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

Um estudo de metodologia para criação de um Depósito de Dados

por
DAPHNIS LOPES VALENTE

Dissertação submetida à avaliação, como requisito
para a obtenção do grau de
Mestre em Ciência da Computação

Profª Dra. Lia Goldstein Golendziner
Orientadora
(*in memoriam*)

Porto Alegre, janeiro de 2001.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Valente, Daphnis Lopes

Um estudo de metodologia para criação de um Depósito de Dados / por Daphnis Lopes Valente. – Porto Alegre: PPGC da UFRGS, 2001.

112f. : il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2001. Orientadora: Golendziner, Lia Goldstein.

1. Extração. 2.Integração. 3. Sistema de Suporte a Decisão. 4. Modelo dimensional. 5. Consulta Analítica. 6. Metadados. 7. Data Mart. 8. Depósito de Dados. I. Golendziner, Lia Goldstein. II. Titulo.

Agradecimentos

Agradeço ao Instituto de Informática da Universidade Federal do Rio Grande do Sul, que ao longo do curso me proporcionou a oportunidade de adquirir os conhecimentos necessários para realizar este trabalho.

Agradeço aos professores e funcionários do Instituto por sua dedicação, e em especial à minha professora orientadora.

Agradeço também ao Professor Clésio Saraiva dos Santos, que me orientou na ausência da Prof[a]. Lia Goldstein Golendziner.

Agradeço, especialmente, à minha maior incentivadora, Susana, minha esposa.

Sumário

Lista de abreviaturas

DBMS	Sistema Gerenciador de Banco de Dados
EIS	Executive Information System
SQL	Structured Query Language
MIS	Sistema de informação gerencial
PC	Computadores Pessoais
DOS	Disk Operating Systems
DD	Depósito de Dados
DM	Data Mart
SSD	Sistemas de Suporte a Decisão
SGBD	Sistema Gerenciador de Banco de Dados
SI	Sistemas de Informação
EMT	Extração, Mapeamento e Transformação
OLTP	On line transactions processing Sistemas orientados a transação
OLAP	On line analytical processing Sistemas orientados a análise
DBA	Database Administrator Administrador de Base de Dados
SW	Software
HW	Hardware
NUMA	Non Uniform Memory Architecture
MPP	Massively Parallel Processing
SMP	Symmetric Multiprocessing
I/O	Input/Output
RI	Retorno de Investimento
SO	Sistema Operacional
CDR	Call Detail Record – Bilhete de central telefônica

Lista de Figuras

Lista de Tabelas

Resumo

Este estudo tem como objetivo analisar as diferentes metodologias existentes para criação de Depósito de Dados (DD) e determinar uma metodologia que melhor atenda às necessidades de uma empresa de telecomunicações, iniciando um projeto de DD e identificando as causas mais comuns de insucesso, a serem evitadas em projetos desta natureza. E para comprovar esta metodologia foi construído um Data Mart utilizando dados da Cia. Rio-grandense de Telecomunicações, com objetivo de análise de qualidade dos bilhetes utilizados nos indicadores de desempenho de centrais bilhetadoras.

Apresenta, também, as arquiteturas possíveis de um Depósito de Dados/Data Mart, suas características e diferenças.

Esta Dissertação de Mestrado é uma contribuição à pesquisa e à análise de metodologias empregadas na criação e manutenção de Depósitos de Dados e a determinação de uma metodologia que atenda às necessidades de uma empresa de telecomunicações.

Palavras-chave: extração, integração, Sistemas de Suporte a Decisão, modelo dimensional, consulta analítica, metadados, Data Mart, Depósito de Dados.

TITLE: "A STUDY OF DATAWAREHOUSE CREATION METHODOLOGIES"

Abstract

The main objective of this study is to analyze different methodologies used in developing a Data Warehouse and also to determine a methodology that will offer the best results when implementing a Data Warehousing for telecommunication segment. It will also identify the most common causes of failure in telecom DW project. And finally as a proof of concept apply the chosen methodology to build a Data Mart, using data from Cia. Rio-grandense de Telecomunicações, resulting in analysis of Call Detail Record quality and telecom exchange performance.

Also shows DW/DM architecture with its characteristics and differences.

This dissertation has as contribution research and analysis of methodologies used in creating and maintaining a DW and to determine a methodology that best suites telecom industry.

Keywords: extraction, integration, Decision Support Systems, dimensional model, analytical query, metadata, Data Mart, Data Warehouse.

1 Introdução

Estrategistas militares utilizam o termo *superioridade incontestável* para descrever o que todo exército gostaria de ter antes de se defrontar com um inimigo. Superioridade incontestável significa possuir a capacidade de combate superior, numérica e tecnológica, e saber efetivamente tirar proveito desta superioridade. Um benefício de ter uma superioridade incontestável e que ela possa fazer com que a maioria dos adversários opte por não entrar em luta na qual acredite não ter nenhuma chance.

Nos dias de hoje, o termo superioridade incontestável se aplica a empresas que tenham alcançado uma posição tão dominante em seu mercado que aparentem ser invencíveis. A Microsoft, a Disney e a Wal-Mart são exemplos que se apresentam nestas condições. Estas empresas, e muitas outras que possuem superioridade incontestável nos seus segmentos de indústria, são capazes de tomar decisões estratégicas e táticas muito rapidamente. O ritmo das mudanças não diminuirá; por isso, a vantagem competitiva pende na direção de empresas que podem acelerar a sua capacidade de tomar decisões.

No ambiente empresarial de hoje, as informações são uns dos bens mais valiosos de que uma empresa pode dispor para combater eficientemente seus concorrentes e defender sua posição no mercado. Cada vez mais os líderes da indústria se sobressaem aos seus adversários, sendo mais rápidos e mais ágeis ao planejar suas estratégias de competição. Conseqüentemente, o objetivo é identificar cada vez mais rápido as oportunidades, planejar suas ações estando de posse do maior número possível de dados, executar suas ações de forma cada vez mais rápida e poder corrigir possíveis enganos antes de seus concorrentes.

Na maioria das organizações os dados se encontram dispersos em vários sistemas, de diferentes tecnologias, armazenados em arquivos, em bases não relacionais e relacionais, sem padronização na identificação dos seus elementos, o que acaba dificultando enormemente sua utilização no suporte à decisão.

A área acadêmica tem desenvolvido muitos trabalhos envolvendo bases de dados heterogêneas com vistas a permitir consultas nestas bases diversas e, após o processo de integração, obter resultados homogêneos conforme solicitado.

Depósito de Dados é uma excelente alternativa ao enfoque tradicional para integração e acesso de dados a fontes de informações heterogêneas. O enfoque DD é especialmente útil quando se necessita de altos desempenhos nas consultas (característica dos sistemas de suporte à decisão), ou quando as fontes de informações são transitórias (bases que sofrem alterações intensas impedem a obtenção de informações de eventos já realizados).

Com o uso de um DD podemos atender às necessidades de empresas que se encontram com muitos sistemas legados ainda em operação, mas que necessitam de informações para apoiar seu planejamento estratégico e ações rápidas. Isso se realiza integrando os dados transacionais de variadas origens, unificando suas formas diversas em uma única e consistente base de dados que permitirá análises e decisões complexas de negócio.

Um DD é semelhante a um depósito físico. Sistemas transacionais criam dados que são carregados no depósito. Alguns destes dados são sumarizados em informações e armazenados no depósito. Usuários de DD fazem solicitações e recebem informações que são criadas com os dados e informações sumarizadas do depósito.

Um DD é tipicamente uma mistura de tecnologias, incluindo base de dados relacionais e multidimensionais, arquitetura cliente/servidor, programas de extração e transformação, interface gráfica para os usuários e muitas outras.

Ferramentas devem ser disponibilizadas de acordo com as necessidades de informação dos usuários, mas com a garantia de obter informações rápidas, confiáveis e consistentes. Um DD é projetado de forma a facilitar a identificação da informação necessária ao usuário e obter estas informações seja através do uso de ferramentas simples, com gerador de relatórios ou planilha eletrônica, ou ferramentas estatísticas complexas.

Três tecnologias fundamentais convergem para formar a infra-estrutura da informação com a finalidade de acelerar o processo de tomada de decisão. O investimento nessas tecnologias ajuda a criar o capital intelectual necessário a que empresas se sobressaiam em relação a suas concorrentes:

- Armazenamento de dados, que é a criação de um repositório de dados completo e preciso.

- Ferramentas de análise, que permitem o acesso e avaliação dos dados. Podem ser multidimensionais, geradoras de relatórios ou ferramentas estatísticas.

- Internet, que melhora a comunicação e colaboração por toda a empresa.

Os exemplos que seguem sugerem algumas de suas áreas de aplicação ou características onde a utilização da abordagem do depósito de dados é adequada [WAT 95].

1. *Reunir dados científicos.* Nestas aplicações, um grande volume de dados heterogêneos pode ser criado tão rapidamente que o processamento de uma pesquisa em tempo real se torna impossível. Além disso, as fontes podem ser esporádicas e incertas, portanto, armazenar os dados em um local seguro e conveniente para futuro processamento é apropriado.

Ex.: NASA-Jet Propulsion Laboratory e Caltech desenvolveram o SKICAT (SKy Image Cataloging and Analysis Tool – Ferramenta de Catalogação e Análise do Céu), um avançado sistema de mineração da informação que automaticamente analisa e cataloga a segunda pesquisa Palomar Sky Survey dos céus do norte. (O SKICAT foi escrito em C, roda sob Unix e emprega algoritmos especiais e um DBMS Sybase). Quando estiver completa, a pesquisa terá catalogado mais de 50 milhões de galáxias, cerca de 2 bilhões de estrelas e 100.000 quasares. A pesquisa produzirá aproximadamente 3 terabytes de dados, que serão reduzidos a um catálogo da galáxia.

SKICAT descobriu nove novos quasares. Com as técnicas anteriores de pesquisa seriam necessários três anos para completar uma descoberta semelhante. Com o SKICAT, os astrônomos do Caltech realizaram o feito em menos de seis meses, utilizando tempo de observação de no mínimo uma ordem de magnitude inferior.

2. *Manter historicamente dados empresariais.* Processar e minerar dados empresariais (calcular o histórico das vendas de todas as lojas de uma grande cadeia de supermercados em um determinado período de tempo) é um trabalho intenso, e é melhor ser executado fora do ambiente normal ("*off-line*") de produção para não afetar as operações do dia-a-dia.

Ex.: O Wal-Mart descentralizou o controle de seus estoques a partir 1989. A mineração de dados serve como base para a estratégia da rede de estoque descentralizado, e neste caso é o processo de localização de dados úteis encontrados nas informações de compras e vendas específicas de cada

região. A empresa utiliza um sistema Intranet que auxilia nas decisões sobre as compras para cada loja. Hoje, aproximadamente 10 milhões de decisões de aquisições são realizadas diariamente na rede com um DD de 101 terabytes.

3. Obtendo informações freqüentemente requisitadas. Pelo armazenamento de respostas em depósitos de dados previamente extraídas e integradas para questões freqüentemente perguntadas, a desvantagem inerente dos sistemas de bancos de dados (ineficiência, atraso no processamento de consultas, etc.) pode ser superada, resultando numa melhora no desempenho e eficiência.

Esta dissertação, além de estudar os fundamentos básicos (capítulo 2 – Conceitos), se preocupou principalmente em analisar as metodologias existentes (capítulo 3 – Estrutura de DD). O capítulo 2 descreve os conceitos de um Depósito de Dados conforme variada bibliografia [INM 96][BAR 96][HAR 96][HAC97] e apresenta em detalhes o porquê da necessidade de utilizar um DD, as características clássicas de um Depósito de Dados, as principais causas de insucesso em projetos deste tipo e os objetivos desta dissertação.

Um dos objetivos deste trabalho é apresentado no capítulo 3 – Estrutura de DD, onde é apresentada a arquitetura de um DD, as várias topologias de um Data Mart, a análise de metodologias existentes e a proposta de uma metodologia que melhor atenda às necessidades de uma empresa de telecomunicações, que gera uma grande quantidade de dados para serem analisados. Exemplificando: uma empresa com dois milhões de terminais telefônicos fixos gera cerca de 25 milhões de registros/dia para chamadas interurbanas, e para chamadas locais (urbanas) um número três vezes superior. Para chegar a uma proposta de metodologia foram examinadas as metodologias de James Martin [Mar 98], de Alan Simon [SIM 98], de Ralph Kimball [KIM 98], de Douglas Hackney [HAC 98], de NCR Corporation [NCR 99] e a metodologia utilizada pela empresa Visible [VIS 99]. Também são apresentados os bancos de dados e ferramentas comerciais dirigidas para criação e uso de um DD.

A necessidade de utilização de uma metodologia que possa contribuir significativamente para o sucesso da construção de um DD é o principal motivador deste trabalho. A existência de grandes bases de dados transacionais em todas as empresas e em especial na área de telecomunicações determina a necessidade da construção de um DD. A escolha da metodologia

adequada para esta construção é primordial para permitir atender às necessidades de toda a corporação nas informações que irão permitir a tomada de decisões.

Esta dissertação também está motivada pela possibilidade de utilizar a metodologia proposta no capítulo 3, visando atender a um problema determinado na Cia. Rio-grandense de Telecomunicações na área de análise da qualidade de bilhetes, para o que foi feita uma prova de conceito com a criação de um Data Mart utilizando ferramentas de extração, base de dados relacional, base multidimensional e ferramenta de acesso.

As conclusões do trabalho e a apresentação de novas direções de pesquisa estão no capítulo 6.

Nos anexos 1 e 2 está a descrição da base de dados utilizada e as transformações realizadas nos dados relativas à prova de conceito do capítulo 4, e finalmente, a bibliografia utilizada.

2 Conceitos

2.1 Conceitos

Os conceitos de Depósito de Dados e Data Mart são apresentados a seguir, por serem estes os conceitos fundamentais deste trabalho.

Vários são os autores que definem Depósito de Dados. A seguir definições que melhor representam este sistema:

Depósito de Dados é uma coleção de dados orientadas a assunto, integrada, não-volátil e temporal, de suporte a decisões gerenciais [INM 96].

Depósito de Dados (tradução do termo inglês *data warehouse*) consiste em um único repositório de dados, extraídos de bases de dados transacionais e/ou dados externos, acumulados ao longo do tempo, integrados, possibilitando a análise massiva de informações, de forma a permitir melhores tomadas de decisões e a descoberta de conhecimento, sem impactar no desempenho dos bancos de dados do mundo transacional [BAR 96].

Depósito de Dados é um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas, *ad hoc*, relatórios analíticos e de suporte a decisão [HAR 96].

Depósito de Dados é uma coleção de técnicas e tecnologias que juntas disponibilizam um enfoque pragmático e sistemático para tratar com o problema do usuário final de acessar as informações que estão distribuídas pela organização [BAR 96].

O mesmo ocorre para Data Mart (DM). A seguir definições que melhor representam este sistema:

Data Mart é um Depósito de Dados especializado contendo um subconjunto de dados, originados do Depósito de Dados, e preparados para atender a uma necessidade particular de uma área do negócio [BAR96].

Data Mart é um conjunto de dados sumarizados derivados dos dados detalhados encontrados no Depósito de Dados [INM96].

Data Mart é um conjunto de informações com foco definido que são projetadas das mesma forma que um DD, mas são implementadas para atender às necessidades específicas de um conjunto de usuários que compartilham características comuns [HAC97].

Data Mart incremental é um conjunto de dados organizados para atender a uma área funcional específica do negócio, baseado em uma arquitetura corporativa definida, preparado para integração, escalável e com

capacidade de subsidiar a construção de um sistema centralizado de Depósito de Dados [HAC97].

2.2 Necessidade de utilizar um DD

Os sistemas de informações, em todas as organizações, evoluíram até a situação atual através de caminhos tortuosos e acidentados. Os sistemas, longe de terem uma visão corporativa, foram criados visando resolver problemas para áreas específicas, agindo localmente e atendendo problemas particulares sem a preocupação com o todo. Isso se aplica perfeitamente a departamentos de uma corporação que tipicamente tendem a enxergar somente os problemas de sua área e buscar sistemas que os resolvam. Por exemplo, Sistemas de Contas a Pagar não se preocupa com a área de Vendas; Mercadologia não se preocupa com área de Manufatura. No entanto, a gerência superior necessita enxergar o todo para poder realizar decisões mais efetivas. Os sistemas atuais apresentam dados fracionados e cada pedaço está em conflito com as outras partes em termos de dados contraditórios ou sobrepostos por terem definições de dados inconsistentes. Como resultado disso, para se de obter estatísticas vitais de todas as áreas da empresa, o trabalho necessário é enorme, principalmente na determinação de sistemas individuais que contêm as partes de dados necessárias, e mais difícil ainda é conciliar as discrepâncias para se obter uma única visão.

O objetivo primário de um DD é resolver este dilema para os gerentes e analistas de cenários e ao mesmo tempo diminuir a carga de trabalho sobre a área de sistemas de informações. Simplificadamente, poderíamos dizer: a solução é obter todos os dados requeridos para suporte a decisão dos sistemas transacionais e arrumá-los em um único repositório de forma consistente, de forma a responder perguntas que necessitam ser respondidas [FLA 97].

As bases de dados destes sistemas normalmente estavam centralizadas em um único banco de dados servindo a toda a comunidade de processamento da informação – de transação, a processamento em lote e a processamento analítico. Na maioria dos casos, o foco primário das bases de dados era transacional. Há poucos anos, surgiu com uma noção mais sofisticada para bancos de dados – uma base que serve às necessidades transacionais e outra que serve às necessidades de informações ou analíticas [INM96]. Isto se deve ao uso de PCs, tecnologias de 4GL, o poder crescente de processamento dos equipamentos de usuários e o decréscimo do custo dos equipamentos.

A necessidade de ruptura dos bancos de dados transacionais e os de informação se deve a vários fatores:

- os dados que atendem às necessidades transacionais são fisicamente diferentes daqueles que servem às necessidades de informações e dados analíticos;

- a tecnologia utilizada no processamento transacional é fundamentalmente diferente daquela utilizada ao suporte das necessidades de informações e dados analíticos;

- o usuário da comunidade transacional é diferente daqueles atendidos por informações e dados analíticos.

Devido a estas e outras razões é que a maneira adequada de construir bases de dados é separar as necessidades dos sistemas transacionais dos sistemas de analíticos.

Dados analíticos aqui são vistos como o processamento que serve às necessidades de gerência no processo de tomada de decisão. Muito conhecido como Sistema de Suporte a Decisão (SSD), o processamento analítico trabalha com grandes quantidades de dados para detectar tendências. Ao invés de trabalhar com um ou dois registros de dados (processamento transacional), quando um analista executa uma consulta muitos registros são processados.

Além disso, é muito raro um analista de SSD atualizar dados. Em sistemas transacionais, dados são constantemente atualizados ao nível de registro individual. Em processamento analítico, registros são constantemente buscados e seus conteúdos analisados, mas pouca ou nenhuma alteração de registros individuais acontecem.

