UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
NÚCLEO DE INFORMAÇÃO TECNOLÓGICA E GERENCIAL
CURSO DE ESPECIALIZAÇÃO EM GESTÃO ESTRATÉGICA DA INFORMAÇÃO

Lucas Herbert de Resende Silva

# USO DE DATA WAREHOUSE NA GESTÃO ESTRATÉGICA DA INFORMAÇÃO: ESTUDO APLICADO A UMA EMPRESA DE TECNOLOGIA MOBILE

Belo Horizonte
2015

LUCAS HERBERT DE RESENDE SILVA

# USO DE DATA WAREHOUSE NA GESTÃO ESTRATÉGICA DA INFORMAÇÃO: ESTUDO APLICADO A UMA EMPRESA DE TECNOLOGIA MOBILE

Monografia apresentada ao programa de Especialização do Núcleo de Informação Tecnológica e Gerencial – NITEG, no curso Gestão Estratégica da Informação da Escola de Ciência da Informação, da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Gestão Estratégica da Informação.

Orientador: Mestre Prof. Eduardo Ribeiro Felipe

Belo Horizonte
Escola de Ciência da Informação - UFMG
2015

**RESUMO**

As organizações contemporâneas convivem com a constante competição causada pela globalização. A gestão da informação é uma resposta das organizações contemporâneas ao ambiente competitivo, que exige agilidade, inovação e capacidade de aprender. O uso do ambiente de *Data Warehouse* e de ferramentas de análise de dados apresentam grande crescimento, pois são utilizados para armazenamento, gestão e disseminação da informação.

Esta pesquisa buscou descrever, a partir de um estudo de caso, o desenvolvimento do projeto de *Data Warehouse* aplicado em uma empresa de tecnologia *mobile,* com atuação no mercado Sul-Americano. Essa organização busca melhorar os instrumentos de análise das informações a partir do ambiente de *Data Warehouse*, armazenando os dados relevantes para extrair padrões de comportamento e de tendências dos seus clientes na utilização de seus serviços e produtos.

O estudo realizado, além de apresentar o ambiente de *Data Warehouse* desenvolvido, demonstra como os usuários podem extrair informação útil, relevante e economicamente importantes. Essas informações são obtidas a partir de múltiplas e volumosas fontes de dados, que proporcionam maior competitividade da empresa no mercado, pois torna possível o monitoramento dos produtos e dos serviços da organização, além de predizer a direções/tendências por meio dos dados produzidos.

**Palavras-chave:** Informação; Data Warehouse; ETL; OLAP

**ABSTRACT**

Contemporary organizations live with the constant competition caused by globalization. Information management is a response of contemporary organizations the competitive environment, which requires agility, innovation and ability to learn. The use of the data warehouse environment and data analysis tools present huge growth because they are used for storage, management and dissemination of information.

This research aims to describe, from a case study, the development of the Data Warehouse project applied in a mobile technology company, with operations in the South American market. This organization seeks to improve the analytical tools of the information from the data warehouse environment by storing relevant data to extract patterns of behavior and their customer trends in the use of their services and products.

The study, in addition to presenting the Data Warehouse developed environment, demonstrates how users can extract information useful, relevant and economically important. This information is obtained from multiple and voluminous data sources that provide greater competitiveness of the company in the market, as it makes possible the monitoring of products and services of the organization, as well as predict the directions / trends through the data produced.

**Keywords:** Data Warehouse; ETL; OLAP; Integration Services; Analysis Services

## LISTA DE ILUSTRAÇÕES

## LISTAS DE ABRIVIATURAS E SIGLAS

| | |
|---|---|
| BD | Banco de dados |
| BI | Business Intelligence (Inteligência de Negocio). |
| CRM | Customer Relationship Management |
| DM | Data Mat |
| DW | Data Warehouse (Armazém de dados ou Deposito de dados) |
| ERP | Enterprise Relationship Management |
| ETL | Extraction / Transformation / Load (Extração / Transformação / Carga). |
| MIC | Microsoft Innovation Center (Centro de Inovação Microsoft) |
| OLAP | On-Line Analytical Processing (Processo transacional on-line) |
| SAD | Sistemas de Apoio a Decisão |
| SGBD | Sistema Gerenciador de Bando de Dados |

**SUMÁRIO**

# 1. INTRODUÇÃO

Diversas aplicações como sistemas Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), sistemas legados e informações não estruturadas, disponíveis na internet, são utilizadas por organizações em busca de maior competitividade no mercado. Tais aplicações produzem e armazenam grandes volumes de dados, gerando diversos repositórios. O desafio das organizações de acordo com Brackett (1996) é criar dados conexos e íntegros, que darão suporte à demanda de informações. Portanto, é necessário desenvolver bons repositórios de dados para que se possa fazer uso eficiente das informações geradas a partir desses dados.

Segundo Barreto (1994), a produção da informação é definida como estrutura significante, operacionalizada por meio de práticas e de processos bem definidos de transformação orientada. Esta é representada a partir das atividades relacionadas à reunião, seleção, codificação, redução, classificação e armazenamento de informação. Apesar de não produzir qualquer conhecimento, por ser estático, o estoque de informação representa uma fonte potencial de conhecimento, imprescindível na transferência de informação.

Apesar das informações armazenadas possuírem a competência para produzir conhecimento, este só se efetiva a partir da ação de comunicação mutua entre o estoque de informação e o receptor (BARRETO, 1994). Conclui-se, então, a grande importância em manipular e proporcionar acesso às informações. Choo (2002) sugere que a sobrevivência das empresas depende da capacidade destas em processar informações sobre o ambiente e transformá-las em conhecimento que lhes permitam adaptar as mudanças.

Para tornar essa rotina mais intuitiva, Kimball (1998) sugere a utilização de ferramentas de *Extraction, Transformation and Load* (ETL). Elas executam com assertividade a padronização das informações existentes nos repositórios de dados. Essas ferramentas também extraem para uma base que possibilita a consulta eficiente, ou seja, um ambiente de Data Warehouse (DW).

Nesse cenário, percebe-se que ferramentas para auxiliar na gestão de dados e de suporte ao processamento analítico ganham espaço por proporcionar melhores resultados no processo de suporte e apoio à tomada de decisão.

Este trabalho procura explorar um estudo de caso em uma empresa de Tecnologia que atua no mercado de telefonia móvel. A empresa analisada desenvolveu um projeto de Data Warehouse, a fim de melhorar os instrumentos de análise da organização, baseado no grande volume de dados que produz diariamente. Objetiva-se, portanto, descrever o desenvolvimento deste projeto.

Para que seja possível atingir os objetivos deste estudo, serão abordados alguns conceitos e definições sobre DW e os aspectos que envolvem este ambiente para que possamos, então, apresentar o estudo de caso proposto.

# 2. REFERENCIAL TEÓRICO

A seguir, será apresentado o levantamento bibliográfico sobre a organização da informação no ambiente de *Data Warehouse*.

## 2.1 HISTÓRICO DO *DATA WAREHOUSE*

A utilização de diversas aplicações não integradas corroboram para que as organizações possuam diversas bases de dados. A história do *Data Warehouse* segundo Inmon (1997), inicia com a evolução dos Sistemas de Apoio a Decisão (SAD), com o objetivo de criar um repositório central de dados históricos para consultas e análises.

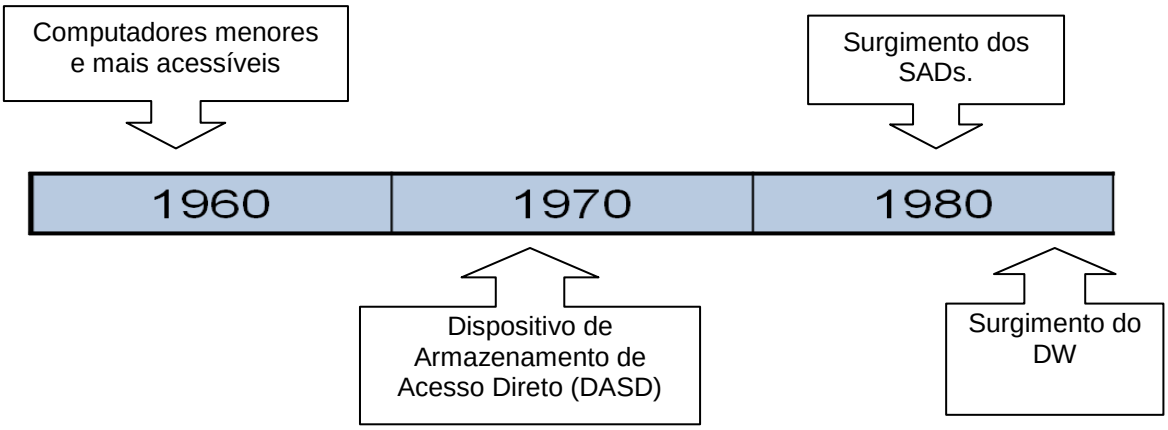
Até o início da década de 80, outras tecnologias começaram a surgir e o computador passou a ter uma nova forma de atuação nas organizações, atuando não apenas com cálculos e previsões, mas como um instrumento de análise. Os SADs surgem para direcionar decisões operacionais.

