UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

RAFAEL GASTÃO COIMBRA FERREIRA

Data Warehouse na Prática: Fundamentos e Implantação

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Marcelo Soares Pimenta
Orientador

Porto Alegre, abril de 2002.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Ferreira, Rafael Gastão Coimbra

Normas para Apresentação de Dissertações do Instituto de Informática e do PPGC / Rafael Gastão Coimbra Ferreira – Porto Alegre: Programa de Pós-Graduação em Computação, 2002

71 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2002 Orientador: Marcelo Soares Pimenta.

1.Data WareHouse. 2.ETL 3.Metadados. 4. OLTP. 5.OLAP I. Pimenta, Marcelo Soares. II. Titulo.

AGRADECIMENTOS

Gostaria de agradecer a Deus,

à minha querida esposa,

à minha família, em especial a meus pais,

à meus amigos,

à meu orientador,

enfim, a todos aqueles que de alguma forma me ajudaram na realização deste trabalho.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

SSD Sistemas de Suporte a Decisão
BD Banco de Dados
DW Data Warehouse
OLE Object Linking and Embedding
OLAP On-Line Analytical Processing
OLTP On-Line Transation Processing
SGBD Sistema Gerenciador em Banco de Dados
SGBDM Sistema Gerenciador em Banco de Dados Multidimensional
SGBDR Sistema Gerenciador em Banco de Dados Relacional
SIG Sistema de Informação Gerencial
SQL Structure Query Language
SAD Sistema de Apoio a Decisão
DTS Serviço de Transformação dos Dados

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Embora o conceito de Data Warehouse (doravante abreviado DW), em suas várias formas, continue atraindo interesse, muitos projetos de DW não estão gerando os benefícios esperados e muitos estão provando ser excessivamente caro de desenvolver e manter.

O presente trabalho visa organizar os conceitos de DW através de uma revisão bibliográfica, discutindo seu real benefício e também de como perceber este benefício a um custo que é aceitável ao empreendimento. Em particular são analisadas metodologias que servirão de embasamento para a proposta de uma metodologia de projeto de DW, que será aplicada a um estudo de caso real para a Cia Zaffari, levando em conta critérios que são encontrados atualmente no desenvolvimento de um Data Warehouse, um subconjunto das quais será tratado no trabalho de dissertação.

Palavras-Chave: data warehouse, banco de dados, OLAP.

Date Warehouse in Practice: Foundations and Implementation

ABSTRACT

Although the concept of Data Warehouse (DW), in its various forms, still attracting interest, many DW projects are not generating the benefits expected and many are proving to be too expensive to develop and to keep.

This work organizes the concepts of DW through a literature review, discussing its real benefit and how to realize this benefit at a cost that is acceptable to the company. In particular methods are discussed to serve as a foundation for proposing a design methodology for DW, which will be applied to a real case study for the CIA Zaffari, taking into account criteria that are currently found in developing a data warehouse, a subset of which will be treated in the dissertation.

Keywords: data warehouse, database, OLAP.

1 INTRODUÇÃO

Desde que as organizações passaram a dar a devida importância a bancos de dados relacionais, todo o hardware e software, as metodologias e ferramentas, se concentraram em mover grande volume de dados para os computadores, de forma tão rápida, segura e barata quanto possível. As empresas passaram por uma fase de processamento de dados, onde os computadores, conhecidos por mainframes (grandes computadores), processavam um volume muito grande de dados. A partir da década de 90, as empresas passaram a considerar que não só os dados processados eram valiosos mas também a informação. A partir dela, se buscava atingir os principais objetivos da corporação. Muitos sistemas de apoio a decisão surgiram para se poder manipular tais informações. Estávamos entrando na era da informação, configurando o que se convencionou denominar de "Era da Informação.

Atualmente as empresas estão passando por uma nova era, onde se procura não apenas gerar um volume enorme de informações, mas aprender a trabalhar com elas, obtendo as respostas necessárias. Por isto este período é denominado "Era do conhecimento", caracterizada pela capacidade que as pessoas tem em ver e conhecer o mundo via informação.

Para [GEN 98] os especialistas da informática tem estimado que somente uma pequena fração dos dados processados em uma empresa está realmente disponível para usuários tomadores de decisões, o que nos leva a conclusão de que as empresas são ricas em grande volume de dados, mas pobres em dados realmente úteis. Para resolver esta questão as organizações estão implantando a tecnologia de Data Warehouse (DW) , provendo as pessoas chaves dentro da empresa com acesso a todo tipo de informação necessária para a empresa sobreviver e prosperar no mercado competitivo.

Por exemplo, numa rede de supermercados, o setor de marketing está preocupado em saber qual é a faixa etária de seus clientes e os produtos mais consumidos por estes. Os dados utilizados para gerar estas informações podem ser extraídos de uma fonte única ou múltipla, permitindo análises e correlações entre as vendas e clientes.

Por definição, um DW é um banco de dados com informações integradas através da coleção de um ou mais sistemas que se tornam a base para a tomada de decisão [POE 98] . Um projeto em DW permite a obtenção de dados (informações) gerenciais a partir de um grande volume de dados operacionais. Organizações que buscam melhorar a habilidade na tomada de suas decisões podem ser prejudicadas pelo grande volume e complexidade de dados disponíveis em seus sistemas de produção operacional. Fazer estes dados acessíveis é um dos desafios mais significantes para os profissionais de informática de hoje. Com respeito a este desafio, muitas organizações escolhem construir um Data Warehouse para elaborar informações a partir dos seus dados operacionais.

1.1 Proposta do Trabalho

O objetivo deste trabalho é organizar e discutir os conceitos de Data Warehouse e de sua real utilização, abordando desde tópicos gerenciais como o real benefício de DW e de como perceber este benefício a um custo que é aceitável ao empreendimento até tópicos tecnológicos envolvidos na construção de DW. Serão aplicados conceitos e tarefas práticas na modelagem de um DW, usando como estudo de caso a modelagem de DW para a empresa Cia. Zaffari, local de trabalho do autor.

O texto da presente dissertação está estruturado da seguinte forma:

Capítulo 2–Conceitos Básicos: apresenta algumas definições relacionadas ao DW, suas características, arquitetura e modelos.

Capítulo 3–Ferramentas para o projeto de Data Warehouse: descreve as principais ferramentas para o projeto de DW, levando em conta seus componentes, sua arquitetura e formatos de apresentação.

Capítulo 4–Metodologias de projeto de Data Warehouse: uma análise: apresenta uma abordagem sobre as metodologias de desenvolvimento para a construção de um DW, considerando as principais características e os problemas de cada uma. Uma análise ainda é feita sobre as ferramentas atuais dos principais fabricantes de DW, mostrando suas soluções, facilidades e vantagens de uso.

Capítulo 5–Proposta da metodologia para o desenvolvimento de Data Warehouse para a Cia Zaffari: descreve uma proposta de metodologia para o projeto de DW, apresentando um conjunto de critérios importantes para o sucesso do projeto e abordando os problemas para a construção de um DW, levando em conta a extração, conversão e migração dos dados.

Capítulo 6–Estudo de Caso: apresenta uma aplicação sobre a metodologia proposta no capítulo anterior, considerando a sua aplicação na Cia Zaffari Comercio e Industria. Os principais aspectos para o empreendimento de DW serão abordados, tais como: um custo aceitável, principais problemas e soluções adotadas, além dos seus reais benefícios.

Capítulo 7 –Conclusões: onde é resumidos a contribuição deste trabalho, os resultados alcançados e as limitações existentes, além da enumeração de alguns temas de pesquisa interessantes (ou em aberto) neste domínio.

2 CONCEITOS BÁSICOS

Neste capítulo, serão abordados conceitos básicos necessários para a compreensão da proposta. Estes conceitos incluem sistemas de apoio a decisão (seção 2.1), Data Warehouse (seção 2.2), processamento e modelagem de dados (seção 2.3), Data Marts (seção 2.4) e arquitetura de dados (seção 2.5).

2.1 Sistema de apoio à decisão

No passado, uma das principais preocupações de uma empresa, estava no armazenamento correto dos dados necessários ao apoio da tomada de decisões. Com advento das novas tecnologias na área de hardware e seu baixo custo de armazenamento, permitiram que as empresas pudessem armazenar grandes volumes de dados, para posterior extração.Desde que as organizações passaram a dar a devida importância a bancos de dados

Atualmente, se pode dizer que a grande "riqueza" de uma empresa, não está no volume de dados armazenados, e nem nas informações geradas pelos sistemas de aplicação comercial (estoque, contabilidades e outros) e sistemas de informações gerenciais, mas pelo conhecimento da informação, isto é, a forma correta de uso da mesma.

Sistemas de apoio à decisão (SAD), são sistemas que ajudam decisores a tomar decisões em situações onde o julgamento humano é uma contribuição importante ao processo de resolução, mas a limitação humana para processar informações atrapalha.

2.2 Data Warehouse

Data Warehouse (DW) pode ser definido como um tipo de SAD, sendo uma fonte de consultas de um empreendimento, uma base de dados analítica que dá apoio a processos decisórios. Segundo [PEK 96], "Um processo é um produto para a montagem e administração de dados provenientes de várias fontes com o propósito de obter uma visão simples e detalhada de parte de todo o negócio".

Um dos conceitos mais simples para a definir de maneira fácil e clara um DW é: "Uma cópia dos dados de transações, estruturada especificamente para consultas e análises", segundo [KIM 98].

Mas existem outros conceitos que podem ser usados para caracterizar um DW, tais como:

- Para [POE 98], "é um banco de dados analítico que é usado como base para os sistemas SAD. É planejado para armazenar um grande volume de dados somente de leitura, provendo acesso intuitivo";

- Para [INM 97], "um conjunto de bancos de dados integrados e baseados em assuntos projetados para suportar as funções dos SAD, onde cada unidade de dados está relacionada a um determinado momento";

- "DW é um processo que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender a necessidade de consultas estruturadas e ad hoc, relatórios analíticos e de suporte a decisão", conforme [HAR 96];

- Para [GAR 98] "é um processo, não um produto, para a montagem e administração de dados provenientes de várias fontes com o propósito de obter uma simples e detalha visão de parte de todo o negócio".

De acordo com [COR 97], o principal objetivo de um DW é de fornecer os subsídios necessários para a transformação de uma base de dados de uma organização, geralmente transacionais, on-line e operacional denominado banco de dados OLTP (On-Line Transation Processing), para uma base de dados maior que contenha o histórico de todos os dados de interesse existentes na organização, denominado de banco de dados OLAP (On-Line Analytical Processing) e também conhecido como DW propriamente dito.

No contexto deste trabalho será adotado como referencial conceitual a definição de [POE 98] para o DW.

2.2.1 Características básicas de um *Data Warehouse*

Um resumo das principais características básicas de um DW será apresentado a seguir, a partir de [KIM 98]. Um maior detalhamento destas características pode ser obtido em [PER 99], [INM 97] e [GRA 98].

2.2.1.1Organização em assuntos

DW é orientado aos principais assuntos, temas ou áreas de negócio do empreendimento. Sistemas comerciais clássicos são organizados em torno das aplicações da empresa. A Figura 2.1 mostra no caso de uma empresa de supermercados, as aplicações podem ser compra, estoque e venda de produtos. Os principais assuntos podem ser fornecedor, análise dos custos e cliente.

Figura 2.1: Exemplo de dados baseados em assuntos.

2.2.1.2Organização em assuntos

De todos os aspectos do DW o mais importante é o fato de ser integrado, no qual ocorre quando os dados passam do ambiente operacional baseado em aplicações para o DW. A Figura 2.2 mostra um exemplo desta integração.

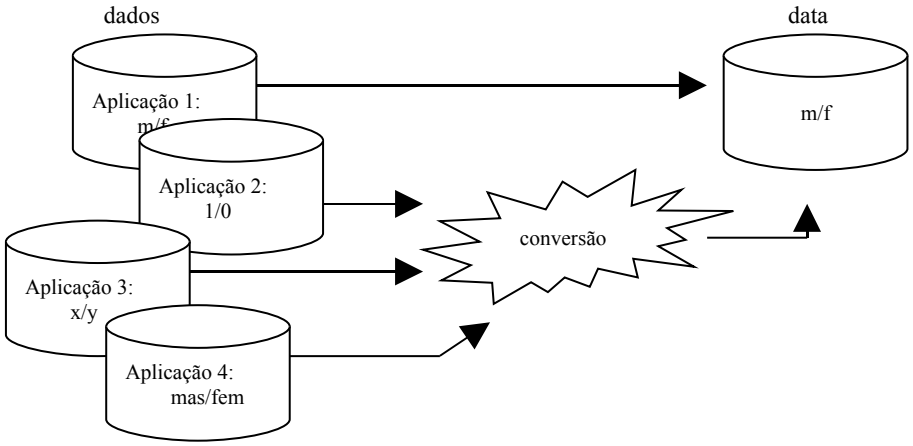


Figura 2.2:Exemplo de dados integrados.

Nesta passagem, os dados fonte de sistemas OLTP são modificados e convertidos para um estado uniforme de modo a permitir a carga no DW. Segundo [INM 97], o processo de introdução dos dados no DW é conduzido de forma que as muitas inconsistências da aplicação sejam desfeitas. Pouco importa, por exemplo, que os dados existentes no DW, usados para representar masculino/feminino, são codificados como m/f ou 1/0. O que realmente importa é que além de a codificação para o DW ser feito, ele deve, ainda, ser feita de forma consistente e independente da aplicação de origem. Caso os dados de uma aplicação estejam codificados como x/y, eles serão convertidos à medida que forem transferidos para o DW. Este processo de transferência é chamado de extração e será melhor definido no terceiro capítulo.

2.2.1.3Não volátil

Os dados após serem extraídos, transformados e transportados para o DW estão disponíveis aos usuários somente para a consulta, logo não podem ser alterados. Já no ambiente de OLTP, os dados normalmente sofrem operações de update, isto é, operações de inserção, alteração e exclusão, além da consulta.

2.2.1.4Variação de tempo

O tempo de tem grande importância no tratamento das informações. São necessários, por exemplo, para se representar: a duração de eventos, o calendário, as previsões e o tempo de vida de documentos ou de operações. Restrições temporais a determinadas atividades são fatores fundamentais para modelar a interação entre as diferentes atividades que podem ser executadas neste tipo de aplicações. É fundamental a capacidade de manipulação do tempo e de informações históricas para o planejamento empresarial, a investigação de relações causais e análise retrospectiva [EDE 94].

No DW o elemento tempo é fundamental. Dados existentes no DW não passam de uma série sofisticada de instantâneos, capturados num determinado momento. É importante salientar que pelo grande volume de dados, a estrutura de chave de um DW sempre contém algum elemento de tempo.

Quando os dados são considerados antigos, passam do detalhe corrente para o detalhe mais antigo. À medida que os dados são resumidos, passam do detalhe corrente para os dados levemente resumidos e a seguir, dos dados levemente resumidos para os dados altamente resumidos.

2.2.1.5 Metadados

Os metadados são os dados que descrevem e caracterizam dados ou conjuntos de dados. A disponibilização de metadados permite que usuários, com pouco conhecimento sobre os dados, possam avaliar a compatibilidade dos dados com suas aplicações, sendo fundamental a existência de metadados adequados tanto para caracterização quanto para a descrição e compreensão dos dados.

Provêm informações sobre a estrutura de dados e as relações entre estas dentro ou entre bancos de dados. São também informações mantidas a cerca do DW em lugar das providas pelo warehouse.

2.2.1.6 Granularidade

É o nível de detalhes dentro do banco de dados DW. Quanto mais detalhe, menor o nível de granularidade, consequentemente, maior o volume de dados armazenado. Esta característica é uma das mais importantes, pois afeta profundamente o volume dos dados que residem no DW e, ao mesmo tempo, afeta o tipo de consulta que pode ser atendida. O volume de dados contidos no DW é balanceado de acordo com o nível de detalhe de uma consulta. A Figura 2.3 mostra um exemplo de granularidade sobre o registro de vendas de uma rede de supermercado contendo 24 lojas.

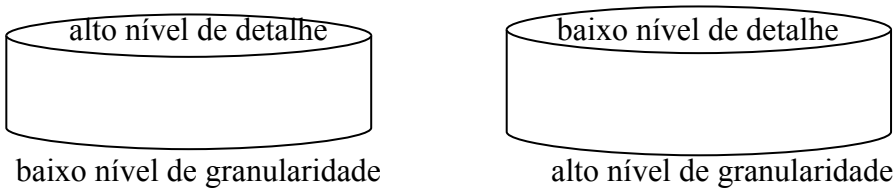


Figura 2.3: Exemplo de granularidade.

No alto nível de detalhe, as vendas registradas são diárias, por loja, num total de 400 mil bytes por mês ou 2 mil registros por mês. Para o baixo nível de detalhe, as vendas são registradas por mês e loja, num total de 2 mil bytes ou 24 registros.

2.3 Processamento OLAP e modelagem de dados

Os dados armazenados em um DW são otimizados para a recuperação através de processamento analítico e devem ser modelados de forma a apresentar os dados em uma estrutura padronizada que permite alto desempenho de acesso. As próximas seções apresentam os principais conceitos referentes ao processamento analítico e quanto à modelagem necessária para suportá-lo.

2.3.1 Processamento OLAP

O processamento analítico On-Line (On-line Analytic Porcessing – OLAP) constitui-se de todas as atividades gerais de consulta e apresentação de dados numéricos e textos provenientes do DW, assim como as formas específicas de consulta e apresentação que são exemplificadas por uma grande quantidade de ferramentas OLAP [KIM 98a]. Sistemas OLAP ajudam analisas e gerentes a sintetizarem informações sobre a empresa através de comparações, visões personalizadas, projeção dos dados e análise histórica em vários cenários sob situações variadas e não uniformes.

Para a representação OLAP, podemos considerar três diferentes abordagens: (a) ROLAP (relacional OLAP – Relational OLAPI), (b) MOLAP (Multidimensional OLAP – Multidimensional OLAP) e (c) HOLAP (Híbrido OLAP – Hybrid OLAP).

ROLAP constitui-se de um conjunto de interfaces de usuário e aplicações que dá ao banco de dados relacional características dimensionais [KIM 98a]. As tabelas de sumários são criadas no SGBD relacional, sendo que nenhum dado é movido para o OLAP Server. As tabelas de sumários, quando necessárias são totalmente deriváveis e seus índices criados automaticamente. A consulta tem baixo desempenho mas permite a utilização do padrão SQL.

MOLAP constitui-se basicamente de um banco de dados multidimensional (Multidimensional database – MDDB) , através de um conjunto de interfaces de usuário, aplicações e banco de dados, com tecnologia proprietária [KIM 98a]. Os MDDB armazenam seus dados em um cubo com várias dimensões. Os dados são trazidos para o servidor OLAP e organizados em arranjos com alto grão de agregação, permitem consultas mais rápidas, quando comparadas a abordagem relacional. Suas limitações para a solução MOLAP referem-se ao fato de serem sistemas proprietários, que não utilizam soluções padrão de banco de dados, a exemplo de que mudanças que podem ocorrer no modelo dimensional provocam a reorganização do próprio banco de dados, pouca escabilidade, além de não suportar a criação de ad hoc de visões multidimensionais.

HOLAP constitui-se de uma nova tecnologia o qual combina as principais características das abordagens ROLAP e MOLAP, com o objetivo de resolver os principais problemas descritos anteriormente, em ambas abordagens. Os dados ficam retidos no banco de dados SGBD, enquanto as agregações ficam no MOLAP. Com desvantagem, torna-se mais lento do que o modelo MOLAP quando a consulta é feita sobre dados básicos.

2.3.2 Modelagem de dados

A modelagem pode ser descrita sobre dois aspectos: (a) a modelagem de dados tradicional e (b) a modelagem multidimensional. Na modelagem tradicional as entidades podem ser objetos (clientes, produtos, lojas) ou transações (vendas, pedidos, notas fiscais). Seus relacionamentos são explícitos, em outras palavras, as entidades se relacionam de forma direta através dos atributos chave. As operações estão direcionadas a dados transacionais, orientada a dados atuais que mudam constantemente.

Já na modelagem multidimensional, as entidades são dimensões que representam resultados para um determinado período de tempo. Os relacionamentos são implícitos, onde as entidades se relacionam indiretamente, através de outra entidade. As operações são direcionadas a dados analíticos, orientada a dados históricos estáveis.

2.3.2.1 Modelagem Dimensional

A principal característica dos sistemas OLAP é permitir a visão conceitual multidimensional dos dados de uma organização. A visão multidimensional é mais natural, fácil e intuitiva para o usuário, permitindo a visão dos negócios da empresa em diferentes perspectivas. Para este tipo de análise é necessária uma modelagem dimensional, uma alternativa para a modelagem entidade relacionamento e contém as mesmas informações [KIM 98a].

