a usabilidade requerida e, por fim, são carregados apropriadamente no DW para o acesso dos usuários. Depois de todos esses processos é disponibilizado a todos os usuários da empresa a leitura do banco. Porém este tipo de ambiente aonde todos usuários têm acesso pode consumir muito tempo e também é muito caro para poder implementar (BALLARD et al, 1998).

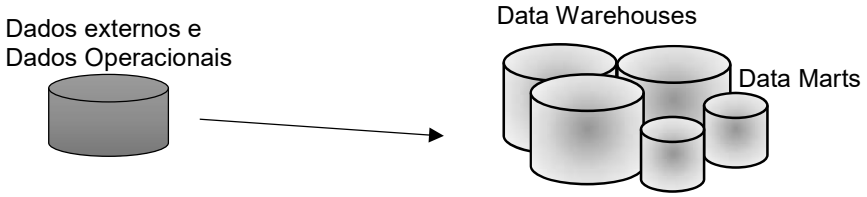

Figura 3. Exemplo de Data Warehouse de Arquitetura Global (Fonte: O autor).

2.6.2 Arquitetura Independente de um Data Warehouse (Data Mart)

Implica-se em uma arquitetura de Data Mart (DM) do tipo independente controlado por um grupo em particular ou departamento e construído exclusivamente para as necessidades dos mesmos. Esta arquitetura requer algumas habilidades técnicas para a implementação, mas os recursos podem pertencer e serem administrados pelo grupo ou departamento.

Este tipo de implementação tem tipicamente o mínimo de impacto nos recursos computacionais e pode resultar numa rápida implementação. Contudo, o mínimo de integração e a falta de uma visão mais global dos dados pode ser prejudicial. Cada DM será accessível apenas pelo grupo ou departamento que pertence como pode ser visto na figura 4.

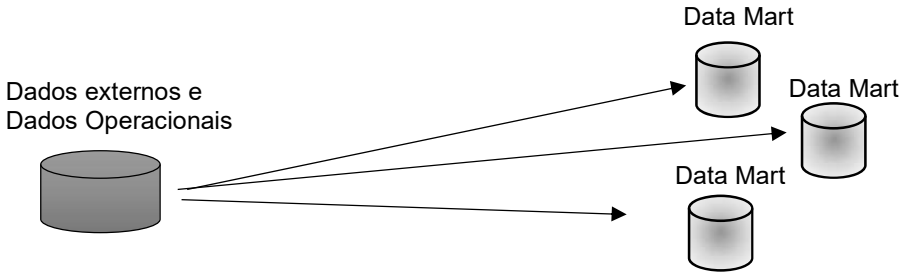

Figura 4. Exemplo de uma Arquitetura Independente de Data Marts (Fonte: O autor).

2.6.3 Arquitetura Interconectada de um Data Mart

Esta arquitetura é basicamente uma implementação distribuída. Embora semelhante com a arquitetura independente, os grupos ou departamentos podem ser integrados ou interconectados, disponibilizando uma maior visão de toda a empresa como é mostrado na figura 5. De fato, ao maior nível de integração eles podem se tornar um DW Global. Assim, usuários de um departamento podem acessar e usar os dados de um outro departamento que esteja em um DM diferente.

Esta arquitetura traz consigo muitas outras funcionalidade e capacidades. Estas opções adicionais podem trazer algumas integrações adicionais que terão requisitos e complexidades como comparada com a arquitetura independente.

Eles podem ser independentes ou controlados por um grupo (departamento) que decide qual será a fonte do banco e como será carregado no DM durante a atualização, ou quem poderá acessar e onde irá se localizar. Eles podem também escolher para prover as ferramentas e habilidades para a implementação do DM eles mesmos (sistema DW).

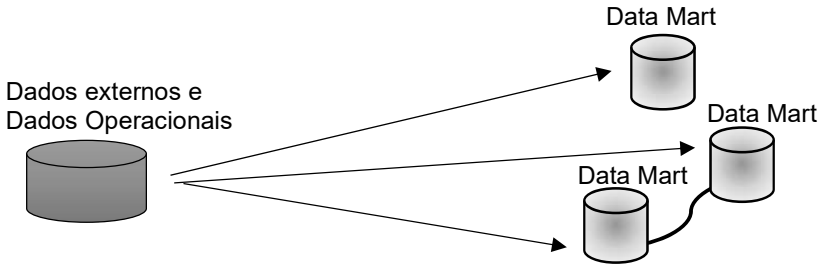

Figura 5. Arquitetura Interconectada de um Data Mart (Fonte: O autor).

2.7 Implementações

A escolha da abordagem da implementação é influenciada por fatores como a infraestrutura computacional, requisitos do retorno do investimento, recursos disponíveis, a arquitetura de seleção, o escopo da implementação e quais grupos são permitidos ter acesso ao banco e a velocidade da implementação.

A seguir são discutidas as duas principais técnicas de implementação e a combinação das duas.

2.7.1 Implementação Top Down

Uma implementação como essa requer maior planejamento e um desenho completo para se dar início ao projeto. Preocupações com decisões a respeito de recursos, segurança, estrutura, qualidade, padrões e acima de tudo a modelagem do banco a ser usados serão necessárias para completar o projeto antes da implementação começar. O custo inicial do planejamento e desenho pode ser significante. Além de consumir muito tempo durante o processo, pode causar atrasos na atual implementação e retorno do investimento. Desenvolver uma modelagem de dados global é também uma longa tarefa, o que tem levado muitas organizações a não optar por esta abordagem. Essa implementação requer que tudo seja planejado antes que iniciado a construção do DW, pois o sistema será desenvolvido completamente sem haver qualquer alternativa de futuras expansões quando finalizado.

Esta implementação pode trazer melhores resultados em cenários de uma organização que centraliza seus sistemas computacionais e outros recursos computacionais. Sendo recomendado sua implementação quando a empresa não possui um DW, ou seja, para uma primeira implantação A Figura 6 demonstra a criação de um DW com uma implementação *Top Down,* onde é criada primeira a infraestrutura da organização (BALLARD et al, 1998).

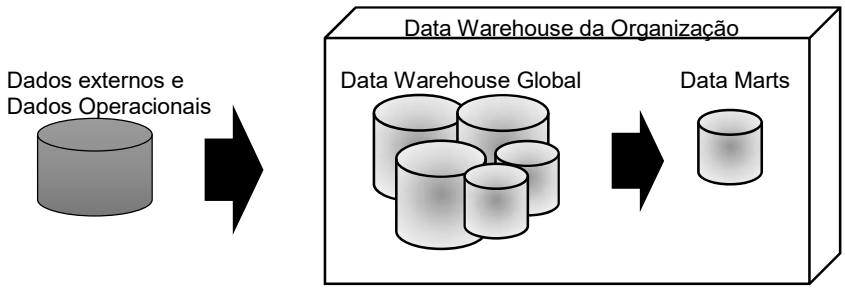

Figura 6. Implementação Top Down. (Fonte: O autor).

2.7.2 Implementação Bottom Up

Envolve o planejamento e o design do DM sem esperar por uma estrutura global de um DW. É construída incrementalmente e inicialmente como um DM que poderá se expandir com as futuras implementações e se transformar em um DW. Tem se tornado uma implementação mais aceita entre as organizações por conta da rápida ação, disponibilização

dos resultados mais rápidos, pois os DM têm um projeto menos complicado que um DW global. Os gastos iniciais também são menores em termos de hardware e outros recursos que são utilizados no DW global.

Quanto mais DM's são criados, mais redundâncias e a inconsistências entre eles podem ocorrer. O cuidado no planejamento, a monitoração e a criação das diretrizes podem ajudar a minimizar este problema. Múltiplos DM's podem trazer com eles umas sobrecargas no sistema pela quantidade de dados que são extraídos das requisições. Se isso um dos focos é integrando o DW em um ambiente mais global, pode haver dificuldades menores se algum planejamento for feito. Algum retrabalho pode também ser requerido como implementação do crescimento e novos problemas serão descobertos forçando a troca de áreas existentes de implementação.

A figura 7 traz a implementação *Bottom Up*, que começa com a criação de um DM e com o tempo se expande para um DW global.

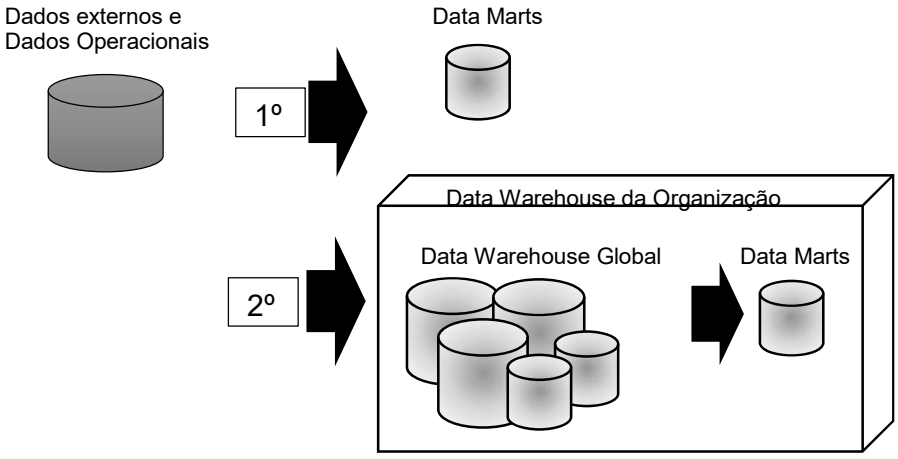

Figura 7. Implementação Bottom Up.(Fonte: O autor)

2.7.3 Implementação Combinada (Top Down e Bottom Up)

Um dos princípios desta implementação é determinar o grau de planejamento e design que será requerido para uma aplicação global para suportar a integração do DM que serão construídos com a abordagem *Bottom Up.* Com os DM sendo implementados e desenvolvidos, é necessário um plano de como será a administração dos elementos do banco que serão necessários para os múltiplos DM. Este poderia ser o início de um DW global mais

estruturado ou simplesmente um banco comum de dados armazenados, acessível por todos os DM.

A decisão de troca entre o armazenamento de fácil acesso e o impacto da redundância deve ser efetuada juntamente com o requisito para deixar o banco em múltiplos DM no mesmo nível de consistência. Um monitoramento cuidadoso de um processo de implementação e gestão dos problemas pode resultar em ganho de melhor benefício de ambas técnicas de implementação (BALLARD et al, 1998).

2.8 Componentes de um Data Warehouse

Cada componente do DW possui uma função específica, o que gera uma grande dificuldade para se definir os papeis e funções de cada componente. Existem quatro componentes distintos como mostra a figura 8, que são: fonte de dados, área de trabalho (*data staging área*), área de apresentação de dados e ferramentas de suporte a decisão.

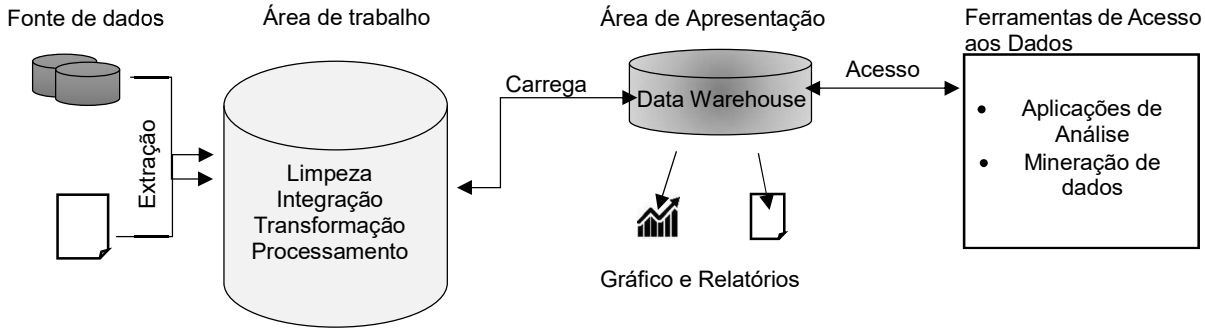

Figura 8. Modelo de Kimball para Componentes de um Data Warehouse. (Fonte: O autor).

O modelo de Kimball (1998) mostrado na figura 8, divide todo os componentes em quatro grupos, sendo que dois primeiros grupos de componentes são responsáveis pela preparação dos dados para serem implementados no DW que seria o terceiro grupo de componentes. O quarto grupo de componentes se compreende as ferramentas auxiliares para enriquecer cada vez mais as consultas do DW.

2.8.1 Fontes de dados

São sistemas de registro que capturam as transações da organização A fonte do sistema normalmente está fora do DW, presume-se, então, que não há controle sobre o formato e o conteúdo dos dados. Cada fonte de dados é frequentemente uma natural aplicação funil, da qual um pequeno investimento tem sido feito para compartilhar dados em comum. As principais propriedades da fonte do sistema são a capacidade de processamento e a disponibilidade (FERREIRA, 2002).

2.8.2 Área de Trabalho (Data Staging Area)

A área de trabalho também conhecida como *Data Staging Area* do DW, age tanto como uma área de armazenamento dos dados como também um grupo de processos comumente chamados de Extração, Transformação e Carregamento (ETL) que é discutido melhor na seção 3.4. Esta área abrange todas as tarefas entre a fonte de dados e a área de apresentação. No DW, dados operacionais brutos são transformados e entregues, em um DW, formatados para as consultas e consumo dos usuários. O *backend* deste sistema aonde são carregados os dados é acessível apenas para profissionais habilitados (KIMBALL,2013).

2.8.3 Área de Apresentação

A área de apresentação de dados é onde os dados são organizados, armazenados e disponibilizados para consulta direta para os usuários, ou outras aplicações de análise. Desde que a área de trabalho conclua sua ação, a área de apresentação são as consultas que podem ser geradas do próprio DW, sendo tudo que a organização visualiza e manipula através de ferramentas de acesso a dados. Isto é o que a área da apresentação com a sua modelagem dimensional se resume.

Normalmente a área de apresentação é referida como uma série de DM integrados. Em sua forma mais simplista, um DM apresenta os dados de um único grupo da organização, por exemplo, em um DW projetado para vendas terá todos os dados das vendas realizadas e nenhum dado a respeito dos clientes, pois o foco será apenas as vendas. Os dados precisam ser apresentados, armazenados e acessados em esquemas dimensionais (KIMBALL, 2013).

Ao contrário das ideias inicias a respeito do DW, os DM modernos podem muito bem ser atualizado, com alguma frequência em certas ocasiões, pois apenas um, no máximo dois, setores da cooperação vão estar trabalhando com ele. Dados incorretos, obviamente, devem ser corrigidos e mudanças nas definições da estrutura, hierarquias, status e propriedades da organização. Muitas vezes, desencadeiam alterações necessárias nos dados originais armazenados nos DM que compreendem o DW, mas em geral, esses problemas são tratados no momento de cargas de atualizações, tem necessidade.

2.8.4 Ferramentas de acesso aos dados

O último componente importante do DW são as ferramentas de acesso a dados. O termo ferramenta está ligado as diversas capacidades que podem ser fornecidas para os usuários dentro da organização para melhorar os resultados obtidos na área de apresentação para uma tomada de decisão analítica. Por definição, todas as ferramentas de acesso a consulta de dados estão alocadas na área de apresentação (FERREIRA, 2002).

Uma ferramenta de acesso a dados pode ser tão simples como uma ferramenta de relatórios ad hoc ou tão complexo como uma sofisticada mineração de dados ou um aplicativo de modelagem. Ferramentas de relatórios ad hoc são poderosas e podem ser compreendidas e utilizadas de forma eficaz apenas por uma pequena porcentagem dos usuários do DW.

A maioria dos usuários irá provavelmente acessar os dados por meio de parâmetros pré-construídos das aplicações analíticas já executadas. Cerca de 80 a 90 por cento dos potenciais utilizadores serão servidos por estas aplicações que são essencialmente modelos já finalizados e não exigem que os usuários construam consultas relacionais diretamente (FERREIRA, 2002). Algumas das mais sofisticadas ferramentas de acesso a dados, como modelagem ou ferramentas de predição, na verdade, podem fazer o *upload* dos seus resultados de volta ao sistema ou para as áreas de trabalho ou apresentação do DW.

3. TÉCNICAS DE MODELAGEM DE DADOS PARA DATA WAREHOUSE

Nos dias atuais Data Warehouse (DW) é a melhor abordagem de análise de dados e para tomada de decisão de organizações, pois proporciona a integração consistente de dados. Contudo pode apresentar problemas complexos e requer uma significativa parcela de tempo e recursos para se implementar.

Os DWs são construídos a partir de uma modelagem, uma abstração e reflexão do mundo real. A modelagem em si, tem como habilidade a visualização do que não poderia ser realizado (INMON, 2012). As duas técnicas de modelagem mais usada são: a modelagem entidade-relacionamento e modelagem dimensional. Em ambientes operacionais, a modelagem entidade-relacionamento tem sido a técnica escolhida. Com o avanço do DW, a necessidade de uma técnica que suportaria a análise de dados em ambientes como grandes organizações surgiu. Mesmo que a modelagem entidade-relacionamento possa ser usada para suportar o ambiente do DW, há ainda a necessidade de uma modelagem dimensional para a analisar os dados. Tradicionalmente, as modelagens de dados devem usar um diagrama Entidade-Relacionamento (ER), desenvolvido como parte do processo de modelagem de dados, como se fosse o canal de comunicação entre a organização e os usuários. O diagrama é uma ferramenta que pode ajudar na análise dos requisitos da organização e o desenho da estrutura de dados da organização. Já a modelagem dimensional ajuda no aperfeiçoamento da capacidade de visualização de perguntas abstratas, aos quais os usuários da organização precisam de resposta.