Em SSD, o tempo de resposta necessário é muito flexível quando comparado a sistemas transacionais tradicionais. Tempo de resposta para SSD é medido de 30 minutos a 24 horas. Tempo de resposta desta ordem de grandeza para sistemas transacionais seria catastrófico.

O número de usuários pertencentes à comunidade de SSD é muito menor que a aquela que serve à comunidade transacional. Usualmente, existirá um número infinitamente menor de usuários de SSD do que os da área transacional.

Diferente da tecnologia que atende ao ambiente SSD, a tecnologia do ambiente transacional deve se preocupar com bloqueio de dados e transações, *deadlock* e outros [INM 96].

Portanto, a necessidade de utilização de um DD se deve à impossibilidade de manter desempenho aceitável nos sistemas OLTP, quando utilizado como base de dados de consultas de SSD, e também pelo baixo desempenho obtido, quando criamos bases replicadas utilizando a mesma plataforma e modelo de dados de sistemas OLTP (altamente normalizado) para uso por SSD.

2.3 Características

A arquitetura de um DD atende às necessidades de uma organização que procura recursos de suporte a decisão flexível e escalonável. Um DD integra dados de vários sistemas OLTP e dados de terceiros em um ambiente comum que toda área de negócio pode facilmente consultar. Os dados são limpos para remover os erros e omissões nos sistemas-origem e carregados no DD, de forma a ter confiabilidade nas informações obtidas pelos usuários. O DD contém dados históricos, de forma que usuários podem examinar tendências e explorar dados históricos detalhados para descobrir padrões desconhecidos. DD é somente para leitura, garantindo aos usuários que os relatórios de hoje serão compatíveis com os do próximo mês. O DD é integrado permitindo uma visão de todas áreas da organização de todos os sistemas OLTP [WID 95].

Uma estrutura clássica de DD é definido pela seguinte série de características fundamentais:

- Base de dados de SSD separada de sistemas OLTP
- Armazenamento de dados somente
- Dados integrados
- Dados limpos
- Temporal
- Não-volátil
- Orientado a assunto
- Metadados

- Modelo dimensional
- Acesso fácil

2.3.1 Base de dados SSD separada

O DD é sempre armazenado numa base de dados separada dos sistemas OLTP numa arquitetura clássica de DD.

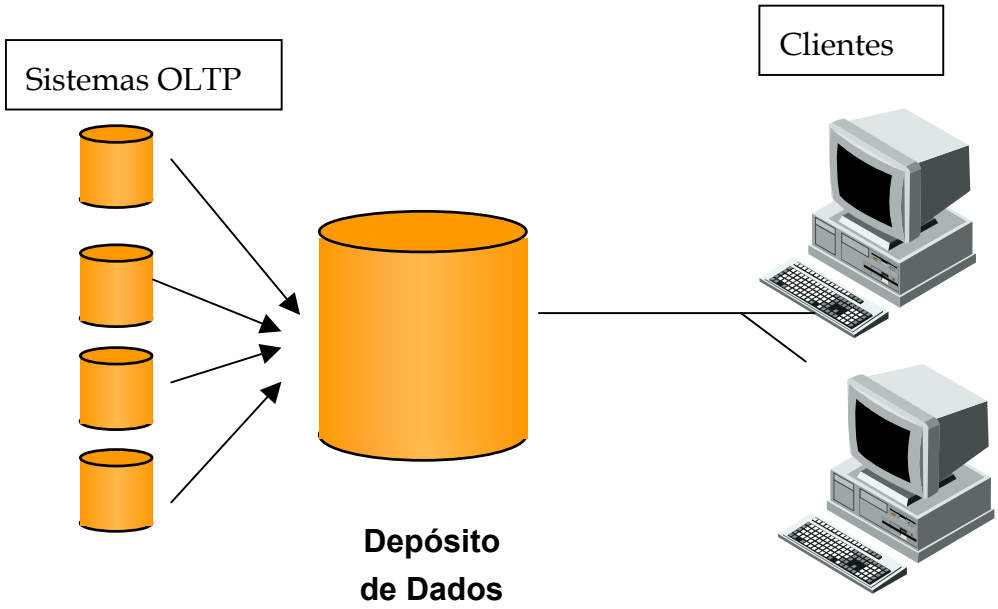

FIGURA 1 – Depósito de Dados

Desta forma teremos ambientes diferentes para os sistemas OLTP e o DD. Normalmente, o DD está instalado em um servidor dedicado, permite que o servidor e a base de dados sejam otimizados, ou afinados, para o melhor desempenho possível. Sistemas OLTP necessitam de otimizações totalmente diferentes de SSD.

2.3.2 Armazenamento de dados

O DD funciona como um repositório de dados. Nenhum dado é criado no DD. Os dados podem sofrer derivações dos dados originais, mas nenhum dado novo é criado no sistema de DD.

2.3.3 Dados integrados

Uma das propriedades básicas de um DD é a integração dos dados de diversos sistemas de OLTP na organização. O DD contém informação de todos os sistemas de OLTP existentes. Isto permite que usuários acessem todas

as informações corporativas disponíveis tanto em nível de transações como descritivas. A habilidade do DD possibilitar o acesso a dados de toda a corporação é extremamente poderoso para os analistas de negócio na tomada de decisões. Esta integração e disponibilidade dos dados é o que viabiliza a criação de um DD, permitindo retorno de investimento de até 400%.

De todos os aspectos do Depósito de Dados, este é o mais importante. A Figura 2 mostra que a integração ocorre quando os dados passam dos bancos de dados transacionais para o DD.

As várias decisões de projeto que os projetistas de sistemas de aplicação fizeram durante vários anos aparecem em milhares de diferentes formas. Não existe consistência na codificação das aplicações, convenção de nomes, atributos físicos, medidas dos atributos e muito mais. Cada desenvolvedor tinha liberdade para tomar decisões.

Quando os dados são colocados no Depósito de Dados, isto é feito de forma que as inconsistências das aplicações sejam desfeitas. Por exemplo, na Figura 2, não interessa se o atributo sexo foi codificado m/f ou 1/0, o que interessa é qual a codificação realizada para o Depósito de Dados, devendo ser consistentes independente da aplicação de origem. Se a aplicação codifica com X/Y, ela é convertida quando é movida para o Depósito de Dados. A mesma consideração de consistência existe para outras características dos dados como estrutura da chave, convenção de nomes e características físicas.

FIGURA 2 – Integrado [INM 94]

2.3.4 Dados limpos

Quando os dados são carregados no DD de um sistema OLTP, ele sofre uma limpeza, retirando erros, irregularidades e outras anomalias. Esta etapa crítica cria ambiantes que usuários podem acessar com confiança. Não se pode utilizar dados oriundos dos sistemas OLTP sem que haja um tratamento dos dados

2.3.5 Temporal

DD contém dados históricos que permitem a usuários examinar tendências e desempenhos históricos. Dados históricos podem ser guardados resumidos ou em detalhe.

Depósito de Dados é variante no tempo. A Figura 3.b mostra como a variação do tempo ilustra este fato, que pode ser de várias formas:

- O horizonte de tempo para o DD é significativamente maior que de um sistema de operação. Um horizonte de 60 a 90 dias de intervalo é normal para sistemas de operação; um horizonte de 5 a 10 anos é normal para um DD.
- Bancos de dados de operação contêm valores correntes – valores corretos para o momento de acesso. Como tal, estes valores podem ser atualizados. Dados de DD são na verdade uma série de fotografias (*snapshots*) tirados em um momento do tempo.
- A estrutura-chave de dados de operação pode ou não conter elementos de tempo, como ano, mês, dia, hora etc. A chave da estrutura do DD sempre conterá algum elemento de tempo.

2.3.6 Não-volátil

Depósito de Dados é uma base de dados em que os dados depois de carregados não sofrem alterações, sofrem atualizações (novos carregamentos), portanto são não-voláteis [INM 96]. A Figura 3.a ilustra a não-volatilidade dos dados. Essa figura mostra que dados transacionais são buscados regularmente e manipulados um registro por vez. Atualização é feita nos dados em ambiente de operação. Mas o Depósito de Dados exibe um conjunto de características bem diferentes. Os dados em um Depósito de Dados são carregados (usualmente em massa) e depois consultados. Mas a alteração dos dados não ocorre no ambiente do Depósito de Dados.

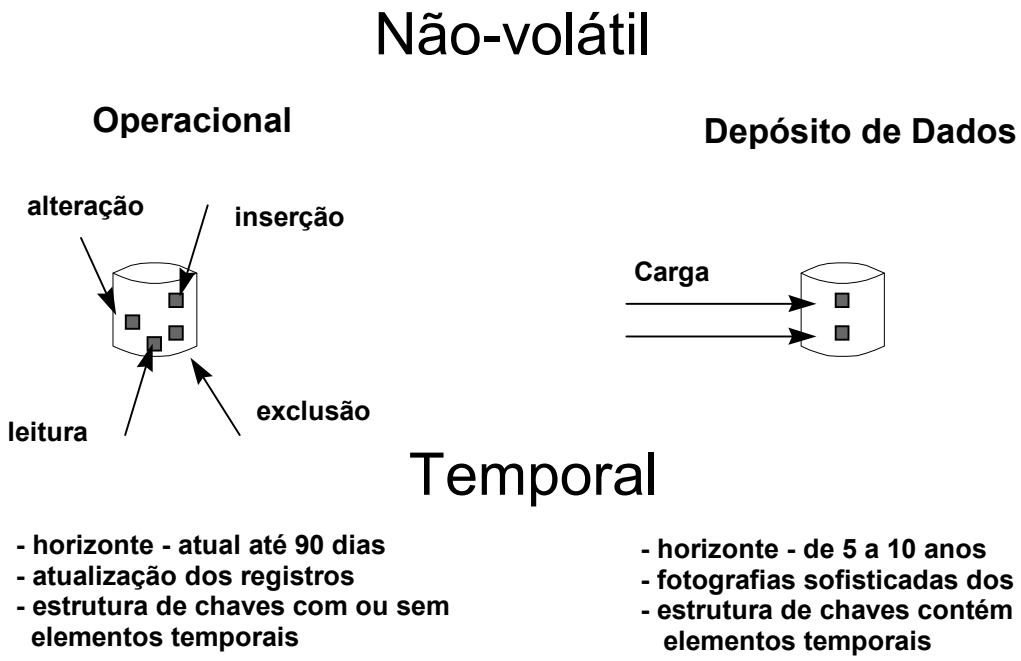

FIGURA 3.a – Não-volátil e 3.b – Temporal [INM 94}

2.3.7 Orientado a assunto

Conforme definição em [INM 96], o Depósito de Dados é orientado a assunto, em contraposição ao sistemas OLTP. Em uma empresa de seguros, os sistemas podem ser de automóvel, saúde e vida. Mas os principais interesses (assuntos) da empresa podem ser cliente, prêmio, apólice e sinistros (veja Figura 4).

Projeto orientado a assunto é totalmente dirigido aos usuários. Os elementos de um DD são otimizados para prover um sistema flexível, fácil de entender e utilizar.

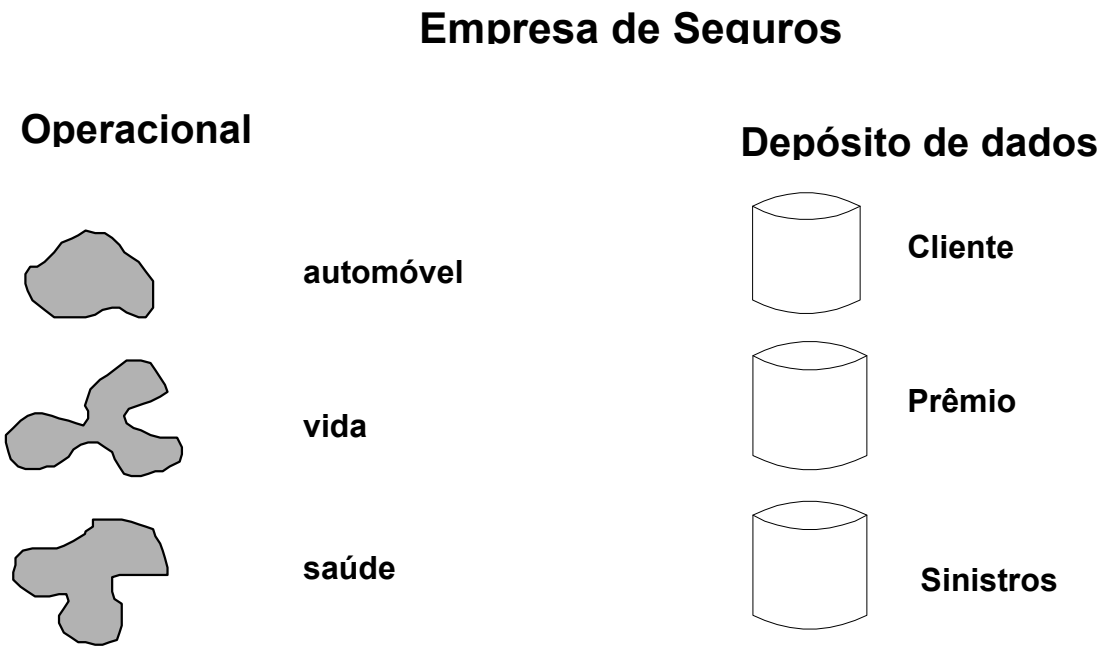

FIGURA 4 – Orientado a assunto [INM 94]

2.3.8 Metadados

Nenhum DD pode obter sucesso sem um conjunto de metadados que seja completo, eficaz e de fácil consulta. Metadados oferece informação sobre o conteúdo, operação, estrutura e gerência de um sistema de DD. Ele permite que os usuários identifiquem rapidamente e compartilhem as informações sobre os dados, permite também que as ferramentas utilizadas para criar, manter e utilizar o DD compartilhem informações [BRA 96].

2.3.9 Modelo dimensional [KIM 96][KIM 98]

DD deve ser projetado visando o benefício do usuário final, não o DBA ou a área de sistemas de informação. Tradicionalmente, esquemas de base de dados apresentam tabelas com inter-relacionamentos complexos com múltiplas uniões circulares entre dois pontos no modelo. Conseqüentemente, uma consulta que une duas tabelas pode ser processada de diferentes maneiras, dependendo no caminho utilizado. Estes tipos de esquemas são difíceis de visualizar inclusive para seus criadores, e muito mais para usuários finais reconhecerem como o modelo de seu negócio.

Esquemas tradicionais de bancos de dados transacionais são projetados para facilitar atualizações de dados; cada entidade lógica que deve ser atualizada tem sua própria tabela, de forma que uma transação afeta muito

pouco o resto da base de dados. Como resultado, o esquema consiste de muitas tabelas, todas conectadas por vários relacionamentos, um para muitos. Embora eficiente para processamento de transações, estes relacionamentos tendem a não ser intuitivos e causam confusão quando se trata de realizar consultas.

A maioria dos usuários utiliza algum tipo de ferramenta de consulta ao invés de SQL para compor consultas, e esta ferramenta permite que se utilize uma base dados sem acompanhamento da área de sistemas de informação. Isso significa que o esquema de um DD deve ser simples o suficiente para facilitar o entendimento e o acesso direto à base de dados.

Para alcançar este esquema, o projetista deve ser flexível e maleável para acomodar as necessidades de todos os grupos de usuários e facilitar a atualização do DD. O projetista de DD deve criar um esquema de DD em que o usuário final possa facilmente entender os termos do negócio. A melhor maneira de visualizar este modelo é através do esquema estrela. Esquemas estrela têm a vantagem de serem simples e intuitivos, mas também fazem uso de novos enfoques de indexação e união de tabelas.

O esquema estrela (multidimensional) (Figura 5), possui uma estratégia que utiliza relacionamentos simples entre todas as partes das informações da base de dados. Uma estrela simples consiste em um grupo de tabelas que descrevem as dimensões do negócio arranjados logicamente em volta de uma tabela imensa que contém os fatos acumulados da empresa. As tabelas menores são as pontas da estrela e a tabela maior é o centro da estrela.

A tabela central, a maior, contendo milhões ou até centenas de milhões de linhas, é conhecida como tabela de fatos. O conteúdo da tabela de fatos são milhões de valores, as medidas do negócio, como transações de vendas ou compras, que são carregadas de sistemas OLTP. As tabelas de fatos podem utilizar até 95% da área em disco necessária para armazenar o DD.

Os fatos armazenados no DD devem representar os valores reais que o usuário tem interesse.

O conteúdo das tabelas dimensões – as pontas da estrela – são os dados do negócio. Cada tabela tem um número fixo de registros, como uma lista de produtos ou serviços, o nome de regiões geográficas e mercados, e dados textuais. Estas tabelas se tornam as colunas nas consultas dos usuários. Tabelas dimensões tendem a utilizar tipos caracteres ao invés de numéricos, de forma que suas linhas são muito mais longas mas em número muito menor, ocupando uma pequena percentagem de espaço em disco [BAR 96].

O benefício do esquema estrela é que realizando pré-joins e a redundância seletiva, o projetista do DD simplifica os dados e aumenta o fluxo de dados necessários para SSD. A melhoria em desempenho se deve à antecipação na modelagem do DD das consultas mais constantes a serem

realizadas pelos usuários, de forma que redundâncias possam ser determinadas com objetivos de obter melhores tempos de respostas nas consultas.

A estrutura de dados estrela é ideal somente porque o ambiente do DD é um ambiente de carga-acesso de dados históricos.