A idéia da modelagem dimensional é representar os tipos de dados de negócio em uma estrutura do tipo cubo de dados. As células deste cubo contêm valores medidos, tais como "unidades vendidas", "lucro" ou "venda liquida" e os lados do cubo definem as dimensões dos dados, a exemplo de "cliente", "produto", "fornecedor" e "tempo".

Um exemplo da representação do negócio em uma estrutura do tipo cubo, seria a descrição que um executivo faz aos processos de sua empresa, como a venda de produtos em uma variedade de lojas e verificar a performance ao longo do tempo, conforme mostra a Figura 2.4. Se pensarmos no negócio em termos de um cubo com nossas dimensões formando a base do cubo, o ponto de interseção das três dimensões dentro do cubo equivale a um ponto de medição para o negócio.

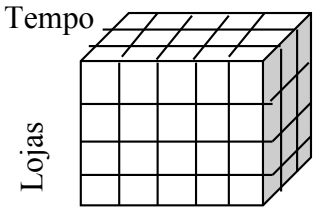
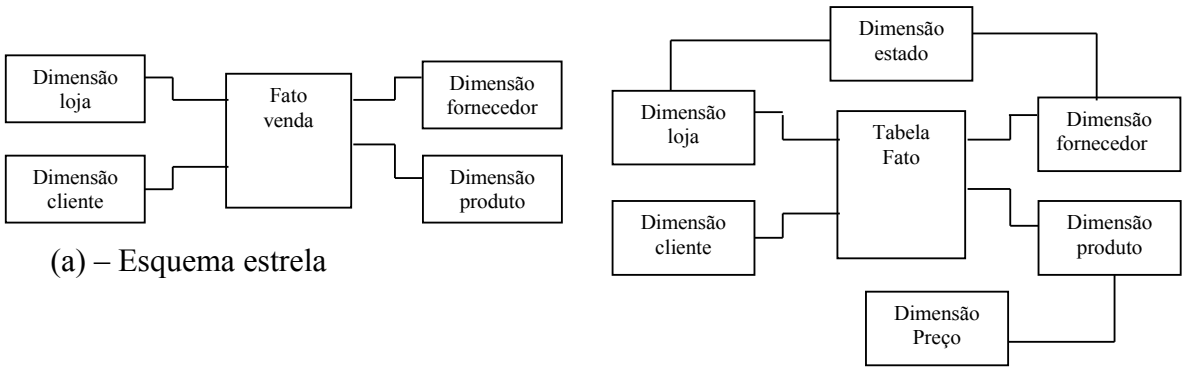


Figura 2.4: Exemplo de um cubo.

Nos banco de dados analíticos que manipulam multidimensões, existem dois tipos principais esquemas que são utilizados: (a) o esquema estrela (star scheme) e o (b) esquema floco de neve (snowflake schema).

O esquema estrela utiliza-se dos mesmos componentes do diagrama entidade-relacionamento, como entidades, atributos, relacionamentos e chaves primárias, existindo basicamente dois tipos de tabelas (entidades) denominadas de "fato" e "dimensão" [KIM 98a]. Este modelo é construído por uma estrutura formada por uma única tabela de fatos (contendo dados numéricos) relacionada com uma ou mais tabelas de dimensão, conforme a Figura 2.5(a).

A tabela fato armazena instâncias da realidade, representando as medidas do negócio, que podem ser mensuradas de forma quantitativa [GRA 98]. Por exemplo, a Figura 2.5(a) mostra a tabela de fato "venda", o qual possui os atributos de valor da venda, quantidade vendida e o custo de venda de um produto relacionada com as tabelas de dimensão "produto" e "cliente", permitindo identificar a quantidade vendida de um produto por um certo cliente. A tabela de fato armazena grande quantidade de dados, possuindo chave primária composta, formada por chaves estrangeiras, através das quais se ligam as chaves primárias das tabelas dimensão.



(a) – Esquema estrela

(b) – Esquema floco de neve

Figura 2.5: Exemplo de um cubo.

Para [KIM 98], as medidas do negócio, também chamadas de métricas ou indicadores, são informações numéricas sobre o negócio e são definidas em três tipos diferentes:

- Aditivas: são as mais freqüentes, podem ser somadas cruzando-se qualquer uma de suas dimensões. Exemplo: lucro liquido;
- Semi-aditivas: podem ser somadas através de apenas uma parte de suas dimensões. Exemplo: quantidade em estoque, uma vez que não faz sentido somá-la através da dimensão tempo;
- Não aditivas: não podem ser somadas por nenhuma de suas dimensões. O exemplo mais comum desse tipo de medidas são valores percentuais.

As tabelas dimensionais armazenam as descrições textuais que ajudam na identificação de um componente (registro) da respectiva dimensão [KIM 98]. Cada tabela dimensão é uma tabela não normalizada que armazena os dados sobre a dimensão [HAR 96]. Armazenam pequena quantidade de dados, quando comparadas a tabela de fatos. Possuem chave primária simples, permitindo a ligação com a tabela de fato ou entre as tabelas de dimensão. Repare que a chave primária da tabela de fato, representada pela Figura 2.6, é composta pelas chaves primárias das tabelas de dimensão com as quais ela se relaciona.

As tabelas dimensionais contêm atributos sobre as entidades que são relacionadas às medidas, ou seja sobre os dados que dão alguma informação sobre a medida.

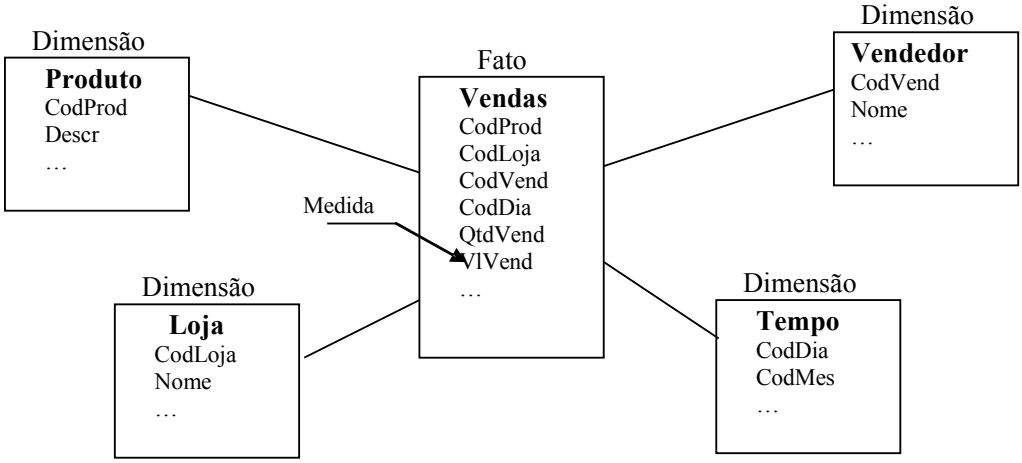


Figura 2.6: Exemplo simples do modelo estrela.

Um exemplo descrito na Figura 2.6 seria com relação à medida "quantidade vendida". Existem vários parâmetros ou dimensões que através dos quais se pode analisá-la: (a) em que loja a compra foi feita? (b) quando ocorreu a venda? (c) que produto foi vendido e (d) quem realizou a venda?

Como se pode ver na Figura 2.6, o esquema estrela é altamente desnormalizado, com o intuito óbvio de reduzir o número de *joins* envolvidos nas consultas. Na verdade, o modelo final de um DW é composto por várias tabelas de fato, contendo diferentes subconjuntos de informações sobre o negócio com diversas tabelas de dimensão, ligadas a uma ou mais tabelas de fato.

O esquema floco de neve, representado pela Figura 2.5(b), é uma variação do esquema estrela os quais as tabelas dimensão são normalizadas, permitindo que se liguem entre si, além da tabela fato. Possui como vantagem no uso deste esquema a economia de espaço no armazenamento, tabelas dimensão menores, mas possui como a principal desvantagem a complexidade sobre o numero de tabelas relacionadas, tornando as consultas complexas.

2.3.2.2 Agregação de Dados

Em aplicações de análise de dados, um dos fatores mais críticos é o tempo de resposta ao usuário devido ao grande volume de dados envolvido nas consultas desse tipo de aplicação. A única maneira de reduzir o tempo de execução das consultas de maneira consistente é pré-navegar ou consolidar os dados em totais e subtotais através das dimensões envolvidas no assunto em questão. Mas de que forma se poderia agregar cada dimensão? Essa é uma questão mais simples do que se parece a princípio, já que é inerente ao ser humano agrupar em hierarquias todas as entidades que o cercam. Agrupamos cidades em estados, regiões e países, produtos em linhas de produtos, meses em trimestres e anos. Apesar das hierarquias não serem partes necessárias das dimensões, as aplicações que refletem negócios do mundo real com um mínimo de complexidade sempre apresentam algumas hierarquias dimensionais a exemplo das listadas acima. Sendo assim, a base para a agregação dos dados será justamente o conjunto das hierarquias existentes.

2.4 Data Marts

Segundo [COR 97], representa um subconjunto dos dados destinados a um usuário específico ou a um conjunto de grupos de usuários, implementados em servidores departamentais de baixo custo, em plataforma Unix ou Windows NT.

Data Marts (DM) podem ser capturados de uma ou mais bases operacionais ou de informações externas. Em alguns casos, os dados armazenados no DM podem também ter sido gerados localmente dentro de um departamento particular ou de uma área demográfica.

2.5 Arquitetura de Dados

A arquitetura de dados para um projeto de DW pode ser dividida em duas partes: (a) arquitetura geral dos dados ou topologias e (b) funcional. Na arquitetura geral dos dados, se permite identificar e entender como os dados fluem através do DW. As arquiteturas de dados mais comuns são:

- Centralizada, caracterizada por um único DW que atende a toda a comunidade de usuários;

- Data Marts dependentes, constitui-se de vários DM ligados a um DW. Cada DM tem um escopo de dados limitados orientados a um tema específico do negócio. Os usuários podem se conectar aos seus DM como ao DW;

- Data Marts independentes, caracteriza-se pela ligação dos usuários aos respectivos DM, as quais fornecem as informações necessárias. Esta arquitetura oferece uma rapidez no desenvolvimento, baixo custo e controle local, ao invés do centralizado;

- Data Warehouse Distribuído, consiste de vários DW interligados através de rede com forte suporte a processamento distribuído.

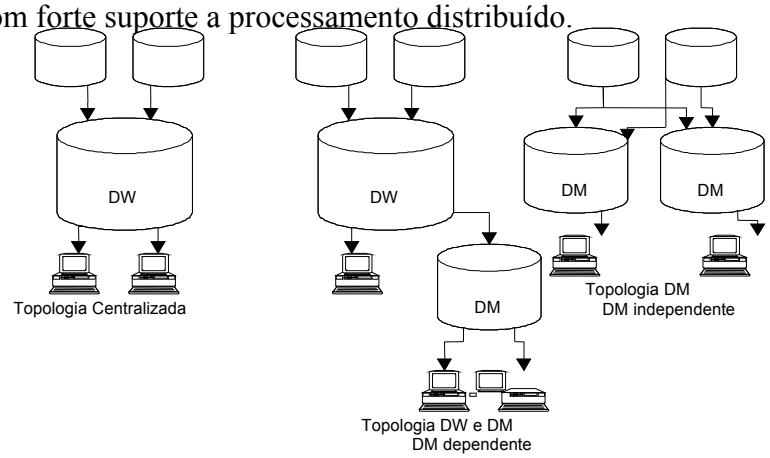


Figura 2.7: Topologias de um Data Warehouse.

A maioria das organizações tende a implementar um DW multi-tier, ou seja, um DW em camadas, que amarram vários DMs integrados a um DW corporativo em uma arquitetura semelhante à Figura 2.7, onde mostra algumas topologias de DW.

Para a arquitetura funcional, o DW, conforme a Figura 2.8, é construído a partir de duas partes distintas. A primeira parte é definida como área interna, onde são feitas as aquisições de dados a partir dos sistemas tradicionais ou de outras fontes quaisquer. O dado é identificado, copiado, formatado e preparado para ser carregado no repositório de dados do DW, que pode ser administrado através de banco de dados relacionais ou multidimensionais. A área de Staging armazena os dados que foram extraídos de fontes externas. A partir daí os dados são tratados, limpos e carregados ao DW.

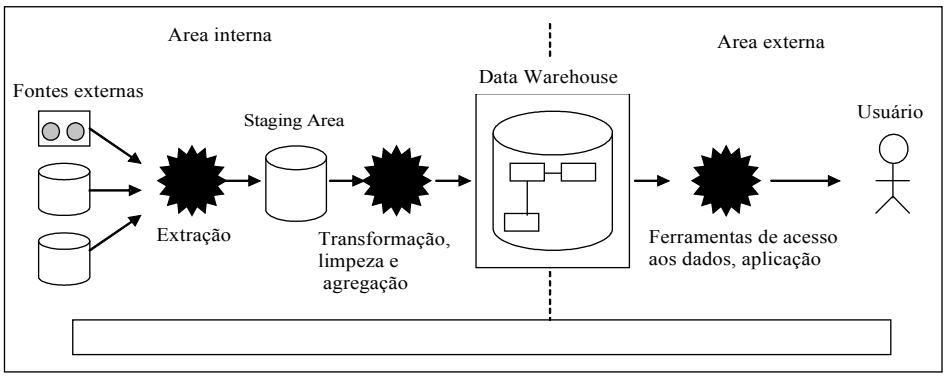


Figura 2.8: Arquitetura de um Data Warehouse.

Segundo [PER 2000], ainda se pode descrever como partes desta área: (a) a carga dos dados, permitindo o armazenamento dos dados transformados no servidor de apresentação, (b) controle dos dados organizados, permitindo o monitoramento sobre o fluxo de dados, através dos metadados, (c) gerenciamento dos recursos da área interna, possibilitando que o DW volte a trabalhar normalmente após a ocorrência do problema.

A segunda parte é definida como a área externa, sendo a interface do usuário com o sistema. É, basicamente, o front-end que é visto e no qual se trabalha, principalmente através de consultas. [PER 2000] Fazem parte desta área: (a) o servidor de apresentação, onde os dados provenientes da parte interna, ficando a disposição dos usuários finais e (b) ferramentas de acesso a dados e geradores de relatórios, permitindo aos usuários finais consultas ad hoc. Tais ferramentas permitem operações que facilitam o acesso aos dados, possibilitando aumentar ou diminuir o nível de detalhes das consultas as tabelas dimensão e fato através dos seguintes recursos:

- Drill-up/drill-down: permite navegar entre níveis de agregação, por agrupar e desagrupar dados progressivamente [POE 98];

- Pivoting: permite agregar duas dimensões para comparar o resultado. Na prática, corresponde a modificação da posição das dimensões em um gráfico ou troca de linhas por colunas em uma tabela [DBM 2001];

- Slice and dice: possibilita ver os dados de diferentes pontos de vista, reduzindo a dimensionalidade dos dados. Slice compreende a extração de informações sumarizadas em um cubo de dados e Dice é a extração de um subcubo ou a intersecção de vários slices;

- Data Mining: é o [DWM 2001] processo de encontrar padrões ou correlações entre milhares de campos em grandes bases de dados. Informações, que aparentemente estão camufladas ou escondidas, permitindo agilidade na tomada de decisões. Maiores detalhes sobre Data Minig estão disponíveis em [FAY 96], [DAM 2000] e [GRA 98].

Se pode ainda obter um maior detalhamento sobre a área externa em [KIM 98a].

3 FERRAMENTAS PARA O PROJETO DE DATA WAREHOUSE

A informação em Data Warehouse se encontra integrada e organizada em áreas de interesse para a empresa. Sua arquitetura se modificará ao longo do tempo, refletindo a natureza iterativa do processo. O processo de DW é inerentemente complexo, caro e longo, requerendo mudanças constantes ao longo do ciclo de vida dos aplicativos operacionais. Este capítulo apresenta um estudo de alguma das principais ferramentas disponíveis no mercado, Microsoft DW (seção 3.1), Oracle DW (seção 3.2), Sybase DW (seção 3.3).

3.1 Microsoft Data Warehouse

A Microsoft, ao longo de vários anos, tem tentado criar uma plataforma para Data Warehouse que possa ser usada para diminuir custos e melhorar a eficiência dos processos relativos ao ciclo de vida de tal sistema. Esta plataforma foi batizada de Microsoft Data Warehouseing Framework, possibilitando a integração de produtos de diversos fabricantes, na tentativa de prover ao cliente uma escolha livre e facilitada na hora de desenvolver sua política para DW. Adicionalmente, existe a necessidade de se ter metadados integrados, das mais diferentes fases do projeto do DW. Finalmente, deve-se prestar atenção nos serviços básicos de gerenciamento, armazenamento, ajuste de desempenho, alertas e notificações.

3.1.1 Componentes do Data Warehouse Framework

Data Warehouse Framework descreve os relacionamentos entre vários componentes usados no processo de construção, uso e gerenciamento de sistema DW Os dois componentes básicos dessa abordagem são: (a) a camada de transporte de dados (OLE BD) e o repositório integrado de metadados. A construção do DW requer um conjunto de ferramentas para descrever o projeto lógico e físico das fontes de dados e seus destinos nos DWs e DMs. Já as ferramentas para usuários finais devem acessar um serviço de diretório que permita a busca de dados apropriados e relevantes para resolver as questões do negócio, além de uma camada de segurança entre usuários e os sistemas servidores.

Microsoft Repository provê a integração para o metadados de todo o ciclo de projeto possibilitando compartilhamento das informações e a integração transparente de várias ferramentas heterogêneas, sem a necessidade de interfaces especializadas entre cada produto.

3.1.2 Capacidades analíticas de OLAP

Microsoft Decision Support Services (DSS) é um aplicativo OLAP, dotado de vastas funcionalidades, componente da Microsoft SQL Server 7.0. Microsoft DSS inclui um servidor na camada intermediária (middle-tier), conforme a Figura 3.1, que permite aos usuários efetuar análises sofisticadas em grande volume de dados, gerando resultados altamente satisfatórios. Um segundo componente do Microsoft DSS é um servidor (do lado do cliente, para cache e cálculos) chamado PivotTable. Este ajuda na melhoria do desempenho das consultas e na redução do tráfego na rede.

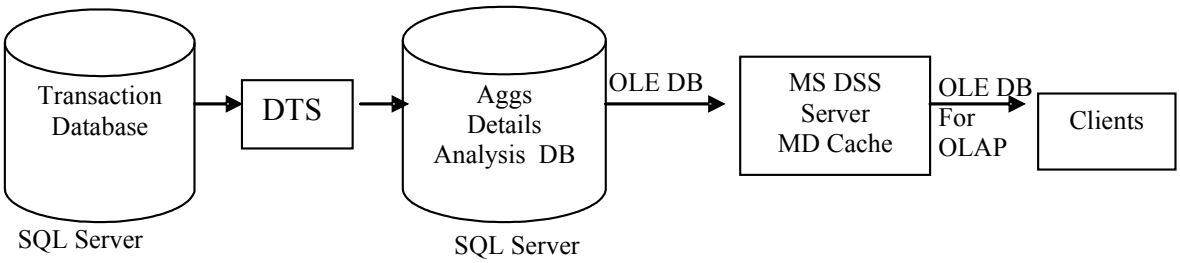


Figura 3.1: Camada intermediária do servidor OLAP.

3.1.3 Arquitetura de serviço de transformação de dados

A definição do serviço de transformação de dados são armazenados no repositório da Microsoft SQL Server ou em arquivos estruturados COM. Os dados operacionais são acessados usando OLE DB, fornecendo acesso a fontes de dados relacionais e a fonte de dados não relacionais. A Figura 3.2 descreve a estrutura de transformação dos dados, onde o data pump executa funções de um script para copiar, validar ou transformar os dados da fonte para o destino. Podem ser criados objetos de transformação padronizados para a busca de dados avançada. Os novos valores para o destino são retornados ao data pump e enviados ao destino através da transferência de dados de alta velocidade.

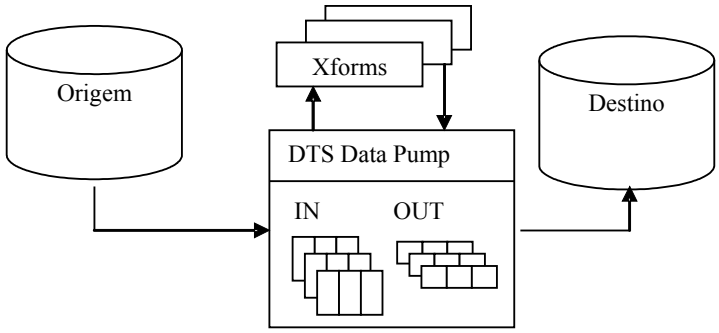


Figura 3.2: Arquitetura de transformação dos dados.