Segundo Inmon (1997), surgiu um paradigma – um único banco de dados que poderia atender, simultaneamente, ao processamento de transações *online* de alta performance e ao processamento analítico ou de SAD. Tal paradigma, assim como o surgimento das ferramentas de extração, transformação e carga de dados, fizeram com que as organizações tratassem o processo de arquitetura de *software* e *hardware* de forma relapsa, produzindo o chamado de “arquitetura de desenvolvimento espontâneo”.

A arquitetura de desenvolvimento espontâneo apresentou dificuldades, como a falta de credibilidade dos dados, os problemas de produtividade e a impossibilidade de transformar dados em informações.

A Figura 1 ilustra os acontecimentos descritos que precederam o surgimento do DW.

**FIGURA 1:** Linha do tempo do histórico do Data Warehouse



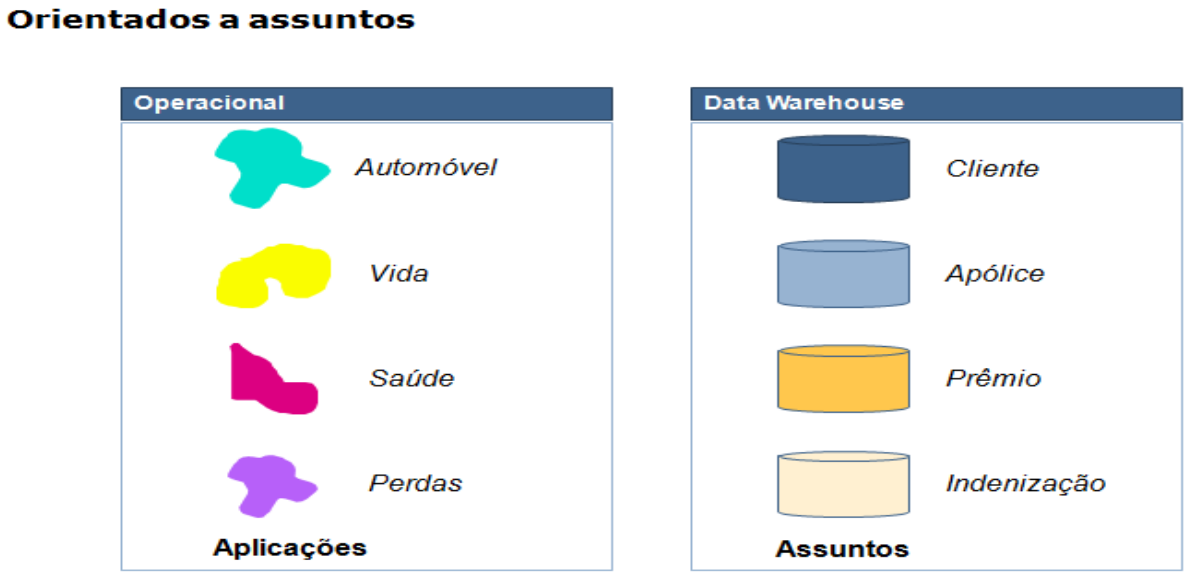

Fonte: Elaborado pelo autor

## 2.2 *DATA WAREHOUSE*

De acordo com Inmon (1997), “um *Data Warehouse* é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”. Para Kimball (1996), *Data Warehouse* é “uma cópia dos dados transacionais estruturados para relatórios e análise”. Isto é, DW armazena os dados operacionais da empresa, gerados pelas aplicações do negócio, integrados e transformados de forma que possamos extrair informações integradas e consolidadas sobre o negócio e seus processos. É importante ressaltar que o DW é a estrutura na qual as ferramentas de processamento analítico utilizam, e não tais ferramentas em si.

Segundo INMON (1997) a primeira característica marcante de um DW é a orientação ao assunto. Os bancos de dados operacionais baseiam-se nas aplicações da empresa, sendo projetados com uma estrutura para um melhor desempenho na aplicação operacional. Já o Data Warehouse atua em outro paradigma, os dados devem estar organizados de acordo com os principais assuntos da empresa, de acordo com o seu ramo ou atividade. A Figura 2 demonstra o DW baseado por assunto/negócio.

**FIGURA 2:** Um exemplo de dados baseados em assuntos/negócios




**Fonte:** Adaptado de INMON (1997)

O fato do DW ser integrado é a segunda característica marcante, tornando este aspecto o mais importante. As inconsistências entre as aplicações de uma empresa, devido aos diversos projetos ocorridos ao longo dos anos, levam a falta de coerência em termos de codificação, padronização de nomes, atributos físicos, entre outros. A carga dos dados no DW é feita de forma a eliminar as inconsistências das diversas aplicações da empresa.

Os dados do DW não são voláteis, ou seja, os dados inseridos em um DW não sofrem alterações ou são excluídos. Já no ambiente operacional, os dados são acessados por muitas pessoas ao mesmo tempo, podendo ser atualizados. No ambiente de DW, os dados são acessados, porém não são atualizados, com o intuito de preservar as análises históricas, que presa pela variância temporal dos dados.

O Banco de dados operacional, por sua vez, é uma foto dos dados no presente momento. Portanto, os dados podem ser atualizados, diferentemente dos dados de um DW, que não podem ser atualizados e sua estrutura chave sempre contém algum elemento de tempo.

## 2.3 FASES DE UM PROJETO DE *DATA WAREHOUSE*

De acordo com Barbieri (2001), um projeto de DW é semelhante a um projeto de desenvolvimento tradicional de sistemas. Contudo, possui diferenças que devem ser observadas atentamente. Os principais passos deste projeto são:

1. Planejamento
2. Levantamento das necessidades
3. Modelagem Dimensional
4. Projeto Físico dos Bancos de Dados
5. Projeto de Transformação
6. Desenvolvimento de Aplicações
7. Validação e Teste
8. Treinamento
9. Implantação

O projeto deverá seguir uma metodologia básica de desenvolvimento de *software*. Reuniões devem ser realizadas com os usuários, onde serão levantadas suas necessidades, definidas as fases do projeto e seus níveis de serviços e produtos liberados. O sucesso de um projeto de DW passa pela escolha dos usuários que definirão as necessidades de informação (BARBIERI, 2001).

### 2.3.1 Planejamento

Nesta fase, objetiva-se definir o escopo do projeto, voltando as atenções para as áreas críticas da empresa e para a necessidade de informações gerenciais (BARBIERI, 2001). O planejamento é a fase onde se define a área departamental e empresarial que será abordada no projeto, a redundância dos dados, assim como os usuários-alvo e a arquitetura tecnológica com a escolha da ferramenta SGBD, ferramentas de *On-Line Analytical Processing* (OLAP), ferramentas ETL e ferramentas para gestão de projetos.

### 2.3.2 Levantamento de Necessidades

A identificação do modelo dimensional é o objetivo desta fase do projeto, assim como a análise da qualidade e da integridade das fontes dos dados. Neste ponto, é importante observar que os dados a serem modelados deverão ser garimpados nos seus vários níveis de detalhe e sumarização. Outro modelo a ser identificado está relacionado com as fontes das informações. Nessa fase, deverão ser registrados os blocos conceituais, suas descrições e a forma de armazenamento, pois diferentes fontes podem ser usadas e relacionadas a diversos sistemas operacionais.

A integridade e a qualidade das fontes devem ser cuidadosamente observadas, além de sua duração histórica. Para isso, é importante a participação da equipe de tecnologia da informação, pois são os grandes conhecedores dos códigos e dados das fontes de informação que servirão de base para o modelo dimensional.

### 2.3.3 Modelagem Dimensional

Segundo Barbieri (2001), esta é uma fase crítica para o sucesso de um projeto de *Data Warehouse*. A primeira observação trata-se da fundamental importância do dado consolidado e/ou sumariado nas dimensões especificas, além dos dados com maior nível de detalhe ou granularidade. A Modelagem Dimensional deverá ser suportada por planilhas e possibilitar a análise dos volumes brutos, visando o processamento para obtenção de informações consolidadas. A definição das tabelas Fato e Dimensão, seus respectivos atributos, o nível de granularidade e agregação dos dados deverão ser realizados nesta fase.

### 2.3.4 Projeto Físico dos Bancos de Dados

Nesta etapa, segundo Barbieri (2001), com as definições de tabelas Fato e tabelas Dimensão, relacionamento, indexação, atributos de tabelas e implantação de regras, serão desenhadas as estruturas lógicas do modelo Dimensional. Desenhos

físicos das estruturas lógicas do modelo dimensional, a estimativa do tamanho da base de dados e, por fim, a criação da base é realizada na fase do projeto físico dos bancos de Dados.