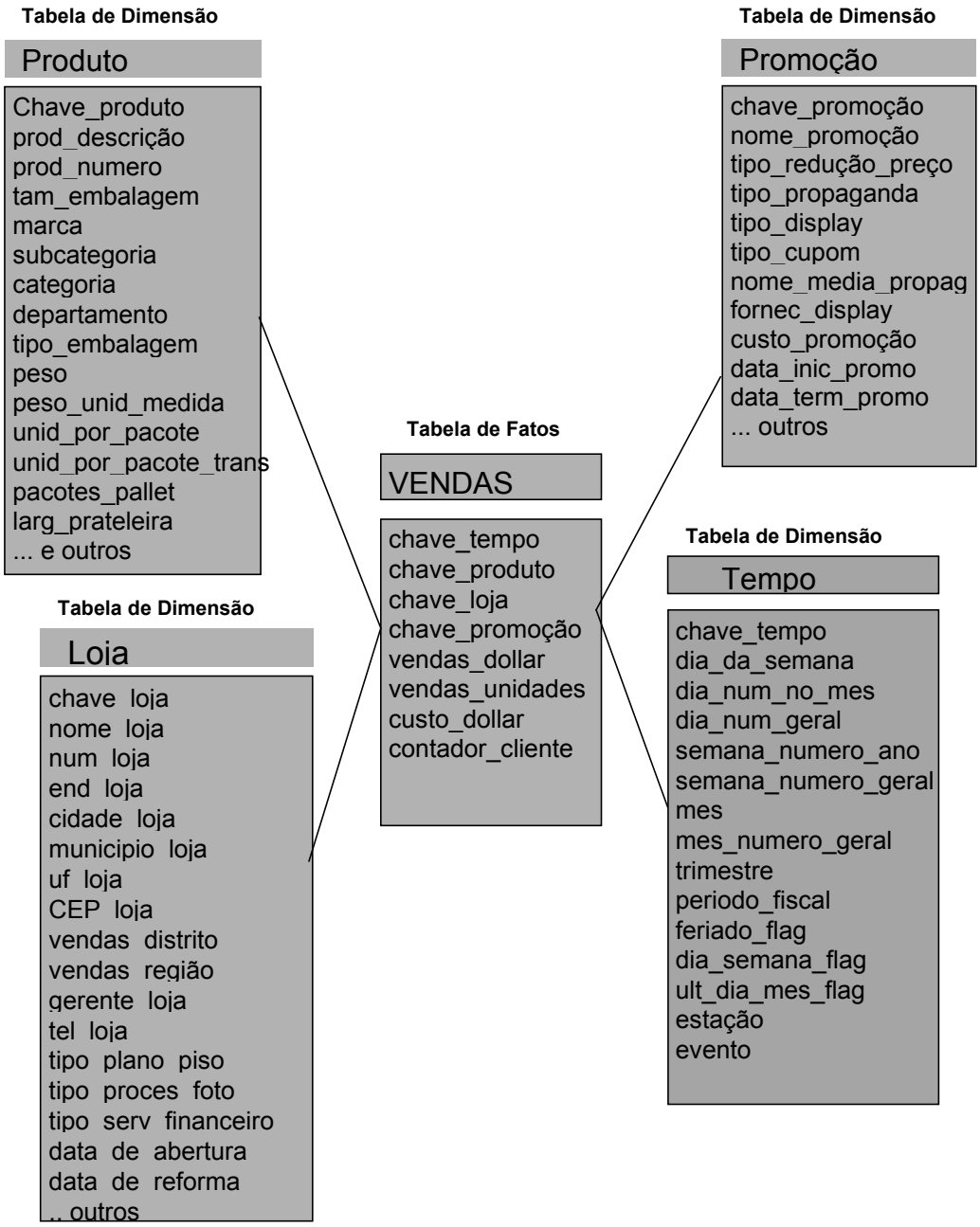

FIGURA 5 – Exemplo de modelo dimensional

A tabela de fatos é onde as medidas numéricas são armazenadas. As medidas numéricas são os valores de venda em moeda, o número de unidades vendidas, custo e contador de clientes. Os melhores fatos são numéricos, continuamente valorados e aditivos.

Uma tabela dimensão Produto típica é a apresentada na figura anterior, contendo em torno de 50 atributos. Cada um destes atributos se torna automaticamente uma possibilidade de utilização como coluna em uma consulta.

A tabela Promoção descreve cada condição de promoção na qual cada produto foi vendido. Condição de promoção inclui redução de preços, colocação nas prateleiras, anúncio em jornais e cupom. Esta tabela descreve fatores que causam alterações na venda de produtos. Uma tabela importante para determinar resultados de vendas relacionados a promoções.

A tabela dimensão Tempo é garantida de existir em todos os DD, porque todos necessitam das informações situadas no tempo, porque as informações são carregadas ao DD em períodos determinados pela aplicação, e recebem necessariamente um selo de tempo.

Uma outra facilidade apresentada pelo modelo dimensional é que normalmente sua estrutura representa a forma de visualizar dos próprios executivos. Assim, uma consulta é feita através da escolha dos atributos departamento, vendas em valores e vendas em unidades.

2.3.10 Acesso fácil

O DD deve ser facilmente acessado por todos os usuários. Ferramentas de acesso de fácil utilização e aprendizado são pré-requisitos para um sistema de DD.

Podemos melhor visualizar as diferenças entre bancos de dados transacionais e Depósito de Dados através de uma forma tabular como se segue:

TABELA 1 – Comparação entre bancos de dados transacionais e Depósito de Dados

Dados transacionais	Depósito de Dados
• Orientados a aplicação	• orientados ao assunto
• detalhados	• sumarizados
• precisos	• representa dados no tempo
• serve a toda a comunidade	• serve a comunidade gerencial
• pode ser atualizado	• não é atualizado
• roda repetitivamente	• roda heuristicamente
• sensível ao desempenho	• não sensível ao desempenho
• orientados a transação	• orientados a análise
• sem redundância	• redundância é um fato
• estrutura estática	• estrutura flexível
• alta probabilidade de acesso	• baixa probabilidade de acesso
• tempo de resposta de 0 a 3 seg.	• tempo de resposta de seg. a horas

2.4 Principais causas de insucesso

O Data Warehousing Institute aponta os dez erros mais comuns na implantação de um Data Warehouse [DAT 99]:

1. começar o projeto com o tipo errado de "patrocínio";
2. gerar expectativas que não podem ser satisfeitas, frustrando os executivos quando da utilização do DD;
3. dizer: “Isto vai ajudar os gerentes a tomar decisões melhores” e outras afirmações politicamente ingênuas;
4. carregar o DD com informações só "porque estavam disponíveis";
5. falhar no objetivo de acrescentar valor aos dados através de mecanismos de desnormalização, categorização e navegação assistida;

6. escolher um gerente para o DD que seja voltado para a tecnologia ao invés de voltado para o usuário;

7. focalizar o DD em dados tradicionais internos orientados a registro e ignorar o valor potencial de dados textuais, imagens, som, vídeo e dados externos;

8. fornecer dados com definições confusas e sobrepostas;

9. acreditar nas promessas de desempenho, capacidade e escalabilidade dos vendedores de produtos para DD;

10. usar DD como uma justificativa para modelagem de dados e uso de ferramentas *case*.

2.5 Objetivos

Este trabalho tem como objetivo principal a análise das metodologias existentes, baseadas principalmente em [HAC 97], [KIM 96], [KIM 98], [INM 96], [NCR 99], [MAR 98], [BAR 96], mas, além disso, pretende apresentar os conceitos, fundamentos e principais componentes de um processo de DD.

Estas informações foram obtidas principalmente na bibliografia apresentada no final desta dissertação, mas também em revistas como: Database Programming and Design, DBMS Magazine, DM Review, Lista de Interesse–dwlist@datawarehousing.com, The Data Warehousing Institute white papers, DB2 Magazine, Teradata Review e do "site" http://pwp.starnetinc.com/larryg/ onde se encontra uma série de informações e links para todos os tipos de ferramentas necessárias e disponíveis para DD na Internet.

Como forma de aplicar a arquitetura e a metodologia sugeridas por este trabalho, foi desenvolvido um Data Mart para uma empresa de telecomunicações utilizando dados de bilhetes gerados pelas centrais telefônicas, com o propósito de analisar a qualidade dos bilhetes que participaram na geração de indicadores de resultado das centrais telefônicas.

Os dados utilizados são uma amostra dos dados em função do tempo de processamento e espaço em disco necessário para os dados totais, mas a solução desenvolvida pode e possivelmente será adotada para uso de análise na área de medições e registro da Cia. Rio-grandense de Telecomunicações.

3 Estrutura de Depósito de dados

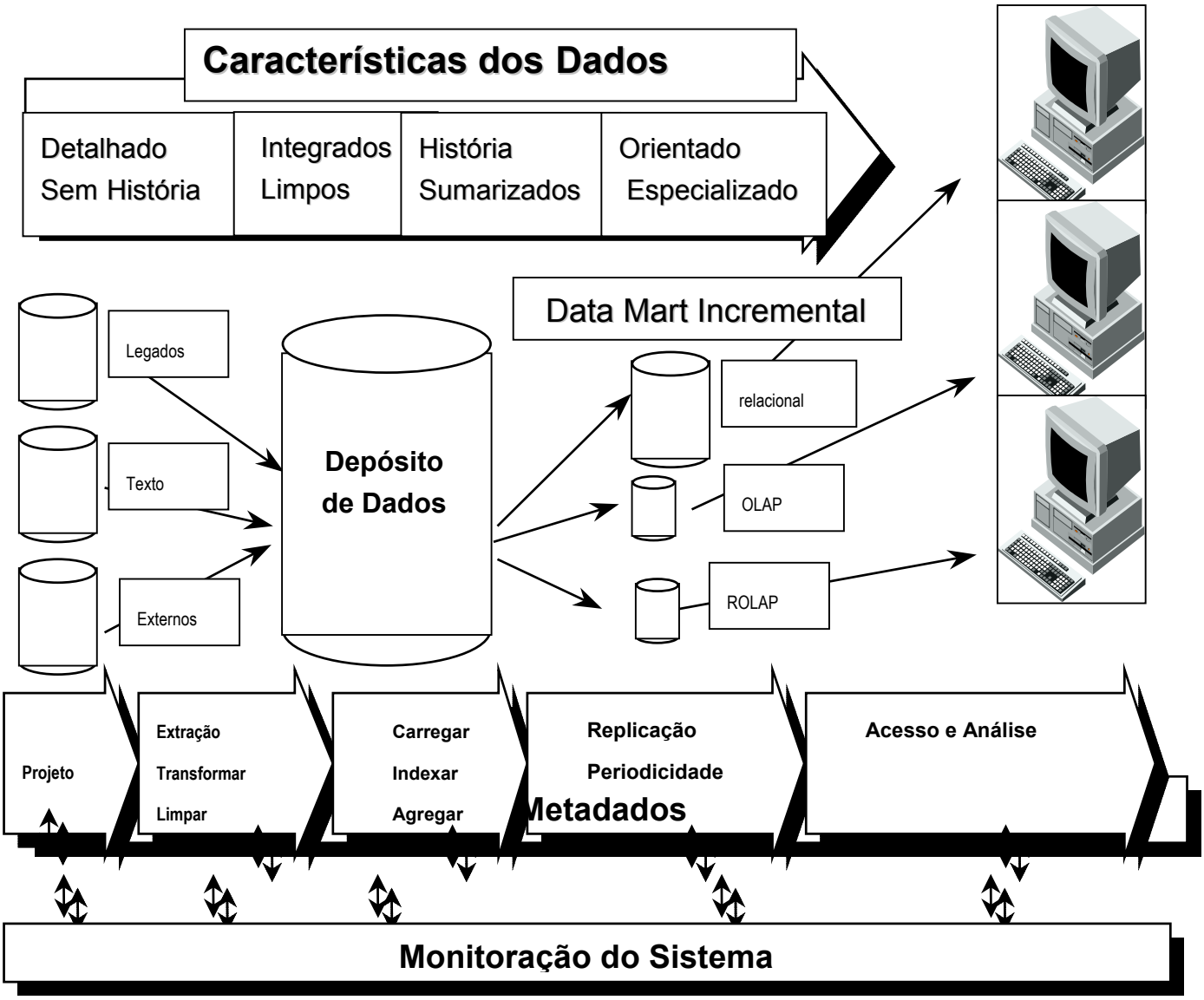

FIGURA 6 – Ciclo de vida de um Depósito de Dados [HAC 97]

A figura acima mostra todas as etapas necessárias para criação de um DD. Utilizando uma metodologia adequada, temos todas estas etapas se repetindo interativamente durante todo o ciclo de vida do DD [HAC 97].

3.1 Arquitetura

Existem diversas arquiteturas possíveis para um DD, mas a que é recomendada neste trabalho é uma arquitetura de Data Marts incrementais. A seguir serão apresentadas vantagens e desvantagens destas arquiteturas.

A primeira arquitetura já é considerada clássica e talvez a mais conhecida. Definida e defendida por William Inmon desde o início dos anos 90

como a única que possibilita o sucesso de um DD [INM 96][HAM 95]. Esta arquitetura defende a utilização de um DD centralizado com os clientes acessando diretamente esta base de dados, por estarem os usuários de SSD localizados normalmente em um único local, e, pelo grande volume de dados necessários, um único repositório é mais adequado e a utilização de dados dispersos em vários locais apresentam problemas de integração e acesso (também conhecido como enfoque Top-Down). Veja na Figura 1 o exemplo desta arquitetura.

Implementações completas de sistemas DD consistem de múltiplas áreas de interesse, compostos de dezenas ou centenas de tabelas de fatos e muitas tabelas dimensões compartilhadas. Estas áreas de interesse (ou assunto) são preenchidas por dezenas ou centenas de programas de extração, limpeza, e transformação. Após o carregamento das tabelas na base de dados configurando os elementos do esquema estrela, o conjunto de dados resumido é criado ou atualizado pela agregação dos valores nas tabelas de fato junto com as diversas dimensões do negócio e hierarquias contidas nas tabelas dimensões. Usuários utilizando uma grande variedade de ferramentas de acesso e tecnologias realizam consultas para obter respostas para problemas e tendências necessárias ao negócio. Estas consultas são disponibilizadas para outros usuários através de uma biblioteca de respostas, permitindo a solicitação de relatórios, análises e outras consultas.

O processo completo – de EMT (Extração, Mapeamento e Transformação), carregamento dos dados e indexação, agregação, replicação e distribuição de acesso – é monitorado e gerenciado pelas ferramentas de sistema. Estas ferramentas de gerência e monitoramento auxiliam a equipe de DD a identificar e corrigir problemas, antecipar e minimizar gargalos e queda de desempenho, padrões de utilização por usuários, e identificar a oportunidade de realizar agregações.

O desafio de implementar sistemas de DD é principalmente pelo custo e prazo de implantação. A perspectiva de investimento de dois a três milhões de dólares, e ainda mais importante, com prazos de 12 a 36 meses para se começar a obter resultados, tem levado a um questionamento da propriedade de começar a utilizar esta tecnologia através de um sistema de DD. Um sistema de DD tem necessariamente que contar com a participação de todas as unidades de negócio da empresa, bem como todos os gerentes de sistemas destas mesmas áreas. Isto irá requerer tempo de todas as áreas e encontraremos várias dificuldades como o estabelecimento de padrões, definições comuns, priorização de atendimentos, etc.

As dificuldades encontradas podem inviabilizar a implantação de sistema de DD, dificuldades do tipo:

- baixo grau de padronização e documentação dos sistemas OLTP;

- dados armazenados em sistemas legados;
- inexistência de modelo de dados padrão;
- alto grau de dispersão de dados em sistemas departamentais.

A necessidade de realizar uma "reengenharia" nos sistemas existentes para implantação de um DD demandará recursos, esforço, investimentos, grande período de tempo e um forte patrocínio do alto escalão da empresa para proporcionar esta base mais sólida necessária para implantação de um sistema de DD. É um grande desafio para uma empresa iniciar e manter um processo desta amplitude, consumindo recursos, humanos e monetários, por um longo período de tempo.

A grande vantagem de utilizarmos esta arquitetura é de garantir alta consistência nos dados, já que os dados de toda corporação estarão disponíveis em um único local, com um metadados também único e com um alto grau de padronização na definição de regras de negócio e de atributos.

A segunda arquitetura é a alternativa para este impasse através da confecção de um Data Mart – uma solução rápida, altamente focalizada, atendendo a necessidades específicas de uma área de negócio [HAC 97] – que permite oferecer os dados necessários, utilizando a mesma tecnologia de um sistema DD, sem a infra-estrutura corporativa necessária para um DD de abrangência total. Uma conseqüência do grande número de insucessos obtidos com a arquitetura anterior, que demanda prazos longos e muitos recursos financeiros para implementar.

A urgência de informações pelos usuários de SSD muitas vezes determina a implementação de Data Marts não integrados, isolados e sem arquitetura. Se vários destes Data Marts não integrados são criados para atender esta "urgência" de conhecimento de problemas de negócio, em pouco tempo estaremos na mesma condição dos sistema OLTP, com um conjunto de Data Marts "legados".

A construção de Data Marts não integrados sem o pré-requisito de uma arquitetura corporativa irá inviabilizar a integração dos mesmos e consumir enormes recursos da área de SI (Sistemas de Informação) para mantê-los. Sem uma arquitetura corporativa de Data Mart que defina áreas de interesse, dimensões comuns, métricas comuns, semânticas comuns, regras de negócio comuns e dados de origem comuns, estes Data Marts se tornarão um problema maior que a solução oferecida. A necessidade de realizarmos, para cada Data Mart, um processo próprio de extração, limpeza, e transformação, utilizando recursos computacionais dos sistemas OLTP, demandará cada vez mais recursos da área de manutenção (veja Figura 7). Toda vez que um sistema de origem se altera, o processo EMT de cada Data Mart não integrado deve ser também alterado. À medida que o número destes Data Marts forem

aumentando, a necessidade de integração também surgirá, mas pela sua forma não padronizada encontraremos grandes dificuldades para tal.

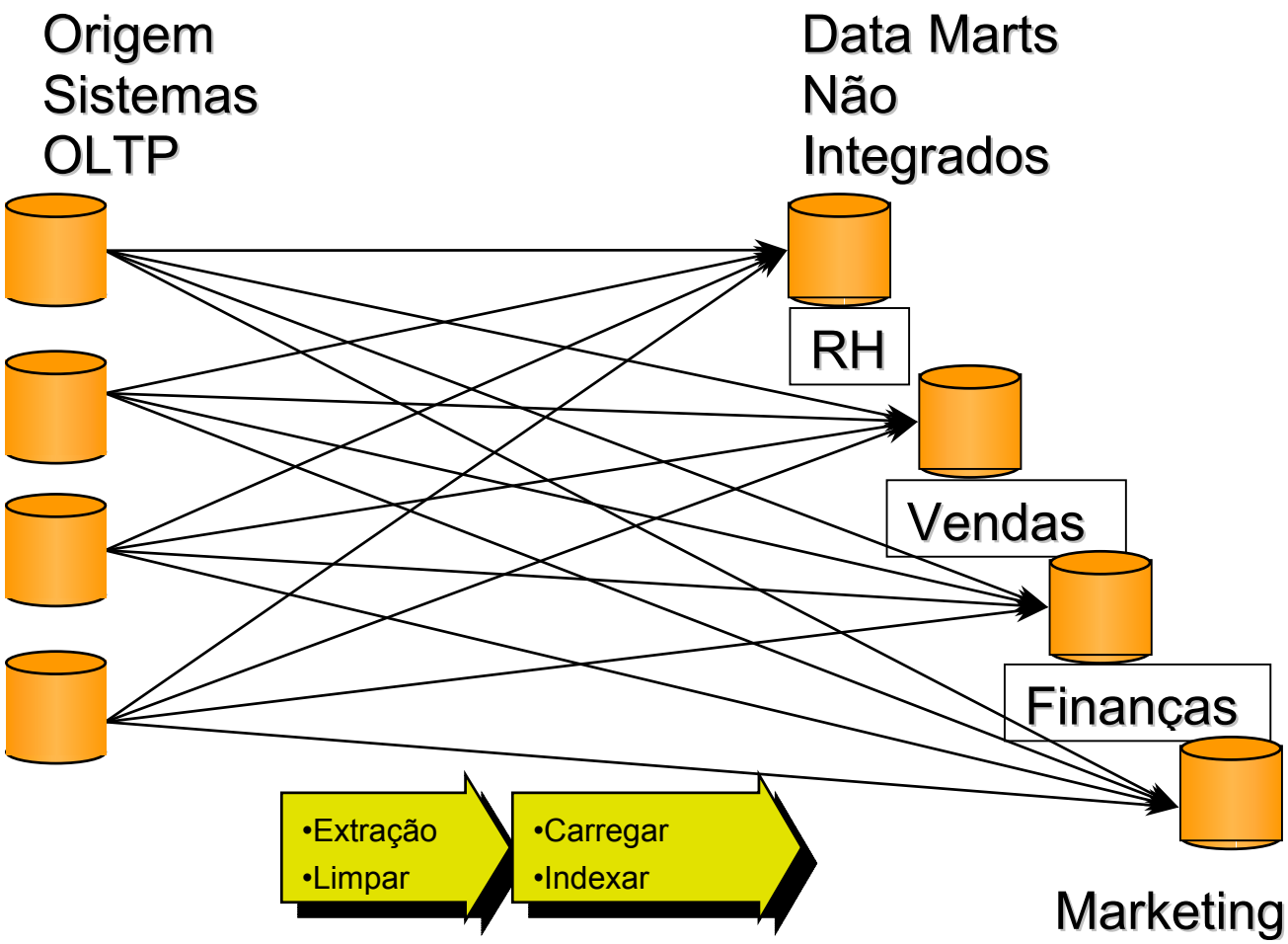

FIGURA 7 – Data Marts não integrados

A terceira arquitetura, a recomendada por este trabalho, é também uma conseqüência de experiências adquiridas com a arquitetura anterior. Nesta, a preocupação da definição de um modelo corporativo dos dados é o primeiro passo para iniciar o processo. Com isso adquire-se uma estrutura que permite que um Data Mart possa ser a semente de um DD corporativo. Começa-se com um Data Mart específico e à medida que outros são criados se estabelece um depósito centralizado destes dados. Esta arquitetura é chamada Data Mart Incremental ou arquitetado (também conhecida como Bottom-Up) (veja Figura 8).