Transformações complexas e validação de dados lógica podem ser implementadas usando um componente activex script. Estes scripts podem invocar métodos de qualquer objeto OLE para modificar ou validar o valor de coluna.

3.1.4 Análise e apresentação dos dados

A Microsoft provê aplicativos para consultas em DW. Por exemplo, Microsoft Access e Excel oferecem recursos para formulação de consultas e análises da informação no DW. Em anexo ao SQL Server 7.0, existe um componente denominado English Query, que permite aos usuários formular consultas utilizando sentenças da

língua Inglesa. Adicionalmente, através do DW Framework, muitos outros produtos tornam-se disponíveis para visualização e análise dos dados.

A filosofia fundamental do Microsoft Data Warehousing Framewowrk é a abertura da solução a outras companhias de software. Através da utilização dos padrões de interfaces para banco de dados ODBC e OLE DB, dúzias de produtos podem acessar e manipular dados guardados em sistemas relacionais. Devido a essas funcionalidades, empresas são capazes de selecionar as ferramentas analíticas mais apropriadas para suas necessidades.

3.2 Oracle Data Warehouse

O Oracle9i Database oferece uma infra-estrutura completa e integrada para análise e data warehouse. As soluções de business intelligence do Oracle9i são menos complexas, menos caras e mais rápidas de implementar. A lista abaixo destaca os recursos do banco de dados Oracle9i e seus compomentes associados, proporcionando uma plataforma de e-business intelligence completa:

- Inovadores recursos de ETL (Extration, Transformation and Load ou Extração, Transformação e Carga) incorporados ao banco de dados atualizados dinamicamente;

- Serviços de OLAP e funções SQL fornecem análise comercial e de mercado rica, robusta e de alta qualidade.

Os recursos de extração, transformação e carga (ETL, Extraction, Transformation and Load) do Oracle9i Database tornam mais fácil integrar dados de muitas fontes diferentes. Os recursos de data warehouse lhe permitem armazenar e acessar grandes volumes de dados com alta performance. A funcionalidade avançada de OLAP (processamento analítico on-line) e data mining o ajudam a detectar tendências e fazer previsões.

3.2.1 Oracle Data Warehouse Framework

A arquitetura Oracle Web é chamada de NetWork Computing Architecture (NCA) que é um framework e que faz a mudança da tecnologia cliente/servidor para Web. Esta arquitetura, conforme a Figura 3.3, tem cinco camadas lógicas que são: (a) source, (b) data, (c) olap/application, (d) publication e (e) presentation.

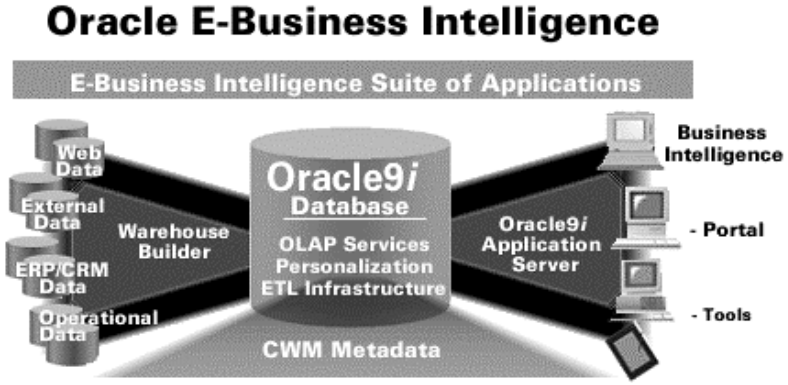


Figura 3.3: Arquitetura Oracle (fonte [MIC 2000]).

A camada Source é onde ficam os vários tipos de dados transacionais, dados de sistemas legados nos maiframes, das aplicações clliente/servidor e também configurar dados externos. Esses dados são transformados e desnormalizados para atingirmos uma ótima performance de consulta. Eles são transportados para as tabelas Oracle do Dataware ou Data Mart.

Já na camada Date, ficam armazenados os dados de apoio a decisão. A Oracle implementa uma estratégica bem interessante onde ela coloca tanto os dados de DW e DM. Os dados de DM para consultas ah hoc, a Oracle utiliza o Express Server com acesso relacional.

É na camada OLAP/Application aonde se realiza toda a lógica do processamento. Além do software que faz o processamento são colocados cartuchos de ligação para integrarem o processamento aos dados.

3.2.2 Processameto analítico on-line - OLAP

A Oracle Express é uma família de produtos voltada para o processamento OLAP, incluído funções de consulta, análise e relatório. Dentre os produtos que compõem a família Oracle Express, se destaca:

- Oracle Express Server: baseado no modelo multidimensional, é otimizado para consultas e análises de dados corporativos tais como vendas, marketing, manufatura ou recursos humanos, sem necessitar de programas ou relatórios especiais dos sistemas de informação;

- Oracle Express Analyzer: é uma ferramenta orientada-objeto de navegação, acesso e análise de dados sumarizados em um DW ou DM Com um conjunto visual de ferramentas desktop para seleção, visualização, análise, anotação e compartilhamento dos dados da corporação, Express Analyser disponibiliza o poder do processamento analítico on-line para usuários de qualquer área, incluíndo vendas, marketing, operações, vendas, manufatura e finança;

- Oracle Express Objects: é um ambiente de desenvolvimento de aplicações orientado-objeto que tem controles data-aware para visualização e manipulação dos dados e suporta desenvolvimento visual como programação orientada a evento.

3.2.3 Arquitetura de serviço de transformação de dados

O banco de dados de Oracle9i introduz uma nova abordagem para entrada de dados integrando isto no próprio banco de dados. A maioria de produtos de ETL não tem o paralelismo e extensa otimização, características existentes no banco de dados Oracle9i.

Esta versão de banco de dados, permite criar um novo paradigma de ETL, possibilitando a eliminação de certas etapas e redefinindo (remodelando) outras etapas para aumentar o fluxo de dados e transformação dos dados, aumentando e escabilidade e eliminando as interrupções. Possui o conceito de ETL toolkit, um conjunto de ferramentas que possibilita aumentar a capacidade de ETL no DW, tais como:

- Change Data Capture: a captura de dado alterado é um importante elemento para otimizar a extração de uma base de dados de origem. Em vez de fazer a extração completa de todos os dados de origem, somente as alterações e inclusões de novos dados são considerados;

- External Tables: várias interações são necessárias entre os dados armazenados na base de dados destino e dados na base de origem. Tabelas Externas do Oracle9i permitem que dados externos possam ser expostos para usuários como qualquer outro dado armazenado em tabelas normais. Como resultado, a tabela externa atua como uma tabela virtual, possibilitando a manipulação e junção com os dados de origem, antes dos dados serem realmente carregados, como mostra a Figura 3.4;

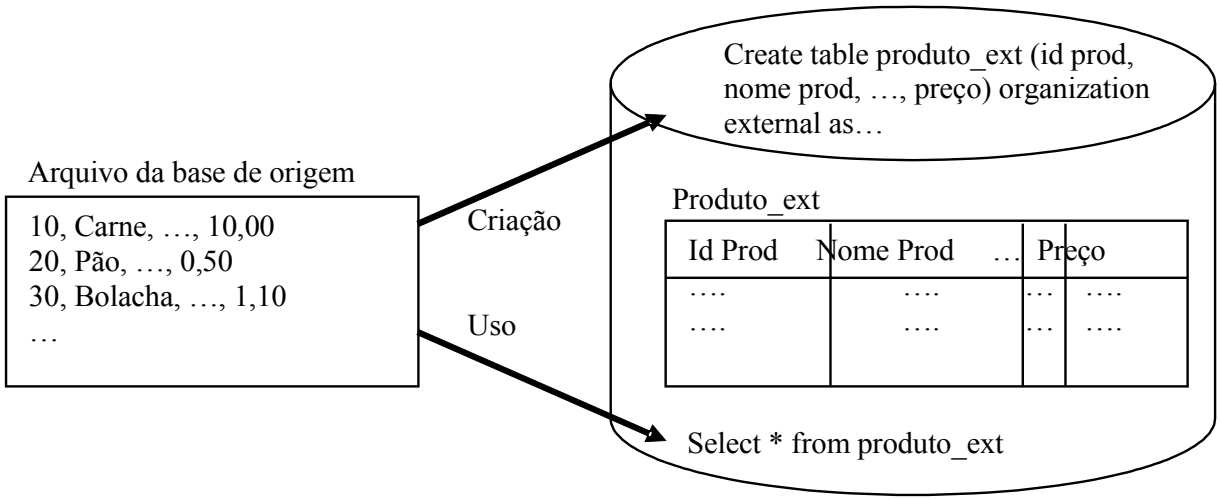


Figura 3.4: Exemplo da Tabela Externa.

- Multi-Table Insert: em certas situações, dados de origem devem ser carregados, baseados na lógica de atributos, em diferentes objetos de destino.Freqüentemente nos ambientes de DW os dados de origens são carregados na mesma tabela destino. Multi-Table Insert oferece um novo comando de SQL para resolver esta questão de transformação e carga, onde dados podem ir para diferentes tabelas de destino, dependendo das regras de transformação do negócio. Como mostra a Figura 3.5, o comando Insert...Select permite que diferentes tabelas de destino participem, dependendo de condições estabelecidas na clausula where do comando Insert;

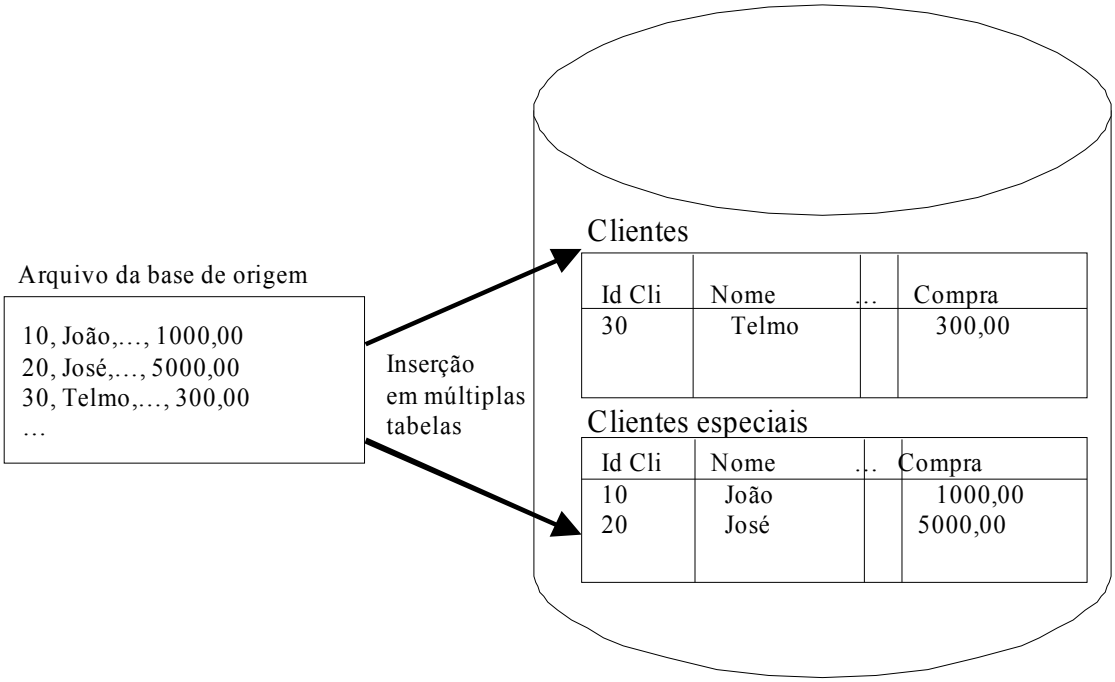


Figura 3.5: Exemplo da Inserção de Múltiplas Tabelas.

- Upserver: depois da carga inicial, o ambiente de DW precisa ser várias vezes atualizado para refletir as mudanças geradas pelos dados de origem. Muitas vezes, os sistemas de origem dos dados não destinguem as recentes inclusões ou alterações durante a extração. Desta forma, o Upserver extende o comando Merge no SQL para oferecer a capacidade de alterar ou inserir condicionalmente linhas na tabela destino. O exemplo abaixo mostra o comando Merge o qual permite identificar quando ocorreu uma alteração ou inclusão, sobre os dados de origem, provocando a mesma operação na tabela destino.

Exemplo: MERGE INTO produtos t USING produtos_origem s

ON t.codproduto=s.codproduto

WHEN MATCHED THEN

UPDATE SET t.preco = s.preco

WHEN NOT MATCHED THEN

INSERT (t.codproduto, …, t.preco) VALUES

(s.codproduto,...,s.preco)

3.2.4 Análise e apresentação dos dados

A Oracle Discoverer é uma ferramenta de consulta e de análise que permite obter informações específicas dos complexos bancos de dados de produção. Projetada tanto para usuário final quanto para especialistas de gerenciamento de sistemas de informação, ela permite que usuários executem análise de negócios, criem relatórios de negócios e gráficos sem qualquer necessidade de conhecimento de programação ou experiência em banco de dados. O objetivo é esconder do usuário a complexidade de banco de dados.

3.3 Sybase Data Warehouse

A Sybase desenvolveu uma nova abordagem de DW permitindo que uma grande quantidade de usuários possam manipular repositórios em escala de terabytes de dados. Estas propriedades são encontradas no Sybase Adaptive Server IQ Multiplex (ASIQ). Com o ASIQ, os processos de churning, dataming e análises feitas por um grande número de usuários passaram a ser uma realidade em empresas de diferentes áreas do mercado. As principais características do ASIQ são descritas a seguir:

- possibilita acesso rápido às informações da empresa;
- finaliza as consultas sem retorno, quando da não liberação dos dados pelo usuário ou por tempo excessivo despendido;
- oferece uma redução na área física de armazenamento de dados de até 80%;
- compatível com as melhores ferramentas de visualização e informação existentes no mercado.

A Figura 3.6 abaixo mostra e estrutura e funcionalidade do IQ Adaptive Server da Sybase, oferecendo um suporte heterogêneo a dados de diferentes origens, incluindo Oracle, DB2, Informix e demais arquivos de origem de aplicação, como planilhas e editores.

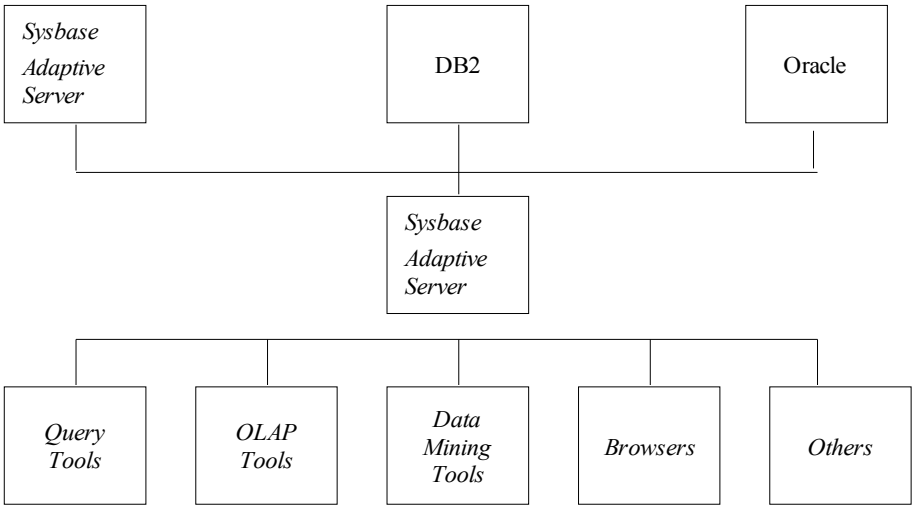


Figura 3.6: Estrutura do IQ Adaptive Server.

3.3.1 Infra-estrutura para o Data Warehouse

A Sybase possui o Industry Warehouse Studio – IWS, uma infra-estrutura de Data Warehouse pré-construída, disponível na forma de um pacote, que consiste em um modelo empresarial vertical para atender às necessidades de setores e clientes específicos. Ë um sistema flexível e aberto que se integra de modo compatível com os investimentos tecnológicos existentes nas empresas. O IWS oferece aplicativos para diversos setores verticais, tais como: serviços bancários, seguradoras, cartões de crédito, telecomunicações. Possui uma visão de escala empresarial, permitindo que as empresas construam DW que possam integrar dados provenientes de diferentes setores, flexibilizando o ambiente empresarial. Possui como principais características:

- Uso de técnicas de modelagem dimensionais, que aperfeiçoam o desenvolvimento do DW, melhorando a escabilidade e tornando os dados mais acessíveis aos usuários;
- Possui arquitetura modular, capaz de suportar as implementações em fases e aquelas que reúnem aplicativos para diversos setores, a exemplo de um banco fundir-se com uma seguradora;
- Um processo simples de personalização para o cliente, usando as ferramentas CASE (Engenharia de Software Auxiliada por Computador) da Sybase, o Power Designer Warehouse Architect e Warehouse Control Center;
- Suporte a uma ampla variedade de plataformas RDMS, tais como IBM DB2, Microsoft SQL Server, Oracle e Informix.

3.3.2 Análise e apresentação dos dados

O IWS oferece uma parceria com os fornecedores de ferramentas de visualização estratégicas, tais como: Business Objects, Cognos, MicroStrategy e SGI. Estas ferramentas podem proporcionar um nível de integração que assegura a implementação mais rápida de módulos empresariais com base no ambiente IWS embutido.

4 METODOLOGIAS DE PROJETO DE DW: UMA ANÁLISE

Este capítulo apresenta um estudo de algumas metodologias de projeto de DW que servirão como fundamentação teórica para a proposta de uma metodologia a ser aplicada no projeto de DW da Cia Zaffari.

Inicialmente serão descritas as propostas encontradas em [MAR 99], [KIM 98a], [POE 98] e [PER 2000]. No final deste, encontra-se uma análise das propostas levantadas e suas limitações, permitindo selecionar as características mais adequadas a serem utilizadas como referencial para o resto deste trabalho.

4.1 A Metodologia de James Martin

A proposta de [MAR 99] apresenta um conceito PACE (Planejar, Ativar, Controlar e Encerrar) no qual os processos que compõem o ciclo de vida do projeto (o qual é referenciado como sendo o mapa de estrada do processo) do DW são iterativos, portanto alguns são repetidos. As fases do ciclo de vida, conforme mostra a Figura 4.1, do projeto de DW são: (a) visão estratégica, (b) avaliação da engenharia da empresa, (c) avaliação do fluxo de valores, (d) caso comercial do DW, (e) projeto e revisão da arquitetura, (f) avaliação da questão comercial, (g) plano de implementação da iteração, (h) projeto detalhado, (i) implementação, (j) transição para a produção e (k) manutenção.

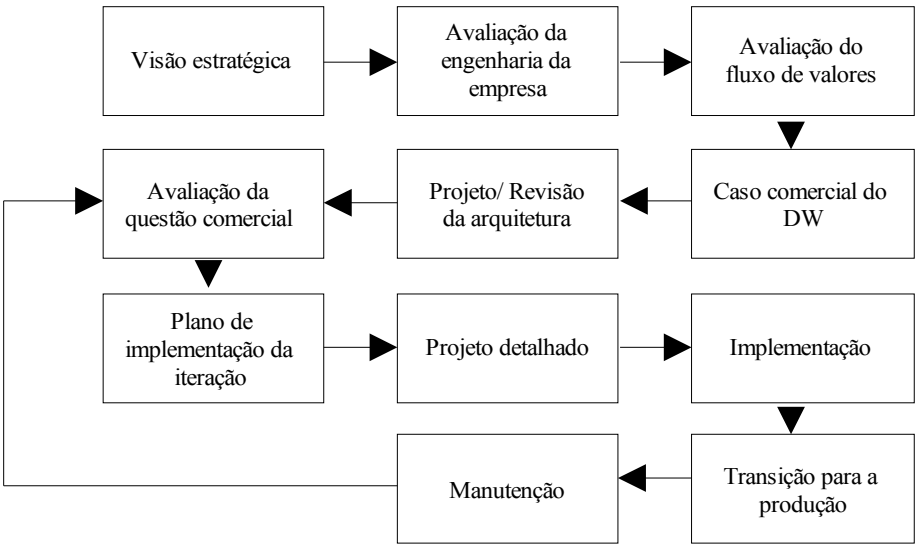


Figura 4.1: Mapa de estrada do processo.

Na fase de *visão estratégica*, também chamada por [MAR 99] de plano de informações estratégicas (SIP- *Strategic Information Plan*), é um processo contínuo que alinha as estratégicas comercial e tecnológica da empresa dentro do mercado. Esse é um pré-requisito para os estágios de engenharia da empresa e re-engenharia do processo comercial. Algumas empresas possuem um SIP pronto, e ele pode servir como semente da qual o projeto do data warehouse se desenvolve.