### 2.3.5 Projeto de Transformação

Para Barbieri (2001), a definição dos processos de transformação do modelo Fonte para o modelo Dimensional deve ser executada nesta etapa do projeto. Os detalhes desse processo serão demonstrados no tópico “O processo de ETL”. Nesta fase, deve ser definido se o processo de transformação será realizado de forma manual ou utilizando-se de ferramentas especializadas.

### 2.3.6 Desenvolvimento de Aplicações

De acordo com Barbieri (2001), nesta fase deve ser desenvolvido o sistema aplicativo, devendo priorizar a interface *web,* por não ser necessária a instalação de *softwares* na máquina dos usuários, apenas a instalação no servidor e cuidados para autorização de acesso. A escolha da aplicação deve levar em conta as interfaces amistosas, geradores de relatórios, condições de visualização de múltiplas formas e importação dos dados para planilhas e processadores de texto. No mercado existem várias ferramentas dedicadas ao desenvolvimento de aplicações OLAP com todas as características recomendadas.

### 2.3.7 Validação e Teste

Nesta fase, o sistema deverá ser testado com o máximo de volume de dados a ser processado. Inicialmente, o sistema deverá ser liberado para um número de usuários restritos e, após análise dos resultados, liberar para o ambiente de produção (BARBIERI, 2001).

### 2.3.8 Treinamento

O grupo objeto de treinamento, segundo Barbieri (2001), deve ser formado priorizando os usuários voltados para o negócio e os gerentes das áreas envolvidas. Já para Lima (2010), deve ser garantido o treinamento da equipe técnica envolvida no projeto das habilidades necessárias para o sucesso. Revisar o papel de cada membro da equipe e treinar, se necessário, nas habilidades: gerenciamento de projetos, arquitetura de DW, modelagem de DW, análise de requisitos e ferramentas específicas.

### 2.3.9 Implantação

No quesito implantação, um rigoroso acompanhamento deve ser seguido no uso das aplicações disponibilizadas. Deverão ser incentivadas sugestões e críticas dos usuários pela equipe do projeto (BARBIERI, 2001).

De acordo com BARBIERI (2001), as informações necessárias para manter o controle sobre os dados armazenados nos metadados técnicos (quando descrevem as características físicas dos dados) e os negócios (que descrevem como usar esses dados), devem estar acessíveis para usuários finais em todos os momentos, pois permitem obter as informações necessárias para uso do sistema. Já para o administrador, possibilitam modificá-los conforme necessidade.

## 2.4 MODELAGEM DIMENSIONAL DE DADOS

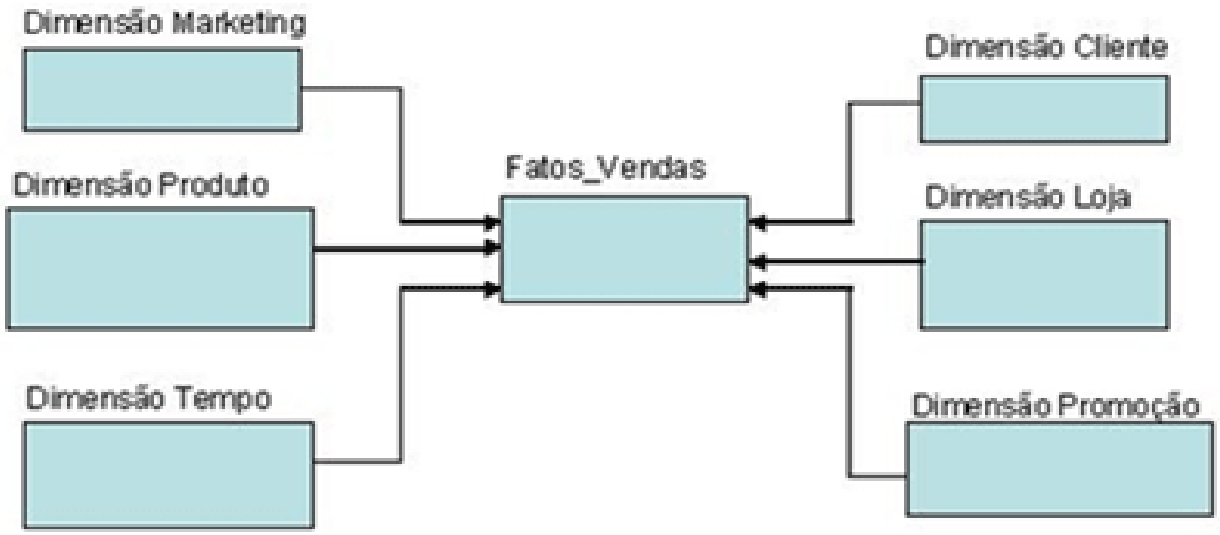
A definição da arquitetura do DW é uma tarefa determinante para o sucesso do projeto. No entanto, a utilização de um modelo de dados inadequado resultará no fracasso do projeto. A modelagem de dados tradicional, desenvolvida nos últimos anos, foi um instrumento de extrema utilidade e de melhor ajuste ao processamento transacional. Porém, este modelo se mostrou inadequado para atender às necessidades de negócio, no que diz respeito a tomada de decisão, pois apresentou imperfeições para os processamentos na ótica dimensional (BARBIERI, 2001).

A estrutura dimensional distribui os campos entre as tabelas de forma diferente, voltando sua estrutura para os pontos de entrada, as dimensões, e menos para os dados granulares, os fatos. Assim, os dados estarão dispostos numa estrutura que remete a uma estrela, apresentando uma estrutura mais sintética, legível e objetiva. Apesar de o modelo dimensional ser mais sintético do que o modelo relacional, na medida em que se forem agregando novas extensões pode se tornar mais complexo (BARBIERI, 2001).

Segundo Barbieri (2001), a produção de um modelo formado por tabelas Fato e tabelas Dimensão deve ser o produto final da modelagem dimensional. As tabelas Fato, como o nome sugere, armazenam vários fatos, ou seja, medidas numéricas associadas ao negócio. Uma ou mais medidas podem ser armazenadas em cada fato, sendo que constituem - valores objetos da análise dimensional. Os dados persistidos nestas tabelas são relativamente estáticos, uma vez que sofrem poucas alterações e são normalmente dados aditivos, manipulados por soma, multiplicação ou outras operações matemáticas. Já as tabelas Dimensão constituem as estruturas de entrada, representam entidades do negócio, como tempo, cliente, geografia, produto etc. Devem ser entendidas como filtros que serão aplicados na manipulação dos fatos. As tabelas Dimensão possuem um número significativo menor de linhas em relação a tabela Fato e sua relação com essa é de 1:N. Suas colunas representam informações e algumas podem também representar sua hierarquia.

Um modelo dimensional conta basicamente com uma tabela de fatos central e tabelas dimensionais ligadas diretamente a elas.

**Figura 3:** Modelo dimensional




Fonte: Elaborado pelo autor

## 2.5 DATA MART

Segundo Barbieri (2001), Data Mart (DM) é um depósito de dados consolidados por assuntos/negócios. Para Inmon (1997), “um data mart é uma coleção de assuntos organizados para o suporte de decisões baseado nas necessidades de um departamento”. Um DM consolida apenas as informações de uma determinada área, diversos DMs podem se unir, formando um único DW.Para os Data Marts, utiliza-se o modelo dimensional (específico para a teoria de DW).

## 2.6 ABORDAGEM DE RALPH KIMBALL – STAR SCHEMA

De acordo com Barbieri (2001), a abordagem de Ralph Kimball (1998) vem de um modelo incremental. Na metodologia *Star Schema*, os DM serão integrados, à medida que o projeto evolui, transformando os dados em tabelas Fatos e em tabelas Dimensões. Com isso, projetos menores e orientados por assunto ou área serão conectados com o passar do tempo, desde que mantida a compatibilidade entre os campos chaves das tabelas.

Esta metodologia traz a desvantagem de possibilitar a produção de vários DM, sem uma perfeita coesão entre estes, além de retrabalho na fase de extração, transformação e carga dos dados, conhecida como ETL.

Centrada na modelagem dimensional, a essência da abordagem de Kimball está na etapa de projeto dos DM, onde se concentram os dados de interesse, passível de manipulação numérica e estatística em tabelas Fatos, e informações de característica descritiva nas tabelas Dimensão.

Ainda segundo Barbieri (2001), a abordagem de Ralph Kimball está sendo aplicado em vários projetos de DW nos EUA, com ênfase em empresas públicas e em grandes empresas de varejo.