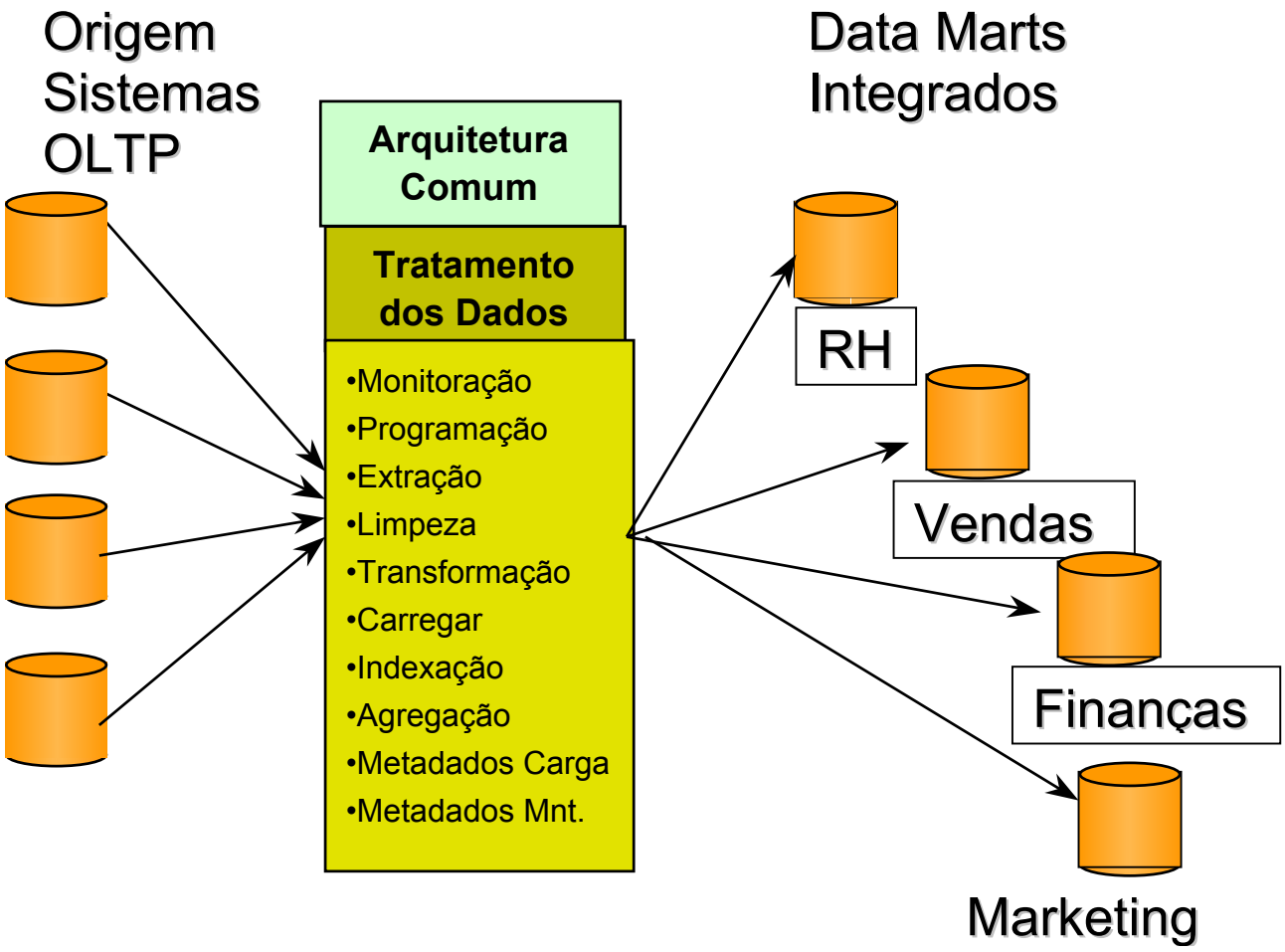

FIGURA 8 – Arquitetura de Data Marts incrementais – Primeira fase

O processo inicia com a criação do primeiro Data Mart, por exemplo, Marketing, e à medida que outras áreas da empresa forem solicitando se cria Data Marts adicionais, mas sempre com uma arquitetura corporativa que vá permitir a integração futura em um Depósito de Dados (veja Figura 9).

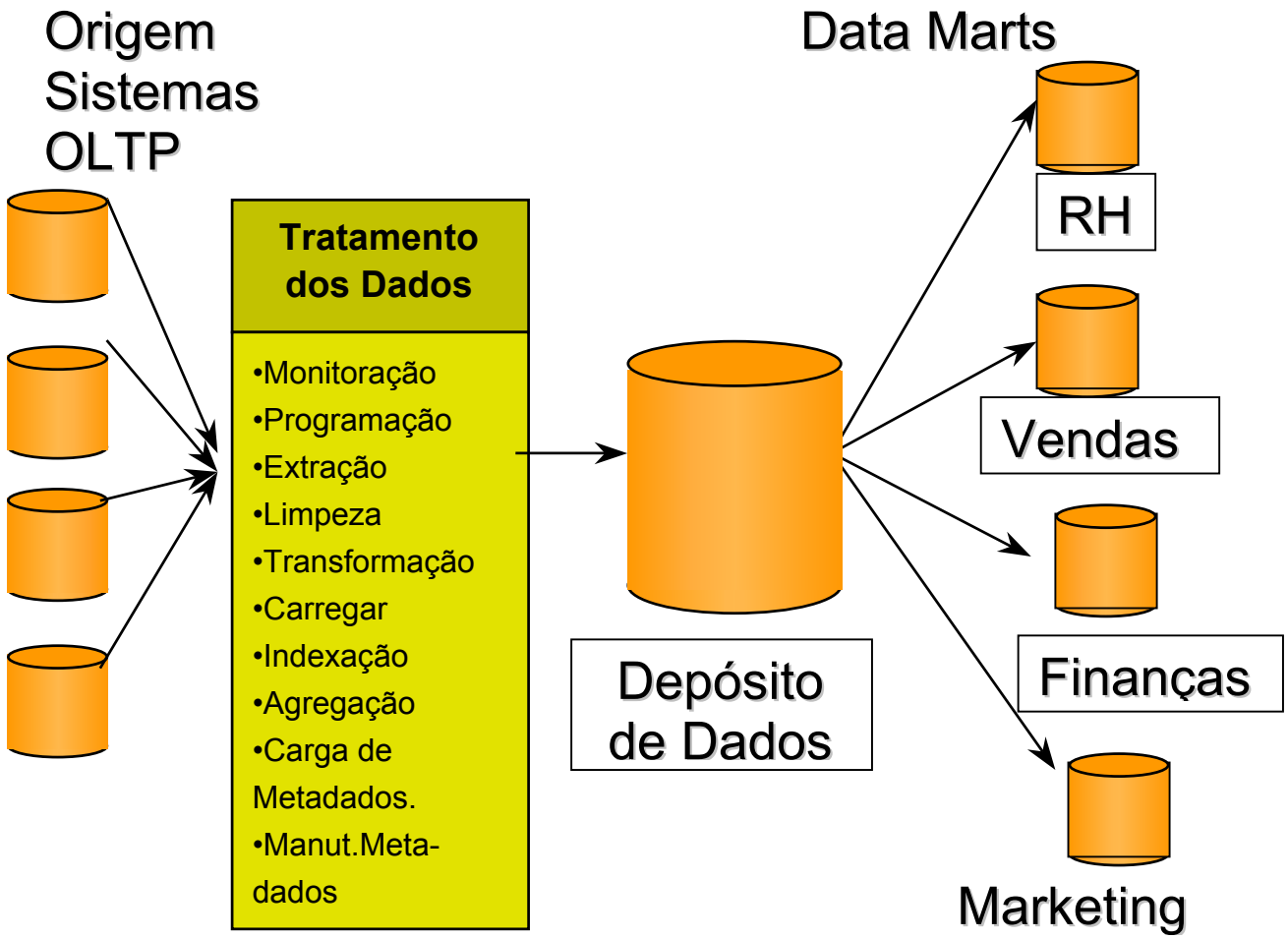

FIGURA 9 – Arquitetura de Data Marts incrementais – Segunda fase

Após a criação de vários Data Marts, o número determinante é decorrente da orientação da empresa e da quantidade de dados transformados. A criação de um Depósito de Dados, sem qualquer descontinuidade no fornecimento de dados aos usuários, é uma conseqüência natural devido principalmente ao volume de dados trabalhados e a necessidade de grande capacidade de armazenamento para séries históricas já obtidas. Um Depósito de Dados permitirá facilmente a criação de Data Marts adicionais e também o acesso direto aos dados sem esforço adicional de tratamento dos dados.

Uma arquitetura corporativa de Data Mart necessita, no mínimo, das seguintes características:

- áreas de interesse corporativas – estas áreas são aquelas que, se um sistema de DD fosse criado, seriam contempladas. Estas definições são realizadas pela alta direção das empresas;

- dimensões comuns – o próximo passo é identificar as dimensões comuns do negócio – a maneira que o usuário gosta de visualizar e analisar suas atividades – para todas as áreas de interesse. Dimensões típicas incluem datas, produtos, cliente, área geográfica, áreas geográficas de venda, promoções e outros. Dimensões comuns permitem o projeto consistente de visões das atividades do negócio em vários Data Marts;
- métricas comuns – em seguida identificamos as métricas utilizadas em cada área de interesse e cada dimensão. Métricas são as maneiras pelas quais a empresa mede suas atividades e processos transacionais. Métricas comuns são reais, unidades, horas, quilos e assim por diante;
- regras do negócio comuns – depois de identificadas as áreas de interesse, dimensões e métricas, temos que estudar a área de interesse para ser criado o Data Mart incremental. Examinamos, em detalhe, as regras de negócio utilizadas para calcular as métricas ou derivar métricas associadas para aquela área de interesse. Neste ponto, devemos documentar as regras de negócio utilizadas para determinar a métrica do Data Mart em construção, mais as métricas similares de outras áreas de interesse. Um exemplo seria a métrica vendas que em Data Mart de marketing seria diferente de uma métrica vendas em um Data Mart de finanças. É muito difícil criar um Data Mart sem utilizar regras de negócio que abranjam toda a corporação, permitindo que as definições de métricas possam ser compreendidas e aceitas por todas as áreas;
- sistema de origem comum dos registros – agora necessitamos identificar qual o sistema-origem de onde os dados serão extraídos. Sem uma fonte comum, estaremos realizando um processo de EMT confuso e errado;
- semânticas comuns – permitem o entendimento e o conhecimento dentro da empresa de definições de termos comuns. Obter um consenso na definição de termos como vendas, retorno de investimento, desconto e lucro líquido é de suma importância para o sucesso de um Data Mart.

Data Mart não é uma idéia nova. Desde do início da utilização de computadores e mais recentemente com o uso de PCs, redes e servidores, foram

criados “Data Marts” não integrados e não padronizados. Antigamente eram chamados sistemas de relatório para usuários ou até de ilhas de dados. Correspondiam a dados retirados de um sistema central e oferecidos ao usuário final.

No conceito atual de Data Mart ele pode atender a milhares de usuários através de múltiplas aplicações, acessando dados resumidos ou detalhados, podendo atingir de poucos gigabytes a centenas de gigabytes.

Data Marts têm as seguintes características:

- arquitetura – Data Marts devem ser construídos numa arquitetura corporativa. No mínimo, esta arquitetura deve identificar as áreas de interesse, dimensões comuns, métricas comuns, semânticas comuns, regras de negócio comuns e dados de origem comuns para organização. Data Marts também são definidos baseados nos princípios e processo de um sistema de DD;
- integrado – Data Marts devem ser integrados com outros Data Marts na organização através de origem, métricas, semânticas, regras de negócio e dimensões comuns;
- escalável – Data Marts devem ser construídos em bases de dados e sistemas transacionais escaláveis;
- usuários homogêneos – Um Data Mart deve atender a um grupo de usuários com interesses comuns, desafios ou outras necessidade homogêneas.

Um projeto de Data Mart que não atenda a estes itens, não será um Data Mart incremental, inviabilizando a sua integração posterior.

3.2 Topologia do Data Mart

- Uma camada – os menores Data Marts são os de uma camada. Neste caso o SGBD reside junto com a própria aplicação em um pequeno servidor. Estes são os casos típicos de ilhas de dados que provocam desinformação na corporação.

- Duas camadas – nesta topologia os Data Marts incrementais residem em um SGBD instalado em um servidor separado dos sistemas OLTP. É comum encontrar esta topologia em ambientes com computador central onde residem os sistemas OLTP (veja figura 10).

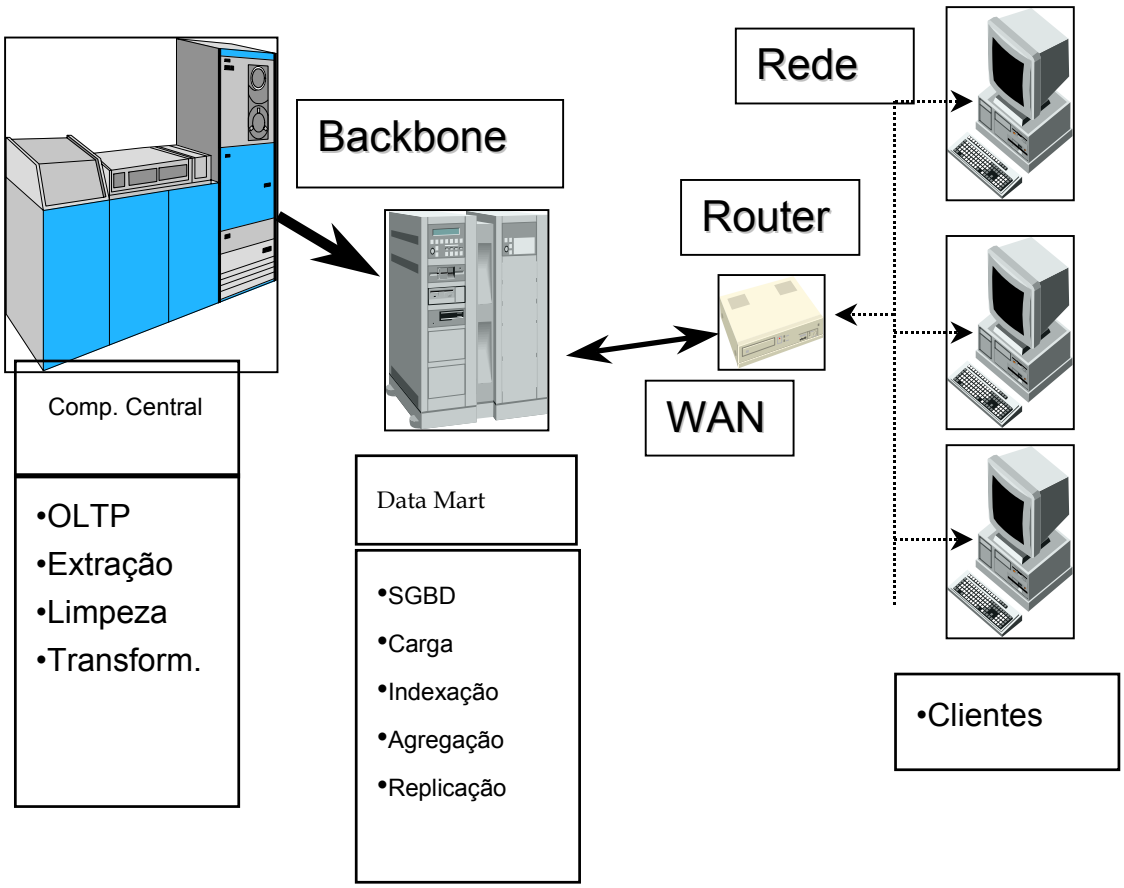

FIGURA 10 – Topologia duas camadas de um Data Mart

- Três camadas – nesta topologia o servidor dedicado de SGBD do Data Mart incremental recebe o auxílio de outro servidor dedicado a ferramentas de extração, limpeza, transformação, carregamento, agregação, replicação, monitoramento, gerência e metadados (veja figura 11).