A fase de *avaliação da engenharia da empresa* (EEA-*Enterprise Engineering Assessment*) desenvolve uma visão em nível de empresa da necessidade de mudança da organização e sua prontidão em aceitá-la. Um data warehouse não é uma solução para tudo. Se uma organização não possui fontes e recursos de dados, um warehouse não poderá ser eficiente. Antes de realizar um projeto de data warehouse, a organização deve decidir se deseja resolver problemas de dados operacionais por meio da re-engenharia comercial, desenvolvimento de sistemas ou planejamento de sistemas de informação. Essa avaliação normalmente é um pré-requisito para a reengenharia do processo comercial ou para uma avaliação do fluxo de valores.

Pela fase de *avaliação do fluxo de valores* (VSA- *Value Stream Assessment*) se podem solucionar problemas comerciais estudando o(s) fluxo(s) de valores de uma empresa a partir de um alto nível por um pequeno período (seis a oito semanas), procurando meios de melhorar os desempenhos gerenciais, operacionais, sociais e tecnológicos. O processo identifica o fluxo de valor predatório – a capacidade exclusiva que lhe permite mover mais rápido e produzir melhor do que seus concorrentes – e suas áreas vulneráveis da fatia de mercado. O conhecimento fornecido pela tecnologia de data warehouse dá suporte para a VSA.

Na fase de *desenvolvimento do caso comercial* se podem identificar as tarefas necessárias para a criação do caso comercial para data warehouse. Nesse ponto, e equipe que justificará, projetará e implementará o data warehouse entra no processo. Eles usam pessoal colateral já desenvolvido pelos consultores ou pessoal interno (entrevistas, sessões de enfoque, análises estatísticas) para documentar: (a) uma estrutura de desmembramento de trabalho de alto nível para o projeto inteiro, (b) uma análise custo/benefício, incluindo um retorno do investimento, se possível, (c) os fatores críticos para o sucesso e (d) os impedimentos típicos do sucesso.

Para a estrutura de desmembramento de trabalho de alto nível para o projeto inteiro, não se precisa incluir as tarefas de baixo nível que seriam usadas no projeto real, mas deverá conter as tarefas e estágios em nível de resumo que refletem os principais esforços.

[MAR 99] ressalta o fato importante de que se ninguém na equipe tiver experiência com projeto de data warehouse, a especificação de tarefas poderá ser muito mais difícil, logo a contratação de um consultor experiente se faz necessário.

Na análise do custo e do benefício, [MAR 99] descreve a importância em se trabalhar com os gerentes comerciais e principais usuários comerciais para identificar e atribuir pesos relativos para os benefícios comerciais de alto nível da implementação de um data warehouse, a fim de dar suporte aos fluxos de valores ou iniciativas estratégicas. A gerência e os principais usuários comerciais podem fornecer os objetivos, os fatores críticos do sucesso e os planos de desenvolvimento futuros para a empresa, junto com uma estratégia para alcançá-los. Um data warehouse projetado efetivamente deverá ajudar uma organização a tomar decisões estratégicas que não podem ser feitas por meio de sistemas de transação em operação.

[MAR 99] apresenta os fatores críticos para o sucesso (CSFs-*critical success factors*) que precisam ser estabelecidos para que o projeto tenha sucesso, citando como fatores críticos: o caso comercial bem definido, projeto de arquitetura sadio, incluindo potencial para crescimento, uma quantidade de dados possível de ser gerenciada, um orçamento aprovado e disponível e um pessoal interno e externo dedicado.

[MAR 99] também cita os impedimentos críticos do sucesso (CSIs-*critical success inhibitors*) que podem impedir ou descarrilar o projeto, tais como:a falta de comprometimento e consciência dos patrocinadores executivos, o impacto de outros projetos estratégicos de tecnologia de informação, a incapacidade de extrair dados de sistemas de origem sem afetar contrariamente o desempenho do sistema de transação, o grau incontrolável de mudança organizacional durante o projeto, a falta de acesso aos dados de origem necessários, atribuições de pessoal em tempo parcial e a falta de uso de padrões e modelos para o gerenciamento de dados.

A fase de *revisão e projeto da arquitetura* define a tecnologia geral e a estrutura do processo. O estágio de revisão e projeto da arquitetura avalia quais partes da arquitetura já existem na organização (uma análise de lacunas). Essa fase desenvolve a arquitetura lógica do DW – o mapa de configuração dos locais de armazenamento de dados de componentes necessários, incluindo um local de armazenamento de dados central da empresa, um local de armazenamento de dados operacional opcional, um ou mais datamarts da área comercial individual (opcional) e um ou mais locais de armazenamento de metadados (contendo dois tipos diferentes de informações de referência de catálogo sobre os dados principais).

Em seguida, as arquiteturas de dados, aplicação, técnica e de suporte são criadas para implementar fisicamente a arquitetura lógica. Os requisitos para essas arquiteturas (explicados a seguir) são analisados cuidadosamente de modo que o DW possa ser otimizado para os usuários. Uma análise de lacuna descobre quais componentes de cada arquitetura já existem na organização e podem ser reutilizados, e quais devem ser desenvolvidos (ou comprados) e configurados.

A arquitetura de dados organiza as origens e locais de armazenamento das informações comerciais e define os padrões de qualidade e de gerenciamento para dados e metadados. Ela define a estrutura na qual os usuários vê-em o significado comercial para os dados nos seus locais de armazenamento, além de oferecer um mecanismo para catalogar os processos de transformação de dados que são necessários para o desenvolvimento do warehouse.

A arquitetura de aplicação é a estrutura de software que orienta a implementação geral da funcionalidade comercial dentro do ambiente do warehouse. Ela controla o movimento de dados da origem para o usuário, incluindo a extração, a limpeza, a transformação, o carregamento, a atualização e o acesso (relatórios, consultas).

A arquitetura técnica oferece a infra-estrutura de computação básica que capacita as arquiteturas de dados e de aplicação. Ela inclui plataforma/servidor, rede, hardware/software/middleware de comunicações e conectividade, SGBD, método cliente/servidor de duas ou três camadas e hardware/software de estação de trabalho do usuário final. O projeto de arquitetura técnica deve atender aos requisitos de tratamento de escalabilidade, capacidade e volume (incluindo tamanho e particionamento de tabela), desempenho, disponibilidade, estabilidade, cobrança e segurança.

A arquitetura de suporte inclui as funções necessárias para gerenciar o investimento de tecnologia de modo eficaz e os componentes do software de DW, tais como: ferramentas e estruturas para backup e recuperação, monitoração de desempenho, gerenciamento de controle/configuração de versão.

[MAR 99] enfatiza que esta fase de revisão e projeto da arquitetura aplica-se à estratégia de longo prazo para o desenvolvimento e refinamento do DW e não é realizada simplesmente numa única iteração.

A fase de *avaliação da questão comercial* (BQA – *business question assessment*) estabelece as áreas de assunto do DW, o escopo das iterações individuais do projeto e a estratégia de implementação em curto prazo, definindo e priorizando os requisitos comerciais, conforme estabelecidos no caso comercial, e outras necessidades de informações que o DW focalizará. Permite medir a qualidade, a disponibilidade e os custos relacionados dos dados de origem necessários em alto nível.

[MAR 99] sugere uma análise do fluxo de valores predatórios priorizados ou a iniciativa estratégica mais importante para decidir quais questões comerciais a implementação terá de responder. As questões comerciais impõem problemas que determinam a direção estratégica. Por exemplo, uma questão comercial para um revendedor poderia ser: "Quais foram os dez itens mais vendidos em todas as lojas da região X no segundo trimestre deste ano fiscal, onde dez mais vendidos são definidos como os dez itens mais altos em termos de receita total por item?"

Se analise cada questão para avaliar sua importância geral para a organização, e depois realize uma análise de alto nível dos dados necessários para fornecer as respostas. Analise-se a qualidade, a disponibilidade e o custo dos dados (para levá-los para o data warehouse). Use-se essa informação para re-priorizar as questões comerciais de acordo com a importância, o custo e a viabilidade (de adquirir os dados exigidos).

Use-se essa análise para decidir sobre o escopo das iterações previsíveis do DW, na forma de projetos de preenchimento de dados. Estima-se, com limites práticos de aquisição de dados, quantas questões comerciais poderão ser respondidas em uma implementação de três a seis meses. Uma questão comercial deverá ser responsável pela análise objetiva dos dados disponíveis. As questões comerciais precisam:

- Enfocar aspectos verificáveis. Uma implementação de DW não pode ser criada para satisfazer a uma necessidade que não possa ser expressa como uma questão comercial. A "consulta mais fácil" possui pouco ou nenhum valor no projeto de uma solução técnica;
- Ajudar a definir o escopo do projeto;
- Orientar diretamente o desenvolvimento do modelo lógico de dados identificando os objetos de dados que oferecem as informações necessárias ao warehouse;
- Definir os testes de aceitação do sistema. O sistema desenvolvido deverá responder a quaisquer comerciais no escopo.

[MAR 99] ainda sugere a realização de sessões de enfoque e entrevistas para desenvolver a lista das principais questões comerciais para as quais o DW deverá fornecer respostas. Identificar as ações que possam ser realizadas por meio de consultas do usuário e na aquisição de dados.

A fase de plano de implementação da iteração permite revisar o plano de projeto original para incluir um plano de trabalho para as tarefas de preenchimento desse ciclo. Caso novos enfoques mudarem o escopo original, a revisão ao escopo será necessária através de novas iterações ao estudo do caso comercial.

Enquanto o que a arquitetura define a estratégia e os processos gerais pelos quais o DW é desenvolvido, a fase de *projeto detalhado* mapeia a implementação desses processos para a iteração. Esta fase desenvolve o modelo físico do DW (esquema de banco de dados), os metadados e atualiza o inventário de dados de origem necessários para a implementação da área de assunto.

A fase de *implementação* pode ser caracterizada pelo fato de que tudo que foi realizado até esse ponto prepara o caminho para uma implementação fácil. [MAR 99] cita que nesta fase é necessário executar algumas etapas, tais como: (a) compra e instalação dos componentes do DW, (b) preparação da base de teste, (c) teste do processo de ETL, (d) testes nos programas de atualização, (e) criação dos módulos de acesso aos usuários, (f) criação e teste das consultas OLAP e (g) registro da aceitação dos usuários.

Para [MAR 99] a fase de *transição para a produção* pode ser vista como um termômetro de que se a execução das fases anteriores foi bem feita, a transição deverá ser rápida e tranquila. Porém, se algo não ficou bem planejado, a equipe do projeto terá que realizar tarefas de desenvolvimento e produção durante a próxima iteração. As principais atividades desta fase são: (a) mover todos os componentes do sistema do ambiente de desenvolvimento para o ambiente de produção, (b) treinar o pessoal de operações e usuários finais, (c) realizar a documentação do sistema operacional e (d) oferecer um *warehouse* que seja totalmente operacional e disponível aos usuários.

A última fase descrita por [MAR 99] é a fase de *manutenção,* necessária uma vez que um DW cresce e muda com o tempo. Podem ocorrer alterações de hardware, software, sistemas de origem adicionais, processos de aplicação adicionais, mudanças no ciclo de atualização, mudanças nas atividades e responsabilidades de controle de versão.

Para a execução das fases, [MAR 99] descreve a necessidade de uma abordagem de quatro áreas comuns a qualquer projeto: (a) Planejar, (b) Ativar, (c) Controlar e (d) Encerrar.

Na área de *planejamento* são estabelecidos os objetivos, o escopo e os padrões do projeto. [MAR 99] enfatiza a necessidade de considerar, nesta etapa, um método para posterior contratação de consultores especializados em projetos de DW.

Para *ativar* o projeto, [MAR 99] descreve que é necessário publicar o projeto, equipar e treinar os membros da equipe. Os usuários comerciais deverão conduzir a tecnologia ajudando a definir os requisitos iniciais e os requisitos para iterações futuras. O pessoal de TI (Tecnologia da Informação) deverá saber das necessidades gerais da comunidade comercial, mas são necessários analistas comerciais para conduzir os requisitos reais de informação com base nas necessidades da comunidade de usuários. Ë importante, no final deste estágio, o treinamento da equipe de projeto sobre as ferramentas que serão utilizadas no decorrer do projeto, como e-mail, gerenciador de projetos e processadores de texto. As ferramentas de DW serão obtidas mais tarde.

Na área de *controle*, [MAR 99] afirma que o mesmo ocorrerá até o final do projeto e se divide em quatro etapas: (a) atribuir tarefas de projeto, (b) motivar os participantes do

projeto, (c) rastrear o processo do projeto e (f) revisar o plano do projeto. A etapa de atribuir tarefas é combinar a cada atividade planejada, o recurso mais apropriado para a sua execução. Como um projeto de DW é contínuo, deve-se trabalhar, na etapa de motivação, para promover o desenvolvimento individual, criar incentivos para o trabalho de equipe, reconhecendo as realizações. A etapa de rastreamento se caracteriza por um acompanhamento sobre o estado do projeto e apresentar as informações para os membros da equipe e responsáveis pelo projeto. Para [MAR 99], a etapa de revisar é a mais importante, pois se deve avaliar cuidadosamente as sugestões e os pedidos do usuário para filtrar os recursos mais interessantes, que podem aumentar o custo final do projeto.

Para [MAR 99], a área de *encerrar* é algo mais que o simples término do projeto de DW, o qual afirma que se deve arquivar o material do projeto, gerar relatório sobre o andamento do projeto, passar os resultados do projeto para o pessoal de operação e suporte e liberar os recursos do projeto para uso em outros projetos.

4.2 A Metodologia de Ralph Kimball

A proposta de [KIM 98a] destaca-se por ser no geral o que mais caracteriza e detalha as diversas fases da metodologia, considerando as fases de: (a) planejamento, (b) levantamento de requisitos, (c) arquitetura funcional, (d) projeto da base de dados, (e) aplicações de usuários finais, (f) auditoria nos dados e (g) uso, suporte e extensão do DW.

A Figura 4.2 mostra o ciclo de vida na qual apresenta as fases são executadas sequencialmente e em paralelo. Ao final de cada uma, [KIM 98a] descreve a necessidade de executar as atividades de revisão e aceitação pelo usuário final, além da revisão do projeto.

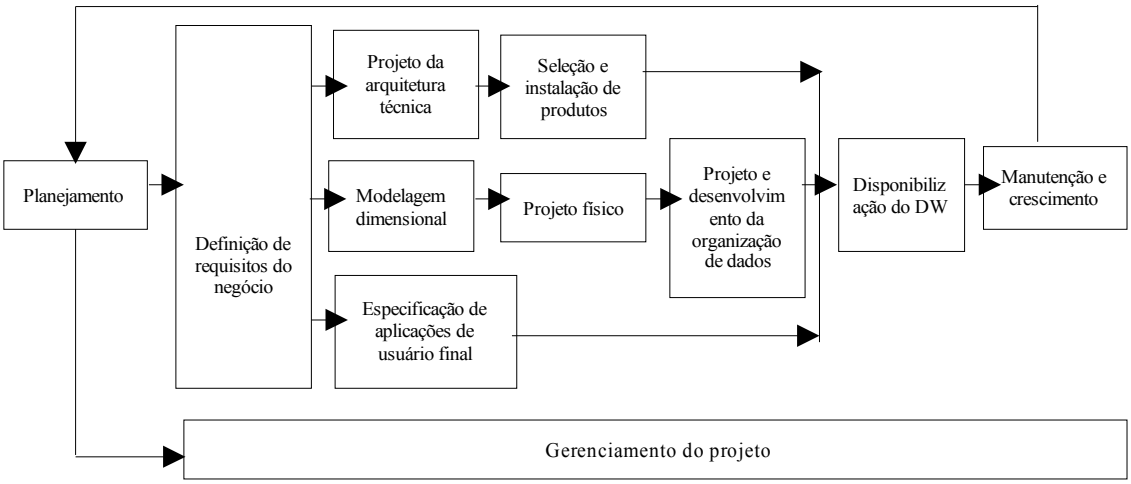


Figura 4.2: Ciclo de vida de [KIM 98a].

O ciclo de vida do desenvolvimento começa com o *planejamento*, que segundo [KIM 98a] pode variar de empresa para empresa. É uma das atividades mais críticas pois a qualidade dos levantamentos e definições afetarão o projeto como um todo.

O principal resultado desta fase é um Plano de Projeto, com os dados levantados tais como: a justificativa do projeto, custos envolvidos, identificação dos integrantes do

projeto, especificação do cronograma do projeto e das diferentes fases e atividades do DW.

Na fase de *requisitos do negócio* são definidos os fatores chaves do negócio do usuário, imprescindíveis para as próximas etapas. Esta fase se caracteriza pela identificação e preparação da equipe de entrevistas, seleção de entrevistadores para cada entrevista, análise dos resultados de cada entrevista e revisão do Plano de Projeto. Segundo [KIM 98a], a fase de *projeto da arquitetura técnica* é uma das mais importantes no projeto DW, onde se define uma arquitetura de alto nível (área interna e externa), e a especificação da infra-estrutura técnica e os respectivos componentes necessários para permitir a criação do DW. Na fase de *seleção e instalação de produtos* consiste na realização de pesquisa, estudo e seleção dos produtos relacionados a construção de um DW, desenvolvimento de protótipos para melhor avaliação das funcionalidades dos softwares e sua aceitação pelo usuário final, além da revisão do projeto.

A fase de *modelagem dimensional* consiste em agregar os dados levantados na fase de requisitos do negócio para desenvolver um modelo de dados. Durante esta fase também é realizada a análise das diversas fontes de dados de modo a identificar aquelas que possuem os dados necessários para atender o modelo de dados, assim como procurar especificar dentre estas as melhores fontes de dados. Essa análise compreende as seguintes atividades:

- Identificação das fontes de dados candidatas, verificando detalhes tais como sistema OLTP que integra, plataforma (e.g.UNIX, Windows NT, etc.), estrutura física dos arquivos (e.g. flat files, Oracle, Excel, etc.), volume de dados (e.g. número de transações por semana, etc.), identificação de chaves primárias e estrangeiras e características dos campos de dados (e.g. tipos de dados, comprimento, precisão, etc.);
- Estudo do mapeamento de dados das fontes candidatas para as tabelas de destino dos dados;
- Estimativa do número de linhas ou registros das fontes de dados candidatas.

Na fase de *projeto físico* são realizadas as seguintes atividades de definição de nomes padronizados para os objetos da base de dados, a execução do projeto físico, criando os objetos da base de dados analítica e o desenvolvimento de um plano inicial de indexação, agregação e particionamento.

A fase *projeto e desenvolvimento da área de organização de dados*, diz respeito a atividades fundamentais a serem realizadas no DW: a extração, transformação e carga de dados. As principais atividades que compreendem esta fase são a criação de uma arquitetura de alto nível que represente o fluxo de dados dos sistemas fonte para a base de dados analítica de destino, o teste e escolha de ferramentas de terceiros ou desenvolvimento de programas específicos para a realização das atividades de organização de dados e o detalhamento da arquitetura de alto nível, determinando, por exemplo, quais tabelas e em que ordem será realizada, a atividade de extração, transformação e carga, quais as atividades de transformação que serão realizadas nas diversas tabelas, etc.

Para uma definição e realização de modificações lógicas e atualização dos registros das dimensões, quando necessário. [KIM 98a] propõem três técnicas básicas quando um registro da dimensão é atualizado: substituição pelo registro mais recente, criação de

novo registro na dimensão ou, por último, criação de um novo campo na dimensão que armazene somente o campo alterado, de forma que sejam armazenados os campos novos e antigos.

[KIM 98a] enfatiza nesta fase ainda a realização do processo de organização de dados com as demais dimensões, assim como a tabela fato, o desenvolvimento de procedimentos que permitam a carga incremental de tabelas fato que sejam muito grandes, utilizando os recursos baseados em novas transações, "logs" de bancos de dados, replicação, realização de múltiplos passos de carga e execução paralela e carga de tabelas de agregados.

Na fase *especificações de aplicações de usuário final* deve-se procurar identificar as áreas prioritárias e, a partir destas, definir um conjunto padronizado de aplicações destinadas aos usuários finais, uma vez que não são todos os usuários que necessitam Ter acesso "ad hoc" aos dados do DW.

Na fase *desenvolvimento de aplicações de usuário final* são desenvolvidas as aplicações necessárias de acordo com levantamentos realizados na fase de "especificações de aplicações de usuário final". A seleção do ambiente de desenvolvimento dos relatórios e o desenvolvimento de procedimentos de manutenção e atualização das aplicações de usuário final, são atividades que compreendem esta fase.