## 2.7 O PROCESSO DE EXTRAÇÃO TRANSFORMAÇÃO E CARGA DOS DADOS

O processo de Extraction, Transformation and Load (ETL), traduzido para extração, transformação e carga, é definida como a coleta, a limpeza e a

transformação de dados de várias fontes e a inserção desses dados em um DW. Esta é a etapa do projeto de DW responsável pela carga de dados e de sua atualização a partir dos dados operacionais. Este processo se resume em extrair os dados das fontes operacionais, transformá-los para o formato de dados do DW, integrá-lo conforme especificações dos metadados e carregá-los no *Data Warehouse*. O processo ETL é uma das fases mais complexas na construção de um sistema DW, pois é nesta fase que grandes volumes de dados são processados (KIMBALL, 1998; INMON, 1997).

De acordo com Lima (2010), esta é uma fase crítica do projeto de DW, por envolver movimentação de dados, mesmo existindo várias ferramentas que auxiliem na execução do trabalho. Esta fase se divide em duas partes, o primeiro passo é definir as fontes de dados e realizar a extração deles. O segundo passo é transformar e limpar os dados, ou seja, eliminar as inconsistências entre eles e garantir a compatibilidade. Além da limpeza, a transformação dos dados é um fator de fundamental importância, pois têm origem de diferentes fontes e podem conter dados iguais com formatos diferentes. Por exemplo, o campo sexo pode ser armazenado como “M” ou “Masculino” quando se refere ao gênero masculino.

Segundo Inmon (1997), o processo de ETL pode parecer um processo simples, dando a impressão de uma programação manual de extrações de dados de um local para outro. Mas não é isso o que acontece geralmente. Ao começar a realizar o serviço manual de extração, os desenvolvedores do *Data Warehouse* se deram conta de que esta tarefa é muito maior e mais complexa do que parecia em um primeiro momento.

O pesquisador Inmon (1997) cita algumas funcionalidades que configuram o ETL como uma tarefa árdua e complexa, como:

- Há uma mudança na tecnologia de dados, é necessário ler os dados no SGBD herdado e gravá-los em um novo SGBD capaz de suportar o *Data Warehouse.*
- Os dados precisam ser reformatados:
    - Datas e medidas necessitam de um padrão. Exemplo: as datas em um banco de dados operacional estão no formato AA/MM/DD e em outro BD estão MM/DD/AA. No *Data Warehouse* foi adotado o padrão DD/MM/AAAA, então, todos os dados precisam ser reformatados para

entrar no DW. Essa tarefa, executada manualmente, se torna muito mais complexa por não ser realizada com conversores automáticos;

- O formato lógico dos dados precisa ser convertido. Formatos de texto EBCDIC para ASCII também devem ser convertidos;

- As chaves dos dados geralmente necessitam ser reestruturadas. Em casos mais simples, é apenas uma questão de adicionar a chave da dimensão tempo. Em casos mais complexos, as chaves nos diversos sistemas herdados são diferentes e precisam ser reestruturadas;
- A seleção dos dados que serão carregados no DW é extremamente complexa, pois podem existem várias fontes para os mesmos dados;
- É necessário eliminar a redundância dos dados. Um mesmo dado é cadastrado em várias aplicações diferentes, por exemplo. Ao ser carregado no DW, esse dado deverá ser único;
- O relacionamento entre os dados em programas herdados não são documentados ou são colocados em uma lógica extremamente otimizada para tais programas. Inmon (1997) coloca este problema como um dos piores a serem resolvidos, pois desvendar os dados atrelados dessa forma é um passo muito demorado e custoso;
- A grande variedade de fontes de dados. Em consequência disso, há igual número de formatos e relacionamentos de dados a serem trabalhados;
- Correção de dados. Quando não confiáveis, os dados de fontes operacionais devem ser corrigidos para que não existam dúvidas quanto à veracidade das informações do *Data Warehouse*. Um simples algoritmo de correção pode resolver, mas talvez sejam necessárias ferramentas de inteligência artificial para tratar tais dados.

De acordo com Lima (2010), deve-se analisar alguns fatores antes do início da fase de extração dos dados, que são apresentados a seguir:

- A extração dos dados do ambiente de origem para o ambiente de *Data Warehouse* demanda mudança de tecnologia, pois os dados são transferidos da bases de dados hierárquicos, uma estrutura relacional, para *Data Warehouse*;

- Pode ser complexa a seleção dos dados, pois vários campos podem ser selecionados dos sistemas transacionais para compor um único campo do DW;
- Os dados devem ser reformatados;
- Deve-se escolher as chaves antes que os arquivos sejam intercalados, quando existirem vários arquivos de entrada. Isso significa que se os diferentes arquivos possuem diferentes estruturas de chaves, deve-se optar por apenas uma chave;
- Os arquivos gerados devem obedecer a ordem das colunas no ambiente de DW;
- Para os campos que não possuem fontes de dados, devem ser fornecidos valores *default.*

Mesmo existindo diversas ferramentas de ETL no mercado, ainda existe a necessidade de criar rotinas de carga como *shell script*, SQL puro ou usando linguagem de programação para atender determinadas situações que poderão ocorrer (LIMA, 2010).

De acordo com Lima (2010), as ferramentas de ETL mais utilizadas no mercado são oferecidas pelos fornecedores IBM, ETI Corporation, Informática, Microsoft, Pentaho (Open Source), Talend (Open Source), Sunopsis e Oracle. Todas possuem seus diferenciais e a escolha deve ser feita de acordo com o caso de cada empresa.

As ferramentas de ETL possuem grande relevância, pois são um poderoso instrumento na geração de metadados e contribuem para a produtividade da equipe. Os benefícios desta ferramenta serão bastante vistosos e o aumento de produtividade é considerável (LIMA, 2010).

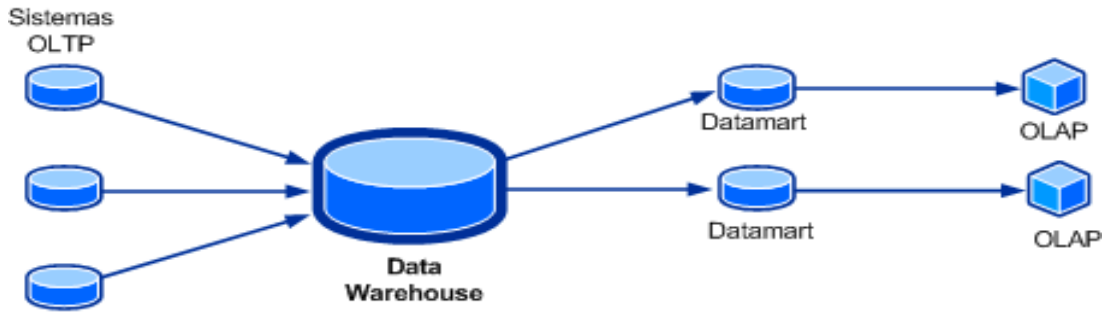
## 2.8 PROCESSAMENTO DE DADOS

Os dados armazenados no DW/DM, segundo Barbieri (2001), estão diretamente ligados à tomada de decisão, visando vantagens competitivas da empresa e suportando as aplicações baseado em On-Line Analytical Processing (OLAP), traduzido em Processamento Analítico On-Line, que é definido por

Brackette (1996), como “processamento que suporta a análise da tendência e projeções do negócio. É também conhecido como processamento de suporte a decisão”.

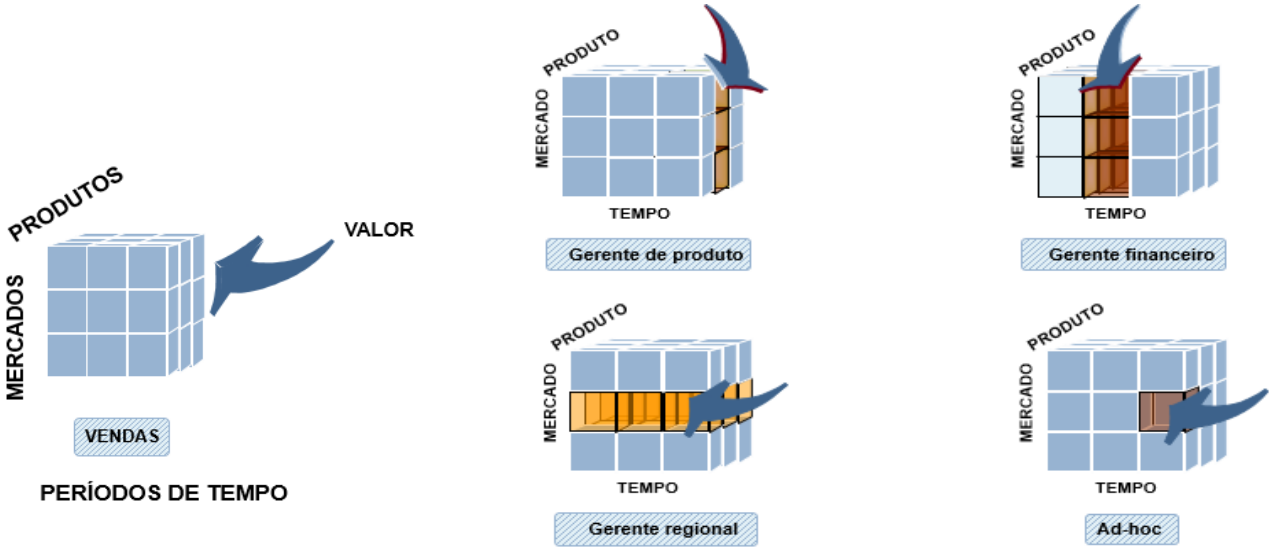
Assim, as atividades de consulta e de apresentação dos dados armazenados em DW/DM são feitas através das ferramentas OLAP e aplicações sobre o negócio, pois analisam as informações, auxiliando os tomadores de decisão a interpretar mudanças na realidade do negócio. Assim, podendo inferir sobre a estratégia e o gerenciamento da empresa (BRACKETT, 1996). A figura 4 apresenta uma arquitetura simples, utilizando DW/DM e OLAP. Já a figura 5 apresenta um exemplo do modelo dimensional em que as aplicações OLAP trabalham.