O debate a respeito dos tipos de modelagem nos dias atuais, está entre os tradicionalistas que promovem a modelagem tradicional com normalização na terceira forma como a única abordagem para DW, e os que proclamam que os modelos ER não são

convenientes para DW porque são muito técnicos e complexos para usuários finais. Para este tipo de pensadores, a modelagem dimensional promove a "salvação", por conta, principalmente, das técnicas que produzem modelos dimensionais de fácil entendimento (modelos estrelas, como o seu oposto o modelo de floco de neve, mostra a estrutura da hierarquia da dimensão no modelo dimensional). Além destas duas técnicas de modelagem a mineração de dados vem também ganhando bastante mercado atualmente como uma possível abordagem para análise de pequenos grupos de dados (KIMBALL,2013).

3.1 Modelagem Tradicional

Para este tipo de modelagem são usados três símbolos gráficos para contextualizar o banco: entidade, relação e atributo.

A entidade pode ser definida como se fosse uma pessoa, lugar ou um evento de interesse para a organização. Representam classe de objetos que são coisas do mundo real que podem ser observadas e classificadas por suas propriedades e características. Mesmo podendo diferenciar-se através das fases de modelagem, geralmente uma entidade tem sua própria definição dentro da organização e um limite definido claramente, onde é obrigatório ser descrito o que está incluído e o que não está. A questão mais crítica em uma entidade é a definição de um atributo de identificação única. Esses identificadores são chamados de chaves candidatas. Encontra-se um único atributo (ou mais de um em alguns casos raros) que descreva melhor a identidade de toda entidade, esse atributo passa ser identificado como chave primária (BALLARD et al, 1998).

As relações são representadas como linhas entre as entidades. Retratam a estrutura de interação e associação entre as entidades no modelo, normalmente é atribuída com uma ação ou substantivos tais como: pertence, possui e tem. A relação entre entidades pode ser definida pela cardinalidade, que define o número máximo de casos que podem ser relacionados com uma outra entidade. As possibilidades de cardinalidade são: um-para-um (1:1), um-para-vários, (1: M) e vários-para-vários (M:N). Em um modelo ER normalizado todas as entidades que tenham qualquer relação de vários-para-vários não é aplicável, pois gerariam muitos problemas. Por conta desse fator, relações deste tipo são dissolvidas em relações do tipo um-para-vários, isto é feito através da criação de uma nova tabela adicionado do lado que possuir a maior cardinalidade (BALLARD et al, 1998).

Os atributos são as descrições das características de cada entidade. O nome do atributo deve ser único na entidade e deve também ser autoexplicativo. Quando um dos

atributos não tem valor, o mínimo de cardinalidade é zero, o que significa que ou é nulo ou opcional. Um banco de dados de pedidos de vendas por exemplo pode começar com um registro para cada linha, mas se transforma em uma complexa teia de aranha.

A normalização é uma técnica que visa eliminar a redundância dos dados e existem várias regras para normalização, mas que o mercado adota até a 3NF. A modelagem relacional em Terceira Forma Normal (3NF) é uma técnica de design que visa eliminar redundâncias de dados. Os dados são divididos em muitas entidades discretas, cada uma das quais se torna uma tabela no banco de dados relacional evitando anomalias nos dados, provendo uma arquitetura mais sólida para a atualização de dados, e reforçar uma longa integridade ao modelo do banco de dados. A indústria, por vezes, refere-se a modelos 3NF como modelos ER (BALLARD et al, 1998).

Este tipo de processo (normalização) é extremamente útil para o desempenho de processamento operacional porque uma transação de atualização ou inserção só precisa ser feita na tabela necessária do banco de dados. Porém, são muito complicados para consulta de dados em DW. Os usuários podem não compreender, navegar, ou lembrar modelos normalizados. Da mesma forma, SGBD não podem consultar um modelo normalizado de forma eficiente; a complexidade supera os otimizadores de bancos de dados, resultando em um desempenho desastroso (BALLARD et al, 1998).

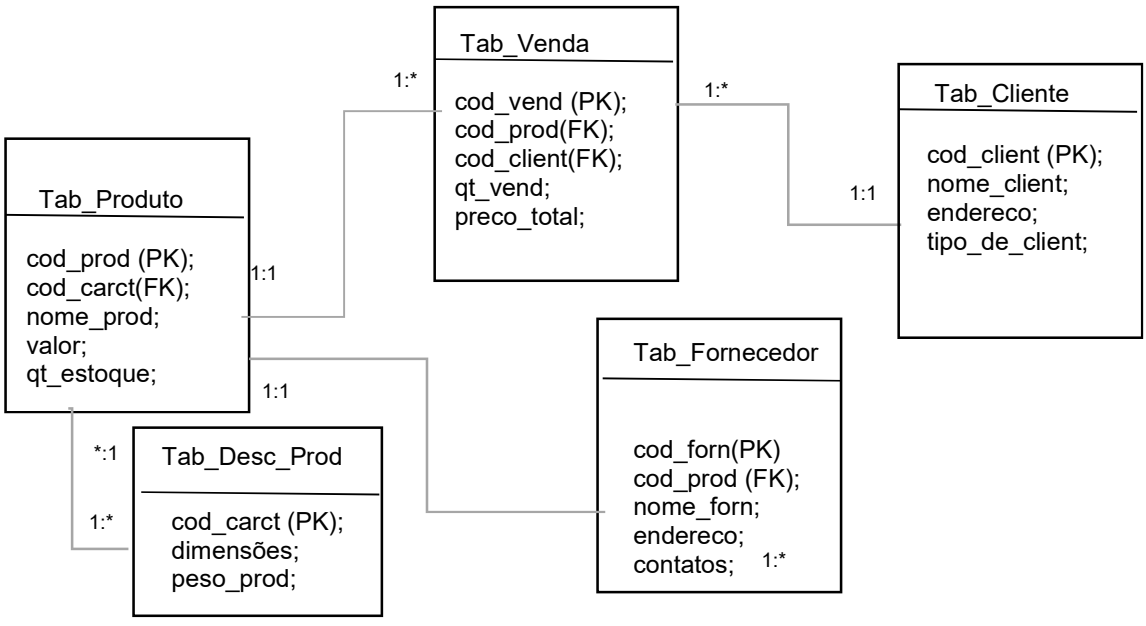

Figura 9. Exemplo de Modelagem de Banco de Dados Relacionais (Fonte: O autor).

Na figura 9, pode-se identificar 5 tabelas distintas: Tab_Produto, Tab_Venda, Tab_Cliente, Tab_Fornecedor e a Tab_Desc_Prod. A tabela Tab_Produto possui 3 relações diferentes com outras tabelas no diagrama ER. Sua relação com a Tab_Venda é uma relação de um-para-vários, pois a mesma foi construída para quebrar uma possível relação de vários para vários entre a Tab_Produto e a Tab_Cliente. Dentro da Tab_Venda encontra-se os seguintes atributos: cod_vend, cod_prod, cod_client, qt_vend e preco_total. Vale destacar que o atributo cod_vend é a chave primária (*Primary Key*) e é o atributo ao qual todos os registros a respeito da tabela Tab_Venda serão identificados. Os outros dois atributos cod_prod e cod_client são chaves estrangeiras (*Foreign Key*) de cada uma das duas relações em que a Tab_Venda tem com as Tab_Produto e Tab_Cliente respectivamente. Outro ponto importante a ser destacado é a Tab_Desc_Prod. Esta tabela nada mais é que as descrições dos produtos. Como referido anteriormente, a normalização visa a menor repetição de dados possível (redundância), por isso uma tabela agregada a Tab_Produto.

A modelagem relacional requer conhecimento técnico para ser aplicada e projetada, porém há uma economia de espaço gera por ela que evita gastos com aquisições de data center (BALLARD, 1998).

3.2 Modelagem Dimensional

A modelagem dimensional é um novo nome para uma técnica antiga que faz os bancos de dados terem uma estrutura mais simples e compreensível. Começando na década de 1970, organizações de TI, consultores, usuários finais e fornecedores foram atraídos a uma simples estrutura tridimensional para coincidir com a necessidade humana fundamental de manter a simplicidade na armazenagem dos dados, a partir dessa necessidade começou a ser desenvolvido a modelagem dimensional. A maioria dos usuários acha que é intuitivo pensar neste processo como um cubo de dados para a visualização dimensional dos dados. As bordas são marcadas por assuntos a serem analisados como, por exemplo, produto, mercado e tempo. Pode-se concluir que o fatiamento ao longo de cada uma dessas dimensões são pontos dentro do cubo onde as medidas para que a combinação de produto, mercado, e hora de um determinado registro sejam armazenados (KIMBALL, 2013).

A figura 10, representa um sistema multidimensional, com as variáveis tempo, mercado e produto. A funcionalidade do cubo se dá, pela flexibilidade de uma análise mais específica, por exemplo, se for analisado os produtos comprados em um determinado

mercado, pode-se restringir ainda mais a resposta para em qual tempo (meses) o produto teve uma alta ou baixa venda.

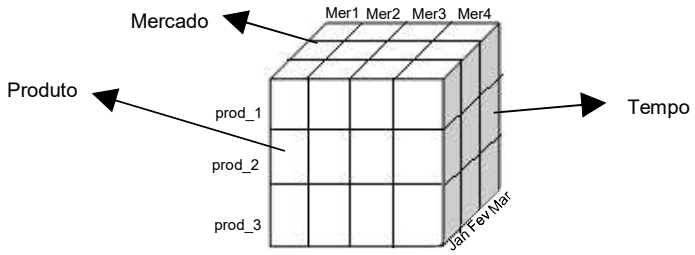

Figura 10. Exemplo de Visualização Dimensional (Fonte: O autor)

Segundo Kimball (2013), com o passar do tempo, o mercado concluiu que a modelagem dimensional é a técnica mais viável para a entrega de dados para usuários de DW. A capacidade de visualizar algo tão abstrato como um conjunto de dados de uma forma concreta e tangível é o segredo da compreensão. Um modelo de dados que começa simples tem uma possibilidade de permanecer simples até o final do processo, e um modelo que começa complicado com certeza é complicado até conclusão do projeto. Logo, modelos complicados serão executados de forma lenta e serão rejeitados pelos usuários.

Este tipo de modelagem é visualizada e conceituada na modelagem de dados como um grupo de medidas que são descritas por um aspecto em comum na organização. É especialmente utilizada para a rearranjar, sumarizar e visualizar o banco de dados, dando o suporte necessário para os dados serem analisados. Modelagem dimensional foca-se em dados numéricos, tais como valores, contadores, pesos, saldos e ocorrências. Possui como conceitos básicos os: fatores, dimensões e medidas (variáveis).

Tanto a terceira forma normalizada (3NF) como os modelos tridimensionais podem ser representados em diagramas ER, porque ambos consistem em tabelas relacionais. A principal diferença entre 3NF e modelos dimensionais é o grau da normalização. Ambos os tipos de modelo podem ser apresentados como Diagrama de Entidade Relação (ER). A modelagem dimensional contém as mesmas informações que um modelo normalizado, mas os pacotes dos dados estão em um formato compreensível pelo usuário, que possui um desempenho satisfatório na consulta e é resistência à mudança (KELLY, 1997).

.

3.2.1 Fatores, Dimensões e Medidas.

O fator tipicamente representa um item na organização, uma transação na organização ou um evento que possa ser usado na análise da organização ou no processo de negócios. No DW, fatores são implementados nas tabelas núcleos em quais todos os dados numéricos são armazenados. A dispersão de dados, ocorre tipicamente no menor nível de granularidade de uma tabela fator, devido à não disponibilidade de uma medida para a combinação das chaves da dimensão (KIMBALL, 1998). Ha três tipos de tabelas fatores*:* transacionais*, snapshots* e acumulados.

Uma tabela fator transacional é a visão mais básica e fundamental da organização. Este tipo de tabela representa um evento que ocorreu em um ponto instantâneo no tempo. Uma linha existe na tabela fator apenas se uma transação ocorreu com um cliente ou produto, por exemplo. Um cliente ou um produto está ligado a múltiplas linhas na tabela fator porque o cliente ou produto pode estar envolvido em uma ou mais transações. Os dados de transação periodicamente são estruturados em um modelo de *framework* dimensional. O nível mais baixo dos dados é o dado dimensional mais natural, suportando análises que não podem ser feitas em sumarizações de dados.

A tabela fator do tipo *snapshot* é descrita como o estado das coisas em uma particular instância de tempo e geralmente inclui fatores semi-aditivos e não-aditivos. *Snapshots* periódicos são necessários para se ver a execução acumulativa da organização em intervalos de tempo regulares. Ao contrário da tabela de fatores transacionais, onde as linhas de registro são carregadas para cada evento ocorrido, o *snapshot* periódico tira uma imagem no final do dia da atividade, da semana, ou mês, depois uma outra imagem no final do próximo e assim por diante (KIMBALL, 2013).

A tabela de fatores acumulados é usada para mostrar atividade de processo que são bem definidos do começo ao fim como, por exemplo, um processo de requisição. Uma requisição move-se através de específicos passos e é processada inteiramente. Com a execução próxima de ser cumprida, a linha de registro associada com a tabela fator é atualizada (KIMBALL, 2013).

Os números que quantificam os fatores são geralmente chamados de medidas. As medidas são os atributos numéricos que representam a performance ou comportamento da organização com relação as dimensões. Estes números também são chamados de variáveis e são determinados pela combinação de membros das dimensões alocados nos fatores, possuindo três classificações:

- Aditivos: medidas que podem ser adicionadas em qualquer dimensão
- Semi-aditivos: medidas que podem ser adicionadas em algumas dimensões
- Não-aditivos: Medidas que não podem ser adicionadas em nenhuma dimensão.

As dimensões são o conjunto de membros ou unidade com o mesmo tipo de aspecto, normalmente representados por uma linha principal. No modelo dimensional, todos os pontos de dados na tabela fator são associados com apenas um membro das outras múltiplas dimensões. A dimensão determina o conceito por trás dos fatores. Muitos processos de análise são usados para quantificar o impacto das dimensões sobre os fatores. Uma característica das tabelas dimensões é que os dados são relativamente estáveis.

3.2.2 Esquemas de Estrutura de dados

Os esquemas (modelo) estrela e flocos de neve são modelos dimensionais, a diferença entre eles está na estrutura físicas. O esquema flocos de neve suporta facilmente manutenções nas dimensões porque eles são mais normalizados. O esquema estrela é de fácil acesso direto e constantemente suporta consultas simples de modo mais eficientes. A decisão de modelar uma dimensão como estrela ou flocos de neve depende da própria natureza da dimensão, como por exemplo a frequência que haverá mudanças, quais os elementos serão mudados e, eventualmente as avaliações comparativas entre a facilidade de usar e a facilidade de manter. Para estruturas complexas é recomendável manter um esquema floco de neve, colocando níveis hierárquicos em tabelas separadas, referenciando a integridade entre os níveis da hierarquia. Serviços OLAP (*On-Line Analytical Processing*) executam o processo de leitura de uma dimensão floco de neve tão bem, ou melhor, que de uma dimensão estrela (SHERMAN, 2014).

O esquema estrela tem se tornado um termo comum pelo seu diagrama se parecer com uma estrela de fato. Esse esquema representa os dados como fatores que são ligados às dimensões. O fator de medida do empenho de processamento da organização ou algum aspecto da organização pode ser, por exemplo, as vendas, lucros ou inventario. Uma dimensão específica à base dos fatores como por exemplo, datas, localizações, produtos e clientes. Este esquema é uma abordagem muito usada em aplicações DW, que por sua vez toma para si aplicações disjuntas e funcionais, para integrá-las colocando os dados em um único banco de dados e armazenando os dados em um formato comum para propósitos de gerar relatórios. A simples estrutura do esquema estrela torna fácil escrever relatórios ad-hoc

que minere os dados e ganhe conhecimento dentro da organização. Contudo a simples estrutura não pode forçar restrições sobre os dados, que é o propósito de aplicações funcionais que são lidadas em operações do dia-a-dia da organização. Este esquema não está apenas limitado à DW, pode também ser usado para aplicações funcionais com muita leitura e pouca escrita no banco de dados.

O esquema estrela pode armazenar os eventos, mas não pode forçar as restrições, tais como a ordem do processamento. Esse esquema pode resolver questões a respeito de: status de cada aplicação (o último tipo de evento que tenha sido processado), a média de tempo para processamento de cada evento como o progresso de aplicações e os funcionários que executam as tarefas mais rapidamente.

Este modelo é uma estrutura básica para o modelo dimensional. Enquanto o modelo tradicional entidade relacionamento tem um estilo preciso e balanceado das entidades e relações complexas entre as entidades, o modelo dimensional é muito assimétrico. Ainda que a tabela fator no modelo dimensional seja ligada a todas as outras tabelas dimensões, haverá apenas uma simples linha de ligação conectando a tabela fator das tabelas dimensões.

Tipicamente tem uma tabela grande central (chamada de tabela fator) e um grupo menor de tabelas (chamados tabelas dimensões) alinhadas em um padrão radial ao redor da tabela fator. O diagrama da figura 11, demonstra 6 dimensões, mas poderia ter qualquer quantidade. A maioria das dimensões são obrigatórias, mas algumas podem ser opcionais também.

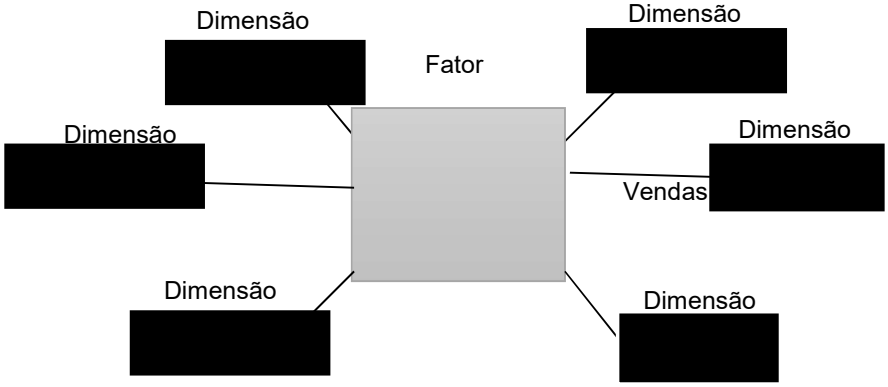

Figura 11.Exemplo de Esquema Estrela (Fonte: O autor).