Para [KIM 98a] a *fase de disponibilização do DW* é composta basicamente pelas seguintes atividades de montar o plano de verificação da infra-estrutura, estratégia de treinamento dos usuários finais, estratégia de suporte ao usuário final, plano de atualização de versão do DW, um teste completo do sistema e a disponibilização do DW propriamente dita aos usuários finais.

Na fase de *manutenção e crescimento do DW* é composta basicamente pelo contínuo suporte e treinamento dos usuários e manutenção da infra-estrutura técnica, além do monitoramento de consultas realizadas pelos usuários finais, desempenho da organização de dados e o contínuo sucesso do DW.

4.3 A Metodologia de Vidette Poe

O trabalho de [POE 98] em relação aos outros já mencionados, apresenta o diferencial que é o detalhamento de variações de esquemas multidimensionais e um enfoque bastante significativo sobre projeto piloto.

Em seu trabalho [POE 98] deixa bem claro a importância de se estabelecer, antes da construção do DW, uma arquitetura de dados adequada, que realmente atenda às necessidades da corporação. A grande importância da abordagem de [POE 98] quanto ao estabelecimento inicial de uma arquitetura de dados deve-se ao fato de que, em decorrência desta escolha, derivar-se-á a infra-estrutura técnica necessária para atender o DW, normalmente composta por tecnologias e itens tais como treinamentos em tecnologias de suporte à decisão, plataformas, bases de dados, ferramentas de conversão de dados, hardware, software, rede local, ferramentas de acesso aos dados, etc. Ressalta-se novamente que a arquitetura de dados está intimamente relacionada com a infra-estrutura, ou seja, os componentes e tecnologias da infra-estrutura a serem utilizados dependerão diretamente da arquitetura de dados escolhida.

A Figura 4.3 descreve as etapas do ciclo de desenvolvimento do projeto de DW, proposto por [POE 98].

A fase de *planejamento* preocupa-se com a definição e/ou clarificação do escopo do projeto, a criação de um plano de projeto, definição dos participantes e responsabilidades, além das tarefas e prazos de conclusão e a montagem de um cronograma, incluindo o prazo final de entrega do projeto. Segundo [POE 98] não é conveniente construir um DW antes de estabelecer corretamente a arquitetura e a infra-estrutura, o que poderá acarretar insucesso no seu desenvolvimento e uso, desta forma, caso a infra-estrutura não esteja pronta, deverão também fazer parte do planejamento do DW.

A fase de *levantamento de requisitos e modelagem* preocupa-se com a definição do negócio, exigências e necessidades dos usuários, e por último, a modelagem do universo de discurso, do qual fazem parte as referidas exigências e necessidades. Esta fase apresenta duas etapas distintas, denominadas de "levantamento de requisitos" e "modelagem".

Na etapa de levantamento de requisitos são realizadas atividades para a identificação do entendimento de como o usuário realiza as diversas atividades de negócio, quais os fatores que dirigem o negócio, quais os atributos o usuário necessita e são absolutamente requeridos e quais são desejados, a estrutura organizacional, quais as ferramentas de acesso aos dados e quais os resultados o usuário espera em suas consultas.

Por sua vez a modelagem de dados refere-se ao processo de tradução dos conceitos dos conceitos do negócio em um modelo dimensional em formato diagramático que inclua fatos, dimensões, hierarquias, relacionamentos, chaves candidatas, etc. [POE 98] apresenta, além dos tradicionais esquemas "estrela" e "flocos de neve", diversas variações, tais como "esquema estrela com múltiplas tabelas fato", "esquema estrela com tabelas associativas", "esquema estrela com tabelas externas" e "esquema multi-estrela", descrevendo a necessidade de adequação de um modelo de dados às condições locais da corporação e seus negócios, apresentando um modelo híbrido.

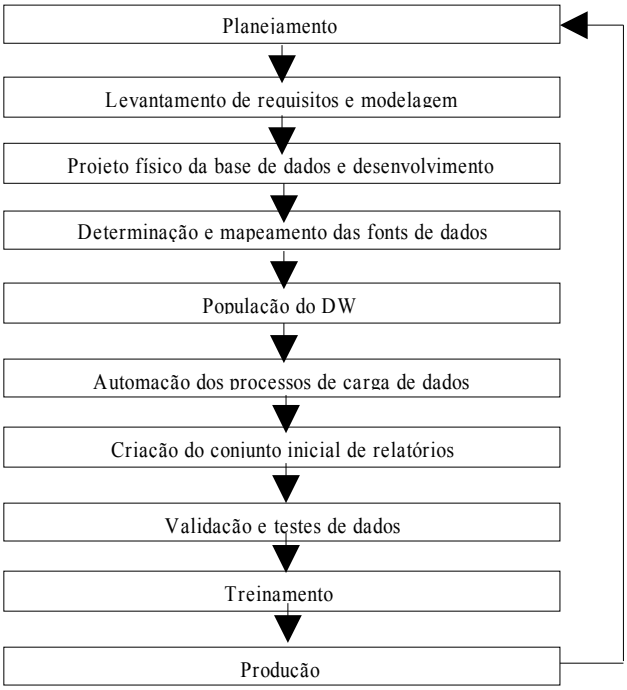


Figura 4.3: Ciclo de vida de [POE 98].

Na etapa de *projeto físico da base de dados e desenvolvimento* são criadas fisicamente na base de dados, as tabelas de fato e dimensão e seus relacionamentos, a desnormalização de dados e as estratégias de criação de índices, de agregação e particionamento.

Segundo [POE 98] a fase de *determinação, integração e mapeamento das fontes de dados* são a que consome mais tempo de desenvolvimento, devido à necessidade de localizar os dados adequados dispersos em sistemas OLTP, analisar e entender os tipos de dados, implementar os processos de transformação necessários e mapear os campos das fontes de dados para os objetos da base de dados, o desenvolvimento de programas que permitam realizar as conversões de dados para cada campo e refinar a estratégia de integração.

Na fase de *população* do DW é feito o desenvolvimento de: programas ou utilização de ferramentas de conversão de dados para integrar os dados, para extrair e mover os dados, carga de dados dentro do DW, desenvolvimento de estratégias de reconsulta e atualização de dados e a execução de programas e procedimentos de extração, transformação e carga de dados.

A fase de *automação dos processos de carga de dados* caracteriza-se pelo agendamento dos processos de extração, conversão e carga de dados, construção de procedimentos de *backup* e recuperação de dados.

Na fase de *criação de conjunto inicial de relatórios* são realizadas as construções de relatórios pré-definidos.

[POE 98] descreve a *fase de validação e teste de dados* como sendo a validação de dados utilizando o conjunto inicial de relatórios e a validação dos dados através de processos padronizados.

Na fase de *treinamento* tem-se a criação de programas de treinamento para atender especificamente a comunidade de usuários, levando em conta o treinamento nas ferramentas de acesso aos dados, obtendo as informações desejadas no DW.

A fase de *produção* inclui as tarefas necessárias para a disponibilização do DW e o correspondente suporte necessário aos usuários finais, a exemplo de aplicações de suporte à decisão.

[POE 98] considera que a interação contínua dos usuários com o DW, após a sua disponibilização através de ferramentas e aplicações, possibilita o surgimento de novas exigências e consequentemente se fazendo necessário a execução, novamente, do ciclo de desenvolvimento do DW, para que as modificações sejam feitas.

Para [POE 98] o desenvolvimento de DW exige que a equipe de profissionais responsável pela sua construção seja integrada por pessoas experientes no assunto. Mas mesmo assim existem ainda muitas empresas que, antes de desenvolverem um DW, constroem inicialmente um Projeto Piloto de modo a evitar um possível insucesso, assim como possibilitar ganho de experiência, dentre outros motivos.

[POE 98] aborda a construção do Projeto Piloto de duas formas distintas: via prova de concepção e via arquitetura e infra-estrutura. A prova de concepção possibilita uma apresentação prática ao comitê diretor da corporação sobre as possibilidades de um DW. Neste tipo de Projeto Piloto, os usuários podem interagir com o sistema obtendo algumas informações de suporte à decisão, possibilitando entender como elas poderão assisti-los na tomada de decisões. Normalmente este tipo de Projeto Piloto é

desenvolvido rapidamente, quando comparado com a segunda forma, uma vez que se apoia sobre um pequeno conjunto de dados. Adicionalmente, não há necessidade de que todos os componentes técnicos e de infra-estrutura estejam disponíveis.

Já na forma via arquitetura e infra-estrutura, por sua vez, é usado para se verificar como todos os componentes do DW trabalham juntos, bem como para entender e ganhar experiência em todas fases do ciclo de vida de desenvolvimento. Esse tipo de Projeto Piloto apoia-se também sobre um conjunto restrito de dados, os quais, entretanto, passam por todas as fases do ciclo de vida de desenvolvimento, incluindo as arquiteturas do DW correspondentes.

4.4 A Metodologia de Alan Perkins

O trabalho de [PER 2000] em relação aos outros já mencionados, apresenta de forma completa e bastante detalhada toda a metodologia de desenvolvimento de um DW. Sua principal contribuição está na elaboração de um protótipo, o qual [PER 2000] conceitua como sendo um "pré-projeto piloto", que se apoia sobre um conjunto restrito de dados, passando por todas as etapas do desenvolvimento, tais como: arquitetura, técnicas de modelagem dimensional, projeto de banco de dados, etc.

A Figura 4.4 descreve as fases do ciclo de desenvolvimento do projeto de DW, que segundo [PER 2000] são divididas em três etapas distintas: (a) experimentação, (b) definição e (c) execução. A experimentação possibilita um contato inicial com as técnicas de desenvolvimento de DW, obter experiência com novas ferramentas, tecnologias, definição de prazos, alem de levantar os requisitos necessários para gerar as informações de suporte a decisão. A etapa de definição permite a redução das incertezas e utilização das experiências obtidas na etapa anterior para a definição do projeto de DW como um todo. Na etapa de execução se desenvolve efetivamente o projeto de DW, considerando os resultados gerados pelas etapas anteriores. [PER 2000] enfatiza em sua proposta que existem dois momentos distintos: a construção de um *projeto piloto*, considerando todas as fases de cada uma das etapas acima descritas e o *protótipo*, montado na etapa de experimentação, servindo como um teste para a equipe de projeto de DW.

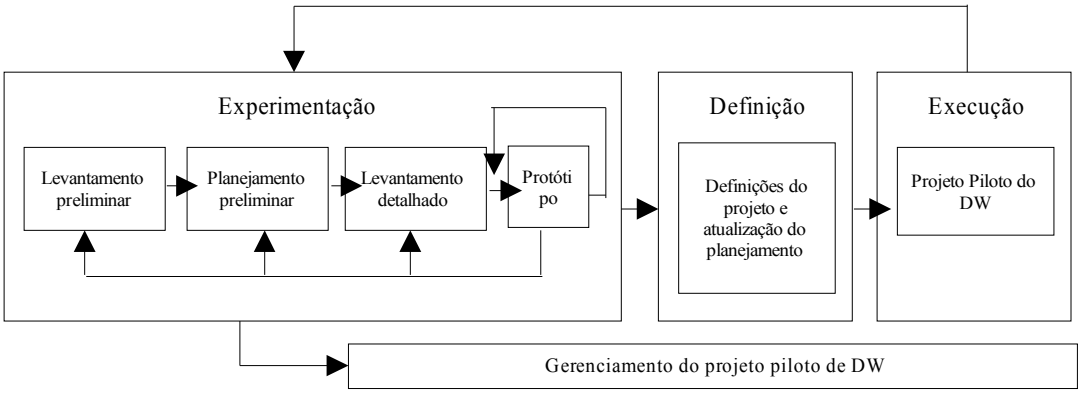


Figura 4.4: Ciclo de vida de [PER 2000].

4.4.1 As fases para cada etapa da metodologia [PER 2000]

A etapa de *experimentação* é composta pelas seguintes fases: (a) levantamento preliminar, (b) planejamento preliminar, (c) levantamento detalhado e o (d) protótipo.

A fase de *levantamento preliminar* consiste no maior levantamento possível de dados a respeito do ambiente organizacional, os usuários finais e o grau de participação deles no projeto. Nesta fase também são realizadas as reuniões de requisito para que se possam levantar os requisitos gerais do negócio da organização e se procura identificar as possíveis fontes de dados, que servirão para a definição das atividades de extração, transformação e carga.

Na fase de *planejamento preliminar* os dados descritos na fase anterior são analisados e repassados para um documento chamado por [PER 2000] de "Plano de Projeto", contendo a estrutura do projeto, análise de sua viabilidade e necessidade, que norteiam o principal objetivo do projeto de DW. Fazem ainda parte desta fase as estratégias de: levantamento de requisitos mais detalhados, integração ou transformação de dados, segurança e atualização dos dados. [PER 2000] enfatiza a necessidade de apresentar o "Plano de Projeto" aos usuários finais do projeto, para possíveis ajustes, aprovação ou até mesmo reprovação.

A fase de *protótipo* é a mais detalhada por [PER 2000], pois representa uma fase similar no ciclo de vida de desenvolvimento de DW. Esta fase é a principal dentro da etapa de *experimentação*, pois sustenta a execução de um pré-projeto piloto através de um conjunto de dados restritos. [PER 2000] sustenta que este protótipo permite estabelecer o contato inicial com as diferentes técnicas de desenvolvimento de DW, ganhar experiência com novas ferramentas e tecnologias e aprender as diversas tarefas que compõem cada fase do desenvolvimento do DW.

A etapa de *definição* permite reduzir as incertezas e a transformar as experiências obtidas nas fases da etapa de experimentação, em definições de projeto que servirão como base para a construção do projeto piloto de DW.

4.4.2 Fase de protótipo

Esta fase, como já descrita anteriormente, apresenta-se como um "projeto piloto", conforme descrito na etapa de "arquitetura e infra-estrutura", proposta por [POE 98], porém detalhando as etapas do projeto de DW. É a fase mais importante da etapa de experimentação, pois permite uma prévia do projeto, seguindo todas as etapas de desenvolvimento do projeto de DW. [PER 2000] enfatiza que esta fase é o principal diferencial sobre as outras metodologias propostas, pois sustenta uma de suas primícias de que um projeto de DW pode ser elaborado por uma equipe sem experiência.

Uma das principais vantagens, sobre este modelo, é que se permite que a fase de protótipo, como um todo, possa ser repetida tantas vezes quantas forem necessárias, dentro das disponibilidades de tempo, pessoal, ferramentas, etc. [PER 2000] enfatiza que o protótipo permitirá o teste aprofundado e repetitivo de produtos e ferramentas a partir das diversas configurações e tecnologias adotadas nos módulos precedentes.

Para que isto possa ocorrer, [PER 2000] descreve que esta fase se divide nos seguintes módulos: (a) planejamento, (b) arquitetura de dados, (c) arquitetura funcional, (d) infra-estrutura, (e) teste de produtos, (f) modelagem dimensional, (g) projeto do BD, (h) execução da arquitetura funcional, (i) aplicações de usuários finais, (j) auditoria nos dados, (k) uso, suporte e extensão e (l) gerenciamento do protótipo.

4.5 Análise das metodologias

As metodologias, apresentadas nas seções anteriores, apresentam características e limitações com relação ao processo de desenvolvimento do projeto de DW. Para esta análise, se enfatiza os seguinte conjunto de critérios: (a) ser completa, considerando todas as etapas do projeto de DW, (b) exigir experiência em projetos de DW anteriores, (c) detalhamento na definição de cada etapa apresentada, (d) apresentação de um anteprojeto para análise teste de tecnologias, (e) iteração completa sobre todas as etapas do projeto de DW e (f) análise dos riscos sobre o projeto de DW.

A tabela 1 descreve uma análise comparativa das metodologias estudadas, considerando os critérios citados anteriormente. É importante lembrar que o resultado apresentado sobre este estudo justifica a necessidade de definição de uma metodologia adequada, como será discutida no Capítulo 5 e analisada no Capítulo 6 a partir de um estudo de caso para a Cia Zaffari, já que as metodologias apresentadas não satisfazem a todos os critérios estabelecidos.

Tabela 1 – Análise das metodologias de projeto de DW.

Critérios	*[MAR 99]*	*[KIM 98a]*	*[POE 98]*	*[PER 2000]*
Completa	Sim	Sim	Sim	Sim
Experiência	Sim	Sim	Sim	Não
Detalhamento	Pouco detalha	Detalha	Detalha	Detalha
Anteprojeto	Não possui	Possui	Possui	Possui
Iteração completa	Não	Não	Não	Não
Análise de riscos	Não	Sim	Sim	Sim

5 PROPOSTA DE UMA METODOLOGIA PARA CRIAÇÃO DE DW

Este capítulo apresenta uma proposta de metodologia para o desenvolvimento de projeto de DW. Pelo fato, já descrito no capítulo anterior, sobre a não adequação de uma metodologia existente, considerando principalmente os critérios sobre a completude, iteração, detalhamento, anteprojeto, análise de riscos e experiência em projetos já elaborados, se faz necessário à apresentação de uma descrição detalhada da metodologia. Inicialmente será apresentada uma visão sucinta da metodologia para posterior detalhamento.

Entre as metodologias estudadas no capítulo anterior, pode-se observar que [POE 98], [KIM 98a] e [PER 2000] são as mais detalhadas e completas, sendo desta forma, utilizados como referencial conceitual para a definição da metodologia de projeto de DW. Vale ainda descrever que a proposta está embasada na metodologia apresentada em [FUR 98] e [BOO 2000], através dos conceitos do Processo Racional Unificado - RUP (*Rational Unified Process*).

5.1 Ciclo de vida da metodologia

O ciclo de vida de desenvolvimento de sistemas clássico não pode ser aplicado ao usuário de um EIS (*Executive Information System* – Sistema de Informações Executivas). Em sistemas tradicionais, a metodologia clássica destina-se ao ambiente operacional, parte-se do princípio que todos os requisitos sejam conhecidos ou desvendados na parte de análise e projeto, o que não acontece no desenvolvimento de um EIS ou DW. O DW funciona segundo um ciclo de vida diferente, começando pela identificação dos dados, sua integração e testes para verificar distorções. Então é feita a codificação dos programas de interface para os dados. Os resultados são analisados pelos usuários e finalmente os requisitos do sistema são compreendidos.

A proposta de metodologia possui um ciclo de vida denominado “ciclo de vida repetitivo”, o qual descreve uma iteração gradativa sobre a execução de todas as etapas do projeto de DW. Esta metodologia possui como principais características:

- É um processo iterativo o qual requer uma compreensão crescente do problema por meio de aperfeiçoamentos sucessivos e o desenvolvimento incremental de uma solução efetiva em vários ciclos;
- Possui flexibilidade para a acomodação de novos requisitos ou de mudanças táticas de objetivos de negócio;
- Permite que o projeto identifique e solucione riscos de início, em vez de posteriormente;

- Encoraja o controle de qualidade e o gerenciamento de riscos, contínuos e objetivos – a avaliação da qualidade é inserida no processo, em todas as atividades e envolvendo todos os participantes, com a utilização de medidas e critérios objetivos;
- O gerenciamento de riscos é inserido no processo, de forma que os riscos para o sucesso do projeto são identificados e atacados no início do processo de desenvolvimento, quando há tempo suficiente para uma reação adequada.

Existem quatro fases no ciclo de desenvolvimento de um projeto de DW, como mostra a Figura 5.1: concepção, elaboração, construção e transição. Cada fase e iteração têm algum foco de redução de riscos e concluem um marco de progresso bem-definido. A consideração do marco de progresso proporciona um ponto no tempo para avaliar como as metas foram alcançadas e se o projeto necessitará ser re-estruturado de alguma maneira para prosseguir.

Para [BOO 2000], a fase de *Concepção* é a fase onde é estabelecido o caso do negócio para o sistema e é delimitado o escopo do projeto. O caso de negócio inclui critérios de sucesso, a avaliação de riscos, estimativa de recursos necessários e um plano para cada fase, mostrando a programação de principais marcos de progresso. Durante a concepção, é comum a criação de um protótipo executável, servindo como teste para a concepção. No final desta fase os objetivos do ciclo de vida do projeto são examinados para poder decidir se deve prosseguir com o desenvolvimento em plena escala.

A fase de *Elaboração* é a fase onde é feita uma análise do domínio do problema, o estabelecimento da fundação de uma arquitetura sólida, o desenvolvimento de um plano do projeto e a eliminação dos elementos de mais alto risco do projeto. [BOO 2000] descreve que as decisões de arquitetura devem ser feitas com uma compreensão de todo o sistema. No final desta fase é examinado o escopo e os objetivos detalhados do projeto, a escolha da arquitetura e a solução para os principais riscos, além de decidir se deve prosseguir com a construção.

Na fase de *Construção* é desenvolvido de maneira iterativa e incremental, um produto completo, pronto para a transição junto a comunidade de usuário. Isso implica uma descrição dos requisitos restantes e de critérios de aceitação, dando corpo ao projeto e concluindo a implementação e o teste do DW. No final desta fase é decidido se o ambiente e usuários estão todos prontos para se tornarem operacionais.