**FIGURA 4:** Arquitetura OLAP




Fonte: Elaborado pelo autor

**FIGURA 5:** Modelo Multidimensional




Fonte: Elaborado pelo autor

# 3. METODOLOGIA

Trata-se de um estudo de caso realizado de forma descritiva e aplicada, elaborado a partir do projeto de *Data Warehouse,* desenvolvido em uma empresa de Tecnologia *mobile* que atua no mercado Sul-Americano. Segundo Goode e Hatt (1979), por meio do estudo de caso se pretende investigar, como uma unidade, as características importantes para o objetivo da pesquisa. O estudo de caso é um meio de organizar os dados, preservando do objeto estudado o seu caráter unitário. Yin (2001) explica que a escolha do método de pesquisa,estudo de caso leva em consideração se a questão principal do trabalho se refere ao como e ao porquê da investigação.

A coleta de dados foi realizada através de observação participante, pois o autor fez parte da equipe como analista e desenvolvedor. Documentos e artefatos físicos também foram utilizados como fontes de dados para base de conhecimento da pesquisa.

O projeto desenvolvido relaciona-se com a Gestão de Dados da organização, que teve a necessidade de criar um banco de dados centralizado, de forma coerente, para se adaptar aos diferentes requisitos do negócio. Tal organização produz grande volume de dados, diariamente, por seus diversos sistemas transacionais para controlar o acesso e utilização de seus serviços, assim como as vendas diárias. Essas aplicações são abastecidas por diversas outras fontes de dados, como arquivos de textos, planilhas eletrônicas e outros bancos de dados.

Na eminente necessidade desta organização em aprimorar seus instrumentos de análise, foram extraídas as informações de seus repositórios de dados para um ambiente de *Data Warehouse* com o objetivo de sanar esta necessidade. O projeto se iniciou com a criação de uma área de extração, uma base de dados intermediária ao DW com dados íntegros, não redundantes e imprescindíveis para as análises, que possuem o intuito de facilitar o tratamento e o armazenamento dos dados no DW. Após a criação dessa base, foi desenvolvido a base de dados do DW, com dados históricos, baseados no negócio e estruturados para relatórios e análises. Para transformação e carga dos dados na base *staging* e no DW, foi utilizada a ferramenta de ETL. Por fim, foram criadas as visões analíticas, utilizando-se da ferramenta OLAP. O projeto será detalhado no capítulo denominado Estudo de Caso.

A partir deste projeto, os analistas e os gerentes de negócios, a área de produtos e a diretoria obtiveram um instrumento de análise das vendas e da utilização dos serviços em relatórios estáticos e dinâmicos, permitindo uma visão geral das tendências do negócio, assim como o cruzamento de informações e análise histórica, proporcionada pelo ambiente de DW.

## 3.1 ESTUDO DE CASO

### 3.1.1 Projeto Proposto

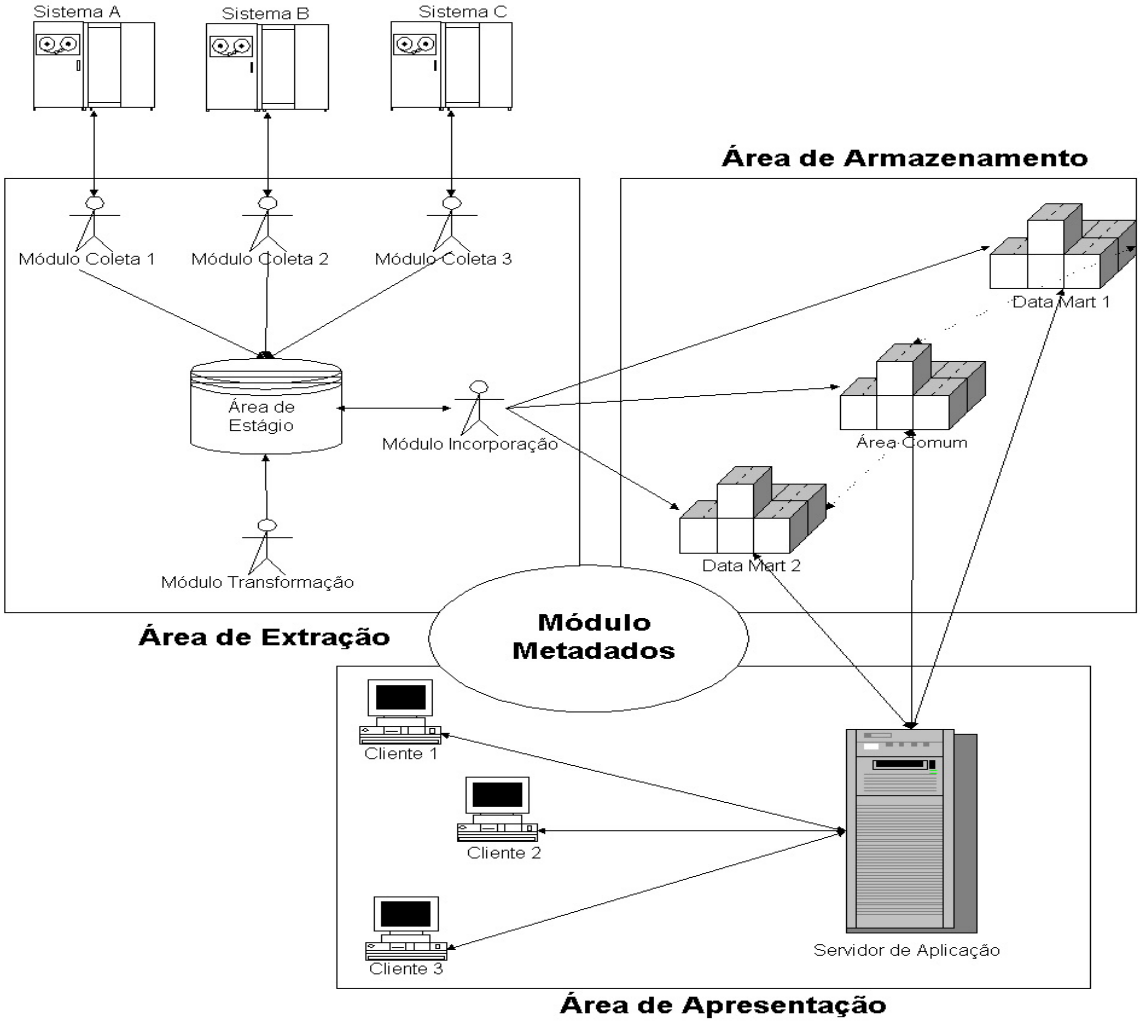
Este capítulo tem por objetivo detalhar como foi o desenvolvimento do projeto de *Data Warehouse*. A motivação em desenvolver o projeto foi a produção de grande volume de dados por diversos sistemas. Logo, obtiveram informações descentralizadas, não íntegras, de difícil e demorado acesso. Assim, o projeto foi desenvolvido em busca de melhoria para o processo de produção de informação, orientada para o armazenamento, a gestão e a disseminação da informação, que representa um estoque potencial de conhecimento, sendo imprescindível para a transferência de conhecimento.

A organização usava sistemas que geravam "toneladas" de relatórios e de dados sem direção, deixando a gerência sem capacidade de usar esses dados e relatórios estrategicamente. Os departamentos geralmente não tinham sucesso em compartilhar informações ou realizavam devagar, pois os relatórios eram disponibilizados de forma tardia. Além disso, sistemas de relatórios sobrepostos forneciam dados, às vezes, sem assertividade. Com isso, a gerência tinha dificuldade em tomar decisões em tempo hábil.

Como solução, a diretoria de Tecnologia da Informação iniciou uma tentativa de identificar os problemas. Foi assim que concluíram a necessidade da implantação do *Data Warehouse*, que como foi abordado anteriormente, é um repositório central de dados históricos, organizado de forma para que o acesso seja fácil e a manipulação para o suporte a decisões seja conveniente. Também se tornou claro que ferramentas de *software* para efetuar o processamento, a exploração, e a manipulação de dados seriam necessárias. Foi configurado, então, um sistema para fornecer dados precisos e em tempo real.