O modelo flocos de neve é uma variação do esquema estrela, onde algumas tabelas dimensões são normalizadas. O resultado da forma dos esquemas tem uma aparência similar

a um floco de neve, por isso o nome. Os atributos redundantes são removidos dos planos e desnormalizados das tabelas dimensões e são alocadas nas tabelas secundárias normalizadas. A grande diferença entre o esquema estrela e o esquema floco de neve é que as tabelas dimensões no esquema flocos de neve podem ser armazenadas de forma normalizada, dessa forma reduzindo redundâncias. As tabelas normalizadas também economizam espaço de armazenamento. Uma vez que grandes tabelas dimensões podem se tornar gigantescas quando a estrutura dimensional e incluída nas colunas. Essa economia de espaço pode ser negligenciável quando comparado ao tamanho da tabela fator.

A decomposição da estrutura de flocos de neve visualiza a estrutura da hierarquia das dimensões. Este modelo é fácil para modeladores de dado entenderem e para designer de banco de dados usarem a análise de dimensão. Entretanto, o modelo flocos de neve se parece mais complexo e tende a fazer os usuários da organização se sentirem um pouco mais incomodados do que trabalhando com o simples modelo estrela (SHERMAN,2014).

Desenvolvedores podem também eleger o modelo flocos de neve como o melhor esquema porque há uma economia de espaço no armazenamento de dados. Considerando uma aplicação onde há uma enorme quantidade de dados em uma tabela subdimensão, pode ser facilmente esperado que salve um bom espaço na tabela dimensão, pois nessa os dados não serão salvos frequentemente, eles serão colocados apenas uma vez numa tabela de subdimensão evitando campus iguais e redundância de dados. Usando este esquema pode se obter uma verdadeira normalização no banco de dados, porém é um pequeno espaço se comparado com o tamanho geral do banco de dados, que geralmente vem de um largo volume de dados da tabela fator. Na Figura 12 é demonstrado um esquema flocos de neve com algumas subdimensões

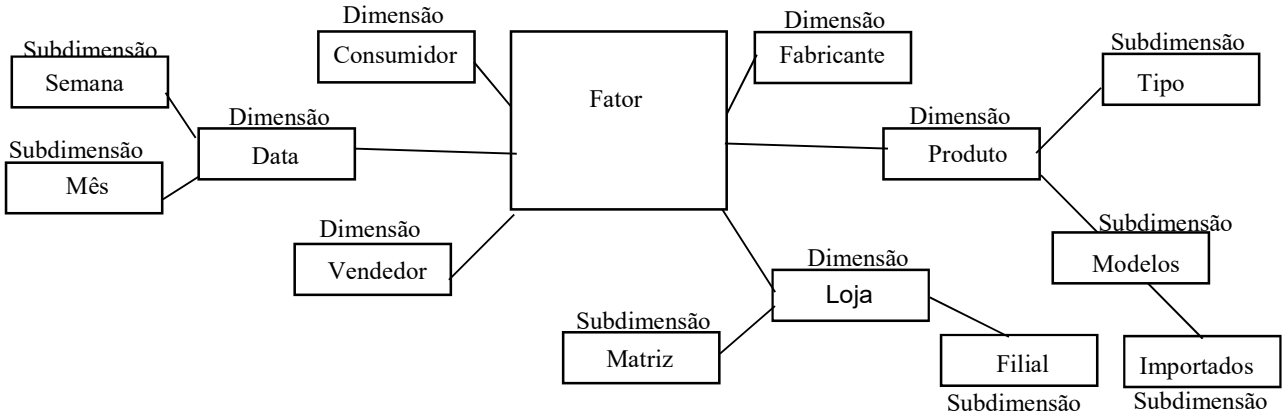

Figura 12.Exemplo de Esquema Floco de Neve. (Fonte: O autor).

3.3 Processamento OLAP e OLTP

On-Line Analytical Processing (OLAP) é um termo muito usado no mercado. OLAP é principalmente usado para analisar dados de negócios coletados de transações diárias, tais como os dados de vendas e dados de transações. O principal objetivo de um sistema OLAP é permitir que analistas possam construir uma imagem abstrata sobre os dados subjacentes, explorando-os a partir de diferentes perspectivas, em diferentes níveis de generalizações, e de forma interativa (HAYES, 2012).

Como um componente de sistemas de apoio à decisão, o OLAP interage com outros componentes, tais como DW e mineração de dados, para ajudar a analistas em tomada de decisões empresariais. OLAP e mineração de dados permitem aos analistas a descoberta de novos conhecimentos sobre os dados armazenados em um DW. Algoritmos de mineração de dados automaticamente produzem conhecimento em uma forma pré-definida, como regra de associação ou classificação. O OLAP não gera tais conhecimentos diretamente, ele depende que analistas humanos o observem e interpretem os resultados da consulta.

Por outro lado, OLAP é mais flexível do que a mineração de dados, no sentido de que os analistas podem obter todos os tipos de padrões e tendências, em vez de apenas o conhecimento de formas fixas. OLAP e mineração de dados também podem ser combinados para permitir a analistas obtenção de resultados de mineração de dados de diferente parte dos dados e em diferentes níveis de generalização. Em uma sessão de OLAP típica, o analista coloca consultas de agregação sobre dados subjacentes. O sistema OLAP geralmente pode retornar o resultado em questão de segundos, mesmo que a consulta envolva um grande número de registros. Com base nos resultados, os analistas podem decidir explorar até as camadas de maior granularidade para que eles possam observar padrões e tendências globais. Ao observar uma exceção a qualquer padrão estabelecido, os analistas podem detalhar diretamente aos dados mais refinados com mais detalhes para estudar os valores extremos (HAYES, 2012).

Online Transaction Processing (OLTP) é um tipo de sistema de informação, que prioriza o processamento de transações, lidando com dados operacionais. Estes tipos de sistemas são identificados pelo grande número de transações que suportam, tornando-os a melhor forma de abordar a aplicação online. As principais aplicações deste método são todos os tipos de sistemas transacionais, como bancos de dados, comercial, aplicações hospitalares e assim por diante (HAYES, 2012).

De uma forma simples, estes sistemas reúnem informações de entrada e as armazenam em um banco de dados, em grande escala. A maioria das aplicações de hoje são baseados nesta metodologia de interação, com implementações de sistemas centralizados ou descentralizados.

No nível de banco de dados, estes sistemas transacionais baseiam a sua operação em multi-acesso, rápidas e eficazes consultas para o banco de dados. As operações mais usadas são INSERT, UPDATE e DELETE, uma vez que estão modificando diretamente os dados, fornecendo novas informações sobre novas transações. Assim, nestes sistemas, os dados são atualizados com frequência, exigindo um suporte eficaz nas operações de escrita (HAYES, 2012).

Um sistema de transação bancária é um exemplo clássico. Há muitos usuários que executam operações em suas contas e o sistema deve garantir a integralidade das ações. Neste caso, existem várias transações simultâneas, sendo a coerência dos dados e operações eficientes o objetivo principal. Segue a Tabela 3 que resume as principais diferenças entre os dois sistemas (HAYES, 2012).

Características	**OLTP** **System Online Transaction Processing** **(Sistemas de Gerenciamento de Banco de Dados)**	**OLAP** **System Online Analytical Processing** **(Data Warehouse)**
Fonte de dados	Dados Operacionais, OLTP são a fonte original de dados.	Dados consolidados; Dados OLAP vem de vários bancos de dados OLTP.
Propósito dos dados	Controlar e executar tarefas fundamentais de negócios.	Ajudar no planejamento, solução de problema e suporte de decisão.
Dado	Revela um *snapshot* do processo continuo do negócio.	Visão Multidimensional de vários tipos de atividades de negócio.
Inserts e Updates	Curto e rápido (Inserts); Atualizações são iniciadas pelos usuários finais.	Periódicas longas execuções de trabalhos em batch para atualizar os dados.
Consultas	Relativamente padronizada e consultas simples, retornando apenas alguns registros.	Constantemente consultas complexas envolvendo agregações.
Velocidades de processamento	Tipicamente muito rápido.	Dependente da quantidade de dados envolvidos, dados *batch* e consultas complexas podem levar muitas horas, velocidade de consulta pode ser melhorada com a criação de índices.
Requisito de espaço	Pode ser relativamente pequeno se o dado histórico está arquivado.	Grande devido a existência de estruturas agregadas e dados históricos; requer mais índices que OLTP.
Desenho do Banco de Dados	Altamente normalizado com muitas tabelas.	Tipicamente desnormalizado com poucas tabelas; uso de esquemas estrela ou flocos de neve.
Restauração e Backup	Backup feito constantemente; Dados operacionais são críticos para executar os negócios, perda de dados pode ocasionar perda monetária significativa e responsabilidade legal.	Ao invés de backups regulares, alguns lugares podem considerar um simples recarregamento de dados OLTP como método de restauração.

Tabela 3. OLTP vs OLAP (fonte: www.rainmakerworks.com)

3.4. Carregando o Data Warehouse (Processo ETL)

Este processo consiste em obter os dados de um sistema externo para um DW e DM. Os dados são capturados pelos sistemas gerenciais e de controle (externos) transformados em um formato utilizável para o DW e finalmente carregado para o DW ou DM. Tanto o modelo como os dados podem ser afetados nesta etapa durante o processo de carregamento (BALLARD, 1998).

A extração é o primeiro passo do processo de obter dados para o DW. Extrair significa ler e entender a fonte de dados e copiar os dados necessários para o DW na área de trabalho para manipulações futuras. Uma vez que os dados são extraídos para a área de trabalho, há muitas transformações possíveis, tais como a limpeza dos dados (corrigir erros, resolver conflitos de domínio, lidar com elementos perdidos, ou padronizar os formatos) reunir dados de diversas fontes, eliminar dados duplicados e atribuir as chaves do DW. Essas transformações são todas procedidas depois para o carregamento dos dados para a área de apresentação do DW. A área de trabalho é dominada por atividades simples de triagem e sequência de processamento. Em muitos casos não é baseado em tecnologia relacional, ao invés disso pode consistir de um sistema de arquivos planos (BALLARD et al, 1998).

Há alguns casos aonde os dados chegam a área de trabalho no padrão relacional de terceira forma normal. Nessas situações, os gerentes dessa área de trabalho simplesmente podem ficar mais confortáveis executando a tarefa de limpeza e transformação usando um grupo de estruturas normalizadas.

O processo final é o de carregamento dos dados, geralmente adota-se a forma de apresentar tabelas com a qualidade necessária para o carregamento dos diversos volumes de dados de cada DM. O alvo do DM deve então indexar o dado recém-chegado para a execução de consultas. Quando cada DM é recém carregado, indexado, abastecido com as propriedades agregadas e a qualidade assegurada, a comunidade de usuários é notificada que o novo dado foi publicado.

3.4.1 Captura dos dados

A fonte de dados para o processo de captura inclui formatos de arquivos, sistemas de bancos de dados relacionais e não relacionais. Os dados podem ser capturados por vários

tipos de arquivos incluindo tabelas, copias de imagens, arquivos de bancos de dados, logs ou arquivos de mensagens (INMON, 2012). Alguns exemplos de captura são:

- Extração de fonte de dados que fornece um *snapshot* estático dessa fonte de dados ou em um especifico ponto no tempo. É suficiente para suportar um modelo de dados temporal. Extração de fonte de dados pode produzir extração de arquivos, tabelas ou imagens.
- Captura de logs permitem a captura de dados de sistemas gerenciadores de banco de dados. Tem um impacto mínimo no banco de dado ou no sistema operacional que está acessando o banco de dados. Esta técnica não requer um entendimento claro do formato de gravação dos logs e razoável nível de programação sofisticada para extrair somente os dados de interesse.

3.4.2 Transformações

O processo de transformação converte a fonte de dados capturada em um formato e estrutura adequado para carregar no DW. Os mapeamentos das características usadas para a transformação da fonte de dados são capturados e armazenados como metadados. Isto define qualquer mudança que seja necessária antes do carregamento dos dados no warehouse. Isso ajuda a resolver anomalias na fonte de dados e produz um alto nível de qualidade para os dados a serem carregados. A transformação do dado pode ocorrer no momento da gravação. As técnicas básicas incluem transformações estruturais, transformações de conteúdo e transformações funcionais (BALLARD, 1998).

A transformação estrutural transforma a estrutura da fonte de gravação do banco. Esta técnica é usada no nível de gravação, e ocorre pela seleção de um grupo de arquivos da fonte de dados, e um selecionado subgrupo da fonte de arquivos, mapeando para diferentes registros ou por algumas combinações de cada.

A transformação de conteúdo muda os valores nos arquivos. Esta técnica transforma os dados no nível de atributos. Transforma os valores dos conteúdos usando algoritmos ou o uso de tabelas de transformação de dados.

A transformação de função cria um novo valor de dado para os registros baseado em dados da fonte de registro. Esta técnica transforma os dados no nível de atributo. Estas transformações acontecem através da agregação ou enriquecimento de dados por cálculos de valores derivados tais como totais ou medias baseadas em atributos múltiplos em diferentes

registros. Enriquecimento combina dois ou mais valores de dados e cria um ou mais novos atributos de um simples ou múltipla fonte de registros que pode ser do mesmo ou diferente fonte (BALLARD et al, 1998).

3.4.3 Carregamento (Aplicação)

O processo de carregamento (aplicação) usa os arquivos ou tabelas criadas no processo de transformação e aplica-os relativamente para o DW ou DM.

Geralmente, as cargas das tabelas possuem dois tipos: total ou incremental. A total, como o próprio nome diz, trata-se da carga completa dos dados toda vez que há a execução de um novo processo de ETL. Nesse tipo de carga todos os dados da origem são extraídos e transformados, recarregando os dados antigos e incrementando com os novos. Já a incremental considera apenas os novos registros dos sistemas operacionais no ETL, inserindo-os ao repositório do DW (KIMBALL, 2013).

3.5 Slowly Change Dimension

Slowly Change Dimension (Dimensões de Mudança Lenta, SCD) são dimensões que mudam lentamente ao longo do tempo. Para aplicações em DW existe a necessidade de acompanhar as mudanças que alguns atributos de dimensão possam sofrer ao longo do tempo. A fim de relatar os dados históricos dessa forma há uma necessidade de implementar um dos tipos de SCD permitindo que o utilizador defina o valor adequado para a dimensão (KIMBALL, 2013). Na tabela 4 são apresentados os 8 tipos de SCD e suas definições.

Tipo de SCD	Ação que ocorre na Tabela Dimensão
Tipo 0	Não há mudança no atributo
Tipo 1	Sobrescreve um valor de atributo
Tipo 2	Adiciona uma nova linha nova no perfil com o dado novo
Tipo 3	Adiciona uma nova coluna para preservar o valor atual e o valor antigo
Tipo 4	Adiciona uma tabela mini dimensao contendo valores de mudança rápida
Tipo 5	Adiciona scd de tipo 4 e
Tipo 6	Adiciona tipo 1 sobrescrevendo os atributos para o tipo 2 (linha) e assim sobrescreve todas as linhas
Tipo 7	Adiciona do tipo 1 com os novos atributos e ainda possui uma visão limitada para valores das linhas e/ou atributo

Tabela 4. Tipos de SCD (Kimball, 2013)

4.MINERACAO DE DADOS E BIG DATA

Este capitulo introduz as definições básicas de mineração de dados e Big Data, além de apresentar as ferramentas utilizadas pelas tecnologias.

4.1.Mineracao de Dados

Segundo Han (2012), a mineração de dados pode ser vista como um resultado da evolução natural da tecnologia da informação. Um caminho evolutivo ocorreu no setor de dados com o desenvolvimento de algumas funcionalidades como a coleta de dados, a criação de banco de dados, o gerenciamento de dados (incluindo o armazenamento, recuperação de dados e transação de processos), dados de análise e compreensão (envolvendo armazenamento de dados e mineração de dados). Por exemplo, o desenvolvimento precoce da coleta de dados e mecanismos de criação serviram como um pré-requisito para o desenvolvimento posterior de mecanismos eficazes de armazenagem de dados, recuperação de consulta e processamento de transações. Com numerosos sistemas de banco de dados oferecendo consulta e processamento de transações, como prática comum, a análise de dados e compreensão dos mesmos naturalmente se torna o próximo alvo.

Simplificando, pode ser afirmado que a mineração de dados se refere à extração ou mineração de conhecimento de grande número de dados (WITTEN, 2011). Muitas pessoas tratam de mineração de dados como um sinônimo para Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases – KDD). A etapa de mineração de dados pode interagir com o usuário ou uma base de conhecimento. Os padrões interessantes são apresentados para o usuário, e podem ser armazenados como uma nova base de conhecimento. De acordo com este ponto de vista, a mineração de dados é apenas um passo

de todo o processo, ainda que seja um fator essencial, uma vez que revela padrões escondidos para uma avaliação. A seguir são descritas algumas das tecnologias utilizadas pela mineração de dados.

4.1.1 Estatísticas

Estatística estuda a coleta, análise, interpretação ou explicação, e apresentação de dados. A mineração de dados tem uma ligação intrínseca com estatísticas. Um modelo estatístico é um conjunto de funções matemáticas que descrevem o comportamento dos objetos em uma classe de destino em termos de variáveis aleatórias e sua probabilidade associada as distribuições. Os modelos estatísticos são amplamente utilizados para modelar dados e classes de dados (WITTEN, 2011).

Métodos estatísticos podem ser usados para resumir ou descrever um conjunto de dados. A estatística é útil para a mineração com vários padrões que acontecem nos dados, bem como para compreender os mecanismos subjacentes à geração de padrões e os motivos que afetam os padrões. A estatística inferencial (ou estatística preditiva) modela os dados de uma forma que respondam por aleatoriedade e incerteza nas observações e é usada para fazer inferências sobre o processo ou um pequeno grupo da população em investigação (HAN, 2012).