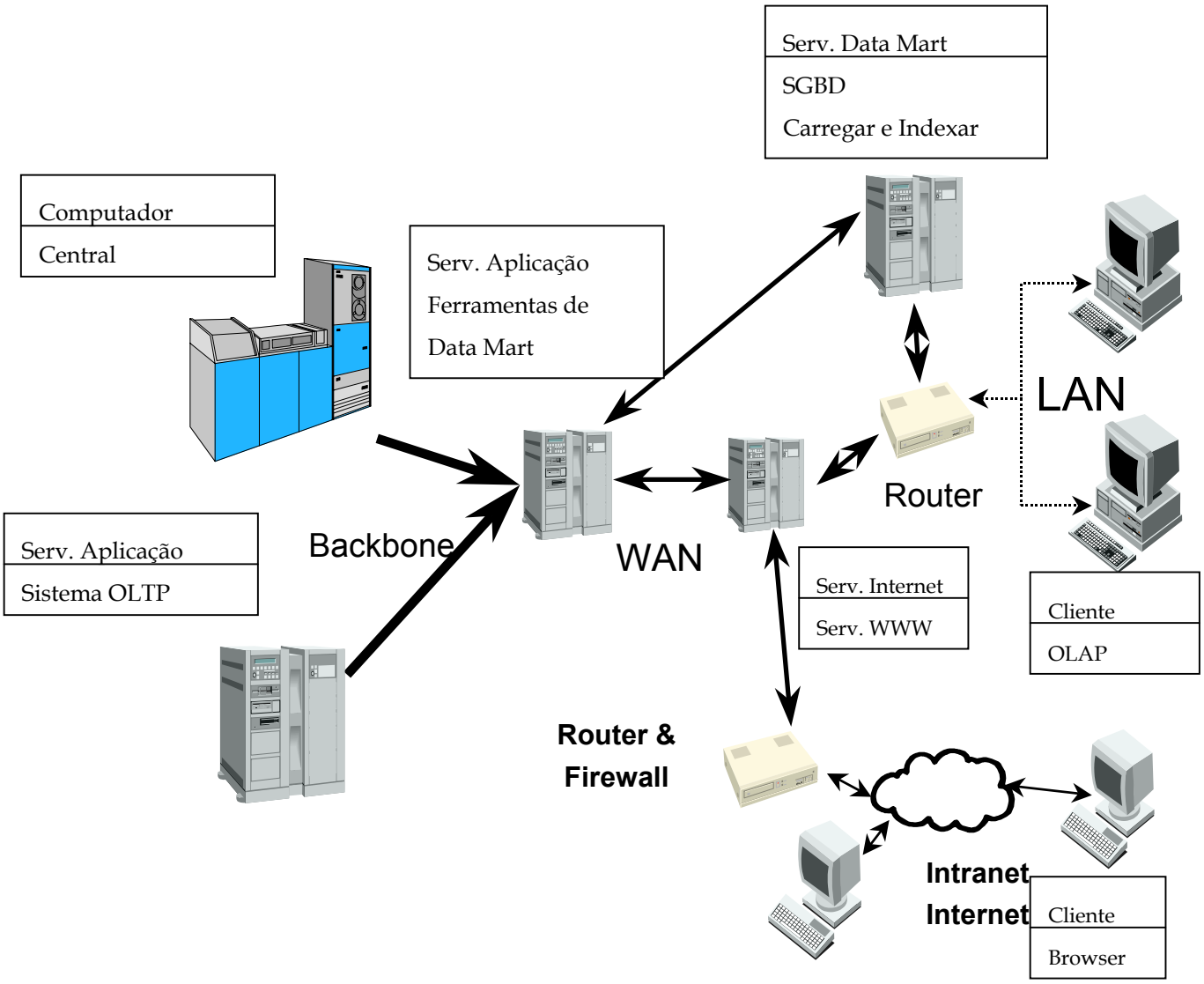

FIGURA 11 – Topologia três camadas de um Data Mart

Devido à situação dos dados contidos nos Data Marts, integrados e limpos, eles são uma fonte natural para aplicações Internet e Intranet. Com as ferramentas existentes para publicação na Internet é possível facilmente publicar informações na intranet para o público interno da empresa. Esta mesma ferramenta de fácil utilização permite disponibilizar a informações, cuidadosamente selecionadas, ao mundo externo da Internet.

A topologia de Internet/Intranet permite que os dados dos Data Marts sejam compartilhados através dos protocolos de baixo nível utilizados pela Internet/Intranet, mas principalmente permite a utilização de clientes universais independentes de plataformas.

Como um dos objetivos de um Data Mart é disponibilizar facilmente as informações para a corporação, com o uso de uma Intranet podemos alcançar este objetivo rapidamente utilizando uma ferramenta poderosa e já conhecida de acesso às informações, o browser (navegador). Por isso foi incluído na figura anterior um servidor Internet/Intranet para disponibilizar as informações por este meio.

3.3 Metodologia

3.3.1 Metodologias analisadas

O objetivo deste trabalho é identificar uma metodologia que permita um alto grau de sucesso em implementações de Depósito de Dados, utilizando uma arquitetura de Data Marts incrementais. A metodologia proposta foi obtida através da análise de [KIM 96], [KIM 98], [Mar 98], [HAC 98], [NCR 99], [VIS 99], e adaptada para melhor atender às demandas de um projeto de DD.

3.3.1.1 Metodologia James Martin

James Martin [Mar 98] propõe uma metodologia para criação de um DD, utilizando a arquitetura baseada em um DD centralizado do qual Data Marts são criados através de *subsets* de dados, com as seguintes atividades:

1. Definir precisamente as informações do negócio que devem estar no DD.
2. Identificar e priorizar as áreas de negócio a serem incluídas no DD.
3. Gerenciar o escopo de cada área de negócio que será implementada no DD de uma forma iterativa.
4. Desenvolver uma arquitetura escalável de forma a ser a fundação do DD e identificar componentes de hardware/software/middleware para implementar.
5. Extrair, limpar, agregar, transformar e validar os dados de forma a garantir consistência e precisão.
6. Definir o nível correto de sumarização que permita o uso de SSD com sucesso.
7. Estabelecer um programa de atualização que seja adequado às necessidades, tempos e ciclos do negócio.
8. Providenciar ferramentas poderosas e amigáveis para o usuário final poder acessar o DD.

9. Educar a comunidade gerencial sobre as reais possibilidades que estão disponíveis com o uso de um DD.
10. Estabelecer um *helpdesk* e treinar usuários para usar eficientemente as ferramentas de acesso ao DD.
11. Estabelecer processo para manutenção, melhorias, e garantias de sucesso na aplicabilidade do DD.

Esta metodologia visa a criação de um Depósito de Dados corporativo centralizado que distribui dados detalhados para a organização, com o uso opcional de Data Marts criados a partir de *subsets* de dados do DD. A utilização de único fornecedor de dados, Depósito de Dados centralizado, não é a melhor forma de iniciar um processo de criação de DD, como já vimos anteriormente.

3.3.1.2 Metodologia Alan Simon

Já Alan Simon [SIM 98] propõe a criação de Data Marts no menor prazo possível (90 dias), sem se preocupar com a integração futura dos mesmos, ou seja, a criação de Data Marts não integrados, através dos seguintes passos:

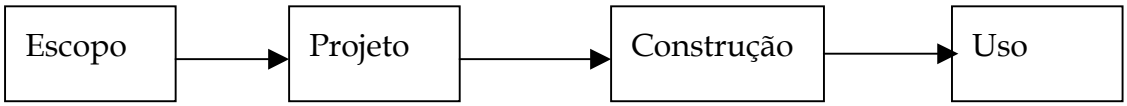

FIGURA 12 - Metodologia Alan Simon

Na etapa Escopo criar e validar a necessidade da área solicitante; Explorar, avaliar e decidir sobre as fronteiras do Data Mart e conduzir explorações preliminares nas tecnologias candidatas para o Data Mart, particularmente ferramentas de usuário final e "middleware" e também o SGBD que o Data Mart utilizar.

Na etapa Projeto será realizada uma série de tarefas paralelas em várias categorias: formalização e finalização de funcionalidades do negócio, que serão obtidas através do Data Mart, análise de dados, modelagem e projeto, análise de processos, modelagem e projeto, condução de pesquisa de sistemas e infra-estrutura, avaliação e seleção de produtos e desenvolvimento de um projeto detalhado para a etapa seguinte de construção.

Na etapa Construção, todas as etapas clássicas de um DD são realizadas – como mapeamento de origem-para-destino, dados físicos, desenvolvimento de transformações, criação de metadados, testes e

refinamentos. Adicionalmente, atividades como treinamento de usuários e preparação de infra-estrutura devem ocorrer simultaneamente, convergindo para o uso do Data Mart.

Na etapa Uso e Administração, o final da etapa anterior já disponibiliza o Data Mart para uso, mas é necessário preparar as ferramentas de acesso para os usuários finais e manter um cronograma de carga de dados para períodos posteriores.

Este método já pressupõe que exista uma equipe técnica treinada e que todos os dados externos já estejam contratados e disponíveis quando do início do projeto.

Este método de criar Data Mart visa somente o menor prazo de entrega para uso do Data Mart, sem realizar análises de requisitos mais profundas nem uma preocupação com o aproveitamento futuro desta implementação, portanto inviabilizando a sua integração com outros Data Marts para um futuro DD.

3.3.1.3 Metodologia Ralph Kimball

Outro enfoque é dado por Ralph Kimball [KIM 98], conforme descrito a seguir:

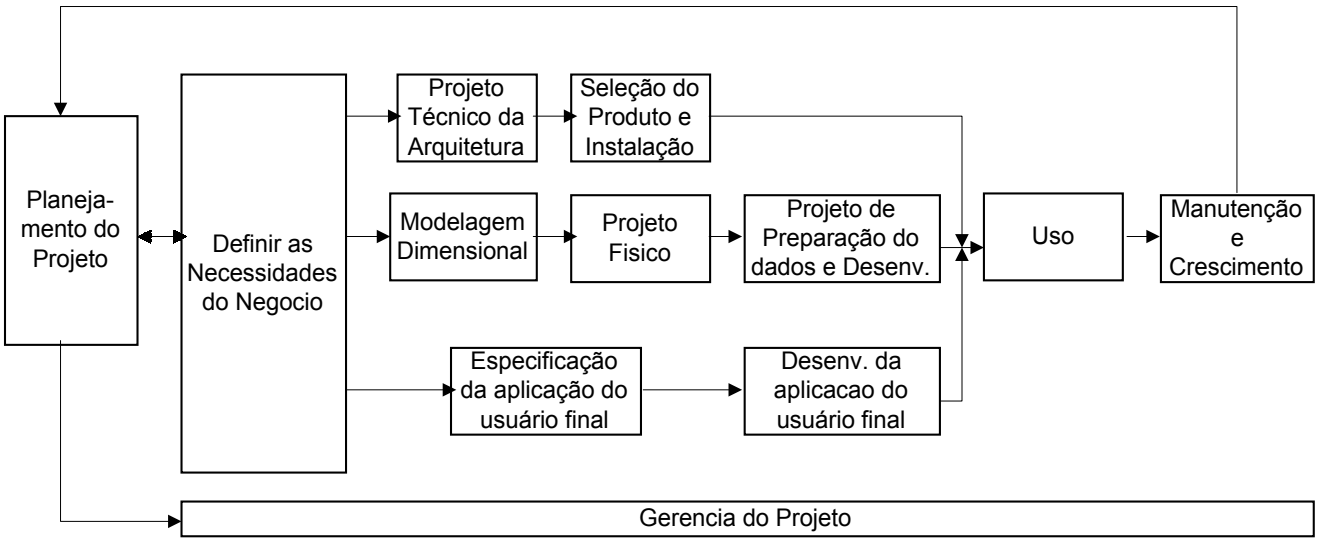

FIGURA 13 – Metodologia Ralph Kimball

Este diagrama apresenta a seqüência de tarefas necessárias para projeto, desenvolvimento e uso de um DD. O diagrama mostra um caminho a ser seguido para se alcançar o sucesso na implementação.

Planejamento do projeto

A metodologia inicia com o planejamento do projeto. O planejamento do projeto abrange a definição e escopo de um projeto de DD, incluindo o prazo de execução e a justificativa para implementação. A partir daí, o planejamento segue para as necessidades de recursos e qualificação da equipe técnica, junto com a designação de tarefas de projeto, duração, e seqüência. O plano do projeto resultante identificará todas as tarefas associadas com o projeto e todas as áreas envolvidas. Este passo é a pedra fundamental para gerência do projeto de DD. Planejamento do projeto é dependente das definições determinadas pelos usuários do DD, como mostra a seta dupla na conexão com a atividade Definir necessidades.

Definir necessidades

O sucesso de uma implementação de DD está ligado diretamente ao conhecimento e entendimento das necessidades dos usuários requerentes do DD. Sem um entendimento claro destas necessidades, o projeto de DD será apenas um exercício para a equipe de projeto. O enfoque necessário para obter o conhecimento das necessidades analíticas dos usuários é bastante diferente dos requeridos por sistemas tradicionais orientados à transação. Os projetistas do DD têm que entender os fatores primordiais que governam o negócio (a solicitação) para poder determinar os pontos críticos a serem atendidos e traduzi-los da melhor forma no projeto do DD. As definições de necessidades determinam três tarefas paralelas dirigidas à tecnologia, dados, e aplicações de usuário final.

Modelagem dimensional

A definição das necessidades requeridas pelo negócio determinam os dados a serem utilizados para resolver os problemas analíticos dos usuários. Desenvolver modelos de dados para atender a estas análises requerem um enfoque diferente do utilizado por sistemas transacionais. A partir daí, se conduz uma análise mais detalhada dos sistemas-origem transacionais. Acoplando esta análise com o entendimento das necessidades, se desenvolve um modelo dimensional. Este modelo identifica a granularidade da tabela de fatos, dimensões associadas, atributos, hierarquia e fatos. O projeto lógico da base de dados estará completo com as estruturas das tabelas e relacionamentos/chaves primárias e estrangeiras. Um plano de agregação preliminar também deve ser desenvolvido. Este conjunto de atividades identifica o mapeamento origem-destino do DD.

Projeto físico

O projeto físico da base de dados tem como objetivo definir as estruturas físicas necessárias para permitir o projeto lógico da base de dados. Os elementos primários deste processo incluem definição de padrões de nomes e o ambiente da base de dados. Indexação preliminar e estratégias de particionamento também são determinadas.

Projeto e desenvolvimento de tratamento dos dados

O projeto de tratamento dos dados é a tarefa a que normalmente não é dada a atenção devida. O processo de tratamento dos dados tem três passos: extração, transformação e carga. O processo de extração sempre expõe problemas de qualidade dos dados que estão enterrados há muito tempo nos sistemas transacionais. Como a qualidade dos dados impacta diretamente na credibilidade do DD, é necessário resolver os problemas de qualidade nesta etapa. Para complicar ainda mais, é necessário projetar dois processos de tratamento de dados, um para a carga inicial e outro para as cargas incrementais.

Projeto da arquitetura técnica

O ambiente de um DD necessita da integração de inúmeras tecnologias. O projeto de arquitetura técnica estabelece a arquitetura geral e visão. Três fatores são requeridos – necessidades identificadas, ambiente técnico atual e a direção planejada de evolução técnica – para simultaneamente determinar o projeto da arquitetura técnica do DD.

Seleção de produtos e instalação

Utilizando o projeto de arquitetura técnica como a estrutura básica, componentes específicos como plataforma de equipamentos, base de dados, ferramenta de tratamento dos dados e ferramentas de acesso aos dados devem ser avaliados e escolhidos. Um processo de avaliação técnica padrão para cada componente deve ser definido para permitir determinar a melhor opção do mercado. Depois de selecionados, os produtos todos devem ser instalados para verificar a compatibilidade com todo ambiente do DD.

Especificação da aplicação do usuário final

Nesta etapa define-se um padrão de aplicações a serem disponibilizadas aos usuários finais, desde ferramentas geradoras de relatórios a ferramentas que permitem *ad hoc*. Estas definições devem ser feitas de uma forma bastante próxima entre a área de SI e áreas usuárias.

Desenvolvimento de aplicações para usuário final

Após a especificação das ferramentas, é necessário o desenvolvimento das aplicações para o usuário final que compreende a configuração da ferramenta com acesso e criação de metadados e relatórios especificados. Normalmente, estas ferramentas apresentam facilidades de configuração e criação, de forma a minimizar o tempo da área de SI e área usuária.

Uso

O uso representa a convergência da tecnologia, dados e aplicações de usuário final acessível de qualquer mesa de usuários do negócio. Planejamento adequado é necessário para que todas estas partes se encaixem apropriadamente. A educação dos usuários deve ser desenvolvida e aplicada. Além disso, suporte ao usuário e estratégias de comunicação devem ser estabelecidos antes que os usuários tenham acesso liberado ao DD.

Manutenção e crescimento

Muito trabalho ainda é necessário após o início do uso do DD. É necessário continuar apoiando os usuários finais com suporte e treinamento. É necessário também atenção na área de operação do dia-a-dia. O uso de métricas de aceitação e desempenho deve ser utilizado para avaliar o sucesso do DD.

Neste momento, em que o DD já está em uso, imediatamente surgirão necessidades adicionais. Alterações devem ser sempre bem-vindas como um sinal de sucesso e não de insucesso. Deve ser definida uma priorização para estas demandas, e vamos ao início do ciclo de vida do DD planejando estas alterações, levando em conta as definições existentes e o ambiente do DD.

Gerência do projeto

A gerência do projeto garante que as atividades que apóiam o ciclo de vida de um DD estejam em sincronismo e na linha. A gerência como apresentada no diagrama ocorre durante todo o processo. Estas atividades visam a monitoração do projeto, acompanhamento de etapas e controle de mudanças para preservar o escopo determinado. E finalmente, gerência de projeto inclui um desenvolvimento de um plano de comunicação que abrange tanto as áreas de negócio como a área de SI. Comunicações constantes são críticas para manter as expectativas das gerências informadas e assim permitir se atingir os objetivos traçados para o DD.

3.3.1.4 Metodologia Douglas Hackney

Douglas Hackney [HAC 98] apresenta uma metodologia em 19 etapas:

1. Identifique o objetivo do negócio.
2. Identifique e quantifique a vontade política de realização.
3. Identifique o patrocinador do projeto.
4. Analise a disponibilidade de execução.
5. Analise as necessidades dos usuários finais.
6. Construa uma arquitetura corporativa de Data Mart.
7. Escolha um Data Mart inicial.
8. Pesquise e implemente ferramentas de DD/Data Mart.
9. Faça o projeto da base de dados objetivo.
10. Construa o Mapeamento, extração, transformação e limpeza.
11. Construa agregação, replicação e distribuição.
12. Pesquise e implemente ferramentas de usuário final.
13. Teste.
14. Treine.
15. Realize o piloto.
16. Desenvolva.
17. Use.
18. Monitore.
19. Mantenha.