O sistema também incluía o recurso de *Dashboard*, que permite aos executivos visualizarem as áreas que merecem atenção em suas unidades de negócio e a sua investigação para identificar os problemas com exatidão, bem como as suas causas. Usando cores diferentes (por exemplo, o vermelho para perigo), um gerente de negócios pode ver em tempo real, por exemplo, quando o tempo de entrega está abaixo do esperado e encontrar, imediatamente, a origem do problema e até mesmo avaliar potenciais soluções. A figura 6 ilustra a arquitetura proposta como solução utilizando ambiente de DW.

**FIGURA 6:** Componentes da Arquitetura Proposta




Fonte: Elaborado pelo autor

Devido ao grande volume de dados e de transações realizadas, na primeira etapa do projeto foi proposta a criação de um *Staging Area* [1](área de extração), uma base de dados relacional para armazenar os dados operacionais existentes nos diversos repositórios da organização. Segundo Kimball (2002), este recurso é responsável por receber as informações dos sistemas transacionais, para então gerar os *Data Marts* de destino. De acordo com Lima (2010), existem casos em que não seria possível processar as extrações e transformações durante a utilização do DW. Então, é recomendado utilizar a chamada *Staging Area* para obter sucesso na execução do processo de ETL. Diante disso, o objetivo desta etapa é criar uma base centralizada, íntegra, padronizada, livre de dados redundantes e não relevantes para os tomadores de decisão.

Em seguida, já com os dados persistidos na área de extração, foi criada a área de armazenamento DW. Este será carregado a partir dos dados armazenados na área de extração. Porém, apenas as informações relevantes ao negócio serão extraídas e carregadas. De acordo com Barbieri (2001), DW trata-se de um banco de dados destinado aos sistemas de apoio à decisão. Os dados serão armazenados em estrutura lógica dimensional, que possibilita o seu processamento por ferramentas OLAP.

Todo o processo de extração, de transformação e de carga dos dados executados dos sistemas de origem para a *Staging Area* foram realizados utilizando ferramentas ETL, assim como da *Staging Area* para o DW, conforme sugere Inmon (1997). Por fim, para aperfeiçoar o processo de consulta e visualização dos dados são utilizadas ferramentas OLAP. Segundo Brackett (1996), estas ferramentas suportam a análise de tendências e projeções do negócio.

### 3.1.2 Ferramentas Utilizadas

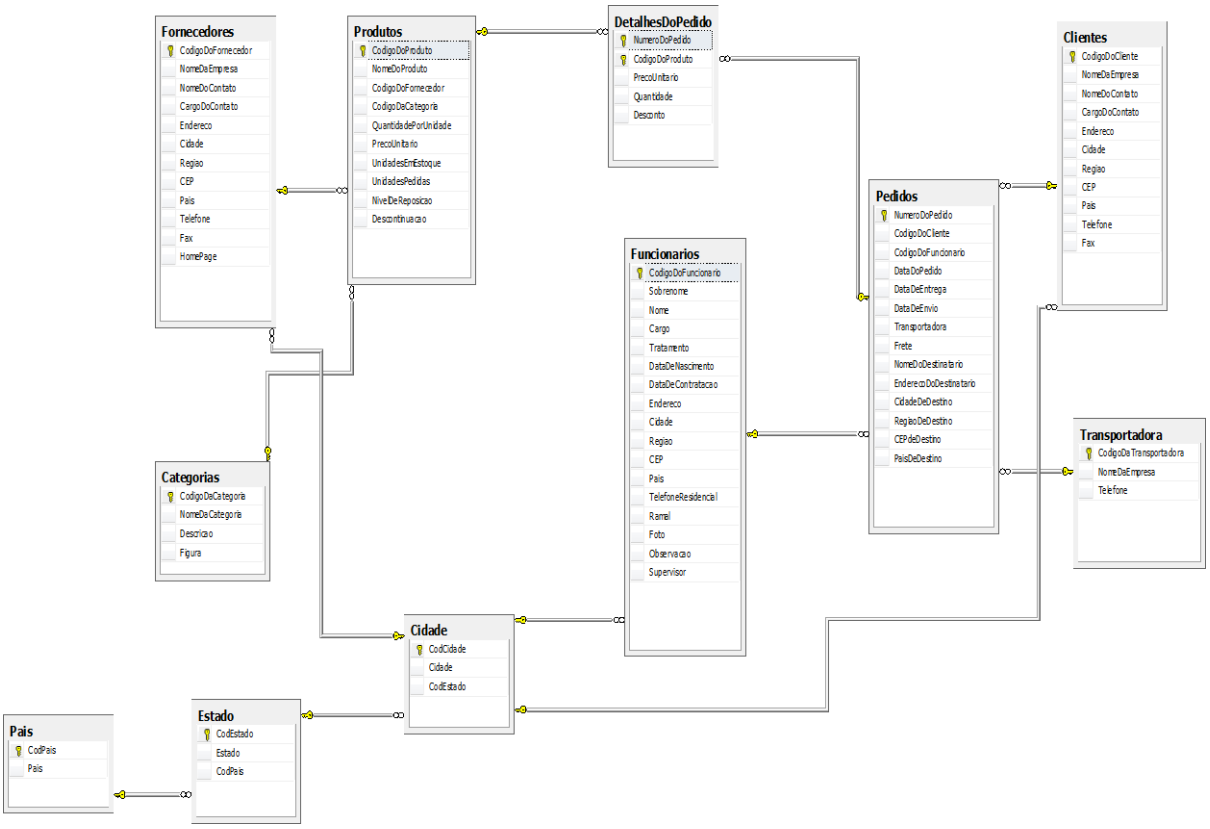
A empresa é parceira Microsoft e utiliza-se de seus produtos e serviços há mais de 10 anos. Além disso, seus funcionários já possuem conhecimento nas ferramentas. Por isso, foi decidido a utilização da ferramenta *Microsoft SQL Server* e seus componentes *Integration Services, Analysis Services e Reporting Services*.

[1] "Área de extração" é um recurso utilizado para transformação dos dados antes de carregados no DW, nos casos em que não é possível realizar as transformações durante a carga do DW.

### 3.1.3 Desenvolvimento do Projeto

O projeto iniciou-se com a criação da *Staging Area*, uma base de dados relacional criada no SGBD SQL Server, destinada ao processamento de extração e transformação das informações à parte do DW. O objetivo ao utilizar a Staging Área é eliminar as diversas fontes de dados existentes na organização, entre elas, planilhas eletrônicas, diferentes bancos de dados e arquivos texto. Além disso, também tem como objetivo transformar e padronizar os dados por meio do processo de ETL em uma janela de tempo na qual o DW não está sendo utilizado, conforme sugere LIMA (2010). A Figura 7 ilustra o modelo relacional da *Staging Area.*