Métodos estatísticos também podem ser utilizado para verificar os resultados de mineração de dados (HAN, 2012). Por exemplo, depois que um modelo de classificação ou previsão é extraído, o modelo deve ser verificado por meio de um teste hipotético de estatística. Um teste hipotético de estatística (algumas vezes mencionada como confirmação de análise de dados) toma decisões estatística usando dados experimentais. Um resultado é chamado estatisticamente significativo se for improvável que tenha ocorrido por acaso. Se o modelo de classificação ou previsão é válido, então a estatística descritiva do modelo aumenta a solidez do modelo.

4.1.2 Aprendizados de Máquina

A aprendizagem de máquina é uma disciplina em crescimento. Investiga como os computadores podem aprender (ou melhorar o seu desempenho) com base nos dados. A principal área de pesquisa é para programas de computador aprenderem automaticamente a

reconhecer padrões complexos e tomar decisões inteligentes com base nos dados. Por exemplo, um típico problema de aprendizagem de máquina é programar um computador para que ele possa automaticamente reconhecer códigos postais manuscritos no correio depois de aprender a partir de um conjunto de exemplos (HAN, 2012).

4.1.3 Bancos de Dados

Sistemas de banco de dados de investigação concentram-se na criação, manutenção e utilização de bases de dados para as organizações e usuários finais. Particularmente, os investigadores estabeleceram sistemas de base de dados em modelos altamente reconhecidos e com linguagens de consulta, processamento de consultas, métodos de otimização, armazenamento de dados, indexação e métodos de acesso. Sistemas de banco de dados são muitas vezes bem conhecidos por sua alta escalabilidade no processamento de grande volume de dados estruturados (WITTEN, 2011).

Muitas tarefas de mineração de dados precisam lidar com grandes conjuntos de dados e em alguns casos com rápidas transmissões em tempo real. Portanto, a mineração de dados pode fazer bom uso de tecnologias escaláveis para atingir alta eficiência e escalabilidade em grandes grupos. As tarefas podem ser usadas para estender a capacidade dos sistemas de banco de dados existentes para satisfazer os requisitos avançado e sofisticados dos usuários na análise de dados.

Sistemas de banco de dados construídos têm capacidades sistemáticas de análise de dados, utilizando instalações de mineração de armazenamento de dados. A integração de dados no DW provenientes de múltiplas fontes e vários prazos, consolida dados no espaço multidimensional para parcialmente materializar e transformar os dados em cubos. O modelo de cubo de dados não só facilita sistemas OLAP em bancos de dados multidimensionais, mas também promove mineração de dados multidimensional (HAN, 2012).

A relação entre a mineração de dados e DW é que o ultimo torna os dados, armazenados mais flexíveis para serem minerados. Uma consulta de mineração de dados executada em um conjunto de dados com um tamanho de terabytes, espalhados por vários bancos de dados em diferentes redes físicas, as consultas se tornam muito mais complexas e lentas para o negócio (HAN, 2012).

4.2. Big Data

Segundo Russom (2011), muitas organizações estão a recolhendo, armazenando, e analisando enormes quantidades de dados. Estas informações são comumente referidas como "Big Data" por causa do seu volume, a velocidade com que é processada, e a variedade de formatos que armazena. Big Data é uma nova criação para a geração de sistemas de gerenciamento de suporte às tomadas de decisão.

As empresas estão reconhecendo o valor potencial dos dados e estão colocando as tecnologias, pessoas e processos em andamento para capitalizar essas oportunidades. A principal maneira para derivar o valor de Big Data, é a utilização de análises. Para apenas coletar e armazenar os dados, não se teria muito valor, pois se trataria de apenas uma infraestrutura de dados qualquer. Os dados devem ser analisados e os resultados utilizados para tomadas de decisão e processos organizacionais, a fim de gerar algum retorno (HURWITZ, 2013).

DWs são construídos em sua maioria usando a tecnologia relacional principalmente para fontes operacionais. Big Data agrega relativamente novos tipos de fontes de dados, como mídias sociais, arquivos públicos, conteúdo disponível no domínio público através de agências ou assinaturas, documentos e e-mails, incluindo ambos os textos estruturados e não estruturados, dispositivos digitais e sensores, incluindo localização com base nos smartphones, tempo e dados de telemetria (HURWITZ, 2013).

O mercado não está acostumado com a coleta de informações a partir destas fontes, que também não estão acostumadas a lidar com grandes volumes de dados não estruturados. Portanto, grande parte da informação disponível para as empresas não é capturada ou armazenada para análise de longo prazo, e as oportunidades para ganhar a informações de valores são perdidas. Muitas empresas não mantêm volumes de dados e, por não percebem qualquer necessidade de mantê-los armazenados. Grandes empresas que querem realmente se beneficiar com Big Data também devem integrar esses novos tipos de informações com dados corporativos tradicionais, e encaixá-los de forma que haja uma hegemonia ao recolher em seus processos de negócios e operações existentes. Há várias abordagens para coleta, armazenamento, processamento e análise em Big Data e o seu foco principal é na análise de dados não estruturados (RUSSOM, 2011).

Dados não estruturados não possuem um modelo de dados pré-definidos, que consequentemente não se encaixam bem em tabelas relacionais. Dados não estruturados

possuem o crescimento mais rápido dos dados (HURWITZ, 2013). Um exemplo poderia ser os sensores, vídeo, documentos, arquivos, e arquivos de dados de e-mail de registro. Existem várias técnicas para resolver este problema de espaço de análise de dados não estruturados. As técnicas compartilham características comuns de escalabilidade, elasticidade e alta disponibilidade.

O MapReduce, em conjunto com o Hadoop Distributed File System (HDFS) e banco de dados HBase fazem parte do projeto Apache Hadoop que é uma abordagem moderna para analisar dados não estruturados. A seção 4.2.3 trabalha melhor a definição do que é MapReduce. Os clusters do Hadoop são um meio eficaz de processamento de grandes volumes de dados, e pode ser melhorado com uma abordagem de arquitetura bem-feita. Como as empresas adotaram o Hadoop para análise de dados não estruturados, uma consideração chave é integrar e interagir com DW e sistemas de banco de dados relacionais (HURWITZ, 2013).

Os governos e as empresas são capazes de integrar os dados pessoais de várias fontes e aprender muito sobre determinados aspectos da população. Tais como, atividades pessoais, lugares visitados, relacionamentos pessoais e as preferencias das pessoas. Essas informações podem ser extraídas muitas vezes das mídias sociais. Isso pode demonstrar uma pesquisa para possíveis melhoraras de serviço nas implementações (e algumas vezes operações de lucros em determinadas empresas), muitas questões e preocupações a respeito da privacidade é levantado (HURWITZ, 2013). Algumas empresas, como Facebook e o Google possuem poucas restrições legais sobre o que as elas podem fazer com os dados retraídos.

Big data e análises estão interligados e muitas técnicas analíticas como, por exemplo a análise de regressão, simulação e aprendizado de máquina, já estão disponíveis há muitos anos no mercado.

Segundo Hurwitz (2013) existem muitas fontes para Big Data, como por exemplo:

- Cada clique de um mouse em um site pode ser capturado em arquivos de log web e podem ser analisados, a fim de melhor compreender comportamentos de compra e influenciar os compradores recomendando produtos dinamicamente.
- Fontes de mídia social como Facebook e Twitter geram enormes quantidades de comentários. Estes dados podem ser capturados, analisados e compreendidos, como o pensamento das pessoas sobre a introdução de novos produtos.
- Medidores de fluxo contínuo de dados sobre o consumo de eletricidade, água ou gás, que pode ser compartilhada com os clientes e serem

combinados planos de preços para motivar os clientes a mudar a rotina de consumo de energia em horários de pico (por exemplo lavar roupa).

- Identificação de rádio frequência (RFID) pode ser colocado em cada pedaço de produto, a fim de avaliar a condição e localização de cada item.
- Dados geo-espaciais (por exemplo, GPS), criados por telefones celulares, que podem ser utilizados por aplicações, como Four Square para ajudar a conhecer os locais vizinhos e para receber ofertas de lojas e restaurantes nas proximidades.
- Dados de imagem, voz e áudio podem ser analisados para aplicações tais como sistemas de reconhecimento facial em sistemas de segurança.

4.2.1. Hadoop

Hadoop é um framework de programação livre, baseado em Java que suporta o processamento de grandes conjuntos de dados em um ambiente de computação distribuída e oferece um armazenamento paralelo. Hadoop é um tipo de sistema de MAD (Magnetismo, Agilidade e Profundidade), o que significa que é capaz de atrair todas as fontes de dados (M vem de Magnetismo), é capaz de adaptar seus motores às evoluções que podem ocorrer nas fontes de dados grandes (A vem de agilidade) e é capaz de suportar análises aprofundadas sobre fontes de Big Data, muito além das possibilidades de ferramentas de análise baseadas em SQL (D vem de Depth que seria Profundidade) (HURWITZ, 2013).

A relação entre o Big Data e Data Warehouse é tratada na seção 4.3 deste trabalho.

4.2.2 Sistemas de arquivos distribuídos Hadoop

O Sistema de Arquivos Distribuídos Hadoop (Hadoop Distributed File System - HDFS) é um sistema de arquivos distribuído escalável que permite o acesso de alto rendimento para os dados do aplicativo e é escrito na linguagem de programação Java (HURWITZ, 2013).

Na figura 13, demonstra-se o modo que um cluster HDFS opera em um padrão de mestre-escravo (master-slave), que consiste de um nó com nome de mestre e qualquer número de nós de dados com o nome de escravo. O nó mestre é responsável pela gestão da árvore de sistema de arquivos, para todos os arquivos de metadados e diretórios armazenados

na árvore, e os locais de todos os blocos armazenados nos nós de dados. Os nós de dados são responsáveis por armazenar e recuperar blocos quando um o nome do nó ou o cliente é solicitado.

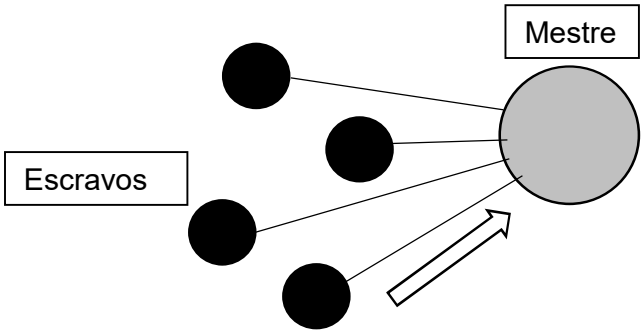

Figura 13. Esquema de nós Hadoop (Fonte: O autor).

4.2.3 MapReduce

MapReduce é um modelo de programação em cima do HDFS para o processamento e geração de grandes conjuntos de dados que foi desenvolvido como uma abstração do mapa e para reduzir presenças primitivas em muitas linguagens funcionais. A captação de paralelização, tolerância a falhas, distribuição de dados e balanceamento de carga permitem que os usuários possam paralelizar grandes cálculos facilmente (HURWITZ, 2013).

O mapa e o modelo de redução funcionam bem para a análise de Big Data, porque são inerentemente paralelos e podem facilmente lidar com conjuntos de dados que estão sendo executados entre várias máquinas. Cada programa MapReduce é executado em duas fases principais: a fase de mapa seguida da fase de redução. O programador simplesmente define as funções para cada fase e o Hadoop lida com a agregação de dados, classificação e passagem de mensagens entre os nós. Pode haver múltiplos mapas e fases de redução em um único programa de análise de dados com possíveis dependências entre eles (HURWITZ, 2013).

A entrada para a fase de mapeamento é feita com dados brutos. Uma função de mapeamento deve preparar os dados para entrada no redutor, mapeando a chave para o valor de cada "linha" de entrada. As saídas dos pares de valor das chaves são dadas por esta função e são classificadas e agrupadas por chaves antes de serem enviadas para a fase de redução.

A entrada para a fase de redução é a saída da fase do mapa, onde existe uma lista dos valores com suas teclas correspondentes. A função de redução deve percorrer a lista e executar alguma operação nos dados antes de emitir o resultado final.

4.3 A Integração De Big Data Com Data Warehouse

Segundo HURWITZ (2013), a maior parte das organizações, incluindo milhões de processamento de bytes de dados diários tem o seguinte requisitos:

- Rápido carregamento de dados
- Rápido processamento de consulta
- A utilização do armazenamento altamente eficiente
- Adaptabilidade forte para os padrões de carga de trabalho altamente dinâmicas

Hadoop pode ser razoavelmente considerado como a evolução de sistemas de DW da próxima geração, com particularidades que dizem respeito à fase de ETL de tais sistemas. O núcleo do Hadoop é um modelo de programação com a estrutura computacional associada e inspirada para mapear e reduzir as linguagens primitivas funcionais (DASAND, 2014). Segundo Hurwitz, (2013) alguns dos postos-chaves da tecnologia baseada em Hadoop são:

- É um armazenamento total, sem qualquer referencial de integridade, isto ajuda na obtenção de um desempenho rápido
- O particionamento é promissor, mas as "colunas" que os dados são particionados não existem mais nos dados mais definidos, são colocadas em parte da estrutura de árvore de diretórios.
- O particionamento divide realmente o arquivo para separar diretórios físicos (às vezes separar máquinas físicas)
- Hadoop é um sistema de gerenciamento do tipo que carrega e já disponibiliza o arquivo (o que significa que os arquivos brutos podem ser copiados para a plataforma)
- Não existe o processo "ETL" para a obtenção de arquivos em Hadoop, eles são copiados e em seguida as regras de transformação são escritas no código.
- Como não existe a abordagem do processo de ETL a transformação é feita depois que os arquivos são carregados.
- Hadoop não é uma ferramenta (ETL), é apenas uma plataforma que suporta a execução de ETL.

Os dados do Big Data, a maioria de formato não-estruturados, são carregados como arquivo para HDFS. Os dados de entrada são implementados e escritos uma vez no sistema. Em seguida, ele é processado por *MapReduce* em duas fases - fase de mapeamento e a fase de redução. Os resultados do processamento são escritos para HDFS. Existem dois tipos de nós HDFS: o nó que armazena os blocos de dados dos arquivos e nó que consistem de metadados (DASAND, 2014). A seguir os passos do *MapReduce* durante o processamento de um trabalho.

- Etapa de entrada: Carrega os dados no HDFS dividindo os dados em blocos e realiza a distribuição para nós de dados do grupo. Os blocos são replicados disponibilizados em caso de falhas. O nó central mantém o controle de blocos e os nós de dados.
- Primeiro passo: enviar o trabalho de *MapReduce* e seus detalhes para o *Job Tracker*.
- Segundo passo: O *Job Tracker* interage com *TaskTracker* em cada nó de dados para agendar tarefas MapReduce.
- Terceiro passo: Processa somente os blocos de dados (*Mapper*) e gera uma lista de pares de valores-chave.
- Quarto passo: *Mapper* classifica a lista de pares de valores-chave.
- Quinto passo: Transfere a saída mapeado para os redutores de forma ordenada.
- Sexto passo: Redutores mesclam a lista de pares de valores-chave para gerar o resultado final.
- Último passo: os resultados são armazenados no HDFS

Os resultados são então recuperados e escritos no HDFS pelo usuário em uma linguagem de programação Java. Em seguida os dados são carregados na tabela dimensão no DW por um mecanismo habitual de carregamento de dados. Como os dados não possuem um esquema fixo de entrada em DW, para lidar com frequentes alterações de esquema na enorme tabela fator, o controle de versão é feito com base em alterações de esquema na tabela fator. Por exemplo, cada linha na tabela fator contém uma identificação para uma versão do esquema utilizado. Portanto podemos saber quais colunas estão disponíveis para essa versão específica, tornando assim as atividades de adicionar ou remover mais praticas (DASAND, 2014).

No diagrama da Figura 14 é apresentado uma possível abordagem do Hadoop e DW trabalhando paralelamente, onde o Hadoop recolhe todos os dados não estruturados e armazena os mesmos, para que depois o DW efetue o processo de ETL e carregue todas as informações que estavam contidas no Hadoop para o sistema DW.

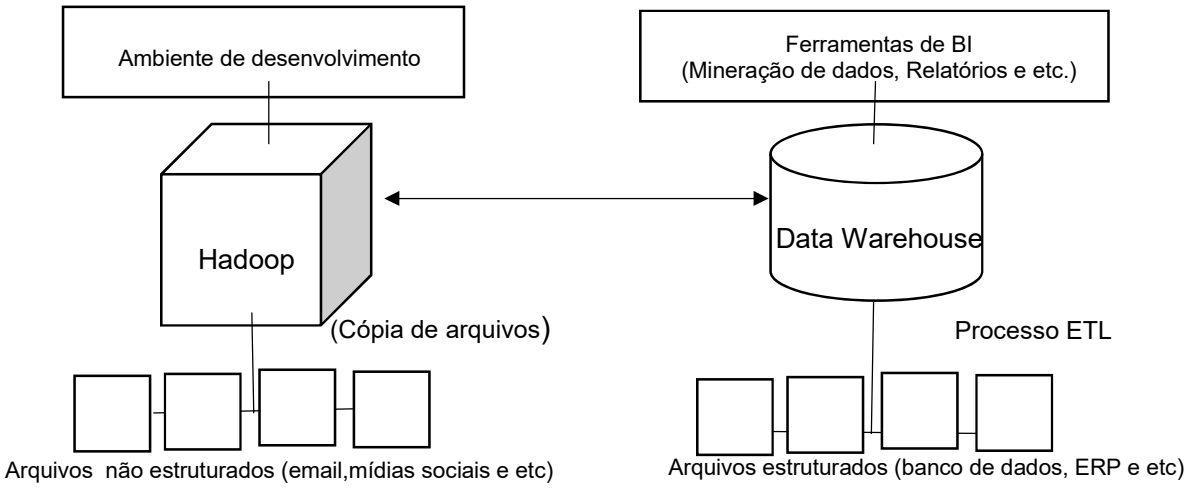

*Figura 14 Exemplo de um Data Warehouse integrado ao Hadoop (*DASAND, 2014*).*

Alguns benefícios desta abordagem incluem:

- Carregamento rápido por meio de cópia de arquivo
- Aplicação da leitura de transformações como se não tivesse referencial de integridade, facilitando a agregação de dados
- Compressão automatizada, em algumas implementações do Hadoop a compactação de arquivos é construída dentro, reduzindo a quantidade de armazenamento necessário para acomodar os dados
- Sem esquemas (até certo ponto) de armazenamento, ainda deve-se definir as colunas e, em alguns casos, os tipos de dados base do arquivo para que o código seja "mapeado" para os elementos. Por outro lado, existem alguns tipos de arquivos que apenas trabalham nativamente (já mapeados), como XML. E ainda outros documentos (como arquivos de texto) que simplesmente trabalham com base em pesquisa interna.
- Sem SQL - Não há limitações que necessite de um SQL padrão.
- Não é necessária para normalizar os conjuntos de dados, de modo ao evitar mesclar os conjuntos de nós.