A fase de *Transição* ocorre quando o software se torna disponível aos usuários. Quando o sistema é colocado nas mãos dos usuários finais surgem questões que requerem algum desenvolvimento adicional, com a finalidade de ajustar o sistema ou corrigir alguns problemas identificados. Essa fase tipicamente é iniciada com uma versão beta do projeto, que depois é substituída pelo projeto de produção.

No final dessa fase é feita uma avaliação para ver se os objetivos do ciclo de vida do projeto foram alcançados e é determinado se deverá iniciar outro ciclo de desenvolvimento. Esse também é um ponto em que as lições aprendidas no projeto deverão ser assimiladas para aprimorar o processo de desenvolvimento e serem aplicadas no próximo projeto.

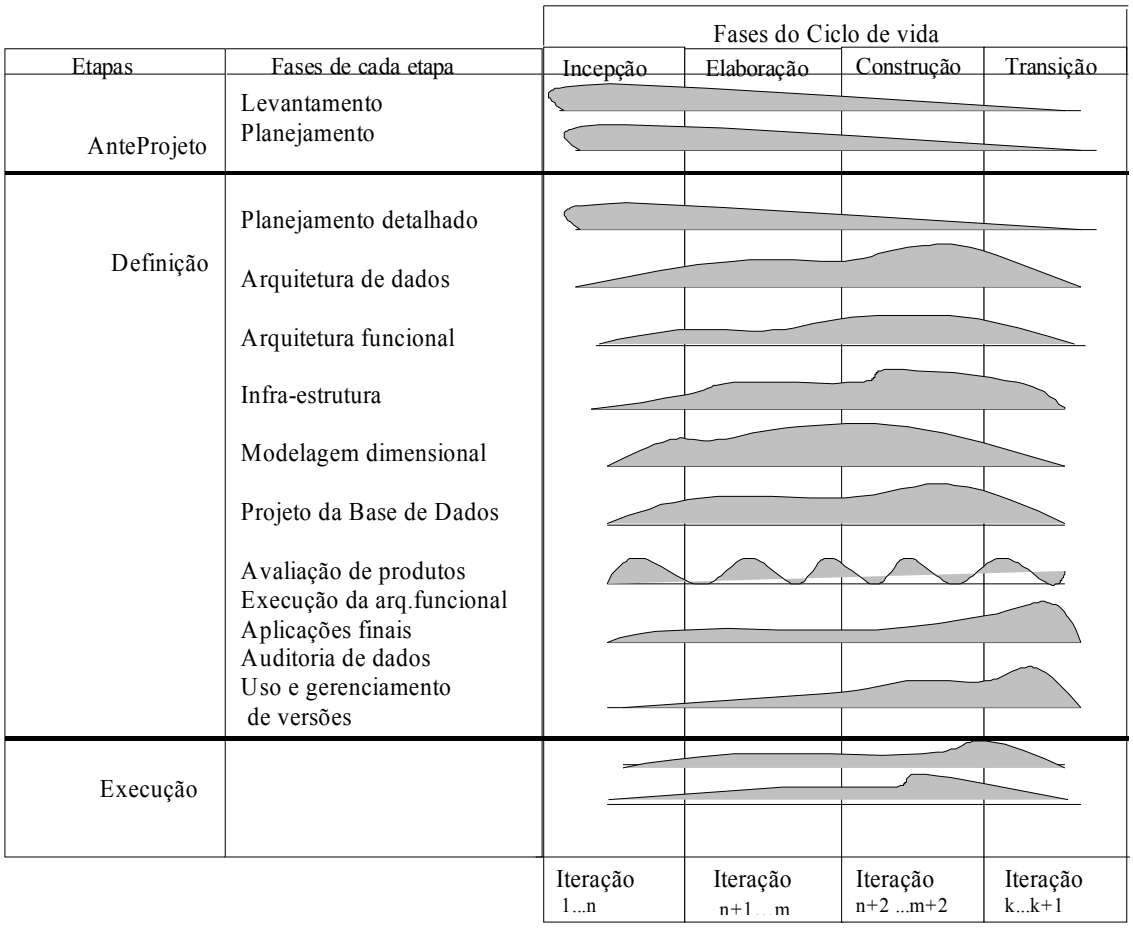


Figura 5.1: Ciclo de vida repetitivo.

5.1.1 Iterações

Cada fase da metodologia ainda pode ser dividida em iterações. Uma iteração é um ciclo completo de desenvolvimento, resultando em uma versão (interna ou externa) de um produto executável que constitui um subconjunto do produto final em desenvolvimento e cresce de modo incremental de uma iteração para outra para se tornar o sistema final. Cada iteração passa pelos vários fluxos de trabalho do processo, embora com uma ênfase diferente em cada um deles, dependendo da fase. Durante a concepção, o foco está na captação de requisitos. Durante a elaboração, o foco passa a ser a análise e o projeto. A implementação é a atividade central na construção e a transição, na entrega.

Desta forma, podemos conceituar a iteração como sendo uma execução completa de as etapas de um projeto de DW (anteprojeto, definição e execução) para cada uma as etapas do ciclo de desenvolvimento, iniciando na fase de incepção e terminando na fase de transição. Para cada fase do ciclo de desenvolvimento podemos ter várias repetições de execução, dependendo do nível de definição do projeto.

5.1.2 Ciclo de desenvolvimento

A passagem pelas quatro principais fases é chamada um ciclo de desenvolvimento e resulta na geração de um projeto de DW. O primeiro passo das quatro fases é chamado

o ciclo inicial do desenvolvimento. A menos que a vida do produto seja interrompida, um produto existente evoluirá para sua próxima geração pela repetição da mesma sequencia de fases de concepção, elaboração, construção e transição. Isso é a evolução do produto, de modo que os ciclos de desenvolvimento posteriores aos ciclos iniciais são seus ciclos de evolução [BOO 2000].

5.2 Fluxo de trabalho da metodologia

A metodologia proposta é composta basicamente de três etapas, ilustradas na Figura 5.2, aqui denominada de fluxos de trabalho, são descritas a seguir.

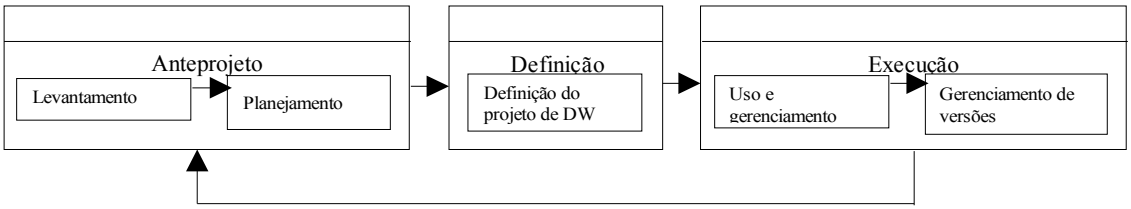


Figura 5.2: Fluxo de trabalho da metodologia.

5.2.1 Anteprojeto

A etapa de *anteprojeto*, algo similar às apresentadas em [POE 98], [KIM 98a] e [PER 2000], permite se fazer o levantamento das necessidades inerentes ao projeto de DW. Possibilita aos diretores e decisores de mais alto nível uma visão de como o DW pode ser viável e valioso para a corporação. Um planejamento não detalhado sobre as possibilidades de um DW também é descrito nesta etapa.

A fase de *levantamento*, considerando os trabalhos apresentados em [DYC 98] e [POE 98], deve ser iniciada através da definição dos recursos técnicos necessários, o entendimento do negócio e as necessidades dos usuários. As principais características desta fase são:

- Identificar o ambiente e influência organizacional, para levantar os fatores "humanos" que cercam o ambiente organizacional, tais como a disposição política da cúpula da corporação para realizar o projeto de DW;
- Realizar reunião para a coleta e o entendimento dos requisitos, junto aos usuários e tomadores de decisão;
- Analisar os requisitos levantados, definindo quais os dados que o usuário utiliza e quais deveriam utilizar;
- Definir o nível de detalhes que o usuário necessita;
- Definição do processo a modelar, considerando os requisitos levantados;
- Determinar as fontes de dados que serão necessárias para realizar as atividades de ETL.

Depois de concluída a fase de levantamento, pode-se realizar a fase de *planejamento*, considerando alguns itens importantes definidos nos trabalhos de [GAR 98], [POE 98] e [KIM 98a], tais como:

- Identificação dos membros da equipe do projeto de DW, assim como de suas funções;

- Levantamento dos fatores que permitem mensurar o sucesso assim como os riscos do projeto de DW;
- Prever as ferramentas necessárias tanto para a efetiva construção do DW, como para as consultas a base de dados analítica;
- Estabelecimento de cronogramas temporais, não detalhados, que permitam a efetiva construção do DW;
- Elaboração das justificativas para o projeto, considerando os investimentos necessários e os seus reais custos, além dos resultados e principais benefícios a serem alcançados.

Esta etapa do projeto se conclui com a elaboração de um Plano de Projeto, contendo os dados levantados e analisados. Deve possuir um título para a identificação do projeto de DW, a definição dos participantes do projeto, tais como os membros da equipe e usuários tomadores de decisão, assim como de suas correspondentes tarefas. Este Plano de Projeto deverá ser atualizado e completado pelas próximas etapas.

5.2.2 Definição

A etapa de definição do projeto inicia logo após o anteprojeto ter sido concluído. Esta etapa possibilita uma definição completa sobre o Plano de Projeto, baseado nos trabalhos de [POE 98], [KIM 98] e [PER 2000], contendo as fases de: (a) planejamento detalhado, (b) arquitetura de dados, (c) arquitetura funcional, (d) infra-estrutura, (e) modelagem dimensional, (f) projeto da base de dados, (g) avaliação de produtos, (h) execução da arquitetura funcional, (i) aplicações finais e (j) auditoria de dados. A Figura 5.3 descreve o fluxo da etapa de definição.

A fase de *planejamento detalhado* destina-se ao aprofundamento dos dados levantados na etapa de "Ante-Projeto". Ë realizado o planejamento das diversas atividades a serem realizadas, devendo conter a definição clara dos objetivos a serem alcançados. Com a identificação dos participantes do projeto, definido na etapa anterior, se atribui a cada atividade os prazos e suas responsabilidades. Mapeia-se uma área de negócio para o desenvolvimento do projeto, incluindo os relatórios necessários a demanda dos usuários, inclusive a sua estrutura de apresentação. Ressalta-se pelo fato de que no ciclo existe o conceito de iteração, novas situações poderão ser impostas ao Plano do Projeto. Desta forma, a execução desta fase apresenta um Plano de Projeto mais detalhado.

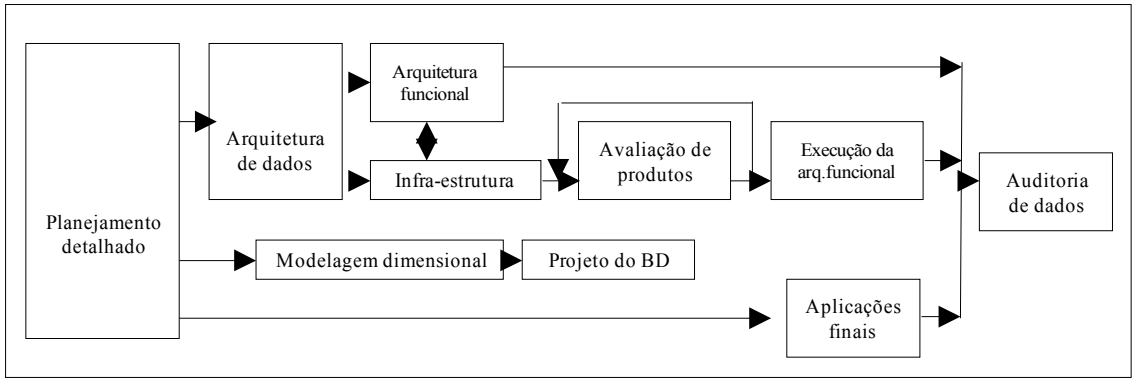


Figura 5.3: Fluxo de trabalho da etapa de definição.

Para uma melhor definição dos dados, esta fase foi subdividida em módulos, provenientes do trabalho de [PER 2000], tais como: (a)atividades preparatórias e (b) requisitos detalhados. As principais características destes módulos são:

- *Atividades preparatórias*: permite descrever as estratégias de entrevista, levando em conta a quem entrevistar, o que perguntar, documentação das reuniões e sequencia das mesmas;
- *Requisitos detalhados*: permitem descrever de forma detalhada todos os requisitos identificados na etapa anterior. Em outras palavras, é a tradução dos requisitos do negócio em objetos físicos na base de dados (e.g. tabelas de dimensão e fato, granularidade, agregações, etc.). Permite ainda identificar as fontes necessárias ao projeto de DW (uma vez definidas na etapa anterior) a sua localização (e.g. externas, internas), as plataformas em que se situam e a estrutura dos arquivos correspondentes, para as tarefas de ETL.

A fase de *arquitetura de dados* destina-se basicamente à escolha da arquitetura que servirá como um mapa de alto nível, através do qual é possível entender como os dados se organizam no projeto. Vale lembrar, como descrito no Capitulo 2, sobre arquitetura de dados, que se deve considerar a forma como os dados fluem através do DW e são utilizados pelos usuários finais, isto é, arquitetura centralizada, DM dependentes, DM independentes ou distribuída.

Na fase da *arquitetura funcional*, o Plano de Projeto deve conter a definição do que se deseja do DW quando estiver concluído. Em outras palavras, deve conter o fluxo dos dados dos sistemas fontes até os usuários e os serviços correspondentes, a exemplo da extração, transformação, carga, consulta, ect [PER 2000].

A fase de *infra-estrutura* permite definir os recursos de hardware e software necessários para dar subsídio na definição da arquitetura funcional.

É na fase de *modelagem dimensional* que se define o modelo de esquema (e.g. estrela, floco de neve) que será adotado ou uma variação "híbrida" [POE 98] necessária para adequá-la junto a corporação, para se obter melhores resultados. A conclusão desta fase se dá pela montagem da modelagem conceitual, considerando os requisitos da área de negócio selecionada.

O *projeto da base de dados* é a fase onde são realizadas as atividades que possibilitem preparar adequadamente a base de dados analítica para receber os dados oriundos da etapa de ETL. São feitas as estimativas com relação ao tamanho do BD, particionamento físico da base de dados, proteção e segurança dos dados, criação de índices, otimização da base de dados, etc [PER 2000];

A fase de *avaliação de produtos* tem como principal objetivo [PER 2000] a definição de quais produtos testados que serão necessários para atender a todos os usuários do ambiente de DW, que vão desde a extração até as consultas. Dependo da situação e necessidades, se pode acrescentar ao Plano do Projeto, a especificação de rotinas que deverão ser customizadas.

A fase de *execução da arquitetura funcional* tem por finalidade realizar atividades para atender às especificações resultantes da fase de arquitetura funcional. Durante a execução desta fase, os dados são identificados das fontes de dados, extraídos, transformados e carregados para o DW, através das ferramentas previamente escolhidas.

Devem-se acrescentar os resultados obtidos no Plano de Projeto, para manter um histórico das execuções realizadas [PER 2000].

Na fase de *aplicações finais*, relatórios e consultas são criadas para atender ao público de usuários tomadores de decisão e que participaram das especificações necessárias ao projeto de DW. Se faz um levantamento do máximo de detalhes referentes à apresentação dos dados aos usuários decisores.

Em linhas gerais, a última fase sobre *auditoria de dados* permite a verificação sobre a qualidade dos dados armazenados, através de atividades de validação dos dados através de consultas, análise nos processos de ETL, estudo dos arquivos de log, etc.

5.2.3 Execução

À medida que novas iterações vão surgindo, esta etapa permite realizar a execução do Plano de Projeto, incluindo as tarefas necessárias para a disponibilização do DW. Esta etapa possibilita uma visão prática de todas as definições de projeto e experiências adquiridas nas fases anteriores. Esta etapa, considerando os trabalhos de [POE 98] e [KIM 98], possui duas fases: (a) uso e gerenciamento e (b) gerenciamento de versões.

A fase de *uso e gerenciamento* permite, mediante o nível de iteração, a simples construção das definições descritas no plano, até a montagem, propriamente dita do DW. Permite a criação de um plano de disponibilização (contendo um plano de verificação da infra-estrutura, estratégia de treinamentos dos usuários finais, de suporte).

Na fase de *gerenciamento de versões*, um plano de atualização de versão do DW pode ser elaborado, para o monitoramento de consultas realizadas pelos usuários, desempenho da organização de dados e o contínuo sucesso do DW.

6 ESTUDO DE CASO: DESENVOLVIMENTO DE UM PROJETO DE DW

Este capítulo apresenta uma aplicação da metodologia para o projeto de DW junto à Cia Zaffari, utilizando o sistema OLTP da área comercial, como estudo de caso, para o levantamento dos dados e informações necessárias ao projeto. Inicialmente será apresentado um breve histórico da empresa, para posterior descrição da metodologia.

Como este projeto desempenha um modelo real da Cia Zaffari, descrevendo dados e informações gerenciais, o modelo proposto foi sumarizado e dados foram simplificados e/ou alterados, de forma a se preservar o sigilo das informações. Além disto, mesmo tendo sido autorizada a execução do projeto de DW pela Cia. Zaffari,, a sua implementação não ficou definida em que momento será feita, e desta forma, os processos relativos à execução não foram realizadas.

6.1 Ambiente da Cia Zaffari

As próximas seções apresentam um breve histórico sobre a Cia Zaffari e descrevem a empresa quanto a sua estrutura operacional e de como os dados estão definidos quanto a sua estrutura e replicação.

6.1.1 Histórico

A Companhia Zaffari é uma empresa que atua a mais de 65 anos, no ramo de comércio no Rio Grande do Sul. Na capital gaúcha, a empresa conta 42 anos de trabalho no varejo de alimentos e auto-serviço, a partir da primeira loja, no ano de 1960. Desta dada até hoje, a Companhia Zaffari montou uma rede de mais de 20 supermercados e hipermercados que atendem a capital e o interior do estado, além de seis shopping centers, um deles em São Paulo. Atua também na industrialização de alimentos, com a indústria de Óleos Vegetais e a fábrica de café e biscoitos Haiti- Plic-Plac.

6.1.2 Estrutura Operacional

Sua estrutura é formada pelas seguintes áreas: (a) administrativa, (b) depósitos e (c) lojas. Na área administrativa, são executadas as tarefas de controle e gestão das compras, preços, pedidos e planejamentos, além das atividades que envolvem toda parte financeira e contábil. Já na área das lojas, situadas nas cidades do estado do Rio Grande do Sul e São Paulo, são concentradas as principais operações de venda dos produtos, além dos recebimentos provenientes dos depósitos ou fornecedores. Os depósitos são as áreas responsáveis pelo recebimento e armazenamento dos produtos entregues pelo fornecedor.

6.1.3 Estrutura dos Dados

Cada uma das áreas, definidas na Seção anterior, também são chamadas de locais e possuem um servidor próprio. Atualmente a companhia Zaffari possui 30 servidores HP-UX e 6 servidores NT. Cada local possui um servidor próprio identificado pelo próprio nome do local (Loja X, Loja Y, Depósito Z). No setor de suporte, localizado na área administrativa, ficam alguns desses servidores, conforme mostra a tabela 2:

Tabela 2 – Descrição dos Servidores.

Nome	*Modelo*	*Descrição*
ZAFFARI	HP K590	Servidor Sybase Zaffari – um dos node do cluster *(MC/ServiceGuard)*.
ZAFFARI2	HP K420	Servidor Sybase Zaffari – outro node do cluster *(MC/ServiceGuard). Open View – Network Node Manager.*
ZAFNT1	HP NetServer 5/100 LH	Servidor WEB, servidor de arquivos e de impressão.
ZAFNT2	HP NetServer 5/100 LH	Servidor *Exchange Server*.
DESENV	HP E55	Servidor da área de desenvolvimento.

Para cada servidor existe um Banco de Dados (SyBase) e várias Base de Dados (ZafA, ZafB, ZafC e outros.) conforme Figura 6.1. Estas Bases de Dados representam os dados que podem estar organizados por sistemas e/ou processos.

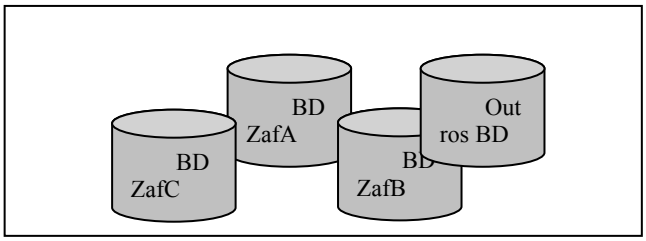


Figura 6.1: Bases de Dados de um Servidor.