3.3.1.5 Metodologia NCR – NCR Corporation [NCR 99]

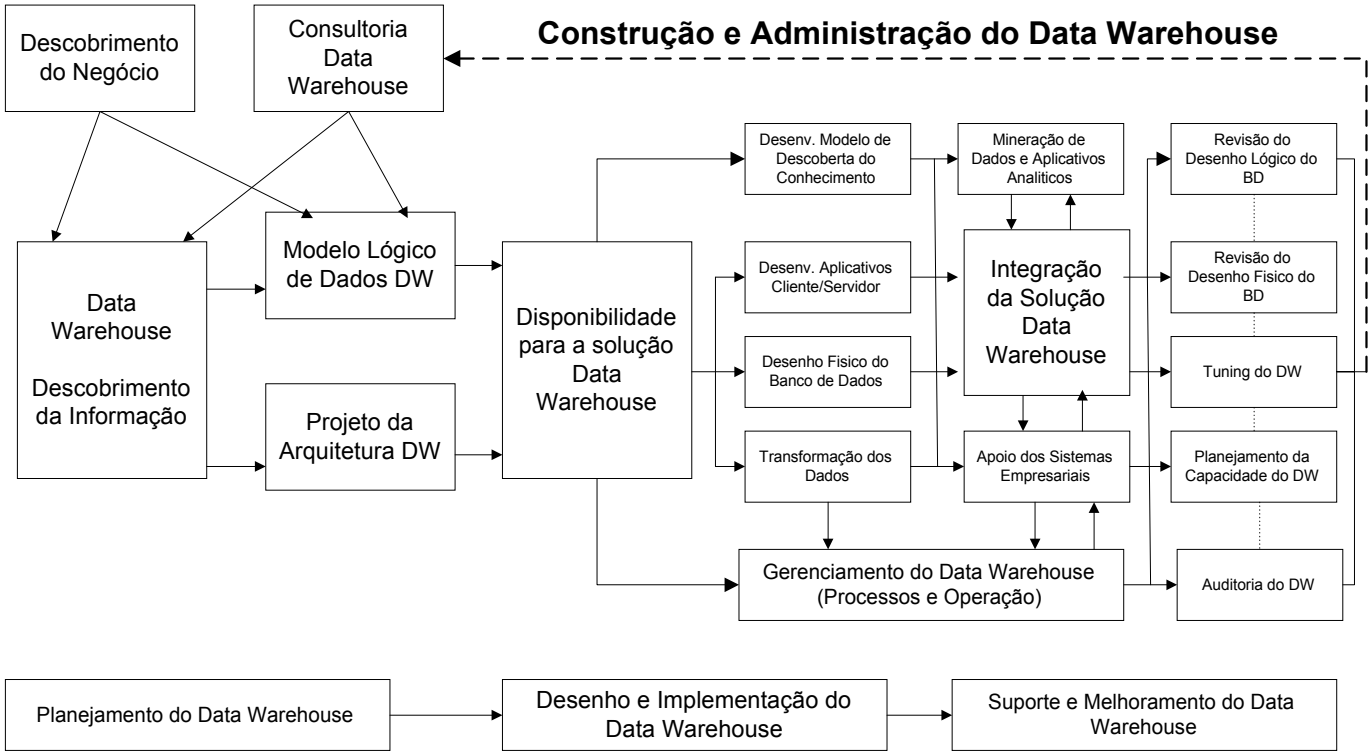

FIGURA 14 - Metodologia NCR

Etapa 1

Planejamento do Data Warehouse

- Descobrimento do Negócio (necessidades) – Determina os problemas de negócios, critérios de êxito, retorno do investimento, regras do negócio, usuários do negócio e Sistemas de Informação, requerimentos dos dados, obtém consenso nas questões chaves do negócio, prioriza as questões e identifica benefícios quantitativos, e estimula a cooperação entre as linhas funcionais do negócio.

- Consultoria de DW – Utiliza o conhecimento de especialistas experientes para analisar o complexo ambiente de DW; examinar o DW de um ponto de vista de negócio e operacional; explorar os benefícios e questões do desenho e implementação de um DW; planejamento, construção, implementação e manutenção de um DW.

- Descobrimento da Informação – Refinamento dos requerimentos de uma solução ao validar e focar as necessidades críticas do negócio; determina como os dados podem ser convertidos em informação para atender a necessidades

críticas do negócio; obter um entendimento dos benefícios do acesso e utilização da informação dos negócios; demonstrar o valor dos dados detalhados, e como podem responder às necessidades críticas do negócio.

- Desenho do Modelo Lógico de Dados – Modelo lógico é crítico para uma solução DW; assegura que as perguntas de negócio sejam respondidas; assegura flexibilidade para suportar melhorias futuras; entregar um modelo com todos os atributos; entregar uma identificação das fontes de dados e documentação de regras de negócio.

- Desenho da Arquitetura de DW – determinar sobre: a localização do DW – centralizado, distribuído, Data Mart; ferramentas para aplicações Cliente/Servidor, métodos de acesso, carga da aplicação, e tipos necessários de aplicação; fontes e integridade dos dados, regras de negócio e relações de dados; administração do DW-hardware, software, rede, suporte, restauração, arquivo e requerimentos de atualização; necessidades e oportunidades do modelo matemático.

Etapa 2

Desenho e implementação do DW

- Disponibilidade para a Solução – Identificar os possíveis problemas na organização que podem impedir uma implementação bem-sucedida avaliando os dados, plataforma tecnológica, aspectos funcionais, educacionais, mudanças necessárias e equipe de suporte ao DW.

- Desenho Físico do Banco de Dados – Traduz o modelo lógico em um desenho otimizado do BD; fornece a construção e provas funcionais do BD.

- Transformação dos Dados – Prepara a localização, extração, acondicionamento, limpeza e carga dos dados; define os processos e atividades requeridas para carregar os dados e manter o DW; fornece um plano operacional para recarga, carga incremental em uma base periódica.

- Desenvolvimento de aplicações C/S – Utiliza os protótipos já definidos conforme a indústria em questão, utiliza ferramentas de desenvolvimento já determinadas para resolver problemas identificados na etapa um; implementa aplicações e/ou interfaces para consulta; fornece treinamento e documentação para assegurar o aproveitamento rápido e completo dos benefícios pelos usuários.

- Descoberta do Conhecimento/Mineração dos Dados – Assistir na definição do problema; identificar as técnicas de análise apropriadas; instruir na utilização das ferramentas e no processo de resolução do problema; criar e modelar os formatos dos relatórios finais; converter dados em informação relevante, competitiva e estratégica; ajudar a perceber os benefícios do marketing "um-para-um".

- Administração do DW – Identificar os procedimentos de carga, manutenção, arquivo e recuperação de dados para assegurar consistência e compatibilidade; implementar os procedimentos de administração de dados, rede, sistemas e operação; implementar os procedimentos para controle de versões, manutenção de hw e sw, relatório de erros e etc.

Etapa 3

Suporte e melhoramento do DW

- Revisão do Desenho Lógico e Físico – Revisão comparativa do modelo lógico contra os requerimentos do usuário com sugestões para melhorar; revisão do modelo físico a partir de sugestões dos usuários.

- Afinação do DW – Identificar problemas relacionados com o desempenho; identificar origens dos problemas reais ou percebidos; fornecer uma análise detalhada da rede, usuários, aplicações, estruturas do BD, utilização de sistema; criar um plano de ação baseado nos resultados das análises.

- Auditoria do DW - Quantificar o retorno de investimento do DW existente, ou projeto de expansão; justificar o investimento, os dados e o treinamento; valorizar os ambientes existentes de DW para as áreas de negócios, operacionais e da administração; comparar um DW existente contra as melhores práticas na indústria; descobrir métodos para melhorar o Retorno de Investimento (RI).

3.3.1.6 Metodologia Visible – Visible Corp.

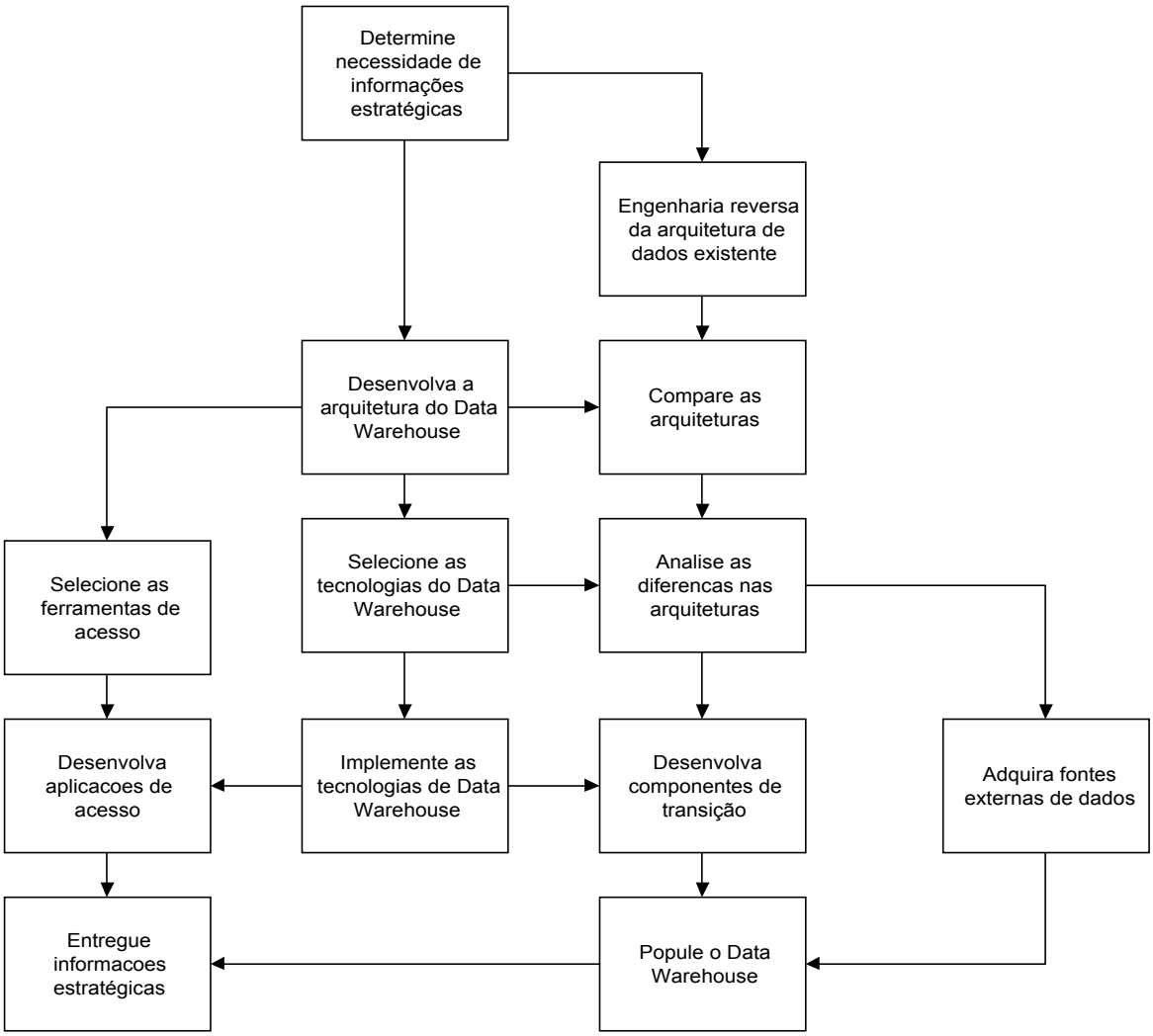

FIGURA 15 - Metodologia Visible

A empresa Visible [VIS 99] apresenta uma metodologia baseada em cinco atividades principais:

1. Estabelecer um patrocinador;
2. Identificar as necessidades da empresa;
3. Projetar a arquitetura do DD;
4. Aplicar tecnologia adequada; e
5. Implementar o DD.

Estabelecer um patrocinador

O primeiro passo é estabelecer um patrocinador para o DD. Estabelecer uma rede de patrocinadores correta auxiliará no sucesso do desenvolvimento e implementação. A rede de patrocinadores deve conter um gerente de DD e duas outras figuras-chave. No topo da rede teremos um executivo com recursos para investir na estrutura de informação. Um gerente de projeto que ficará entre o executivo e o gerente de DD para gerenciar e manter o projeto andando dentro do cronograma.

Um aspecto importante no estabelecimento do patrocinador é assegurar que toda a empresa compreenda o objetivo de um DD, seu potencial, e o plano corporativo de implementação. O plano de engenharia do DD deve ser desenvolvido compreendendo todas as etapas necessárias.

Identificar as necessidades da empresa

Identificar as necessidades da empresa é um componente importante no ciclo de vida para qualquer sistema de informação, e é crucial na implementação de um DD. Quando se desenvolve sistema transacional, normalmente somente teremos um único patrocinador ou grupo de usuários com uma visão clara do que necessitam, como o sistema deve parecer, como deve funcionar. Mas no desenvolvimento de um DD, sempre teremos múltiplos usuários em potencial, cada um com diferentes idéias de como o DD deve ser e o que deve prover. Devido à falta de um foco único, identificar com precisão a necessidade da empresa é crítico para o sucesso de um projeto de DD.

Para alcançar este propósito, é necessário analisar os planos de negócio da empresa e as definições para os sistemas de informação que irão formar a base para o projeto. Além disso, é necessário entrevistar os gerentes-chave do projeto e examinar a documentação pertinente para determinar estas necessidades.

Também é necessário entrevistar indivíduos nas áreas atingidas para complementar um primeiro relatório sobre o andamento do projeto.

Com o início do projeto, iniciam-se reuniões com os potenciais usuários do DD para realizar um refinamento das necessidades já levantadas.

<u>Determinar o ciclo de avaliação</u>

Definir o período de avaliação utilizado pela empresa inclui descrever os ciclos ou períodos de tempo em que as avaliações/resultados são medidos. Trimestres, meses ou horas são apropriados para capturar os dados a serem utilizados nas avaliações? Qual o período histórico que deve ser armazenado? Estas perguntas têm várias respostas dentro de uma empresa. Uma empresa de seguros necessita de décadas de dados atuariais para avaliação. Uma empresa de telecomunicações, por outro lado, necessita de medidas por hora e pode manter históricos de apenas algumas semanas ou meses.

Validar o ciclo de avaliação

Após identificar e definir as necessidades da empresa, é de grande valia comunicar a toda a empresa estas definições. Um dos ótimos subprodutos da criação de um DD é estabelecer com todas as áreas da empresa uma forma única de tratar os dados e suas medidas.

Resolver conflitos de dados

Um modelo bem definido de DD não pode conter homônimos, sinônimos, e outros conflitos de dados. A razão destes conflitos de dados existirem deve-se à existência de significados diferentes para termos de grande uso na empresa, e um deles é o cliente. Para área de finanças o cliente é a organização ou indivíduo que recebe a conta. Para área de vendas pode ser o contato que a empresa mantém. Para área de engenharia clientes seriam outras áreas da própria empresa. Para área de telecomunicações um termo controverso é o de central telefônica.

Construir um modelo empresarial

As entidades de um DD são aquelas que, a qualquer momento no tempo, indicam aos usuários do DD o desempenho da empresa. Definir de forma unívoca e clara todas as entidades do DD, descrever a forma que é utilizada e também definir fórmulas derivativas, agregações e períodos de tempo, são atividades críticas para capturar um entendimento claro das medidas utilizadas pela empresa. O resultado é um modelo de arquitetura empresarial, que apresenta as necessidades da empresa como as entidades do DD e suas regras, estabelecendo uma documentação e fonte de comunicação do conteúdo do DD (metadados).

A Visible propõe a utilização de "Universal Model", uma série de modelos de dados de negócio pré-definidos para iniciar este processo.

Projeto de arquitetura de DD

Após a definição e documentação das necessidades da empresa se inicia o projeto de arquitetura de DD. Este processo envolve a participação dos usuários em sessões especiais para auxiliar nesta atividade.

É neste momento que se cria o metadados, dividido em dois tipos diferentes:

1. Metadados estrutural – é utilizado para criar e manter o DD. Ele descreve toda a estrutura e conteúdo do DD. O bloco básico deste metadados é o modelo que descreve as entidades, as suas características, e como elas se relacionam entre si, e quais as entidades que estão agregadas. Também identifica a lógica de integração e transformação para mover os dados-origem para o

DD. E por último define o cronograma de atualização e arquivo para cada entidade. Metadados estrutural inclui também métricas de desempenho para programas e consultas de forma que usuários e desenvolvedores saibam o tempo de resposta. O DBA do DD também utiliza estas informações para afinar o SGBD.

2. Metadados de acesso – é uma ligação dinâmica entre o DD e a aplicação do usuário final. Geralmente contém as medidas empresariais e o dicionário de termos contidos no DD. Também incluem a localização e descrição dos servidores do DD, bases de dados, tabelas, dados detalhados, e sumarizações junto com as descrições dos dados originais. Prove também regras para *drill up*, drill down e visões sobre hierarquias de assuntos como produtos, mercado e clientes. Além disso, contém todos aspectos de segurança.

Uma parte crítica na definição da arquitetura do DD é o momento de realizar a engenharia reversa e armazenar as definições dos dados originais e destinos no metadados.

Identificar os registros dos sistemas transacionais (OLTP)

A definição da arquitetura do DD também envolve a identificação correta dos dados-origem transacionais que irão popular o DD. Este esforço também determina a necessidade de integrações e transformações. Identificar os registros-origem é uma forma de validar as medidas empresariais resultantes das entrevistas realizadas.

Aplicar a tecnologia correta

Somente depois de definir as necessidades da empresa e projetar a arquitetura do DD, deve a empresa começar a selecionar a tecnologia do DD. A chave para determinar a tecnologia correta, além de determinar a plataforma de hardware e software, inclui desenvolver/adquirir os programas para carga dos dados no DD, implementar controle de acesso e selecionar uma ou mais ferramentas de acesso para os usuários.

Determinar a plataforma de hw/sw

Algumas considerações sobre a determinação da plataforma de hw e sw:

- Qual a quantidade de dados que será colocado no DD e quanto que a plataforma pode acomodar economicamente? Qual a escalabilidade da plataforma? É otimizada para DD? Aceita o sw escolhido para o DD?
- No sw as escolhas começam com o sistema operacional, sw de desenvolvimento, e sistemas gerenciadores de base de dados. A

estrutura e o tamanho do DD determinaram algumas destas necessidades.

Desenvolver programas de integração e transformação

Integração e transformação são programas necessários para extrair dados transacionais e realizar a carga inicial e cargas subseqüentes.

- Um programa para carga inicial é necessário quando o volume de dados inicial é tão grande que não pode ser transferido sem impactar no OLTP origem.
- Programa separado para carga de dados históricos é também interessante, pois só acontecerá uma vez, e normalmente requer o uso de dados arquivados em fitas.
- E um programa adicional para cargas de atualizações.

Segurança

O DD é uma fonte somente de leitura para a empresa, portanto não é necessário aos desenvolvedores se preocuparem com controlar acessos e direitos de alteração. Mas o acesso a informações estratégicas deve ser protegido de uso não autorizado.

Interface de usuários

Os critérios mais importantes na definição para escolha de interfaces de usuário são as informações necessárias e o nível de conhecimento informático do usuário. Ferramentas mais simples para usuários que necessitam de dados extremamente sumarizados e ferramentas mais complexas para usuários que necessitam de dados mais detalhados. E um critério final seria o suporte a acesso a metadados pela ferramenta/interface escolhida.