5. METODOLOGIA A SER UTILIZADA

O objetivo deste capítulo é apresentar a metodologia desenvolvida no estudo de caso apresentado no capitulo 6. Para o desenvolvimento do sistema, dois aspectos serão fundamentais, o primeiro aspecto é a identificação das variáveis a serem utilizadas e suas classificações. Já o segundo aspecto é o processo de atualizações das dimensões do Data Warehouse.

Com isto a metodologia proposta nesse capítulo é baseada na metodologia de Herdem (2002) que foca apenas nos tópicos de modelagem conceitual, projeto prático e projeto físico como pode ser observado na Figura 15.

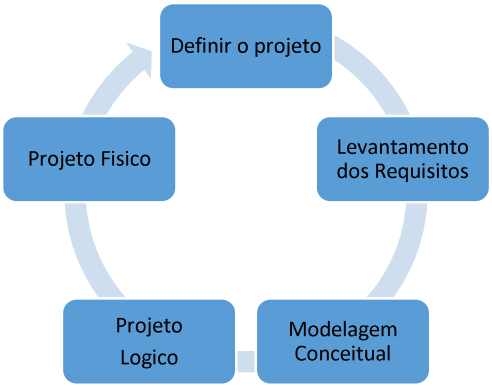

Figura 15. Ciclo da Metodologia baseada em Herdem (Fonte: O autor).

Em sua metodologia, Herdem (2002) utiliza a mesma abordagem de Kimbal (1998), a abordagem dimensional, a qual, ambos acreditam ser a mais eficaz em um projeto de DW. Por este tipo de abordagem, o usuário alcança um melhor entendimento do domínio da realidade a ser modelada.

As etapas de modelagem explicadas a seguir são: definição do projeto, levantamentos de requisitos, modelagem conceitual, projeto logico e projeto físico.

5.1 Definições do projeto

Como qualquer projeto, deve existir um componente que será responsável pela gerência do mesmo. A responsabilidade desse componente é estabelecer planos gerais do projeto e deve ser reconhecido por todos os membros do trabalho a fim de estabelecer todos os parâmetros e atestar que as necessidades do usuário serão levadas em conta e analisadas quando o projeto iniciar.

Os escopos já são pré-definidos nessa etapa para fins de levantamentos dos recursos necessários, para construção de regras e para que as futuras etapas do projeto estejam embasadas.

5.2 Levantamentos dos Requisitos

Com o escopo definido, os recursos necessários tanto para o desenvolvimento do projeto, como para manutenção devem ser definidos. Existem inúmeras ferramentas para desenvolver sistemas de DW, algumas soluções de banco de dados são de código aberto e não é necessário comprar uma licença para usá-la.

Nesta etapa do projeto deve-se tentar economizar o máximo de recursos financeiros (não gastando com licenças de softwares muito caros), pois recursos serão necessários para a implementação de clusters para o armazenamento dos dados históricos.

5.3 Modelagem Conceitual

Nessa etapa a empresa é analisada e é levantadas as entidades e os relacionamentos do modelo dimensional, essas relações entre os dados devem ser consideradas para fornecer um certo modelo inicial (como se fosse um protótipo). Posteriormente, o modelo será alterado e estendidos, seguindo o crescimento e a mudanças dos requisitos.

Outro aspecto que deve ser considerado é o envolvimento de toda a empresa diante do DW, não tendo só envolvimento apenas dos técnicos de TI.

5.4 Projeto Lógico

Com o escopo definido, os requisitos especificados e a modelagem discutida, o próximo passo é transforma-lo em um modelo lógico que será implementado no sistema de banco de dados através da modelagem dimensional.

Nesta etapa define-se o esquema a ser utilizado (esquema estrela, esquema flocos de neve ou algum esquema derivado dos dois já existentes). Este esquema permite que os requisitos sejam classificados em dois processos: o que está sendo analisado e o critério da avaliação para o que está sendo analisado.

5.5 Projeto Físico

A implementação final do DW na maioria dos projetos fica sob responsabilidade do Administrador de Banco de Dados (DBA). Os principais aspectos a serem considerados no projeto nesta etapa são:

- Indexação,
- Materialização das visões,
- Particionamento,
- Nível de redundância dos dados,
- Sintonia dos parâmetros do banco de dados,
- Processo de atualizações.

Já foi destacado a diferente sintonia do DW e sua carga no sistema, com a de um sistema transacional, pois ambientes de DW são construídos para vários acessos simultâneos no menor tempo possível sem se preocupar com a redundância dos dados. Tais parâmetros, se forem configurados e ajustados de forma correta sobre o modelo, pode ser alcançada uma melhora de 20% a 25% de desempenho (HAYES, 2002).

Ao final destas etapas o DW estará construído e disponível para os usuários realizarem suas consultas e relatórios.

6. ESTUDO DE CASO: IMPLEMENTANDO UM AMBIEMTE WAREHOUSE PARA PESQUISAS DO IBGE.

Seguindo a metodologia detalhada no capítulo 5, este capítulo está dividido em 6 seções, das quais 5 tratam das etapas para implementação de DW e a última da análise dos dados.

6.1 Ambiente e Estabelecimento do Pré-projeto

A primeira seção define o ponto de partida do projeto, provendo a responsabilidade de cada etapa, e os elementos principais para iniciar o projeto.

6.1.1 Ponto de Partida

Os dados utilizados nessa etapa do trabalho foram retirados diretamente de uma planilha disponibilizada no site do IBGE e são de responsabilidade de seus autores. Será utilizada a "Pesquisas de Informações Municipais", com os dados mais recentes de 2014, como fonte de dados desde projeto.

Essa pesquisa, segundo o próprio IBGE, é efetuada periodicamente, através de um levantamento pormenorizado de informações sobre a estrutura, a dinâmica e o funcionamento das instituições públicas municipais, em especial prefeituras compreendendo, também, diferentes políticas e setores que envolvem o governo municipal (IBGE, 2014).

O acesso a planilha eletrônica pode ser feito através do site do IBGE (www.ibge.gov.br) acessando o menu banco de dados e a opção para download dos arquivos desta pesquisa.

O arquivo utilizado é um arquivo em formato xls contendo 11 planilhas (Dicionário, Recursos Humanos, Comunicação, Educação, Saúde, Direitos Humanos 1, Direitos Humanos 2, Segurança Pública, Segurança Alimentar, Vigilância Sanitária, Variáveis Externas) o total de colunas é de 1039 (porém 11 colunas são repetidas, pois referenciam cada linha da planilha) e o total de linhas é de 5571 (contando com o cabeçalho). Esta planilha será a fonte de dados para a construção do DW deste estudo de caso.

6.1.2 O Foco do Projeto

Com a crescente demanda de servidores municipais, torna-se um trabalho árduo manter uma organização quantitativa e a distribuição desses funcionários pelos municípios dos 26 estados e o Distrito Federal. Portanto, neste trabalho, é criado um sistema de DW para coletar a quantidade de funcionários municipais nos setores de educação, saúde, administração da prefeitura e para outros serviços como vigilância sanitária.

Um exemplo é manter cada secretaria dos departamentos dos municípios, traçando o perfil de cada gestor para futuras analises proporcionais de categorizações de níveis escolares.

6.1.3 Responsabilidades de Gerência do DW

As responsabilidades são dividias em dois níveis, o primeiro nível do administrador do DW e de segundo nível do usuário do sistema.

Responsabilidades de primeiro nível, tratam as camadas físicas do DW, como o servidor que é armazenado os dados, os sistemas operacionais, como também o SGBD e todos os demais recursos computacionais que serão necessários para o desenvolvimento e a manutenção desse sistema.

Os usuários (responsabilidades de nível dois), são encarregados de gerar os relatórios e efetuar as análises dos dados através dos recursos computacionais disponibilizados. Além de ser de responsabilidade do usuário manusear o banco, também é de responsabilidade destes usuários estabelecer um canal de comunicação entre o setor de TI para que ambos determinem horários para as atualizações periódicas (normalmente é o ETL automatizado).

6.2 Levantamento de Pré-requisitos

Para a construção de um sistema de DW existem três componentes essenciais, o primeiro é um banco de dados instalado, o segundo um software ETL para fazer a limpeza, organização e compilação dos dados até o servidor de banco de dados e o terceiro, são os dados, uma vez que já é de conhecimento a necessidade de um banco ou fonte de dados para a criação do DW.

6.2.1 Softwares Utilizados

Nesta seção apresenta-se os três softwares utilizados na exportação dos dados, armazenamento e análise dos dados.

6.2.1.1 MySQL Workbench 6.3

MySQL Workbench é uma ferramenta visual e unificada voltada para arquitetos de banco de dados, desenvolvedores e DBAs. Fornece modelagem de dados, desenvolvimento SQL e ferramentas de administração abrangentes para configuração do servidor, administração de usuários, backup e muito mais. Estando disponível para Windows, Linux e Mac OS X.

Existe uma versão comunitária de código aberto e a comercial, nesta é possível ter acesso aos bancos salvos na nuvem. MySQL Workbench oferece ferramentas visuais para criar, executar e otimizar consultas SQL. O Editor de SQL fornece destaque de cores para diferentes sintaxes, digitação automática (*auto complete*), reutilização de trechos SQL e histórico de execução de SQL. O Painel de Conexões de banco de dados permite a administração as conexões de banco de dados padrão, incluindo MySQL *Fabric*. O Navegador de Objetos oferece acesso instantâneo aos esquemas de bancos de dados e objetos

MySQL Workbench fornece interface visual para administrar ambientes MySQL e ganhar uma melhor visibilidade em bases de dados. Os desenvolvedores e DBAs podem usar as ferramentas visuais para a configuração de servidores, administração de usuários,

realizando backup e recuperação, inspecionar dados de auditoria, e visualização de saúde banco de dados.

Mais informações podem ser retiradas no site do desenvolvedor **(**MYSQL, 2016).

6.2.1.2.Pentaho

Pentaho Data Integration (conhecido também como Kettle ou PDI) é uma suíte de softwares com ferramentas responsáveis para o auxílio de tarefas de extração, transformação e carregamento, mais conhecido como processos ETL.

Pode ser utilizado na migração de bases de dados entre aplicações ou outras bases de dados. Tarefas como essa e outras atividades se tornam mais simples com a interface gráfica do PDI, que conta com um recurso de clicar e arrastar para o design do ambiente, além de possuir capacidades de ETL poderosas.

O Pentaho também possui as versões empresariais e de código aberto, e todo seu material como manuais e o próprio programa podem ser conferidos no site do desenvolvedor (PENTAHO, 2016).

6.2.1.3 Tableau

Tableau Desktop é baseado em uma tecnologia desenvolvida na Universidade de Stanford que permite clicar e arrastar dados para serem analisados, se conectar facilmente a bancos de dados do DW e produzir diversos painéis gráficos para uma análise mais crítica dos fatos. A tecnologia é baseada num sistema flexível que do suporte ao raciocínio das pessoas para desenvolver as soluções dos problemas visualmente. Muda a forma continua entre as visões que seguem o fluxo natural da forma automática de pensar possibilitando que usuários não somente do departamento de informática possam analisar e extrair informações do banco de dados criando ricas visualizações. É uma alternativa de rápida resposta para o BI tradicional, uma vez que não se exige nenhum tipo (TABLEAU, 2016).

6.3 Modelagem dos dados

Analisando os dados da planilha do IBGE, pode se constar que o foco principal são os recursos humanos de cada município. Possuem a descrição a respeito da quantidade de

pessoas envolvidas na administração pública, de hospitais, escolas e prefeituras. Tal cenário viabiliza uma modelagem do DW no que tange a administração para a modelagem do DW é a respeito da administração desses funcionários públicos nos setores do município, como também assuntos complementares como o perfil do órgão gestor do departamento. Um exemplo dessa modelagem conceitual pode ser analisado na Figura 16.

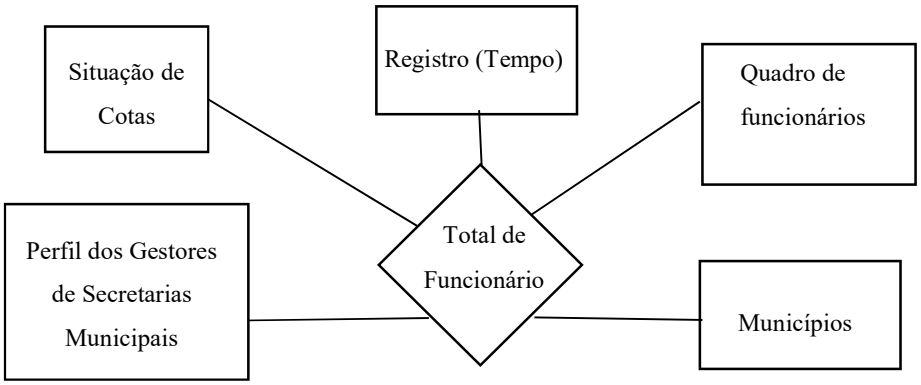

Figura 16. Modelo Conceitual do Sistema (Fonte: O autor)

No modelo conceitual (Figura 16) as medidas numéricas podem ser classificadas de acordo com a demanda das mesmas em um o sistema computacional, porém alguns estudiosos apontam que não se deve usar atributos numéricos quantitativos nas tabelas dimensões (representado pelos retângulos no modelo da Figura 16). Outros já apontam que para tomar essa decisão, deve ser considerado a frequência que o atributo será atualizado, pois a tabela fator (representado pelo losango no modelo da Figura 16) não pode sofrer interrupções por conta do seu relacionamento com todas as dimensões.

A tabela fator terá as quantidades totais de funcionários e também a quantidade total para cada área de atuação básica do município (educação, saúde e vigilância sanitária). Além dos quatro campos de medidas, cinco chaves estrangeiras serão adicionadas para fazer a relação das tabelas dimensão entre elas e com a tabela fator.

Antes de desenvolver a tabela fator é necessário identificar qual será a arquitetura administrada entre as três anteriormente discutidas na seção 2.5. Por se tratar de um trabalho acadêmico, a modelo estrela terá o melhor perfil, pois é mais utilizado e muito mais discutido, além de satisfazer melhor o sistema. Outro ponto importante que não pode ser esquecido, é a arquitetura para o sistema DW, como o sistema terá apenas um administrador a melhor solução é um DW Global, possibilitando acesso a todas as informações de forma simples.

6.4 Projeto Lógico

Nessa etapa do projeto é utilizado o programa MySQL para criar o esquema estrela do banco de dados do DW.

Deve-se lembrar que por existir a redundância em DW, o tempo de resposta para as transações serão menores, mas em compensação o espaço de armazenamento cresce com grande facilidade. Na modelagem deve-se agregar o máximo possível de fatores por tabelas semelhantes para que se tenha a menor quantidade de tabelas possível, ajudando assim a reduzir o tempo de resposta. Neste estudo de caso o objetivo do sistema DW é fornecer informação de funcionários públicos nos diversos departamentos em cada município.

O próximo passo é desenvolver o diagrama estrela para que a visualização e a percepção dos relacionamentos fiquem mais claros. Na Figura 17 é demonstrado as cinco tabelas dimensões e a tabela fator (uma cópia das tabelas e do diagrama estará disponível no Apêndice A).

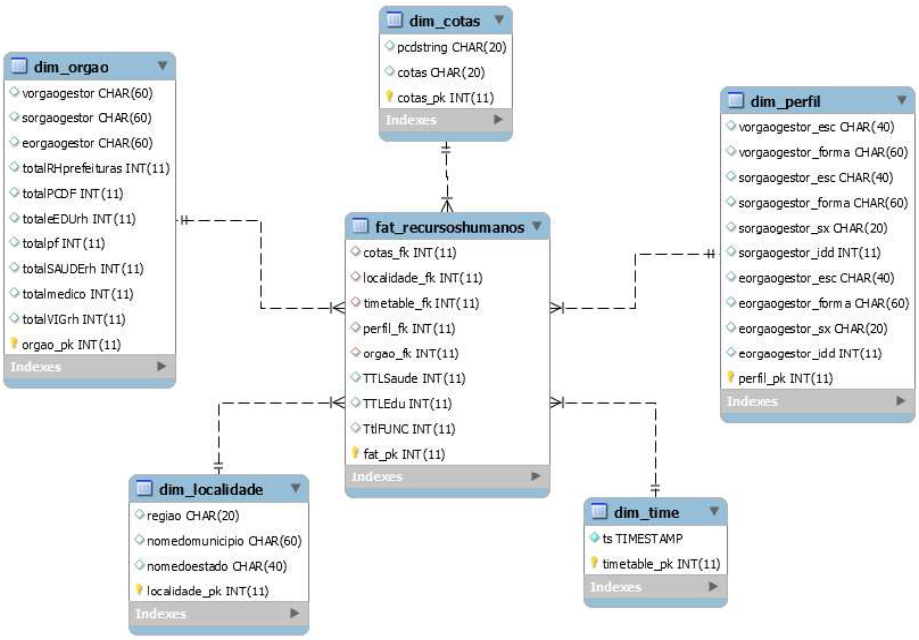

Figura 17. Esquema Estrela do Sistema (Fonte: O autor)

Não existe um número mínimo de dimensões para a implementação de um DW, muito menos é necessário que a tabela fator tenha fatores de medidas, em alguns casos as tabelas

fatores podem possuir apenas as chaves estrangeiras e nenhum outro dado a mais, este tipo de tabela é chamado de tabela *factless.* A tabela fator deste projeto possui três medidas de somatória de cada departamento e o número total de todos os funcionários por município.