6.1.4 Transferência de Dados

Além do servidor de dados, cada local possui um servidor de replicação (SR) responsável pelo envio e/ou recebimento dos dados processados para outros servidores, tais como o pedidos de loja para o depósito e dados que ficam centralizados no servidor central (área administrativa).

O Servidor de Replicação utiliza dois objetos: (a) *Replication Definition Administration* (RDA) e (b) *Subscription Administration* (SBA). O RDA descreve uma tabela que pode ser replicada para muitos servidores. Nesta parte, se identifica dados referentes a tabela que irá replicar (e.g. o nome da *replication definition,* o nome da tabela primária, a localização da tabela primária no qual vai ser copiada, o nome e o tipo de dado das colunas da tabela primária, etc.).

A SBA é usada para replicar os dados da tabela primária de um banco de dados de um servidor. Define o nome da *replication definition* para a tabela que está sendo replicada e o Banco de Dados onde a tabela será copiada. A clausula *where* permite especificar as linhas (registros) que poderão ser replicados. O padrão da replicação é sobre toda tabela.

Quando se cria uma *subscription*, existem registros de ajuste da *subscription* especificadas que devem ser copiadas da tabela primária para a tabela de replicação. Este processo é chamado de *subscription materilization*. Depois da materilização ser definida, o SR distribui para os servidores de réplica, assim como se muda no servidor primário.

O servidor de replicação está definido para a transferência on-line, fazendo com que os dados sejam enviados imediatamente após a sua transação. A Figura 6.2 ilustra como seria a estrutura de replicação dos dados entre os servidores.

6.2 Etapa de "anteprojeto"

Com já descrito no capítulo 5, é na etapa de "anteprojeto" que se executa das fases de levantamento e planejamento, necessário para o andamento do projeto de DW. As próximas seções apresentam e detalham cada uma destas fases.

6.2.1 Levantamento

Nesta fase foram levantadas, de forma sucinta, as principais características que cercam a companhia Zaffari. Os dados foram coletados de forma a que se pode escolher um processo do negócio para modelar. Um processo do negócio é uma operação importante na organização, suportada por algum tipo de sistema legado de onde é possível coletar dados para o DW. No geral foram realizadas as seguintes atividades:

- Identificação dos fatores "humanos" que cercam o ambiente organizacional, tais como o interesse e expectativa de gerentes e administradores em geral com o desenvolvimento e implantação do projeto de DW;
- Capacidade e disposição de investimentos da corporação no projeto de DW;
- Reuniões para a identificação dos requisitos necessários para a definição do processo do negócio a se modelar;
- Levantamento de dados e informações sobre os sistemas OLTP da Cia. Zaffari, através do estudo dos documentos e relatórios;
- Levantamento de informações analíticas de natureza tática e estratégica da Cia. Zaffari, através de relatórios, planilhas e entrevistas junto a coordenadores, gerentes e potenciais usuários do sistema.

6.2.1.1Resultados do Levantamento

Em consequência dos levantamentos realizados nesta fase, se pode apresentar diversas particularidades a respeito do ambiente de desenvolvimento do projeto:

- Apesar de ter sido autorizada a execução do projeto de DW pela Cia. Zaffari,, como já mencionado anteriormente, a sua implementação não ficou definida em que momento será feita, desta forma nenhum patrocinador de alto nível e influente foi designado. Este fator refletiu-se mais tarde na dificuldade da obtenção de certos dados e informações, especialmente de natureza gerencial;
- Pelo fato da não definição exata do cronograma de implantação do projeto de DW, constatou-se que não seria necessário nenhum investimento financeiro para o estudo de caso;

- Com a divulgação do projeto junto aos usuários da gerência e coordenação, criou-se uma grande expectativa dos potenciais usuários, especialmente na área de compras do setor comercial, os quais puderam enxergar claramente os benefícios e contribuíram para a realização do trabalho proposto.
- Usuários do setor comercial estão preocupados com a logística de organização e com o abastecimento das gôndolas;
- O processo a ser modelado, considerando as necessidades dos usuários envolvidos, será no movimento diário de itens;
- Quanto à fonte de dados, os sistemas ditos convencionais (estoque, crediário e outros) da Cia Zaffari utilizam a maioria das vezes fontes de dados internas, cujos dados provêm de arquivos independentes denominados de Tabelas, Cadastros e Movimentos.

6.2.2 Planejamento

Considerando os dados levantados na fase anterior, nesta fase foram definidas as principais características que completam o plano do projeto, com já visto na Seção 5.2.1. No geral foram realizadas as seguintes atividades:

- Identificação dos membros da equipe de projeto e suas funções;
- Entrevistas e reuniões para identificar os fatores que permitem mensurar o sucesso assim com os riscos do projeto;
- Levantamento das ferramentas necessárias para construção do DW, como para as consultas a base de dados analítica;
- Definição do cronograma não detalhado para a efetiva construção.

6.2.2.1Resultados obtidos no Planejamento

Como resultado das definições feitas nesta fase, se pode obter o Plano de Projeto, contendo: os dados levantados e analisados até o momento, identificado por um título para a identificação do projeto de DW e a definição dos participantes do projeto, tais como os membros da equipe e usuários tomadores de decisão, assim como de suas correspondentes tarefas. Este plano ainda apresenta:

- Os principais fatores que permitem mensurar o sucesso do projeto de DW foram identificados como sendo a necessidade de disponibilização de relatórios pré-formatados de acordo com as necessidades dos decisores e no menor tempo possível e relatórios gerenciais que integrem dados e informações de anos distintos;
- À exceção do Sistema Gerenciador de Banco de Dados Sybase (SGBD) e das ferramentas de desenvolvimento (e.g. linguagem de programação e CASE) a Cia Zaffari não dispunha de nenhuma linguagem de software ou ferramenta específica para o desenvolvimento de projeto de DW (e.g. ferramentas de ETL, consultas, etc.), e como já descrito, não havia a disponibilidade de aquisição devido a falta de investimentos para este estudo de caso;
- Quanto aos prazos de desenvolvimento deste estudo de caso, para a aplicação do modelo de metodologia proposta para o projeto de DW, foi estimado inicialmente como sendo em 2 meses, levando em conta os objetivos propostos.

Os resultados obtidos após diversas reuniões e levantamentos de dados, foram apresentados ao grupo de usuários finais para que possíveis correções e adequações ainda fossem verificadas e possíveis sugestões e melhorias fossem consideradas ao plano do projeto. Ao final desta etapa foi necessária uma aprovação do plano do projeto perante a cúpula de gerentes e coordenadores participantes, para a continuidade do projeto.

6.3 Etapa de "definição"

Para o modelo proposto foi utilizada a etapa de definição apresentada na Seção 5.2.2, sendo executadas as seguintes fases:

6.3.1 Planejamento detalhado

Esta fase se constitui especificamente no aprofundamento dos levantamentos realizados na etapa de definição e atualização do Plano do Projeto. No geral foram realizadas as seguintes atividades, as quais foram organizadas em dois módulos distintos, conforme já abordados.

6.3.2 Módulo de "atividades preparatórias"

Quanto a este módulo, verificam-se as seguintes atividades preparatórias:

- Estudo de bases de dados, análise dos arquivos de dados, sistemas OLTP, inclusive o código fonte de programas;
- Análise de relatórios e documentos disponíveis, levando em conta os requisitos preliminares levantados nas fases anteriores;
- Na definição de quem entrevistar, levando em conta as definições feitas nas fases anteriores, foram entrevistados os gerentes e coordenadores do setor de compras;
- Quanto ao que se perguntar, foi questionado tudo referente a proposta que foi modelada, por exemplo, que tipos de informações são importantes na transação das vendas de uma loja, etc;
- Os tempos das entrevistas variaram em média de 30 a 60 minutos, entre os gerentes e coordenadores do setor comercial.

6.3.3 Módulo de "requisitos detalhados"

Levando em conta a execução do módulo anterior, verificam-se , neste módulo, as seguintes definições:

Cada uma das lojas é um supermercado moderno, levando em conta o conceito de shopping, composto por uma variedade de departamentos, incluindo mercearia, congelados, laticínios, açougue, verduras e legumes, padaria, flores, bens duráveis, bebidas e toda a área de bazar. Cada loja trabalha, dependendo do tamanho, com aproximadamente 50 mil produtos individuais em suas prateleiras. Os produtos individuais são chamados aqui, como unidades de estoque (UNE).

Cerca de 30 mil das UNEs provêm de fornecedores externos e apresentam códigos de barras nas embalagens. Estes códigos de barras são chamados de EANs e

representam o mesmo grão que uma UNE individual. Cada variação de embalagem de um produto possui um a UNE diferente e, portanto, uma EAN diferente.

As 20 mil UNEs restantes provém de setores como açougue, verduras e legumes, padaria ou flores e não possuem códigos EAN reconhecidos em âmbito nacional. Entretanto se deve atribuir um número UNE a esses produtos e se colocar etiquetas para a leitura óptica.

Considerando que o processo a ser modelado é no movimento das vendas, o banco de dados permitirá ver em detalhes quais produtos estão sendo vendidos em que lojas, a que preço e em que dias. Usuários do setor comercial, além da preocupação com a logística de organização e com o abastecimento das gôndolas, estão preocupados com a venda dos produtos, assim como a maximização do lucro em cada loja.

Ficou definido que a granularidade dentro do banco de dados é por venda do cliente (transação). Uma vez que quanto mais detalhes, menor o nível de granularidade, consequentemente, maior o volume de dados armazenado, foi prevista uma grande área de armazenamento. O fato de se poderem identificar os clientes, por transação de venda, possibilita buscar informações a respeito da logística de comportamento de compras, por exemplo, o perfil de clientes numa loja, produtos mais vendidos na região, produtos menos vendidos, produtos indispensáveis, etc.

O lucro resulta principalmente de cobrar o máximo possível por um produto, reduzir os custos de aquisição e custos indiretos do produto e, ao mesmo tempo, de atrair o maior número possível de clientes por meio de uma política de preços altamente competitiva. As decisões administrativas mais significativas que podem ser tomadas em tempo real relacionando-se com promoções e política de preços.

Tanto o setor comercial, como o setor de marketing, gasta muito tempo ajustando preços e lançando promoções. As promoções em um supermercado incluem reduções temporárias de preços, anúncios e encartes em jornais, display nos supermercados. A maneira mais objetiva e eficaz de criar um aumento no volume (quantidade do produto vendida) é diminuir o preço drasticamente.

Uma redução de 50 centavos no preço de toalhas de papel, especialmente quando conjugada a um anuncio e a um display, pode aumentar em 10 vezes o volume de vendas de toalhas de papel. Infelizmente, uma redução de preço como esta não é sustentável porque provavelmente as toalhas estarão sendo vendidas com prejuízo. Pode-se concluir que a visualização de todos os tipos de promoção consiste em uma parte importante da análise das operações do setor comercial.

6.3.4 Arquitetura de dados

Dentre as arquiteturas apresentadas no Capitulo 2, se definiu pela utilização da arquitetura centralizada, uma vez que os dados propostos na modelagem deste estudo de caso ficam centralizados no servidor da Administração e pela facilidade de sua implementação.

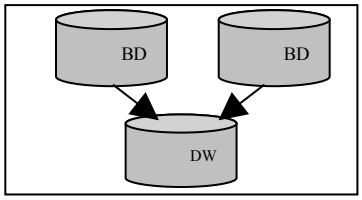


Figura 6.2:Topologia Centralizada.

Para os registros das vendas, as lojas digitalizam os códigos de barras diretamente para o sistema de ponto-de-venda (PV). Os PVs ficam na entrada da loja e é onde os clientes fazem as suas compras. Os dados registrados de uma transação de venda são acumulados e transferidos de tempos em tempos para uma tabela temporária do banco de dados local, para que possam replicar para o servidor da Administração.

Para o processo de controle das O conceito de RPD, se dá sobre a forma no qual os dados estão dispostos em Banco de Dados(BD). Esta replicação ocorre via Servidor de Replicação (SR), no qual gerencia a atualização das Tabelas dos BDs. No caso da Cia. Zaffari, a estrutura para o processo de replicação do ponto de venda (PV) pode ser representada conforme Figura 6.2.

Cada local possui um servidor contendo o Banco de Dados Sybase (BD), o Servidor de Replicação (SR), o Servidor de Backup (SBK) e o Arquivo de Log (Log). Em cada servidor, os Banco de Dados possui seus dados originais e dados replicados, oriundos de outros Bancos de Dados. É claro, que isto vai depender de definições físicas sobre cada Loja e/ou Depósito (e.g. a estrutura física da Loja X pode ser composta do Depósito (CD) e da própria Loja X).

Na Figura 6.3 os dados sobre a venda do servidor da Loja X são replicados para o servidor o ADM, pelas rotas B1 e B2, no qual retornam uma mensagem de status (envio de ok ou não). Observe que nas rotas A1 e A2 representam a replicação de outras Lojas para a Administração, uma vez elas também participam deste processo.

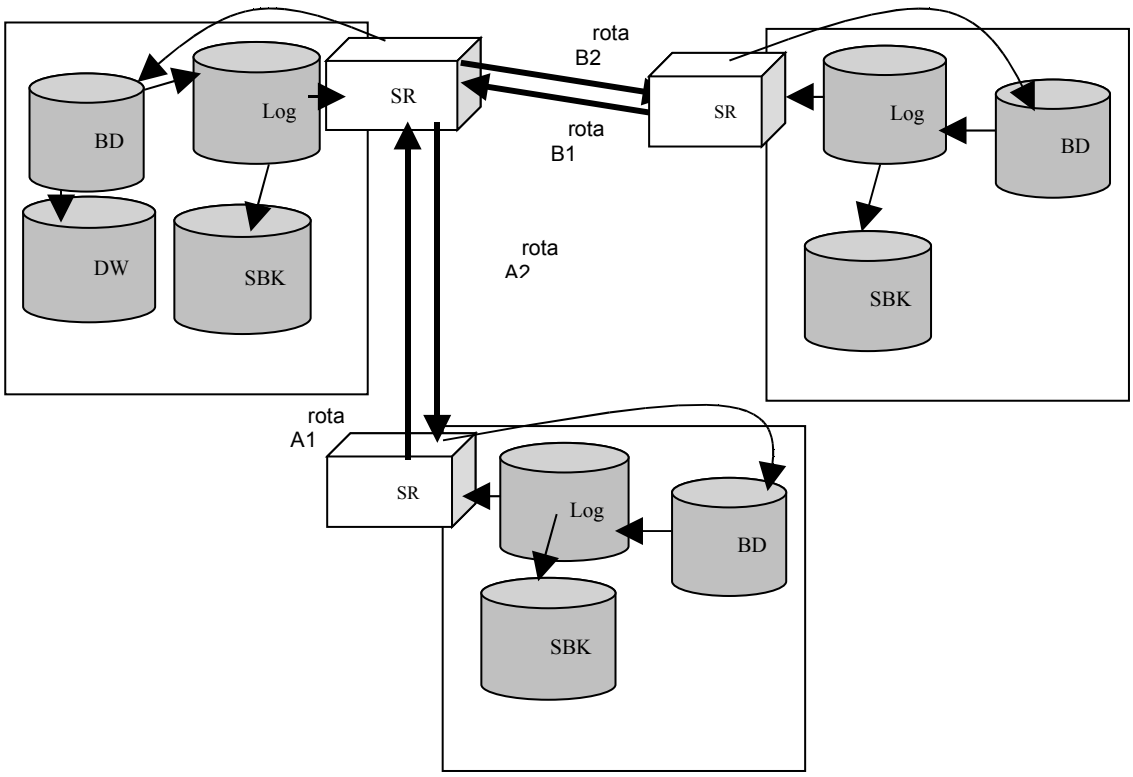


Figura 6.3: Estrutura de Replicação dos dados.

Outra definição importante, é que os dados do BD, a cada transação, são copiados para o arquivo de Log e limpos de hora em hora, após serem transferidos para o Backup.

6.3.5 Arquitetura funcional

Para esta fase foram consideradas as propostas de [KIM 98a] e [PER 2000] para a definição da arquitetura funcional interna. Esta arquitetura foi adaptada ao projeto proposta a Cia Zaffari, conforme mostra a Figura 6.4, permitindo identificar os objetos a serem criados na base de dados, influenciando nas fases de infra-estrutura e projeto de banco de dados.

Como já apresentada na Seção 2.5 (Figura 2.8), a área interna pode ser classificada em: (a) área de sistemas fontes, (b) área de organização dos dados (*Staging)*, onde os dados de fontes tradicionais (sistemas e outras) são copiados, formatados e armazenados e (c) a área do DW (servidor de apresentação), onde os dados tratados são carregados.

6.3.6 Infra-estrutura

Para a execução do modelo proposto, será utilizado a arquitetura de servidor HP K580 e sistema operacional Unix. Será utilizada para a construção de programas de transformação e carga, para a arquitetura funcional, a linguagem de programação que atualmente é utilizada para a construção das aplicações convencionais da própria Cia Zaffari.

6.3.7 Modelagem dimensional

Nesta fase foram definidas as tabelas necessárias para atender as definições da arquitetura funcional, além de modelar as tabelas de dimensão e fato.

Levando em conta as definições sobre os esquemas multidimensionais apresentadas na Seção 2.3.2.1 (Figura 2.5), foi definida a utilização do esquema “estrela” pelo fato de ser mais eficiente na recuperação de dados e informações, além da facilidade de compreensão do modelo. O esquema “floco-de-neve” não foi utilizado pelo fato de não haver relacionamentos “muitos para muitos” entre as tabelas dimensão e por apresentar a desvantagem do aumento da complexidade da arquitetura.

Por definição, a tabela de fatos em uma estrutura dimensional é de natureza altamente normalizada e é a maior tabela do banco de dados dimensional. As tabelas dimensionais, por definição, são quase sempre geometricamente menores que a tabela de fatos. Qualquer estimativa realista do espaço em disco necessário para o DW pode efetivamente ignorar as tabelas de dimensão.

Considerando o processo de registro de vendas, cada loja deve gerar um relatório completo de todas as vendas de produtos de cada cliente. Levando em conta o grande volume de vendas diárias, tomou-se como abordagem a transferência a cada transação individual de venda para o servidor da administração e desta forma executar o resumo diário. Esta transferência é feita através de uma programação nos PVs, para que de tempos em tempos façam o processo de atualização e envio.

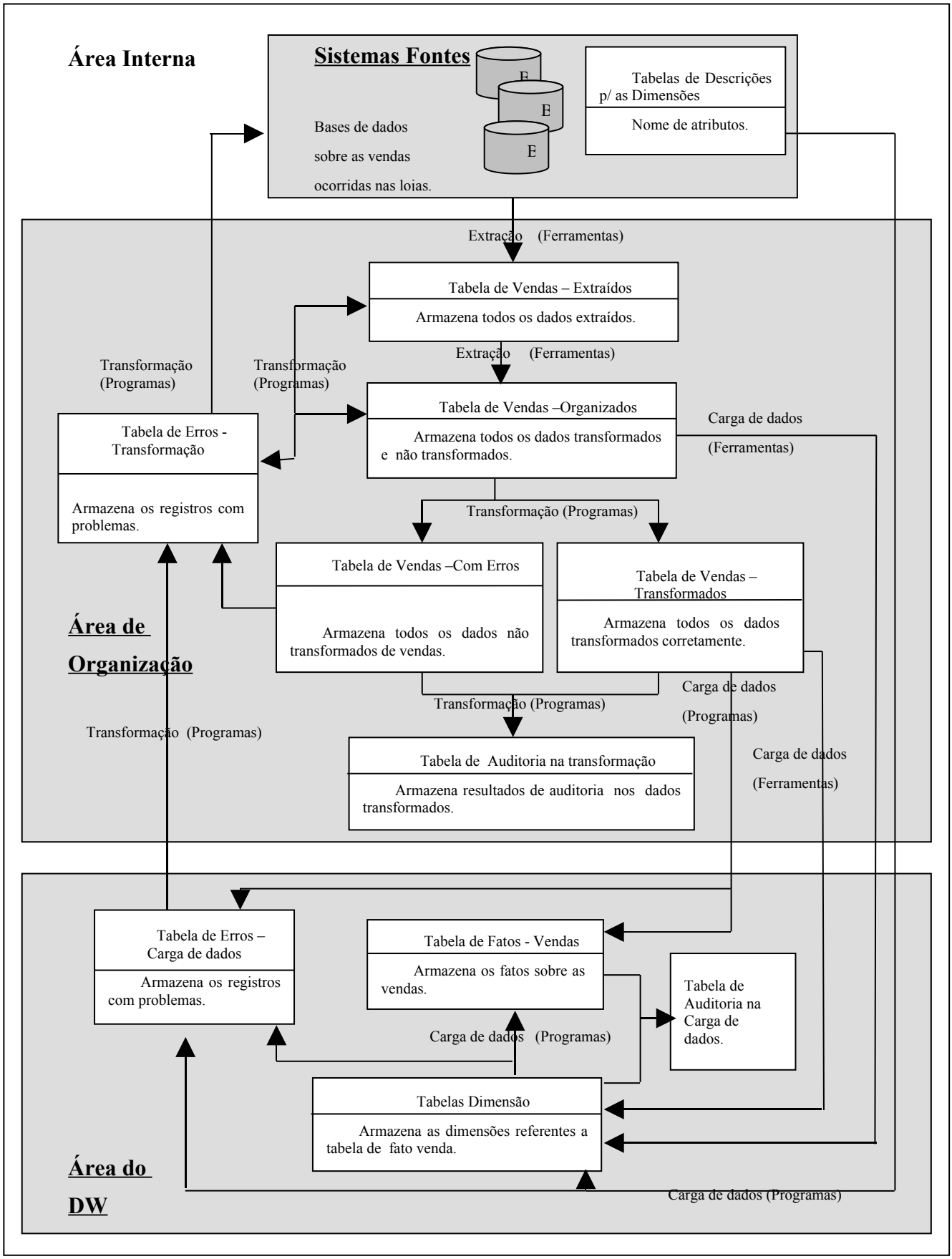


Figura 6.4: Arquitetura Funcional.