Implementar o DD

Um programa de implementação do DD inclui a carga de dados preliminares, implementar as extrações e transformações, customizar as ferramentas de acesso, desenvolver um padrão de consultas e relatórios e um treinamento completo dos usuários.

3.3.2 Metodologia proposta

Para realização da metodologia proposta é importante que se utilize um método adequado para esta tecnologia. Na figura 16 vemos o método em espiral para desenvolvimento de um DD.

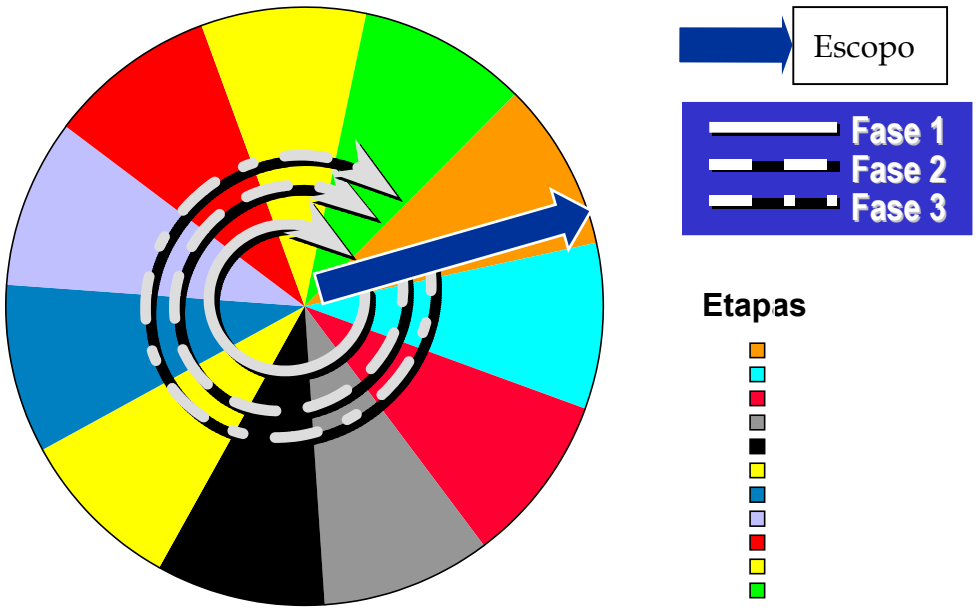

FIGURA 16 – Método espiral de desenvolvimento

Os métodos tradicionais, seqüenciais, não são apropriados para a criação de um Depósito de Dados devido à necessidade de interatividade no seu desenvolvimento. Durante o ciclo de vida de um DD acontecem constantes revisões das necessidades dos usuários, com ampliações de escopo a ser alcançado, demandando a aplicação da metodologia para cada nova fase a ser implementada. Estas alterações já são reconhecidas como naturais num projeto deste tipo, e com o método espiral atende a esta característica.

Para facilitar a implementação, esta metodologia foi dividida em 12 etapas, que são as seguintes:

1. Identificar a origem da solicitação, um patrocinador, riscos e taxa de retorno.

   A construção de um Data Mart somente deve ser feita para resolver um problema grave da corporação, e este problema deve ser específico e com escopo determinável. Como por exemplo: a empresa está perdendo clientes para os concorrentes ou promoções falham por razões desconhecidas.

   Um patrocinador de alta posição deve adotar o projeto para que não haja descontinuidade no processo.

2. Avaliar as necessidades dos usuários e identificar as funcionalidades desejadas.

   As seguintes perguntas devem ser respondidas na avaliação de necessidades:

   - Qual a missão e responsabilidade desta unidade de negócio?
   - Quais são os desafios desta unidade de negócio?
   - Quais os sistemas OLTP e SSD que estão em uso?
   - Quais são especificamente os problemas com o sistema atual de SSD?
   - Que tipo de usuário deve ser atendido, para simples consulta, para emissão de relatórios, para analistas financeiros e para executivos?
   - Quais são as regras de negócios e semânticas para esta área de negócio? Elas são consistentes com todas as outras áreas?
   - Definir regras de negócios comuns, semânticas, definições, fatos, métricas, dimensões e atributos.

   As seguintes perguntas devem ser respondidas na identificação das funcionalidades:

   - Assunto e área de negócio do Data Mart
   - Sistemas-origem (De onde vem o dado detalhado?)
   - Granularidade dos fatos (Qual o nível de detalhe necessário?)
   - Por quanto tempo temos que armazenar dados históricos e que nível de detalhe?
   - Quais são as dimensões e atributos desta área de negócio? (Quais são as colunas em um relatório?)
   - Necessidades de agregação multidimensional (Qual a combinação de visões são solicitadas pelos usuários?)
   - Tabela histórica (O usuário necessita acompanhar as mudanças em dimensões e atributos?)
   - Identificar os fatos, métricas, dimensões e atributos.

3. Projetar uma arquitetura corporativa de Depósito de Dados de longo prazo no papel.

   Definir a arquitetura do DD em um a dois dias de trabalho

   Projetar todos os componentes de uma aplicação DD:

   - Base de dados de origem
   - Ferramentas de extração, limpeza, transformação e carga
   - Repositório central de metadados

- Ferramentas de administração e manutenção de um DD
- Ferramenta de modelagem
- DD para dados detalhados
- Data Marts incrementais múltiplos
- SGBD destino
- Ferramentas de acesso e análise para usuários finais
- Integração de componentes de DD com metadados central

4. Definir necessidades funcionais para o projeto inicial.
   - Identificar área inicial do negócio a ser utilizada
   - Conduzir reuniões de planejamento de necessidades para definir funcionalidades, tarefas, fases, plano do projeto, projeto piloto, programação, orçamento, recursos, conhecimento etc.
   - Objetivo das reuniões é obter um consenso sobre a abrangência do projeto
   - Após término do item anterior, iniciar imediatamente a implementação do Data Mart inicial

5. Pesquisar e selecionar componentes de DD e ferramentas.

   Componentes a serem selecionados:
   - Ferramentas de extração, transformação e carga
   - Arquitetura de metadados que permita a integração entre metadados de ferramentas diferentes
   - Base de dados destino (relacional, multidimensional e híbridas)
   - Ferramentas de acesso e análise para usuário final
   - Servidores de base de dados, sistemas operacionais e redes (LAN, WAN).

6. Projetar a base de dados destino.

   Projetar modelo lógico de dados para base de dados destino:
   - Dados de fatos, dimensões e atributos definidos na etapa 2
   - Projetar modelo lógico de dados para todas as entidades destino
   - Modelos de dados podem ser importados ou desenhados com componentes de modelagem de dados

Projetar modelo físico de dados através da denormalização do modelo lógico

- Esquema estrela para bases relacionais
- Matriz multidimensional para base de dados multidimensionais

7. Construir o mapeamento de dados, extração, transformação e regras de limpeza.

   Utilize relatórios atuais para definir necessidades iniciais de entidades

   Identifique dados de origem para cada entidade definida e mapeei dados de origem para dimensões e atributos

   Defina regras de limpeza dos dados

   Defina regras de extração e conversão

   - Alteração de nomes, mudanças de chaves
   - Mudança física de atributos
   - Filtros, padrões, tabelas de referências
   - Seleções de múltipla escolha

8. Construir agregações, sumarizações, particionamento e distribuição.

   Identificar agregações mais utilizadas

   - Por exemplo: Vendas por região, por cliente, por produto, por mês

   Implementar agregações, sumarizações utilizando interface gráfica.

9. Construir projeto piloto para um Data Mart incremental, utilizando um exato *subset* da arquitetura corporativa desenvolvida no item 3.

   Crie Data Mart inicial como piloto para comprovar os conceitos concebidos

   Implemente o Data Mart inicial exatamente como um subconjunto da arquitetura do DD

   - Incluir ferramentas de extração, transformação e carga
   - Fazer as primeiras transformações bem simples
   - Criar o Data Mart inicial para uma área de negócio
   - Selecionar uma ferramenta OLAP simples para o usuário final

   Termine o projeto piloto em três meses após a data de início

Tenha certeza do rápido retorno dos investimentos realizados no Data Mart inicial

Resolver os problemas de integração – Data Marts incrementais têm em comum regras do negócio, semânticas e definições

Integração no nível de um repositório de metadados centralizado

Todos os componentes do DD são dirigidos pelo metadados centralizado

Escolha somente componentes para o DD que tenham padrão de interconexão comum

Área de SI deve padronizar a criação de Data Marts conforme arquitetura desenvolvida

10. Construir um Data Mart incremental adicional.

Identifique áreas de negócio que necessitem Data Marts

Analise as necessidades dos usuários (dimensões, atributos e adicionais)

Defina arquitetura de novos Data Marts

Defina necessidades funcionais dos Data Marts adicionais

Projete base de dados destino para Data Marts adicionais

Construa mapeamento dos dados, regras de extração e transformação

Construa agregações, sumarizações e partições

Garanta que todos os componentes se integrem no metadados central

Prontifique Data Marts adicionais a cada três meses se for necessário

11. Amplie para um DD central utilizando a arquitetura corporativa.

Grande DD servindo como origem dos dados de múltiplos Data Marts

DD centralizado armazena dados detalhados

Dados detalhados são necessários quando usuários assim solicitam

Dá suporte a análises consolidadas, relatórios, consultas

Ambiente complexo e alto custo de desenvolvimento

12. Mantenha e administre o DD.

O processo de administração de um DD é contínuo.

Use ferramentas de administração de DD para criar usuários, permitir acessos, segurança, monitorar acessos e padrões de uso
Bloqueie consultas demoradas e coloque para execução noturna
Monitore ações da corporação de forma a antecipar as necessidades de consultas e poder realizar um planejamento para executá-las.
Utilize ferramentas de administração para ajustar estruturas físicas da base de dados para melhorar o desempenho.

Na tabela 2 é apresentado um quadro comparativo das metodologias apresentadas

TABELA 2 – Quadro comparativo das metodologias

Metodologias	Data Mart incremental	Modelo de dados dimensional	Orientado a área de telecomu-nicações
Metodologia James Martin	Não	Não	Não
Metodologia Alan Simon	Não	Sim	Não
Metodologia Ralph Kimball	Opcional	Sim	Não
Metodologia Douglas Hackney	Sim	Sim	Não
Metodologia NCR	Não	Não	Não
Metodologia Visible	Não	Não	Não
Metodologia proposta	Sim	Sim	Sim

3.4 Sistema de Gerência de Banco de Dados e Servidores

3.4.1 SGBD

Para prover um nível de desempenho adequado para um DD, os SGBD devem dispor de capacidades para processamento paralelo, escalabilidade, particionamento de dados e administração sistêmica [KIM 98].

Processamento paralelo aumenta o desempenho e a escalabilidade dos SGBD. Em processamento paralelo, as operações do banco de dados como consultas, carga de dados, indexação, arquivo e restauração que são realizadas com um grande volume de dados são divididos em pedaços e processados em paralelo em servidor multiprocessado.

Escalabilidade de um SGBD significa que um sistema permite a adição de recursos computacionais e usuários sem afetar a disponibilidade dos dados e aplicações. Um SGBD escalável suporta servidores SMP e MPP (veja a seguir) e tem a habilidade de adicionar servidores ou nós, discos, e memória quando necessária. Para ter uma vida útil adicional, o SGBD deve ser projetado para suportar novas tecnologias (por ex.: NUMA), quando estiverem comprovadas as suas utilidades.

Particionamento dos dados permite que tabelas possam ser distribuídas entre vários discos de forma que o I/O seja realizado em paralelo. Particionamento também permite que um sistema seja expandido e reconfigurado para maior espaço de disco, sem degradar o desempenho do sistema ou sua disponibilidade, características críticas para um DD em constante crescimento. Administradores de sistema também podem utilizar particionamento para realizar operações tipo arquivamento e restauração no nível de tabela, ao invés de toda a base de dados, uma importante funcionalidade quando se está mantendo um DD.

Administração de sistemas é uma característica de SGBD que permite que para bases de dados muito grandes se monitore a configuração do sistema e a atividade do SGBD sem perturbar os usuários e, portanto, aumentando a disponibilidade do sistema.

3.4.2 Servidores

Como os servidores não funcionam sem um sistema operacional (SO), na aquisição do HW o SO o acompanha. No ambiente de computadores centrais, na verdade, só temos uma opção. No ambiente de sistemas abertos, todos os fornecedores têm a sua própria versão de UNIX.

As principais plataformas de HW, portanto, se enquadram em três categorias:

- Computadores Centrais – Os computadores centrais não são a melhor escolha para DD, principalmente devido ao seu alto custo de operação. Estes servidores apresentam custo de administração, hardware e programação normalmente mais altos que os de sistemas abertos, em parte devido ao ambiente robusto para suporte a processamento transacional, que não é crítico para DD. Casos de implantação em computadores centrais usualmente se deve por já terem nascido nesta plataforma ou por existir excesso de capacidade na configuração existente.

- Sistemas Abertos – Servidores de sistemas abertos ou UNIX são a escolha para a maioria dos DD de tamanho médio (100 a 500 GB) a grande (>500GB). Unix é robusto o suficiente para suportar aplicações de produção e foi adaptado para processamento paralelo há mais de 10 anos. Se o servidor escolhido for UNIX deverá ser criada uma equipe para atender às demandas do DD, desenvolvimento e gerência.

- NT – Servidores Windows NT é a categoria que mais cresce no mercado de servidores, mas somente agora conseguiu atender a DD de tamanho médio, pois até pouco tempo só atendia às de tamanho pequeno (<100GB).

Outro ponto a analisar é a arquitetura de processamento paralelo, uma necessidade em aplicações de DD. Hoje, basicamente, são encontradas três alternativas no mercado, como segue:

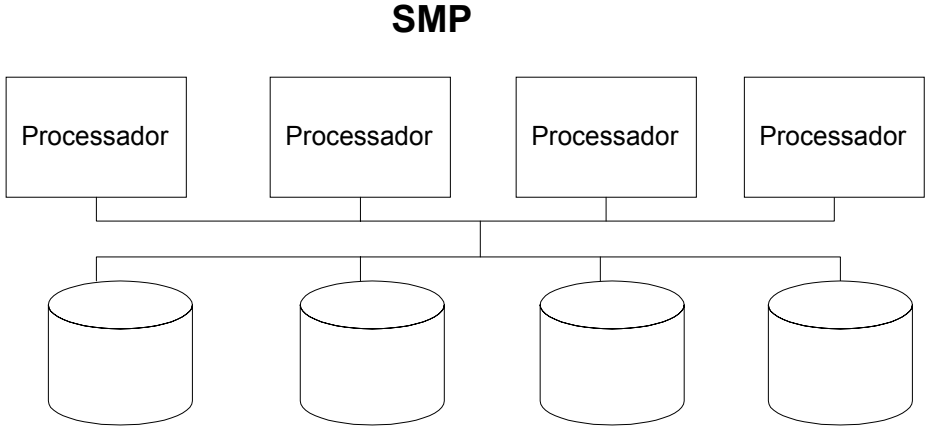

FIGURA 17 - SMP

- SMP – Symmetric Multiprocessing – Esta arquitetura é uma maquina única com múltiplos processadores, todos gerenciados por um sistema operacional e todos acessando a mesma área de memória e disco. Uma máquina com 8 a 32 processadores, um SGBD paralelo, memória grande (2 ou + gigabytes), disco rápido, pode atender a um DD de tamanho médio. A desvantagem desta arquitetura é de ser uma única máquina, portanto um único ponto de falha, e outro ponto é o caminho de acesso a memória e disco, que é única entre todos os processadores e pode se constituir num gargalo.

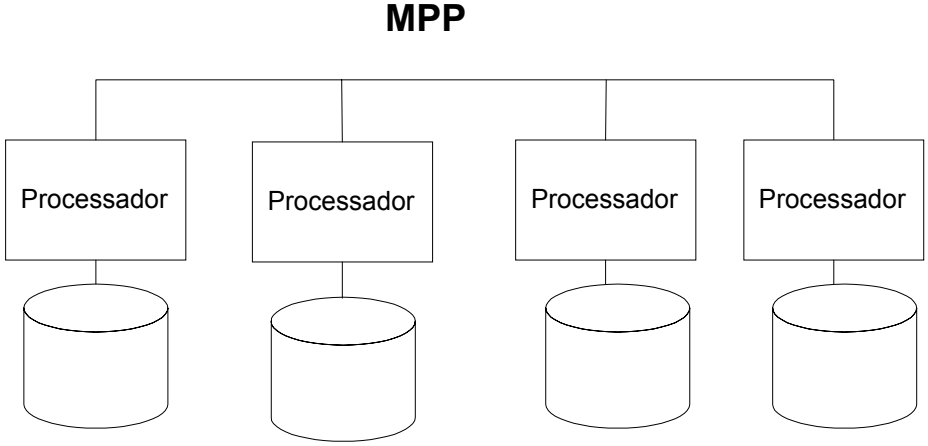

FIGURA 18 - MPP

- MPP – Massively Parallel Processing – Sistemas MPP são basicamente um conjunto de computadores relativamente independentes, cada um com seu próprio sistema operacional, memória, disco, todos coordenados através do envio de mensagens entre eles. A vantagem do MPP é a habilidade de conectar centenas de nós de máquinas e aplicar um problema utilizando o enfoque de forca bruta. Nesta arquitetura também pode haver um congestionamento na tarefa de coordenação, quando o problema é difícil de dividir em pedaços bem segmentados. MPP são tipicamente encontrados em soluções de larga escala (> 1 Terabyte) ou em aplicações intensas como mineração de dados. Novamente o SGBD tem que ter sido projetado para fazer uso desta arquitetura.

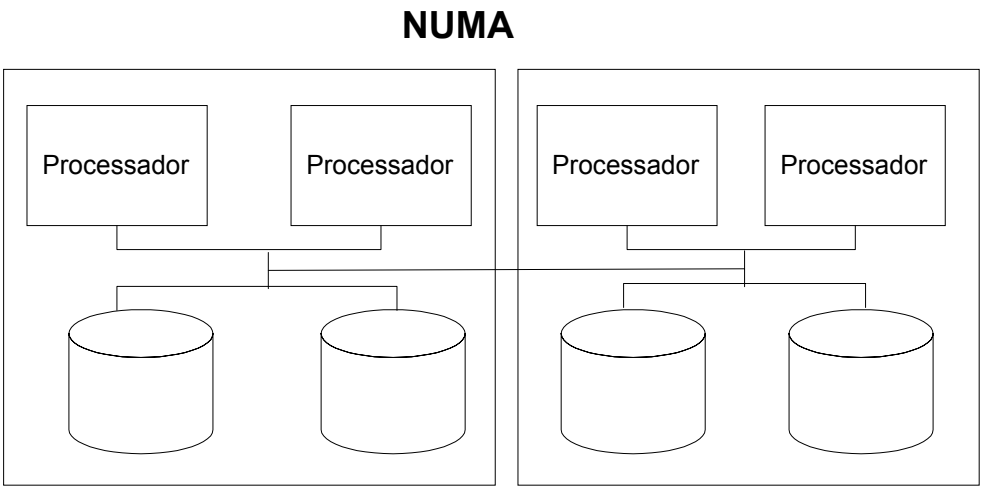

FIGURA 19 - NUMA

- NUMA – Non Uniform Memory Architecture – NUMA é essencialmente uma combinação de SMP e MPP de forma a combinar a flexibilidade de discos compartilhados do SMP com a velocidade do MPP. Esta arquitetura é relativamente nova, e pode ser viável para DD. NUMA é conceitualmente similar à idéia de *clustering* de máquinas SMP, mas com conexões mais justas, mais largura de banda e melhor coordenação entre os nós. Se for possível segmentar seu DD em grupos independentes e colocar cada grupo no seu próprio nó, então a arquitetura NUMA pode ser a solução [KIM 98].