A tabela dimensão "dim_perfil" recebe as informações a respeito dos gestores das organizações responsáveis por cada um dos departamentos (saúde, vigilância sanitária e educação), as características são o grau de escolaridade (vorgaogestor_esc, sorgaogestor_esc e eoragaogestor) a formação deles (vorgaogestor_forma, sorgaogestor_forma e eoragaogestor_ forma) outros atributos são idade e gênero.

A "dim_orgao", é uma das principais dimensões e armazena os números de funcionários de cada setor (saúde, educação e vigilância sanitária, prefeitura), total de pessoas com deficiência física, quantidade de médicos e professores municipais.

A "dim_time" possui apenas a sua chave primária e uma função em TIMESTAMP, nesta dimensão serão gravadas as datas das mudanças do DW de forma a obter um banco de dados com histórico de transações de processos, mudanças de dados, e entre outros.

A dim_cotas refere-se a quais municípios possuem a inclusão de pessoas com deficiência física e cotas para pessoas pardas ou negras em seus funcionários.

Com as dimensões definidas, o próximo é transformar e carregar os dados para o DW.

6.4.1 Processo ETL

Depois de ter realizado a modelagem, o desenho e ter criado um esquema estrela para no banco, o próximo passo será popular o DW. Já foi citado na seção 3.4, que para a implementação de um DW é necessário que, alguns dados estejam disponíveis para modelados e para carregá-los aos DW.

Neste estudo de caso o processo de extração foi simples uma vez que os dados já estavam de certa forma organizados na planilha do IBGE, caso não estivesse, alguns recursos externos seriam necessários para capturar os dados para a população do DW.

Uma vez que já se tem a fonte de dados em um arquivo CVS o próximo passo é carregar o arquivo ao Pentaho a melhor opção por ser um arquivo de CVS, seria usar o recurso Input CVS. A interface do Pentaho possui a tecnologia "*drag and drop"* (arraste e solte) tornando a tarefa muito mais rápida e flexível. O input precisa de configurações como, por exemplo, a classificação de cada atributo, tamanho e formato. Existe quatro botões de ação como mostra a Figura 18, o que mais se destaca é o "Get Fields". Esse botão irá adicionar

todas as colunas que possui no arquivo com o nome dos atributos originados da primeira linha da planilha. Além disso o próprio Pentaho oferece a opção para ler as linhas do arquivo e classifica-las automaticamente pelo tipo do atributo (coluna).

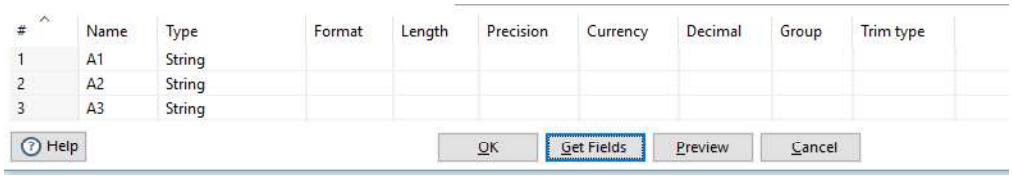

Figura 18. Pentaho Input passo (Fonte: O autor)

Mesmo reconhecendo os atributos de maneira automática, erros podem acontecer pela ocorrência de caracteres indesejados ou linhas com informações vazias. Por este motivo é importante filtrar o dado e analisa-los de forma que os mesmos entrem nos padrões do projeto.

Inúmeras abordagens podem ser realizadas com esse programa, a mais comum é separar a transformação das dimensões da tabela fator.

6.4.2 Carregando os dados nas tabelas dimensão

Na Figura 19, representa a transformação das tabelas dimensões. Este modelo de transformação de tabelas de dimensão utiliza somente um input (arquivo de entrada) e diversos Output (arquivos de saída) que estão ligados às suas tabelas correspondentes no servidor de banco de dados (nesse caso no MySQLServer). Primeiro é decidido qual elemento se tornará a chave primária e a sua relação com a tabela fator, pois muitas vezes as restrições das chaves estrangeiras é o que fazem o serviço se tornar lento.

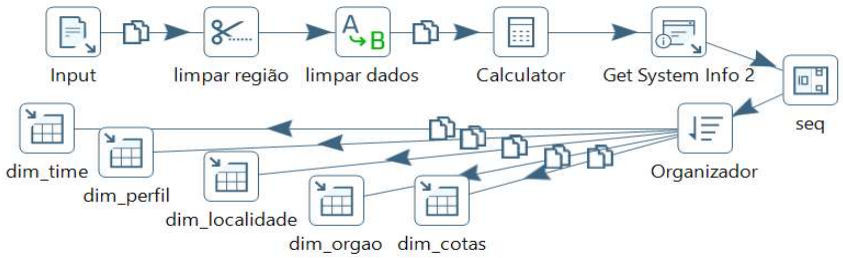

Figura 19. Transformação de Dimensões (fonte: O autor)

As tabelas dimensão possuem na maioria atributos qualitativos, isto é, *strings,* o que acaba provocando muitas linhas do banco de dados sem conteúdo, tornando-se difícil o ajuste. A preocupação que se tem com o processo de transformação, não se limita somente as imperfeições nos dados, mas também às relações das chaves primárias com as chaves estrangeiras.

No estudo de caso, como a planilha do IBGE já traz um identificador, torna-se mais simples extrair esse fator como chaves primárias e estrangeiras. Outra abordagem que poderia ser realizada seria a de organizar as linhas do arquivo (no Pentaho existe um passo com essa finalidade) e utilizar um gerador de sequência para criar as chaves. É importante realizar este tratamento e definir as chaves, pois facilita a geração da tabela fator.

O fluxo dessa transformação dos dados é iniciado por um arquivo de entrada, conforme apresentado na Figura 19. Neste projeto essa entrada é um arquivo em formato de uma planilha eletrônica CSV do Excel (Input). Os próximos dois passos são para o tratamento das *strings* problemáticas. É possível optar por remover uma *string* de um campo ou substituí-las por um argumento, neste caso várias regras foram criadas para cada variável *string* impondo que a cada sinal de "-" deverá ser substituído pela expressão "sem dados".

O elemento "*calculator*" realiza as operações matemáticas necessárias, tais como realizar a soma de todos os funcionários de determinado departamento. Já "*get system info*" fornece a data atual da transformação com base na data configurada no sistema computacional que foi instalado o Pentaho. O modulo "*seq*", adiciona as chaves naturais em cada tupla, essa chave é útil porque o mesmo elemento pode ter um registro duplicado no sistema por causa de alterações e está chave terá numeração única diferenciando cada registro.

Por fim, quando as variáveis terminarem de ser tratadas e estruturadas elas serão organizadas em ordem crescente da primeira coluna e distribuídas para cada tabela dimensão do DW.

Não é um procedimento lento, no caso todo esse processamento de dados não demora mais que alguns segundos, afinal são apenas 5570 linhas sendo processadas. Depois de finalizado, o processo pode ser analisado na aba "*metrics*", como pode ser visto na Figura 20, onde é possível acompanhar o processamento em números: dados de entrada, dados lidos, dados escritos e dados de saída. Em caso de erros a linha do processo ficara vermelha juntamente com o processo, e na aba "*logging*" aparece uma mensagem mais detalhada sobre o problema.

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time
1	CSV file input	0	0	5570	5571	0	0	0	0	Finished	2.6s
2	Calculator	0	5570	5570	0	0	0	0	0	Finished	2.6s
3	Sort rows	0	5570	5570	0	0	0	0	0	Finished	2.8s
4	Database lookup 3	0	5570	5570	5570	0	0	0	0	Finished	4.8s
5	Database lookup 4	0	5570	5570	5570	0	0	0	0	Finished	6.2s
6	Database lookup 5	0	5570	5570	5570	0	0	0	0	Finished	6.6s
7	Database lookup 6	0	5570	5570	5570	0	0	0	0	Finished	6.0s
8	Database lookup	0	5570	5570	5570	0	0	0	0	Finished	6.8s
9	fact	0	5570	5570	0	5570	0	0	0	Finished	8.4s

Figura 20. Janela metric do programa Pentaho (fonte: O autor)

6.4.3. Carregando os dados na tabela fator

Com o carregamento dos dados para as tabelas dimensões realizadas, é preciso carregar dados para a tabela de fatores. O fluxo do processo está demonstrado na Figura 21, a partir de uma entrada de dados, são realizadas operação de agregação dos valores dos funcionários em variáveis de soma e é para isso que a etapa "calculator" entra como um dos primeiros passos do projeto da tabela fator.

Neste momento as linhas são organizadas pela primeira coluna da tabela em ordem crescente. Chega-se a etapa de combinação das chaves primárias e estrangeiras dos arquivos. Todos as etapas Database lookup buscam determinadas dimensões com critérios para localizar campos semelhantes que serão usados para numerar as chaves estrangeiras na tabela fator.

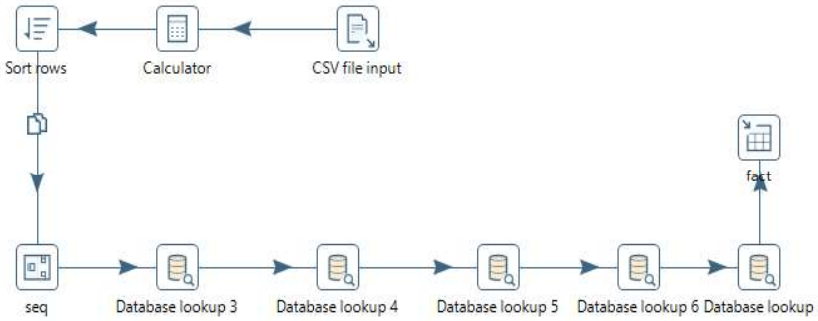

Figura 21. Transformação da Tabela Fator (fonte: O autor)

A maior dificuldade nessa etapa, é a relação das chaves primarias com a sua respectiva chave estrangeira localizada na tabela fator. Como o número de dados a serem carregados é menor o tempo gasto é consequente menor que o tempo gasto com as tabelas dimensão.

6.5 Projeto Físico

A última etapa do projeto, é a implementação da fonte de informações do DW. Esta implementação pode ser feita de duas maneiras, através de um banco de dados relacional, ou de planilhas do Excel.

É fundamental analisar e definir o comportamento relacionado as futuras atualizações. Os sistemas de DW não são projetados para receber dados constantemente como um banco de dados relacional, as atualizações de dados no sistema devem acontecer em intervalos de tempo que nenhum usuário esteja utilizando o DW. Por conta de atualizações ou apenas o carregamento de carga no DW, é importante definir os tipos de *Slowly Change Dimension* (SCD) a serem implementadas, esse assunto é discutido na seção 3.5.Um modelo para atualizar as dimensões, de acordo com a literatura de Kimball (2014) podem ser classificados em 7 tipos de SCD. O Pentaho suporta aplicar os tipos 1 e 2 que são originalmente os tipos mais usados.

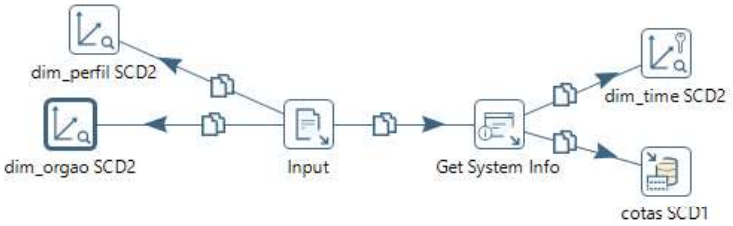

Figura 22. Modelo para SCD 2 e 1 (fonte: O autor)

As melhores opções para poder trabalhar com SCD no Pentaho são os módulos *dimension lookup/update* ou *combination/update*. As dimensões dim_perfil, dim_orgao e dim_time (Figura 22) serão sempre atualizadas com novas linhas e seus dados não serão modificados, pois há uma necessidade de se manter um histórico para futuras comparações. Tanto o dim_perfil como o dim_orgao representa entradas que podem ser constantes no DW, pois quase sempre o quadro de profissionais municipais vem crescendo. Porém a tabela dim_cotas apenas sofrera alterações nos registros, ou seja, novas linhas não serão adicionadas e sim modificadas.

6.6 Analise dos dados de forma dinâmica

BI é uma das áreas apoiadas por DW que mais crescem. E com isso várias ferramentas surgiram e foram melhoradas. O Tableau é um software para análise OLAP de dados para DW, é de fácil entendimento e vem ganhando espaço no mercado. O sistema disponibiliza uma versão de teste, disponibilizada gratuitamente no endereço http://www.tableau.com/.

6.6.1 Interface do Software

Com o Tableau é possível gerar vários tipos de gráficos, além da possibilidade de construir *dashboards* ou até mesmo *story points* para apresentações empresariais. Não são necessários conhecimentos avançados de informática, assim o programa pode ser usado por qualquer de usuário. O primeiro passo pós instalação é o cadastro de e-mail para ativar o programa (*trial free*), daí a conexão ao banco de dados e a escolha do repositório. Para concluir a etapa de configuração do banco de dados é necessário unir as tabelas para uma melhor análise dos dados, como consta na Figura 23.

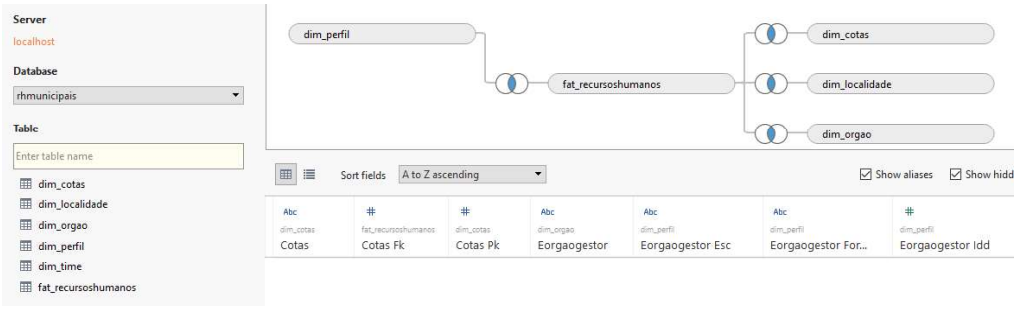

Figura 23. Tela de configuração do Data Warehouse. (Fonte: O autor)

Na Figura 24, do lado direito encontra-se as configurações dos tipos de gráficos, são muitas possibilidades, porém para a utilização de alguns gráficos em específico são necessárias algumas informações específicas. Por exemplo, gráficos de linhas somente são disponibilizados para o uso se uma das variáveis que foi selecionada seja conhecida por ser uma variável do tipo tempo.

Ao lado esquerdo da Figura 24 exibe-se todas as variáveis do DW disponibilizadas como *dimensions* (Dimensões) e *measures* (Medidas) agrupados pelas tabelas a que pertencem. Mesmo algumas dimensões sendo apenas *string*, é possível usá-las como um atributo uma vez que você pode contar a frequência que ela ocorre no cenário gerando assim um atributo numérico.

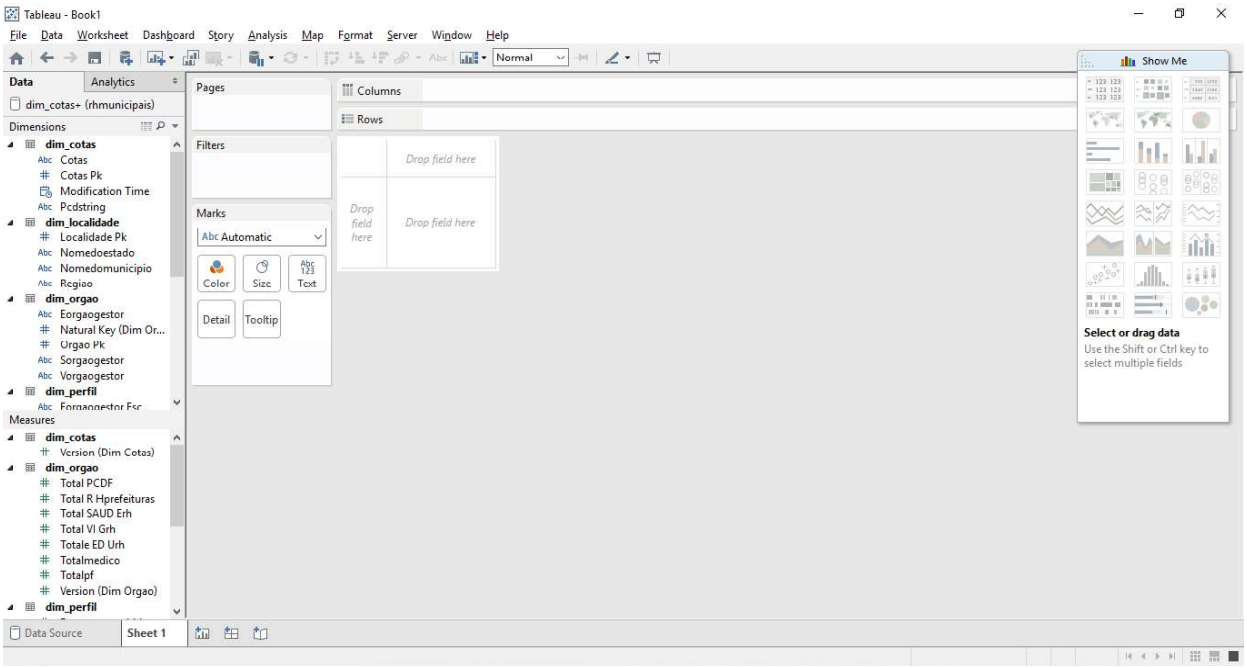

Figura 24. Área de trabalho do Tableau (Fonte: O autor)

6.6.2 Gráficos

Nessa última seção descreve-se as análises dos dados de modo a descobrir padrões existentes na planilha do IBGE, através de gráficos gerados a partir das informações armazenadas no DW.