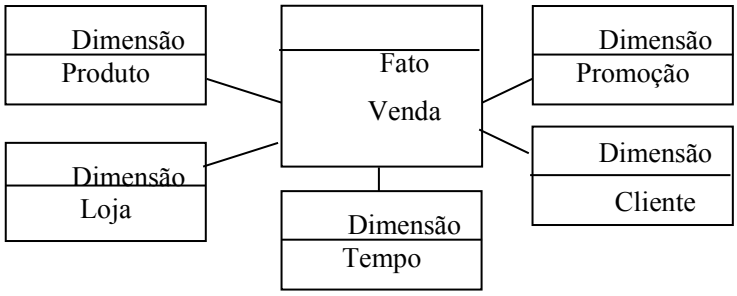


Figura 6.5: Modelo Estrela para o modelo proposto.

Levando em conta as necessidades levantadas para este modelo, definiu-se que o modelo proposto será composto pela tabela de fato "venda" e pelas tabelas de dimensão: produto, cliente, tempo, promoção e loja . A Figura 6.5 descreve o relacionamento existente entre a tabela fato e as dimensões.

Uma vez que se propõe a modelar a transação de venda por cliente , a tabela de dimensão tempo é necessária para que se possam separar os dados por dias úteis e feriados, por períodos fiscais, por estações ou ainda por eventos. Cada registro da tabela de dimensão tempo representa um dia e ao contrário das demais tabelas de dimensão, pode ser construída com antecedência, podendo gerar de cinco a dez anos de registros. Os campos que representam a tabela de dimensão tempo são: chave-tempo, dia-da-semana, numero-do-dia-do-mês, numero-em-dia, mês, trimestre, período-fiscal, indicador-fiscal, temporada, evento.

A tabela de dimensão produto descreve cada UNE, o qual é identificado pelo código de barra EAN. Um produto é representado por uma estrutura aqui chamada de "hierarquia de mercadoria" , composta de grupo, classe, seção e departamento. Desta forma, os atributos que representam a tabela de dimensão produto são: chave-produto, nome-produto, EAN, embalagem, data-da-estrutura, grupo, classe, seção, departamento, peso, tipo-dieta, embalagem-por-palete e outros.

A tabela dimensão loja descreve cada loja da rede de supermercado da Cia Zaffari, e pode ser considerado como sendo um elemento geográfico. Os atributos que representam a dimensão loja são: chave-loja, nome-loja, numero-loja, endereço-loja, cidade-loja, município-loja, estado-loja, cep-loja, região-loja, gerente-loja, fone-loja, tipo-planta-loja, data-abertura, data-ultima-reforma, numero-de-habitantes, renda-per-capita e outros.

A tabela dimensão promoção descreve todas as condições de promoção aplicadas à venda de um produto na rede de supermercado da Cia. Zaffari. Condições de promoção incluem reduções temporárias de preços, terminais promocionais, anúncios de jornais e cupons. Os atributos da tabela dimensão promoção são: chave-promoção, nome-promoção, tipo-de-preço, tipo-anuncio, tipo-cupom, nome-mídia, fornecedor-promoção, data-inicio-promoção, data-término-promoção e outros.

A tabela dimensão cliente ascende para todos os clientes que efetuam uma transação de venda, gerando um cupom de venda. Esta dimensão é importante pois possibilita uma análise de comportamento de compras, identificando o perfil de compra, por

região e loja. Os atributos da tabela dimensão cliente são: chave_cliente, nome-cliente, endereço-cliente, fone-cliente, e-mail-cliente, e outros.

A tabela de fato venda representa as medidas do negócio, que são mensuradas de forma quantitativa, tais como os atributos de valor da venda, quantidade vendida e o custo de venda de um produto relacionada com as tabelas de dimensão "produto" e "cliente", permitindo identificar a quantidade vendida de um produto por um certo cliente. A tabela de fato armazena grande quantidade de dados, possuindo chave primária composta, formada por chaves estrangeiras, através das quais se ligam as chaves primárias das tabelas dimensão. Desta forma, os atributos chave que representam a tabela de fato venda são: chave-tempo, chave-cliente, chave-produto, chave-promoção, chave-loja. Os atributos fatos (medidas) são: valor-da-venda, quantidade-vendida, custo-da-venda, e são considerados fatos aditivos pelo fato de que podem ser somados cruzando-se qualquer uma de suas dimensões.

6.3.8 Projeto da base de dados

Nesta fase, a participação do administrador de banco de dados é fundamental para a definição de uma alocação de espaço de disco necessário para armazenar os dados da base analítica, assim como para as atividades de organização de dados para a construção de um DW. Para o cálculo de estimativa do numero de linhas de itens das transações individuais em uma das lojas, se identifica a receita bruta da rede como um todo, por exemplo, quatro bilhões ao ano. Além disto, se mensura o preço médio de um produto vendido, por exemplo, o valor de dois reais. Considerando estes dois parâmetros, pode-se afirmar que existirão dois bilhões de linhas de itens, por ano (4 bilhões divididos por 2 reais). Pode-se concluir que, o numero básico de linhas de item nas transações de venda de um negócio pode ser estimado dividindo-se a receita bruta do negócio pelo preço médio do item.

Levando em conta o cálculo sobre as transações de venda, o tamanho do DW pode-se dividir o numero de linhas de item, por ano, para toda a rede de lojas (2 bilhões) por 365 dias e pelo numero de lojas, por exemplo, 20 e desta forma, teremos o valor de 11 mil itens por dia e loja.

6.3.9 Avaliação de produtos

Para esta fase, se devem testar as ferramentas necessárias para o projeto e construção do DW. Por motivos de não se ter uma aplicação mais prática para o estudo de caso, uma vez que a Cia Zaffari não aprovou a execução do projeto, esta fase fica apenas descrita com uma definição mais conceitual. Como um dos critérios para viabilizar e justificar o uso desta metodologia, é a necessidade de experiência em projetos de DW, uma empresa de consultoria e desenvolvimento se fez necessária, e desta forma, produtos serão apresentados e analisados, em tempo de aprovação da execução do projeto, propriamente dito.

6.3.10 Execução da arquitetura funcional

Esta fase, assim como a fase anterior, não foi aplicada. É a fase que mais consome tempo, pois se constitui da efetiva execução da arquitetura funcional, levando em conta a criação de rotinas para a transformação e organização, alem das tabelas auxiliares destinadas a armazenar informações sobre os dados errados ou rejeitados.

6.3.11 Aplicações finais

Esta fase representa a construção de relatórios previstos na fase de "planejamento detalhado", levando em conta as ferramentas que deverão ser escolhidas para esta função.

6.3.12 Auditoria de dados

Esta fase se constitui na utilização intensiva dos relatórios gerados na fase de "aplicações finais", pela equipe do projeto com o objetivo de verificar se os resultados obtidos estarão corretos e consistentes, além disto, um acompanhamento na fase de extração, organização, transformação e carga se fazem necessária, para se obter uma melhor performance.

6.4 Etapa de "execução"

Como já recomendado anteriormente, apesar de ter sido autorizada a execução do projeto de DW pela Cia. Zaffari,, a sua implementação não ficou definida em que momento será feita. Desta forma a execução do Plano de Projeto de DW não pode ser executada e este fator refletiu-se mais tarde na dificuldade da obtenção de certos dados a respeito desta fase. Apesar deste fato, um aspecto que ficou evidenciado, durante a montagem do Plano de Projeto, foi a facilidade e rapidez na obtenção de informações para o projeto de DW.

6.5 Considerações finais

Pode-se observar que dentre os resultados obtidos com a aplicação da metodologia proposta, confirmou-se a grande demanda por informações gerenciais ainda não dispostas, e que pelo grande volume de dados, a necessidade de implantação do projeto de DW. No geral foram obtidas as seguintes considerações:

- Consegue-se, a partir da metodologia proposta, se fazer um levantamento completo a respeito do movimento de venda, por cliente e desta forma, apresentar ao grupo de usuários os principais resultados obtidos;
- Comprova-se que o esquema "estrela" apresenta uma maior eficiência na recuperação de dados e informações quando comparado com o modelo "floco de neve";
- Pelo fato de que a tabela de fatos é de natureza altamente normalizada, os esforços para normalizar qualquer uma das tabelas de dimensão em um banco de dados dimensional, simplesmente para se obter espaço em disco, são uma perda de tempo;
- Desenvolver um DW não é diferente de se desenvolver um projeto de tecnologia de informação. Ë necessário planejamento, definição de requerimentos, projetos e implantação;
- Podem-se identificar questões não comumente existentes no âmbito gerencial, tais como: (a) vale a pena estocar tantos tamanhos diferentes de determinados produtos? , (b) Quais produtos são "canibalizados" quando se promove um determinado produto, em outras palavras, quais produtos sofrem uma queda de venda quando se promove um outro? (c) Quais produtos são os mais vendidos,

por lojas, (d) Qual é o perfil dos clientes, por regiões e (f) Quais os produtos indispensáveis em uma loja?

- Não se pode identificar na tabela de fatos, quais produtos que não são vendidos, uma vez que ela representa a transação sobre as vendas, por cliente, dia e loja;
- Um DW quase sempre precisa de dados expressos no nível de menor grão de cada dimensão, não porque as consultas queiram ver registros analíticos individuais, mas porque as consultas precisam aprofundar-se no banco de dados de maneira precisa;
- A definição do nível de granularidade permite determinar as dimensões primárias da tabela de fatos;
- Levando em conta a definição por uma arquitetura centralizada, o DW projetado no estudo de caso pode parecer propenso a se tornar um DM, uma vez que possui características quanto ao foco de usuários finais serem de uma área da empresa (setor comercial);
- Por existir um monte de dados históricos para suporte a decisão que raramente são usados, a empresa pode reduzir o armazenamento de dados, baseada em algum critério, o se justifica o fato de que projetos que começam como um DW, algumas vezes evoluem para Data Marts;
- O processo de iteração permitiu se ter várias repetições de execução das etapas de um projeto de DW (anteprojeto, definição e execução) para cada uma as etapas do ciclo de desenvolvimento, iniciando na fase de incepção e terminando na fase de transição.

7 CONCLUSÃO

Este trabalho foi feito com o objetivo de propor uma metodologia prática de projeto de DW, e desta forma desenvolveu-se um trabalho de pesquisa que estabeleceu como metas principais: (a) estudo das principais metodologias de DW existentes, (b) a proposta de uma metodologia de desenvolvimento para projeto de DW e (c) a avaliação da adequação e consistência da metodologia proposta, através de um estudo de caso real na Cia Zaffari.

A metodologia proposta, depois de aplicada mostrou-se adequada e consistente na solução dos problemas com relação à completude, detalhamento e iteração, critérios que motivaram este trabalho e que foram citados no capítulo 5.

Levando em conta que estes critérios definidos não foram totalmente satisfeitos por outras metodologias de projeto de DW apresentadas no capítulo 4, a metodologia proposta sustenta-se sobremaneira sobre o critério de iteração, baseada no conceito apresentado pela metodologia RUP. Nele, o projeto de DW elimina as incertezas e riscos de insucesso, através da realização de uma série de execuções sobre o ciclo de vida da metodologia proposta, até a obtenção de um Plano de Projeto consistente para ser aprovado e implantado.

A tabela 3 apresenta o resumo das metodologias de projeto de DW estudadas, assim como as principais características da metodologia proposta.

Tabela 3 – Resumo das metodologias de projeto de DW e proposta.

Características	*[MAR 99]*	*[KIM 98a]*	*[POE 98]*	*[PER 2000]*	*Proposta*
Completa	Sim	Sim	Sim	Sim	Sim
Experiência	Sim	Sim	Sim	Não	Sim
Fases da metodologia	- visão estratégica; - avaliação da engenharia; - avaliação do fluxo de valores; - avaliação da questão comercial; - projeto/revisão; - caso comercial do DW; - plano de implementação da	- planejamento; - definição de requisitos do negócio; - projeto de arquitetura técnica; - modelagem dimensional; especificação de aplicações de usuário final; - seleção e	- planejamento; - levantamento de requisitos e modelagem; - projeto físico da base de dados; - determinação e mapeamento das fontes de dados; - população do DW;	- levantamento preliminar; - planejamento preliminar; - levantamento detalhado; - protótipo; - definição do projeto e	- levantamento; - planejamento; - planejamento detalhado; - arquitetura de dados; - arquitetura funcional; - infra-estrutura; - modelagem dimensional;

	iteração; - projeto detalhado - implementação; - transição p/ produção; - manutenção.	instalação de produtos; - projeto físico; - projeto e desenvolvimento da organização de dados; - disponibilizar do DW; - manutenção e crescimento.	- automação dos processos; - criação do conjunto inicial de relatórios; - validação e testes de dados; - treinamento; - produção.	atualização do planejamento; - projeto piloto do DW	- projeto de BD; - avaliação de produtos; - execução da arquitetura funcional; - aplicações finais.
Detalhamento das fases	Pouco detalha	Detalha	Detalha	Detalha	Detalha
Arquitetura de dados	Não	Acesso a dados (Rolap e Molap)	Centralizada, DM dependentes e independentes	Centralizada, DM dependentes e independentes	Banco de dados integrado a um DW e [POE 98]
Arquitetura funcional	Não	Ára interna, servidor de apresentação, área externa e metadados.	Integração de dados, transformação de dado, arquitetura de dados e metadados	Ára interna, servidor de apresentação, área externa, serviços e metadados	[KIM 98ª]
Anteprojeto	Não possui	Possui	Possui	Possui	Possui
Auditoria de dados	Não detalha	Detalha	Detalha	Pouco detalha	[KIM 98 a] e [POE 98]
Iteração completa	Não	Não	Não	Não	Sim

A metodologia proposta foi concebida de forma a ser genérica o suficiente para ser aplicada em vários ambientes organizacionais e vários domínios de problema. No entanto, a validação desta generabilidade não foi realizada pois só foi aplicada em um único estudo de caso. Além disto, não se pode atingir a etapa de execução na sua totalidade devido a fatores já citados na seção 6.2.1.1, o que dificultou numa avaliação dos resultados esperados no estudo de caso.

Ao encerrar a execução do projeto junto a Cia Zaffari, utilizando a metodologia proposta neste trabalho, pôde-se concluir que os resultados obtidos neste projeto, além de atender as necessidades gerenciais, também despertou o interesse do público envolvido no projeto sobre a tecnologia de DW.

Como perspectivas de trabalho futuro para a metodologia proposta, sugere-se:

- A realização de um estudo mais detalhado sobre a tecnologia de *Data Mining*;
- O acompanhamento da evolução da equipe do projeto em sua experiência nas necessidades metodológicas correspondentes;
- Análise de tendências mais específicas através da tecnologia de DW;

- Integração de banco de dados heterogêneos para a formação do DW, levando em conta a utilização de padrões de comunicação e acesso como CORBA e ActiveX;
- A realização de testes de campo mais aprofundados, procurando-se apresentar aos tomadores de decisão, informações com as quais normalmente não se trabalha.

E mais amplamente em relação a área de DW, sugere-se:

- Estudo sobre *Data Warehouse* orientado a objetos, pois ainda não existe nenhuma implementação nesta área, levando em conta a modelagem, armazenamento, métodos de acesso e carga de dados em um banco de dados orientado a objetos;
- Descoberta de conhecimento em banco de dados, uma nova área que integra o DW com informações externas de modo que essas informações possam gerar informações mais complexas e detalhadas ao usuário;
- Um estudo sobre as técnicas de compressão de dados, considerando que a base de dados de um DW é muito grande, como se poderia efetuar a compactação e descompactação de forma a não se ter perda na performance sobre as consultas.

REFERÊNCIAS

[BOO 2000] BOOCH, Grady; RUMBAUGH, James; JACOBSON, Ivar. **UML – Guia do usuário**. Rio de Janeiro: Campus, 2000.

[COR 97] CORBELLINI, Humberto. **Um estudo sobre a tecnologia data warehouse.** 1997. Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[DAM 2000] Data Mines for Data Warehouse. Disponível em: <http://www.datamining.com/dm4dw.htm> Acessado em : 28 de junho de 2000.

[DBM 2001] DBMiner E1.1. **User Manual – For Windows NT/95.** DBMiner Technology Inc. Disponível em: <http://db.cs.sfu.ca/DBMiner/dowload2> Acessado em:10 de setembro de 2001.

[DWH 2000] Data Warehouse HomePage Disponível em: <http://www.datawarehousing.com/ > Acessado em: 05 de julho de 2000.

[DWM 2001] Data Warehouse – Data Mining Disponível em: <http://www.datawarehousing.inf.br > Acessado em:01 de setembro de 2001.

[DWP 2000] Data Warehouse papers & Articles Disponível em: http://www.datawarehousing.com/papers.asp Acessado em:01 de julho de 2000.

[EDE 94] EDELWEISS, Nina; OLIVEIRA, José P.M. **Modelagem de Aspectos Temporais de Sistemas de Informação**. IX Escola de Computação. 1994 , Recife.PE.

[FAY 96] FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory, PANDHRAIC, Smyth. **From data mining to knowledge discovery:** an overview. Advances in knowledge and data mining. Califórnia: AAAI Press, 1996.

[FUR 98] FURLAN, José Davi. **Modelagem de objetos através da UML –The Unified Modeling Language**. São Paulo: Makron Books do Brasil, 1998.

[GEN 98] GENERINI, Adelize O. **Conceitos e soluções.** Florianópolis: 1998. v.1.

[GRA 98] GRAY, Paul; WATSON, Hugh J. **Decision Support in the Data Warehouse.** New Jersey: Prentice Hall, 1998.

[HAR 96] HARJINDER, G; RAO, P.C. **The official design the data warehousing,** : Que Corporation, 1996.

[INM 97] INMON, Willian H. **Como construir o data warehouse.** Rio de Janeiro: Campus, 1997. v.1.

[KIM 98] KIMBALL, Ralph. **Data warehouse toolkit.** São Paulo: Makron Books, 1998.

[KIM 98a] KIMBALL, Ralph, REEVES et al. **The data warehouse lifecycle toolkit:** expert methods for designing, developing and developing data wharehouse**.** New York: Jonh Wiley & Sons, 1998.

[MAR 99] MARTIN, James. **Microsoft SQL Server 7.0 – Manual Prático.** Rio de Janeiro: Campus, 1999.

[MIC 2000] Microsoft Servers SQL 7.0 - **Data Warehousing Framework**. Disponível em: <http://www.microsoft.com/SQL/techinfo/datawareframe.htm> Acessado em: 15 de junho de 2000.

[MOR 97] MORAES, Rodrigo L. **Um survey sobre a tecnologia data warehouse.** 1997. Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[PEK 96] PERKINS, Alan. **Developing a data warehouse.** Visible Sustems Corporation. 1996.

[PER 2000] PEREIRA, Walter A. L. **Uma metodologia de inserção de tecnologia de data warehouse em organizações.** 2000. Trabalho Individual II (Mestrado em Ciência da Computação) – Instituto de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

[POE 98] POE, Vidette, KLAUER, Patricia, BROBST, Stephen. **Building a data warehouse for decision support.** New Jersey, Prentice Hall PTR. 1998.

[SYB 2000] Sybase Products. Disponível em: <http://www.sybase.com/products.html> Acessado em: 07 de julho de 2000.

[VAL 96] VALENTE, Daphnis L. **Estudo sobre armazém de dados.** 1996. Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.