**FIGURA 7:** Modelo relacional da *Staging Area*

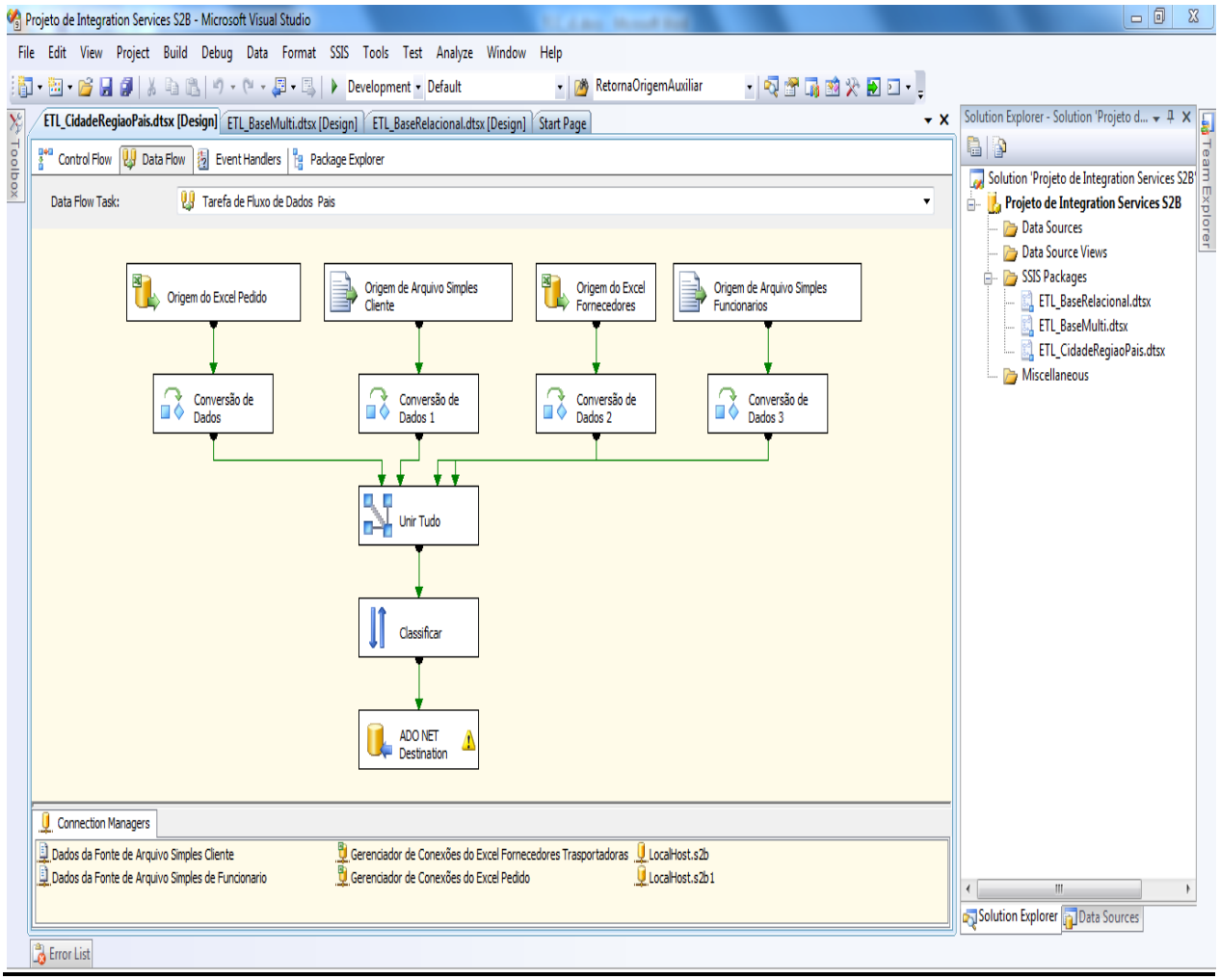



Fonte: Elaborado pelo autor

Concluída a criação da *Staging Area*, foi iniciada a extração e a transformação dos dados das diversas fontes e carga dos mesmos na base *Staging* utilizando o processo de ETL. Conforme Inmon (1997) sugere, foi utilizado uma

ferramenta de ETL para executar esse processo, por ser uma das tarefas mais árduas e complexas de um projeto de DW. A ferramenta utilizada foi a SQL Server Integration Services (SSIS), conforme razões já mencionadas anteriormente. A Figura 8 ilustra um dos mais de 50 processos de ETL desenvolvidos.

**FIGURA 8:** Exemplo do processo ETL utilizando a ferramenta SSIS




Fonte: Elaborado pelo autor

Conforme LIMA (2010) sugere, foi dividido o processo de ETL em duas partes. Primeiro, foram definidas as fontes de dados que seriam extraídas. Em seguida, foram analisados os campos que seriam transformados e limpos.

Durante o processo de ETL, foram convertidos, padronizados e eliminados os dados duplicados entre as fontes existentes. Nesta fase, procurou-se garantir que toda a informação estava correta e consistente, para que dados incorretos não

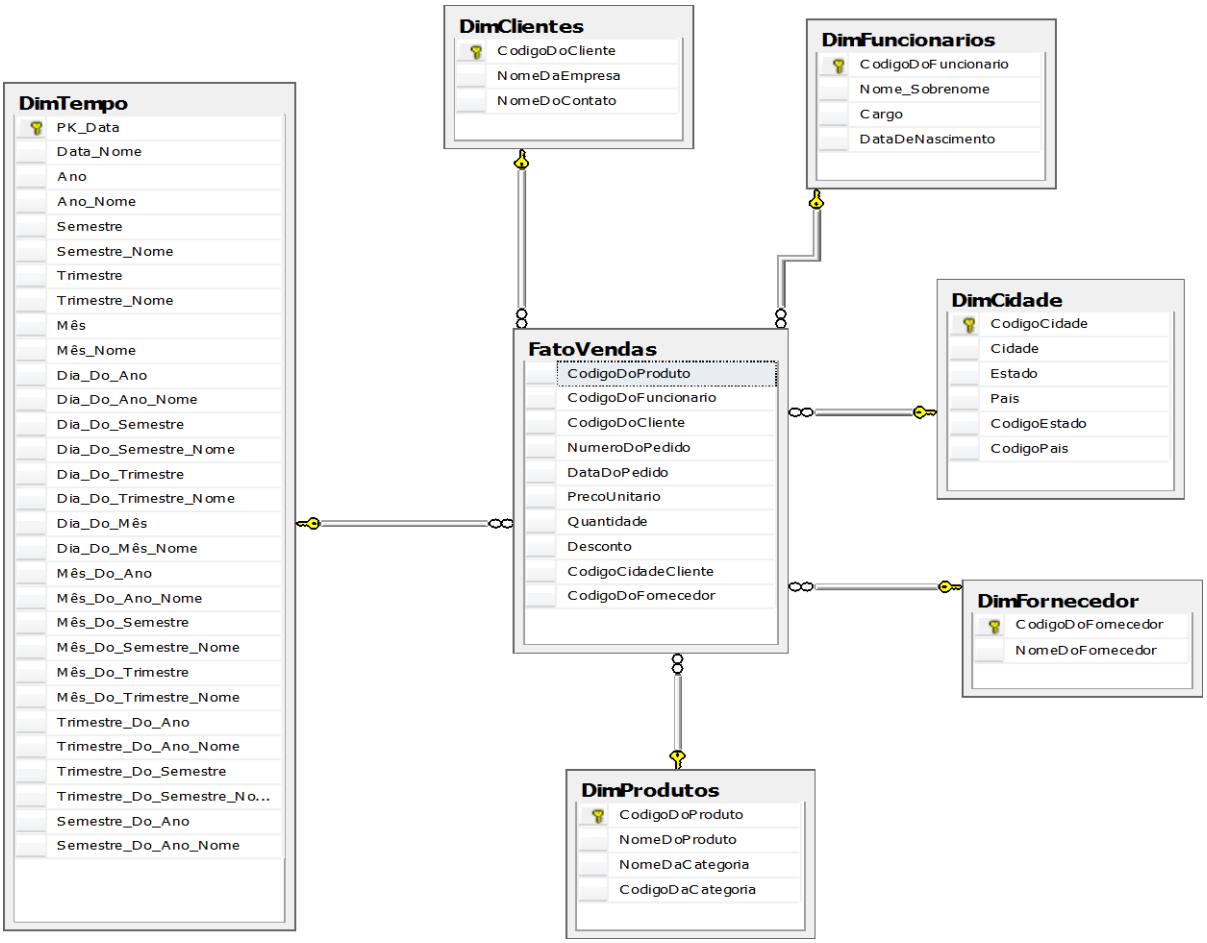
pudessem conduzir erros críticos de tomada de decisão. Dada a importância de detecção de erro, são explicitados alguns objetivos estabelecidos para o processo ETL:

- Comparar o número de registros entre os dados das fontes e o número de registros carregados para o DW;
- Comparar os valores únicos de determinados atributos entre as fontes e os dados carregados para o DW;
- Fazer um esquema de dados para perceber as limitações dos valores atribuídos;
- Validar os conteúdos de cada atributo, ou seja, não permitir por razões de codificação, o limite de caracteres entre cada esquema relacional (fonte e destino) não resulta na falha do fluxo de dados;
- Transformação de Dados - Nesta fase, o objetivo é assegurar que os dados são transformados corretamente, de acordo com as regras de negócio especificadas;
- Validar o processamento correto de campos no ETL, tais como chaves estrangeiras;
- Verificar sempre se os tipos de dados presentes na base são os esperados;
- Garantir a qualidade dos dados, o processo ETL rejeita ou substitui valores por defeito, corrige ou ignora dados e reporta dados inválidos;
- Realizar as conversões dos dados corretamente;
- Nos casos de atributos NULL, procura-se sempre inserir valores equivalentes a "não identificado";
- Identificar valores duplicados e corrigir o problema;
- Os carregamentos são efetuados com volumes de dados pequenos para garantir o bom funcionamento;
- Efetuar operações simples com junções para validar a performance em volumes de dados muito grandes.

Com os dados carregados na *Staging Area*, foi realizada a modelagem Dimensional para o DW, de acordo com a abordagem de Ralph Kimball, modelo *Star Schema*. A Figura 9 ilustra o modelo Dimensional desenvolvido. Para o desenvolvimento do modelo Dimensional, foram levantadas as entidades, os fatos e

as métricas que fazem parte do negócio da empresa. A partir desse levantamento, foram definidas as tabelas Dimensão, a tabela Fato e as métricas que seriam analisadas. Segundo Barbieri (2001), a modelagem dimensional proporciona aos usuários compreender os dados numa forma mais próxima de seu entendimento, proporcionando maior interação dos usuários com o projeto.