A figura 25 traz um gráfico de barras com o número de municípios por estado, onde se constata que o estado com maior quantidade de municípios e consequentemente ira exigir mais secretarias de setores municipais será o estado de Minas Gerais, seguido de e São Paulo e Rio Grande do Sul. O estado com menos municípios é Amapá. Para o desenvolvimento desse gráfico foi utilizado variáveis da dimensão dim_localidade.

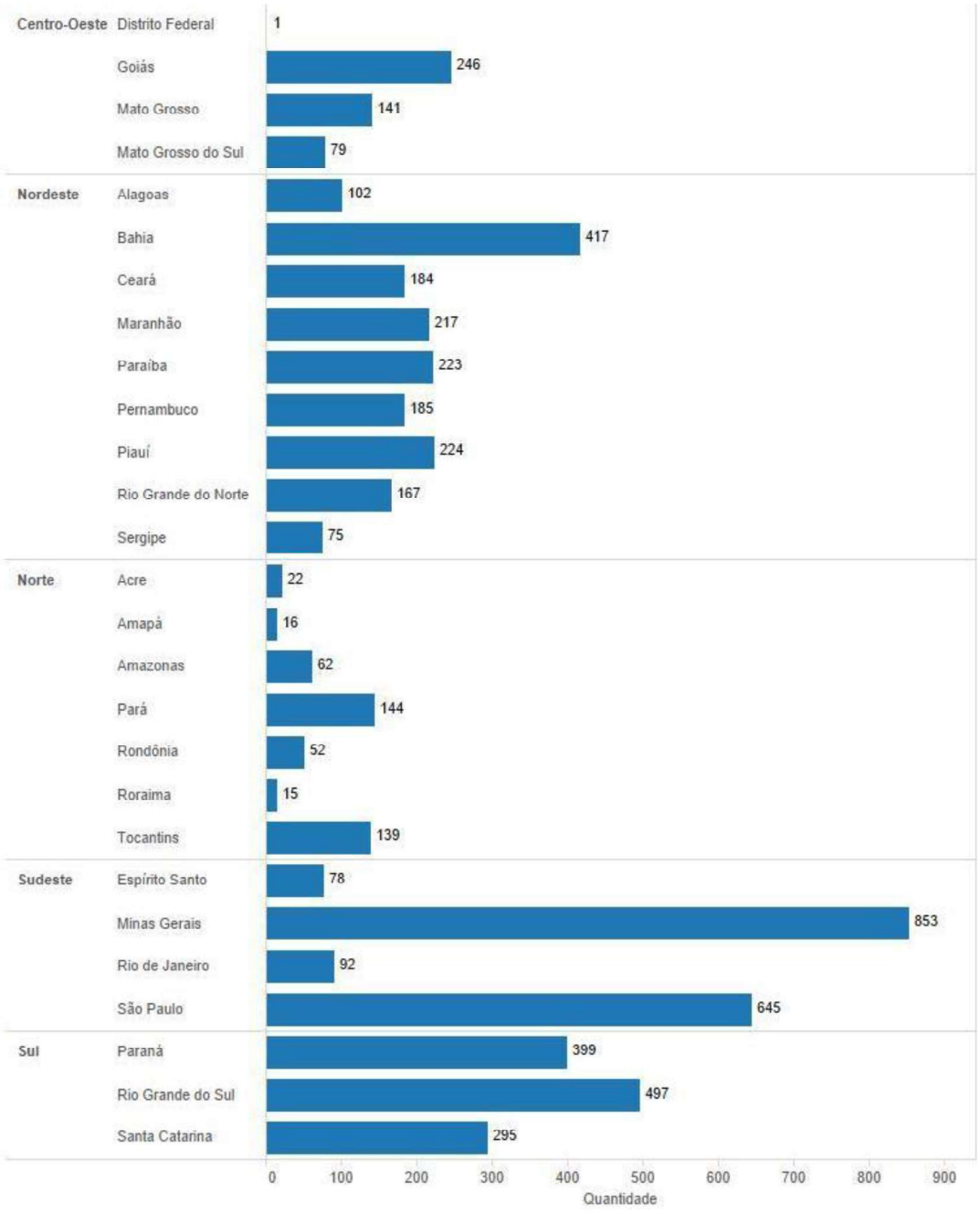

Figura 25. Total de Municípios (Fonte: O autor)

Pela figura 26, com tabelas relacionadas a dispersão de municípios. Evidencia uma dispersão menor de municípios são região Norte e Centro-Oeste. Em uma média total de funcionários por região, as maiores parcelas são na região Sudeste (4.306.981 funcionários municipais) e Nordeste (3.671.031 funcionários municipais).

Média

Regiao	Saúde	Educação	Prefeitura	Total
Centro-Oeste	159,476	250,726	7,254	995,432
Norte	147,142	313,285	2,614	1,061,144
Sul	214,287	441,699	2,006	1,534,571
Nordeste	562,361	1,060,874	2,009	3,671,031
Sudeste	690,814	1,185,231	5,722	4,306,981

Variância

Regiao	Saúde	Educação	Prefeitura	Total
Sul	313,155	862,470	4,024,675	11,944,147
Nordeste	787,715	804,014	4,037,883	13,632,059
Norte	637,621	1,306,789	6,831,132	20,050,245
Centro-Oeste	3,825,029	6,798,570	52,621,426	138,895,728
Sudeste	3,345,621	12,886,522	32,744,675	122,759,413

Total

Regiao	Saúde	Educação	Prefeitura	Total
Centro-Oeste	159,476	250,726	585,230	995,432
Norte	147,142	313,285	600,717	1,061,144
Sul	214,287	441,699	878,585	1,534,571
Nordeste	562,361	1,060,874	2,047,796	3,671,031
Sudeste	690,814	1,185,231	2,430,936	4,306,981

Desvio Padrão

Regiao	Saúde	Educação	Prefeitura	Total
Centro-Oeste	1,956	2,607	7,254	11,785
Norte	799	1,143	2,614	4,478
Sul	560	929	2,006	3,456
Nordeste	888	897	2,009	3,692
Sudeste	1,829	3,590	5,722	11,080

Figura 26. Cálculos Estatísticos por região.

A quantidade de Professores e Médicos é destacada na Figura 27, com um gráfico de barras. Nota-se uma quantidade maior de professores de escolas municipais à médicos em todas as cidades.

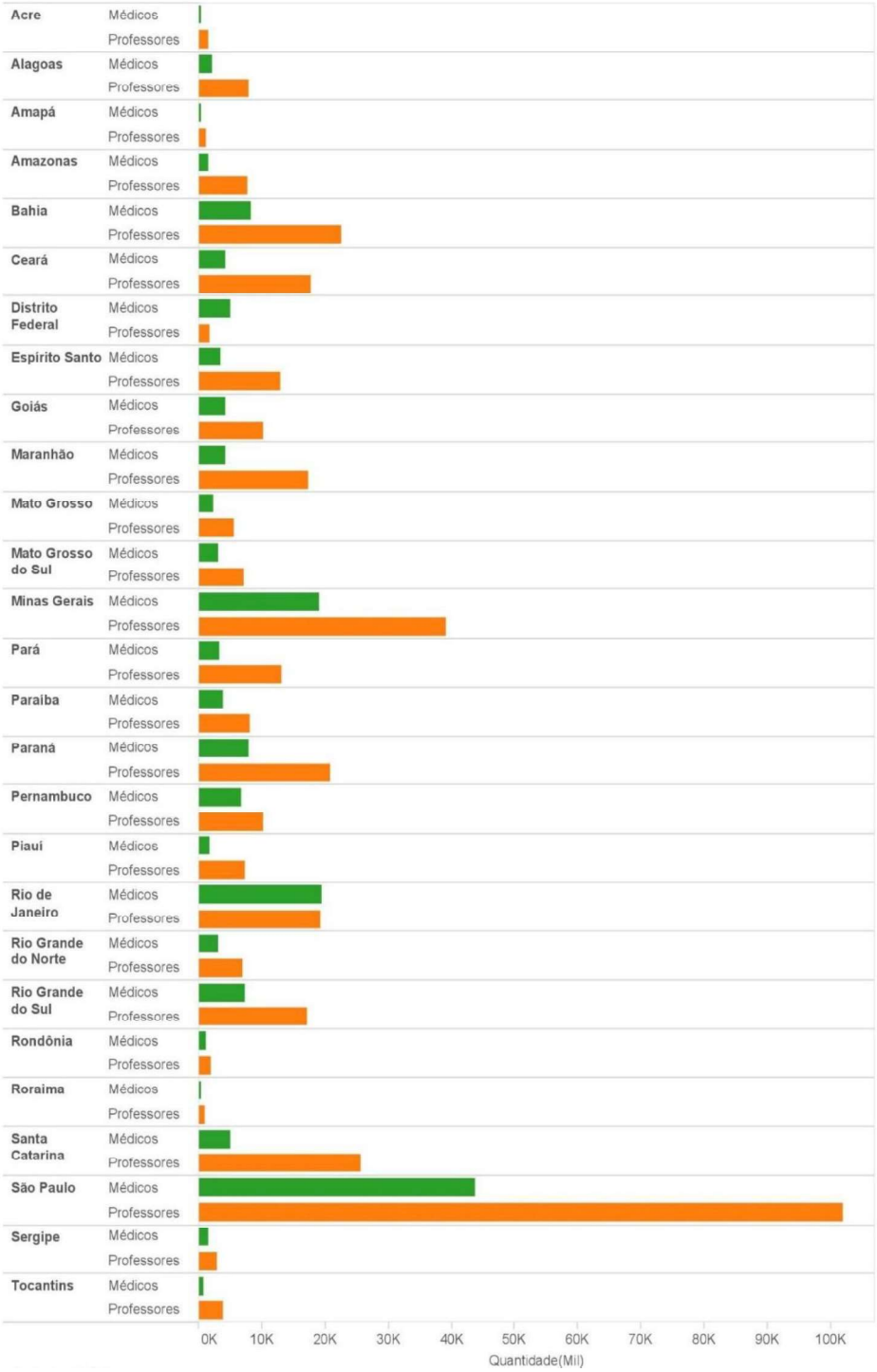

Figura 27. Médicos e Professores por estado (Fonte: O autor)

Na área de gestão da educação e da saúde municipal o gênero predominante é o feminino entre idades de 25 a 70 anos, porém é com 35 anos a idade de maior incidência para a saúde e entre 45 a 50 para a educação como demonstra a Figura 28 com um gráfico de linhas.

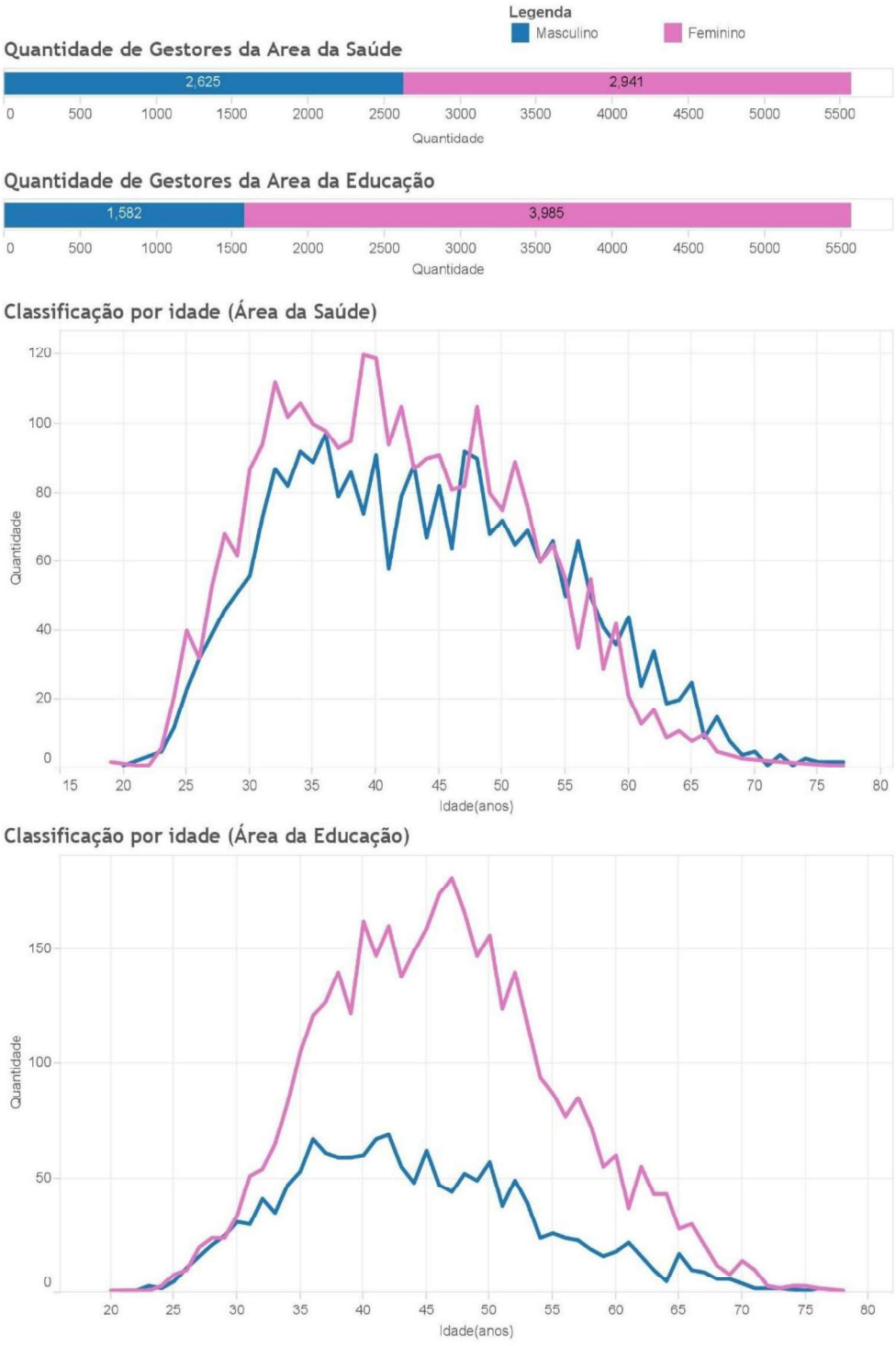

Figura 28. Gestores da Educação e Saúde por gênero (Fonte: O autor)

O nível de escolaridade dos gestores municipais de educação, saúde e vigilância sanitária são apresentados na figura 29. Eles são respectivamente pós-graduação, ensino superior completo para os dois últimos.

Figura 29. Grau de Escolaridade e (Fonte: O autor)

A área majoritária de formação dos gestores municipais de educação é a pedagogia, de saúde é enfermagem e vigilância sanitária não possuem dados suficientes para a análise. Essas informações podem ser verificadas na Figura 30.

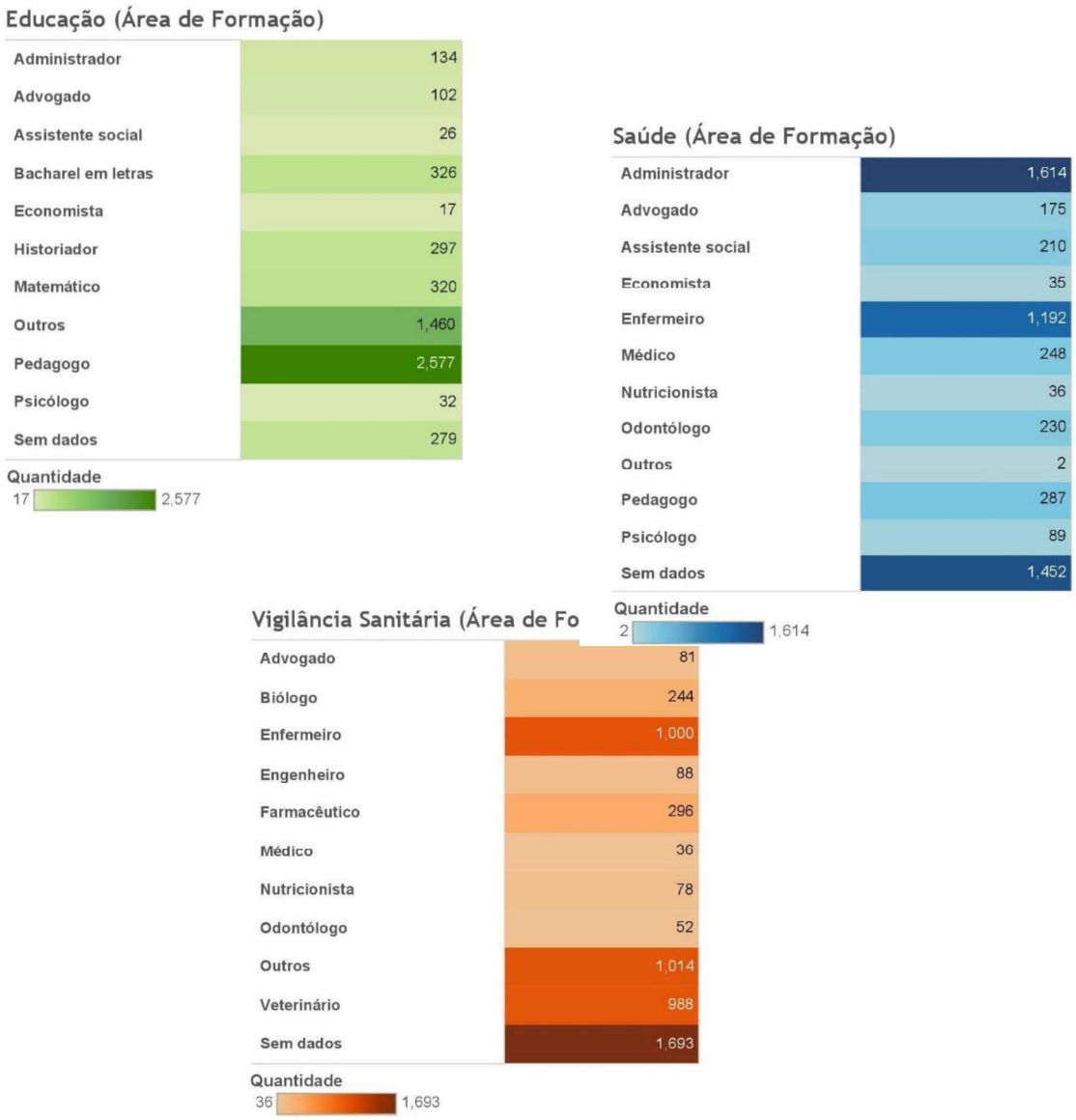

Figura 30. Formação dos gestores municipais (Fonte: o autor)

A criação de cotas raciais para concursos públicos vem aumentando gradativamente devido as novas diretrizes, porém pouquíssimos municípios aderiram a essa norma. Para pessoas com deficiência física os estados com um maior índice de funcionários municipais são os estados de São Paulo, Goiás e Minas Gerais. Os números de ambas políticas (políticas de cotas raciais e quantidade de deficientes físicos) podem ser visualizadas na Figura 31.

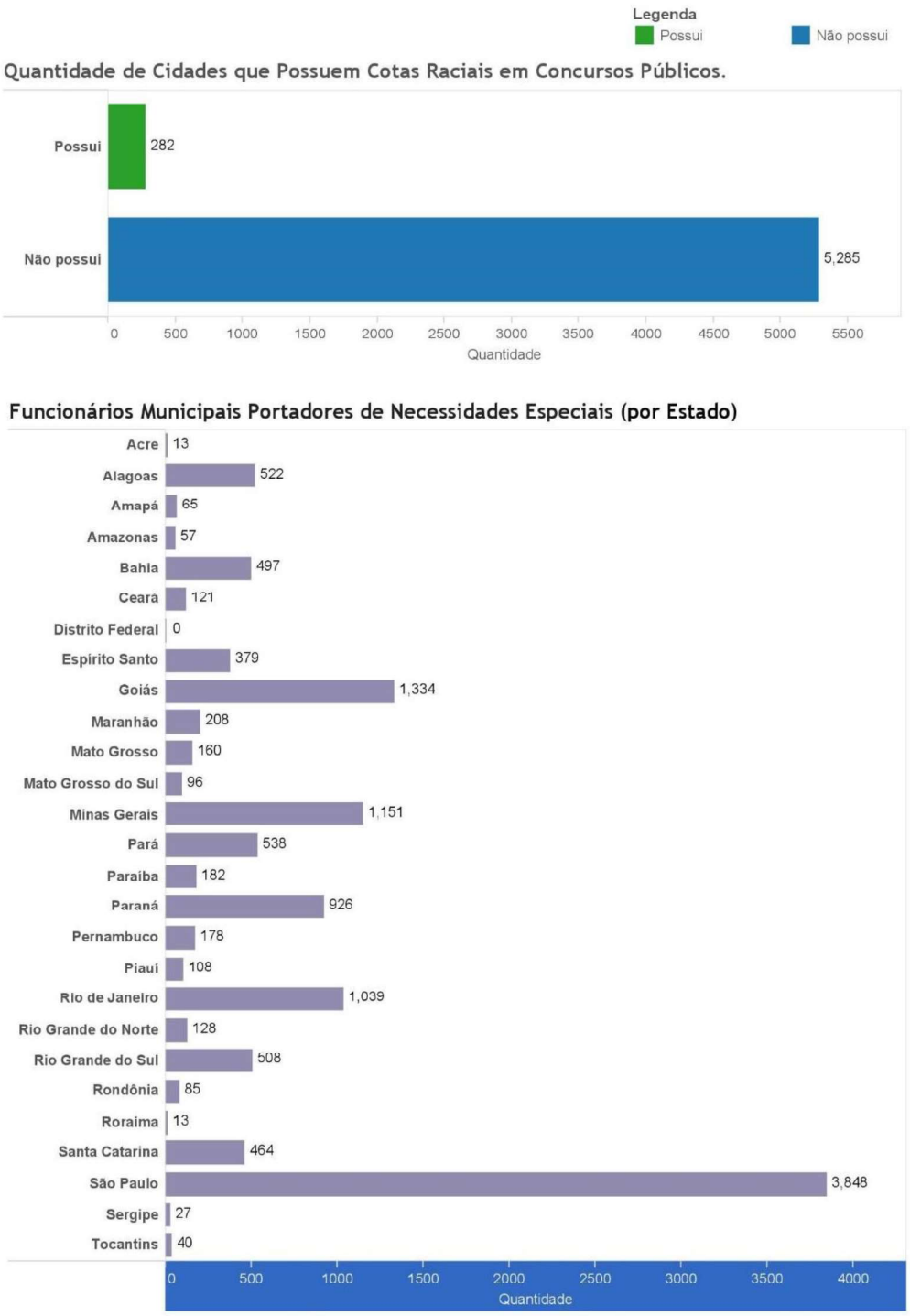

Figura 31. Inclusão Social (Fonte: O autor)

Para uma análise especifica, foi selecionado só municípios de Inconfidentes, Borda da Mata, Bueno Brandão, Ouro Fino, Jacutinga e Pouso Alegre para analisar o quadro de funcionários locais. Inconfidentes é a cidade com menos funcionários em todos os setores, seguido de Bueno Brandão, Borda da Mata, Ouro Fino, Jacutinga e por ultimo Pouso Alegre como mostra na Figura 32.

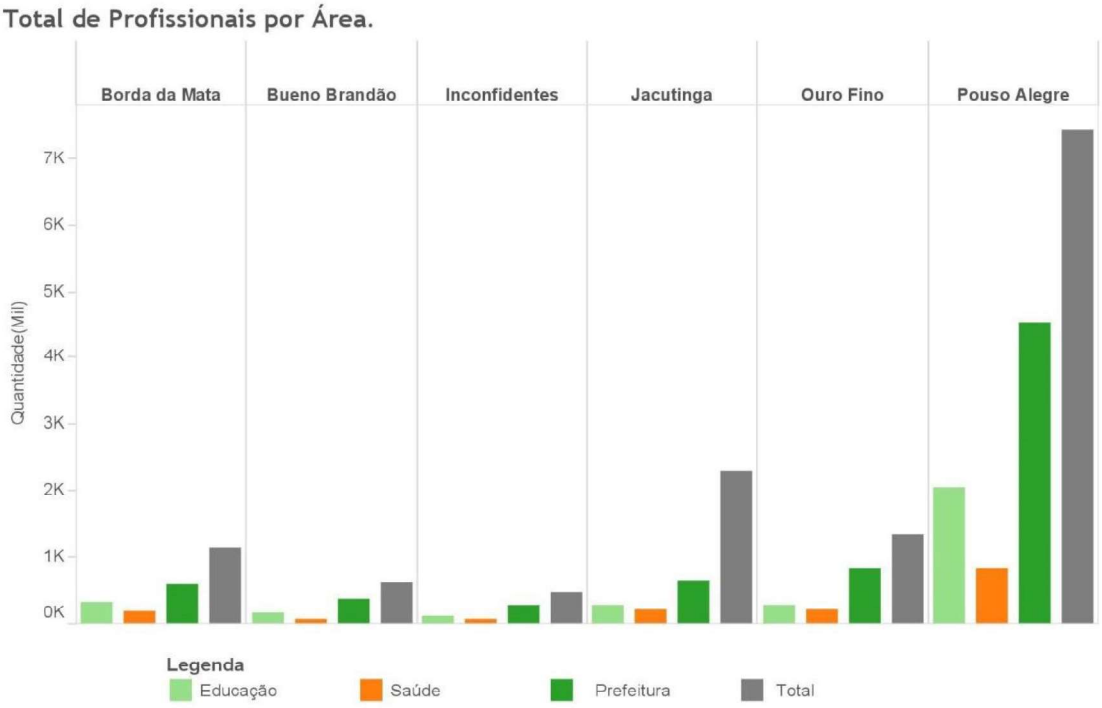

Figura 32. Quadro de funcionários municipais de Inconfidentes e Região (Fonte: O autor)

Na figura 33, se encontra a classificação dos representantes municipais na Secretaria da Educação, Secretaria da Vigilância Sanitária e Secretaria da Saúde.

Perfil do Gestor de Educação*

	Ensino superior completo	Pós-graduação	
	Outros	Matemático	Pedagogo
Borda da Mata			1
Bueno Brandão			1
Inconfidentes		1	
Ouro Fino	1		
Pouso Alegre			1

Perfil do Gestor de Vigilância Sanitária

	Ensino superior completo		Ensino superior incomplet..	Pós-graduação
	Advogado	Veterinário	Sem dados	Veterinário
Borda da Mata				1
Bueno Brandão		1		
Inconfidentes			1	
Jacutinga			1	1
Ouro Fino		1		
Pouso Alegre	1			

Perfil do Gestor de Saúde

	Ensino fundame..	Ensino superior ..	Ensino superior i..	Pós-graduação		
	Sem dados	Nutricionista	Sem dados	Advogado	Enfermeiro	Médico
Borda da Ma..				1		
Bueno Bran..	1					
Inconfidentes					1	
Jacutinga		1	1			
Ouro Fino						1

*Não há registro a respeito do gestor de educação da cidade de Jacutinga

Figura 33. Perfil dos Gestores de Inconfidentes e Região (Fonte: O autor)

De forma geral analisando todos os gráficos, desenvolve um projeto de melhoria de gestão para a inclusão de mais cotas raciais em cada município e a presença de mais médicos deve ser avaliada também. Uma possível meta de melhor formação para cargos de gestão de setores nos municípios deve ser requisita pelas prefeituras, além de formações especificas para cada secretaria, como por exemplo para a secretaria de saúde deveria ser profissões na área da saúde como também deve ser na área de vigilância sanitária

7.CONSIDERAÇÕES FINAIS

Data Warehouse é uma solução desenvolvida em razão do tempo levado para o processamento dos dados durante uma pesquisa em bancos de dados relacional. Enquanto bancos relacionais tem grande preocupação com o espaço de armazenamento, Data Warehouse é focado na velocidade, no tempo de resposta. Essas diferenças os tornam necessários para sistemas corporativos, apoiando área operacional e gerencial. Ambos são necessários para um sistema corporativo.

A semelhança entre o DW e o banco de dados relacional está no tipo de arquivo que ambos tratam, dados estruturados. Já Big Data se difere, uma vez que é capaz de tratar qualquer tipo de dado ou arquivo. Um DW, por muitas vezes é alimentado por banco de dados existente, organizados pelo processo de ETL, tornando-os de mais fácil leitura, consulta e acesso.

Um ponto importante sobre DW é a sua capacidade de modelar os dados em uma perspectiva multidimensional. Tal perspectiva provê uma visão dinâmica facilitando a extração de informações e a tomada de decisão. Ferramentas OLAP auxiliam na construção e preenchimento dos “cubos de dados” que auxiliam nas soluções de problemas e situações que antes não era resolvida devido ao limite do banco de dados.

A segunda parte de destaque do Data Warehouse é a sua capacidade de armazenamento de dados históricos, vital para operações de BI como, por exemplo, a mineração de dados. Um projeto de BI que atua sobre um DW normalmente possui 4 camadas: a camada 1 é a extração dos dados, a camada 2 a análise e modelagem dos dados, a camada 3 o processo de limpeza dos dados e a inserção do mesmo no banco de dados, por fim a camada 4, a mineração efetiva, que trabalha com os dados de forma integra, sem a possibilidade de encontrar dados “sujos”.

As aplicações de DW estão totalmente interligadas com os sistemas computacionais e com a rede de internet. Existe uma imensa quantidade de usuários na rede produzindo e

consumindo informações. Eles podem produzir um canal de comunicação com as informações que estejam do outro lado do mundo, sem a necessidade de muito esforço, apenas utilizando alguns sites de busca.

Com toda essa atividade diária na internet, muitas empresas estão migrando seu mercado para ela com a intenção de persuadir novos clientes e principalmente manter os já antigos. Além da Internet ser um importante canal para ligar o cliente à empresa, a mesma é uma riquíssima fonte de informação para aplicações de DW. Diversas informações podem ser extraídas de servidores de acesso, cada link visitado pelo cliente no servidor pode ajudar a entender o comportamento de consumo e predizer quais possíveis produtos serão consumidos, através das informações atuais e do histórico armazenado no DW.

Empresas precisam de um suporte para tomada das decisões, pois muitas variáveis estão em jogo, o relacionamento que a empresa tem com o cliente, a relação do marketing com o cliente e também a relação do produto com o cliente. Uma maneira simples e eficaz de se relacionar todas as variáveis, a fim de se obter respostas confiáveis é a encontrada nos princípios do DW.

A respeito do caso de estudo abordado neste trabalho, pode-se concluir que foi viável e eficaz trabalhar em um ambiente de DW para extrair informações de uma planilha complexa do IBGE, e possibilita comparações ou umas predições.

Uma vez que o IBGE sempre publica pesquisas similares, usuários de empresas trabalhando com diversas ferramentas de DW e BI podem utilizar as informações como vantagem empresarial Apesar de simples, este estudo de caso sinaliza a possibilidade de se explorar dados e extrair informações ricas e diversificadas. Um trabalho futuro poderia estender e focar em dados de uma instituição de ensino, um bom exemplo seria uma instituição de ensino que possuiria 4 dimensões (estudante, professores, funcionários, aulas) e todas as dimensões estariam interligadas por um fator central (que seria a própria instituição). A instituição pode se beneficiar com o projeto em termos que o mesmo vai ajudar a solução de problemas como evasões de curso, gerenciamento dos profissionais, além do rendimento acadêmicos dos alunos.

Haverá sempre inúmeras maneiras de trabalhar com Data Warehouse devido aos diferentes tipos de abordagens que podem ser utilizados.Com novos projetos, a dinâmica de trabalhar com Data Warehouse começa a parecer menos complexa em cada etapa do projeto. Por essas razões Data Warehouse se torna um tópico interessante, e deve cada vez mais, ser explorado nas diversas áreas de TI.

REFERENCIAS

BALLARD, Chuck et al. **Data Modeling Techniques for Data Warehousing.** Eua: Red Books Ibm, 1998. 216 p. Disponível em: <http://www.redbooks.ibm.com/redbooks/pdfs/sg242238.pdf>. Acesso em: 14 jan. 2016.

DASAND ,T.K. MOHAPATRO, Arati. A Study on Big Data Integration with Data Warehouse. **International Journal Of Computer Trends And Technology (ijctt).** Eua, p. 188-192. mar. 2014.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3. ed. Waltham, Massachusetts: Morgan Kaufmann, 2012. 740 p.

HAYES, Scott; GUNNING, Philip**. Tunning Up for OLTP and data warehousing**. DB2 Magazine.Vol 7 Num 3, 2002.

HERDEM, Olaf. **A Design Methodology for Data warehouses**. Oldemburg Research and Development Institute for Computer Science Tools and Systems (OFFIS). Oldenburg, Germany.

HURWITZ, Judith S. et al. **Big Data for Dummies**. Haboken, New Jersey: John Wiley & Sons, Inc., 2013. 339 p

INMON, W. H.. **Building the Data Warehouse.** 3. ed. Eua: John Wiley & Sons, Inc., 2002. 428 p

INMON, William; STRAUSS, Derek; NEUSHLOSS, Genia. **DW 2.0: The Architecture for the Next Generation of Data Warehousing**. Estados Unidos: Morgan Kaufmann Digital, 2012. 400 p.

KELLY, Sean. **Data Warehousing in Action**. Estados Unidos: Wiley, 1997. 334 p.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit** 1. ed. Estados Unidos: John Wiley And Sons, Inc., 1998. 539 p.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit:** The Definitive Guide to Dimensional Modeling. 3. ed. Estados Unidos: John Wiley And Sons, Inc., 2013. 600 p.

MYSQL. **MySQL Workbench.** Disponível em:<http://dev.mysql.com/doc/workbench/en/>. Data de acesso 23 de março de 2016.

PENTAHO. **Data Integration** Disponível em: <http://www.pentaho.com/product/data-integration>. Data de acesso 23 de Março de 2016.

RUSSOM, Philip. **Big Data Analytics**. Estados Unidos: Tdwi Research, 2011. 34 p.

SHERMAN, Rick. **Business Intelligence Guidebook:** From Data Integration to Analytics. Estados Unidos: Morgan Kaufmann, 2014. 550 p.

TABLEAU. **Tableau Desktop** Disponivel em:< http://www.tableau.com/support/desktop>. Data de acesso 23 de Março de 2016.

TURBAN, Efraim et al. **Business Intelligence:** Um Enfoque Gerencial para Inteligência do Negócio. Porto Alegre: Artmed Editora S.a., 2009. 252 p.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A.. **Data Mining:** Practical Machine Learning Tools and Techniques. 3. ed. Burlington,Massachusetts: Morgan Kaufmann, 2011. 665 p.

APÊNDICE A: Diagrama Estrela do Projeto.