**FIGURA 9:** Modelo dimensional: esquema estrela do *Data Warehouse*

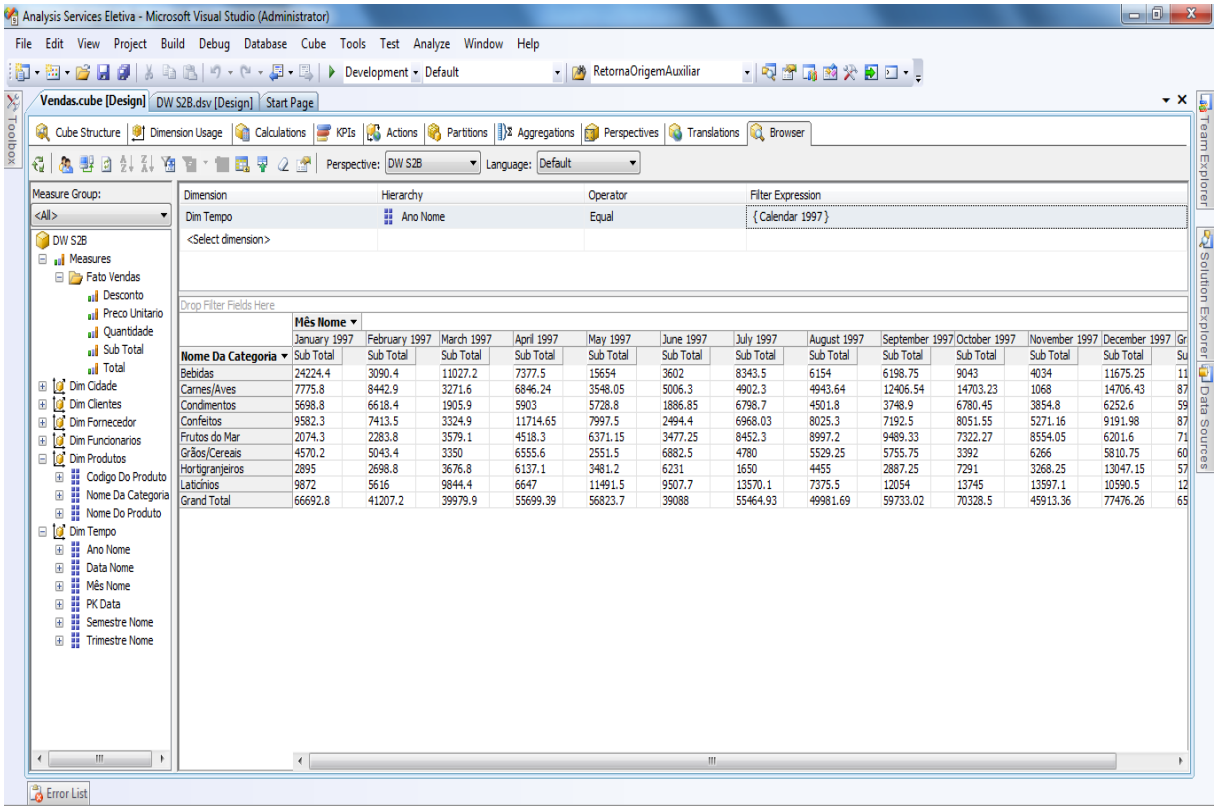



Fonte: Elaborado pelo autor

Após a modelagem, foi criada a base de dados *Data Warehouse* no SGBD SQL Server. Esta base DW foi carregada com os dados da *Staging Area*, por meio de um segundo processo de ETL, novamente utilizando a ferramenta *Integration Services*. Durante o processo de ETL, foram filtrados os dados inúteis, ou seja, configurados os dados relevantes para as necessidades de análise com objetivo de facilitar o acesso e consulta aos dados da organização.

Após carregar os dados no ambiente de *Data Warehouse* foi utilizada a ferramenta Analysis Services, uma ferramenta OLAP, capaz de criar visões de múltiplas perspectivas, denominadas Cubos, a fim de proporcionar uma melhor análise dos dados de vendas e verificar tendências do negócio. A Figura 10 ilustra a utilização da ferramenta *Analysis Services.*

**FIGURA 10:** Exemplo da ferramenta *Analysis Services*




Fonte: Elaborado pelo autor

Ao utilizar o projeto, os executivos e os gerentes da organização obtiveram um sistema de ferramentas para apoiá-los no processo de tomada de decisão. Esses profissionais consideram tendências extraídas a partir dos dados armazenados no *Data Warehouse* e consultados utilizando as ferramentas OLAP, podendo, assim, determinar correlações não óbvias entre dados antes não observados.

A gestão da informação, utilizando o DW, ao proporcionar a disseminação e possuir potencial para gerar conhecimento através do estoque informacional, se

mostrou muito importante na obtenção de vantagens competitivas no aperfeiçoamento de produtos e serviços para os gerentes da organização. A figura 11 exemplifica a utilização dos painéis de informação que disponibilizam os dados armazenados no DW.

**FIGURA 11:** Exemplo de visualização dos dados a partir de um DW


Fonte: Elaborado pelo autor

Sempre existiram dados, informações e conhecimentos na organização, o que há de novo nesse processo é a gestão do estoque informacional a partir do projeto de DW desenvolvido. Saber gerir, armazenar, disseminar e aplicar o conhecimento é fundamental para obter sustentabilidade e vantagem competitiva, por meio de melhorias e inovações. Por isso, é necessário que gestores estejam abertos à criação de um ambiente favorável ao apoio de práticas que levem à formação de conhecimento como capital intelectual ativo nas instituições.

# 4. RESULTADOS

O projeto apresentou resultados fascinantes, em poucos dias, na área de gestão e disseminação da informação. Os sistemas implantados ajudaram a definir o verdadeiro retorno dos investimentos aplicados em canais de divulgação na mídia. Com isso, observou-se que algumas mídias não apresentavam o retorno imaginado. Outro resultado relevante foi a melhoria no processo de repasse financeiro aos parceiros comerciais. Eram necessários três funcionários no período de cinco dias para executar esse processo e, após a implantação do projeto, um funcionário executa a atividade em um dia. No geral, a organização conseguiu aumentar a eficiência de seus produtos e processos.

Após o retorno positivo do projeto, muitos setores adotaram as ferramentas e os processos de gestão da informação. Os gerentes e diretores passaram a utilizar painéis em seus escritórios para acompanhar os indicadores da empresa (gerenciar despesas, compras etc) e a equipe de suporte e operações passaram a utilizar painéis que acompanham o desempenho em tempo real dos produtos (assinaturas, cancelamentos, tarifações etc). O acesso às informações estratégicas demonstrou que as ferramentas de análise de dados podem melhorar a disseminação de conhecimento e de resultados na tomada de decisão.

Outros benefícios foram a redução do custo em armazenamento de dados, a redução dos incidentes nos sistemas de informação, além da melhora ao entender o comportamento dos clientes desta organização. O projeto de gerenciamento e disponibilização de informação ajudou a empresa de Tecnologia móbile a alcançar altas margens de lucro no mercado. Além disso, a participação no mercado está aumentando consideravelmente, devido o conhecimento dos tomadores de decisão, adquirido por meio das informações estratégicas disponibilizadas em tempo oportuno.

## REFERÊNCIAS

BARBIERI, Carlos. **BI – Business intelligence**: modelagem e tecnologia. Rio de Janeiro: Axcel Books, 2001.

BARRETO, Aldo de Albuquerque. **A questão da informação**. Revista São Paulo em Perspectiva, Fundação Seade, v 8, n4, 1994.

BRACKETT, Michael H. **The Data Warehouse challenge**: taming data chaos. USA: Wiley, 1996.

CHOO, Chun Wei. **Information management for the intelligent organization**: the art of scanning the environment. Medford, NJ: Information Today Inc., 2002.

Goode WJ; Hatt PK. **Métodos em pesquisa social**. 5. ed. São Paulo: Companhia Editora Nacional, 1979.

Yin R. **Estudo de caso: planejamento e métodos**. 2. ed. Porto Alegre: Bookman, 2001.

INMON, William H., **Como construir o Data Warehouse**. 2. ed. Rio de Janeiro: Campus, 1997.

KIMBALL, Ralph. **Data Warehouse Toolkit:** técnicas para construção de Data Warehouses Dimensionais. São Paulo: MAKRON Books, 1998.

KIMBALL, Ralph. **The Data Warehouse Toolkit:** the complete guide to dimensional modeling (second edition), Wiley, 2002.

LIMA, Carlos Alberto Lorenzi. Disponível em: http://litolima.wordpress.com/2010/01/13/etl-extracao-transformacao-e-carga-de-dados/. Acessa em: 20 nov. 2014.

LIMA, Carlos Alberto Lorenzi. Disponível em: http://litolima.com/2010/01/20/gerenciando-um-projeto-de-dw-bi/. Acesso em: 20 nov. 2014.