3.5 Ferramentas

3.5.1 Integradas para criação de DM/DD

3.5.1.1 Microsoft SQL Server 7 e Microsoft Excel 2000

Microsoft SQL Server 7 é constituído de um conjunto integrado de softwares que atende parcialmente às necessidades de uma solução DD, mas através de parceiros e do MS Excel 2000 como ferramenta de acesso complementam-se para formar uma solução [MIC 99].

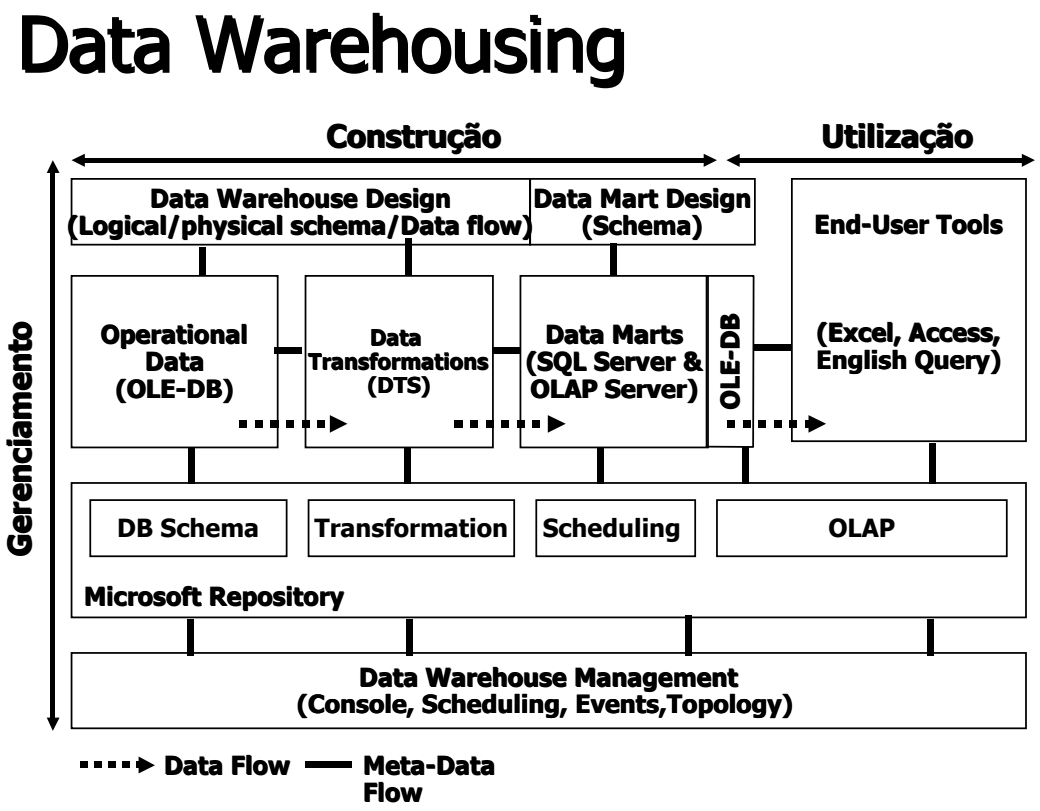

FIGURA 20 – Microsoft Data Warehousing Framework

Este conjunto está incorporado no servidor de banco de dados SQL 7, integrando uma ferramenta gráfica para projeto do Data Mart; uma ferramenta gráfica para extração dos dados transacionais; uma ferramenta OLAP de geração de cubos; um repositório de metadados no próprio SQL e complementando como ferramenta de acesso o MS-Excel 2000. Para permitir o acesso via Internet dos dados ao servidor OLAP é necessário utilizar o sistema operacional NT Server versão 4 com service pack 4, Internet Information Server, NT option pack 4 e o navegador Internet Explorer 4 com programas adicionais (*add-ins*).

O conjunto de softwares é:

- Microsoft SQL Server 7
- Data Transformation Services
- OLAP Server
- SQL Server Enterprise Manager
- MS Excel 2000

Microsoft SQL Server 7 – Provê uma base de dados relacional; escalabilidade até um milhão TB; bloqueio de linha.
Data Transformation Services (DTS) – DTS provê um conjunto de serviços que auxiliam a construção de um DD ou DM. Facilita a importação de dados de qualquer base, transformando-os e exportando-os para qualquer outra base de dados. Disponibiliza *wizards* para uso de VB, Java e outros códigos de escrita a fim de agilizar a transformação dos dados. A linhagem dos dados é automaticamente escrita fora do Microsoft Repository, permitindo independência de bases de dados origem e destino.
OLAP Server – Provê um servidor OLAP flexível que pode ser MOLAP, ROLAP ou HOLAP; agregações inteligentes; fácil de utilizar; serviço de pivot table; permite a criação de cubos virtuais.
SQL Enterprise Manager – Provê autogerenciamento dinâmico; gerenciamento de múltiplos sites; ferramentas de análise de desempenho.
Repositório – Provê um local no SGBD para armazenamento de metadados de todas operações realizadas para criação e manutenção de um DD.

A instalação do sw é bastante simples, com auxílio de especialistas em diversas situações.

Vantagens: Baixo custo; compatibilidade com toda linha Microsoft; fácil de utilizar; componentes incluídos no SQL atendem muito bem ao *low-end* do mercado de Data Mart.

Desvantagens: Baixa escalabilidade; único sistema operacional MS-Windows NT; orientado para Data Marts.

3.5.1.2 Oracle Data Mart Suite

Oracle Data Mart Suite é um conjunto integrado de software necessário para rapidamente e simplificadamente implementar um Data Mart. Os produtos neste conjunto pretendem atender a todo o ciclo de vida de um Data Mart, do projeto à construção, análise e gerência[ORA 98].

Este conjunto está baseado no servidor de banco de dados Oracle 8, integrando uma ferramenta visual para o projeto do Data Mart; uma ferramenta gráfica para extração dos dados transacionais; uma base de dados de alto desempenho e escalável que serve como repositório para ambos dados e metadados; um servidor Internet para acesso via Intranet; uma ferramenta para consulta, relatórios e análise, e uma documentação e tutorial *on-line*.

O conjunto de softwares é:

- Oracle Data Mart Designer (ODMD)

- Oracle 8 Enterprise Edition
- Oracle Enterprise Manager
- Oracle Data Mart Builder
- Oracle Discoverer
- Oracle Reports
- Oracle Web Applications Server

Oracle Data Mart Designer – Utilizado para projetar os Data Marts e armazenar os modelos em um repositório para posterior referência pelo Oracle Data Mart Builder. Permite o uso de engenharia reversa para obter o modelo de dados das bases transacionais através de uma interface gráfica chamada Data Schema Diagrammer, e utilizá-los para criar um esquema para o Data Mart, normalmente um esquema estrela.
Quando o projeto de Data Mart estiver pronto e armazenado no repositório, o ODMD gera os scripts SQL para gerar as estruturas físicas dos dados, tabelas e índices, na base de dados Oracle 8.
Oracle 8 Enterprise Edition – Provê uma gerência específica de dados para sistemas de suporte a decisão. Data Marts requerem técnicas diferentes de processamento devido à forma com que os dados são acessados. Oferece também a possibilidade de utilizar o paralelismo desta versão, melhorando os tempos de resposta e a capacidade de efetuar *drill-down*. Seu otimizador está preparado para consultas complexas e para consultas em esquemas estrela. Além disso, provê índices *hash* e *bitmapped* para melhorar o desempenho das consultas.
Oracle Enterprise Manager – Provê uma solução completa para gerência de sistemas, com aplicações para gerenciar o armazenamento, esquemas, arquivo lógico e físico, inicialização e finalização da base de dados, e privilégios de acesso aos usuários do Data Mart. Utiliza uma interface gráfica para realizar estas rotinas, mas também aceita scripts feitos pelo administrador.
Oracle Data Mart Builder – Provê a tecnologia necessária para realizar as tarefas de extração e transformação, gerenciar o fluxo de dados da origem para o Data Mart. As definições são feitas com auxílio de uma interface gráfica, chamada de plano de fluxo de dados, que determina como os dados devem ser processados enquanto flui da origem para o destino.
Oracle Discoverer – É uma ferramenta de análise e relatório para usuário final que permite navegar através do Data Mart. A ferramenta é construída em uma arquitetura voltada para o usuário, permitindo a abstração dos dados e uso de termos de negócio. Provê também de mecanismo automático para criar e

manter tabelas sumarizadas, e redireciona as consultas para estas tabelas se for mais rápido desta forma. Prevê a eliminação destas tabelas por falta de uso pelo usuário através de estatísticas. Está preparada para utilizar os novos índices existentes no Oracle 8 como índices *bitmapped*.
Oracle Reports – Provê uma ferramenta para criação de relatórios.
Oracle Web Application Server – Provê uma plataforma de aplicação Internet segura e escalável que trabalha com todos os navegadores populares.

Acompanha um manual com um tutorial para facilitar o uso das ferramentas através de um estudo de caso, permitindo que passo a passo seja criado um Data Mart.

O estudo de caso utiliza uma empresa revendedora de hw e sw para clientes em todo mundo. O Data Mart vai ser criado para atender à área de Vendas e Marketing, com a função de determinar onde os seus lucros estão diminuindo, quais os componentes mais lucrativos e quais podem melhorar.

Seguindo os seguintes passos:

- Requerimentos e projeto
- Construção
- Extração, transformação e transporte (carga)
- Acesso a base de dados
- Criação de relatórios
- Gerência

A instalação dos sw é fácil (somente para Windows NT), terminando com todos os módulos instalados e as bases iniciais do tutorial também instalado.

Vantagens: o produto tem alta integração e cobre todos os passos de uma implementação. Para empresas que já têm como base de dados corporativa o Oracle, fica bastante fácil obter dados destas bases pela facilidade de interconexão e uso de ferramentas comuns, além de já dispor de DBA treinados para estas bases.

Desvantagens: ferramentas somente funcionam para bases destino e repositório Oracle; custo alto; para bases não relacionais a importação é feita somente via formato texto.

3.5.2 Ferramentas de projeto e modelagem

3.5.2.1 Sybase – Datawarehouse Architect

Ferramenta desenvolvida para projeto, criação e manutenção de sistemas de DD, com suporte ao modelo dimensional estrela e *snowflake*. Trabalha com vários SDBD de mercado e com integração maior com produto da própria Sybase IQ, uma base de dados criada especificamente para armazenar DD.

3.5.2.2 Prism – Logic Works – Erwin

ERwin é um produto que integra definições conceituais de negócios e, a partir de modelos de dados construídos com rapidez, gera automaticamente os scripts do banco de dados, incluindo regras de negócios, triggers, stored procedures, entre outros recursos que tornam o desenvolvimento rápido e construtivo.

3.5.2.3 Anubis-Constructa

Ferramenta para criação de modelo dimensional, com geração automática de modelos estrela, *snowflake* e normalização parcial de dimensões, criando tabelas de fatos e dimensões; geração de esquema para vários SGBDs de mercado; customização de esquema e ajuste de desempenho.

3.5.3 Extração, transformação e carga de dados

Existem várias ferramentas de extração no mercado sendo que as mais antigas utilizam programação, normalmente COBOL, já as de segunda geração fazem uso de API´s para acesso a base de dados ou através de arquivos-texto.

As tarefas de extração, limpeza, transformação e carga de dados em um depósito de dados demandam uma quantidade enorme de tempo. Segundo [INM 94], a maior parte dos esforços na construção de um DD são consumidos nestas tarefas.

Os produtos oferecidos no mercado procuram automatizar processos que teriam de ser feitos manualmente ou utilizando ambientes de programação de mais baixo nível. De fato, não existe uma ferramenta única capaz de oferecer suporte aos processos de extração, limpeza, transformação e

migração dos dados: diferentes ferramentas especializam-se em questões específicas.

O grande desafio por trás da alimentação de dados das fontes para o Data Warehouse não é técnico, mas gerencial. Muitos dos processos envolvidos – como mapeamento, integração e avaliação de qualidade – ocorrem de fato durante a fase de análise e projeto do Data Warehouse [MOR96]. Especialistas afirmam que identificar fontes, definir regras de transformação e detectar e resolver questões de qualidade e integração consomem cerca de 80% do tempo de projeto. Infelizmente, não é fácil automatizar estas tarefas. Embora algumas ferramentas possam ajudar a detectar problemas na qualidade dos dados e gerar programas de extração, a maioria das informações necessárias para desenvolver regras de mapeamento e transformação existe apenas na cabeça dos analistas e usuários. Fatores que certamente influem na estimativa de tempo para estas tarefas são o número de fontes e a qualidade dos metadados mantidos sobre estas fontes. As regras de negócio associadas a cada fonte – tais como validação de domínios, regras de derivação e dependências entre elementos de dados – são outra fonte de preocupações. Se estas regras tiverem de ser extraídas do código fonte das aplicações, o tempo para mapeamento e integração pode dobrar.

Um cuidado adicional tem que ser tomado com as alterações nas bases de origem. Sempre que modificações relevantes são feitas nas fontes de informações, estes novos dados são extraídos e traduzidos para o modelo de dados do Data Warehouse, onde são integrados com os dados já existentes.

A detecção e extração das modificações dependem das facilidades disponíveis pela fonte. Se esta fonte for sofisticada, como um SGBD relacional que possui *triggers*, este processo é relativamente fácil. No entanto, em muitos casos a fonte não dispõe de recursos avançados para detecção e captura das modificações (sistemas legados, por exemplo). Nestes casos, existem basicamente três alternativas para detectar e extrair modificações:

1. A aplicação que utiliza a fonte de informação é alterada de modo a enviar notificações de alteração para o Data Warehouse. Esta alternativa requer que o código existente seja modificado. No entanto, na maioria dos casos esta opção é impraticável devido à complexidade do código e ao grande tempo necessário para sua alteração.

2. O arquivo de log do sistema é analisado de modo a obter as modificações relevantes. O problema com esta alternativa é que normalmente são necessários privilégios de DBA para acessar o log e muitos administradores relutam em prover este acesso, pois coloca em risco a segurança do sistema.

3. As modificações são determinadas através da comparação da carga corrente da fonte com uma carga anterior. O problema com esta alternativa é que ela não é muito escalonável, ou seja, à medida que os dados-fontes aumentam é necessário um número muito maior de comparações. Deste modo, torna-se absolutamente necessário implementar este algoritmo do modo mais eficiente possível.

4. Toda e qualquer alteração nas bases transacionais são comunicadas pelos DBAs responsáveis ao DBA do DD, processo bastante difícil de implementar quando as bases-origens se encontram distribuídas (inclusive em países diferentes), mas factível quando as bases se encontram em um único local.

Nas abordagens onde são mantidos o DD centralizado e Data Marts por departamentos ou funcionais, há a necessidade de se estabelecer uma estratégia que coordene a entrega de novos dados a todos os bancos. É preciso considerar a incorporação de um servidor de replicação na arquitetura de distribuição dos dados. Um servidor de replicação é uma aplicação sofisticada que seleciona e particiona dados para distribuição para cada Data Mart, aplicando restrições de segurança, transmitindo uma cópia dos dados para os locais adequados e criando um log de todas as transmissões armazenadas no metadados.

Uma ferramenta completa de extração deve gerar os programas necessários para extrair, transformar e mover os dados de qualquer sistema de dados existentes para qualquer outro sistema independente de hardware e plataforma de software.

3.5.3.1 De primeira geração, produtos geradores de código:

Têm como características: a extração/transformação/carga ocorrem no servidor ou computador central; dados detalhados necessitam de programação manual em COBOL ou C; programa de extração é gerando automaticamente em código-fonte; código-fonte é compilado, programado para rodar em modo de lote; dados são extraídos de arquivos-fontes e processados no servidor ou computador central; dados são gravados em arquivos intermediários; programas não suportam utilização de processamento paralelo e geram muito pouco metadados automaticamente.

As ferramentas comerciais mais representativas encontradas no mercado são as seguintes:

- Prism Solutions Inc.

Prism Change Manager – automatiza a manutenção de DD através da captura, transformação e carga de dados modificados para o DD. Captura dados modificados que foram utilizados como fonte para o DD, aceitando bases como DB2, Enscribe, IMS e Oracle. Utiliza como fonte histórica de fitas, e captura somente registros relevantes alterados via execução desatendida e protegendo a performance dos sistemas transacionais. Prevê a lógica necessária para extração e transformação a ser aplicada nos dados alterados dos históricos de fita e posterior carga no DD. Automatiza a integração, sumarização, filtragem e conversão de dados para DD de acordo com as especificações desenvolvidas no projeto do DD. Logo depois, gera um programa para anexar, substituir, inserir ou apagar uma ou mais linhas no SGBD pretendido. Bancos de dados objetivo utilizados por esta ferramenta são DB2, Informix, Oracle, Rdb, Sybase e Teradata. Documenta o processo de manutenção através da coleta técnica de metadados do processo de atualização. O metadados reflete a variância tempo e a transformação realizada na atualização do DD.

• Evolutionary Technologies International

ETI* Extract Tool Suite – Permite aos usuários automatizar e disparar a migração de dados entre ambientes de armazenamento distintos. Este produto permite a manipulação de dados, tecnologia de geração de códigos e metodologia de implementação, permite aos usuários populares manter um DD, migrar para novas arquiteturas, integrar sistemas heterogêneos e migrar dados para novas bases de dados, plataformas e aplicações. Permite a migração de qualquer plataforma, sistema operacional e SGBD para outros ambientes, inclusive sistemas proprietários. Um utilitário de metadados permite ao usuário acessar, exportar e unir metadados. Inclui também capacidades de versão, interface gráfica que permite mover dados através de um simples apontar-e-clicar; uma facilidade de armazenamento interna orientada a objeto; uma facilidade que permite o uso concorrente de múltiplos usuários; proteção de integridades de dados e um utilitário flexível com opções de consultas como análise do impacto de mudanças. Executa em sistemas baseados em UNIX, incluindo Sun Solaris, IBM AIX e HP-UX, e pode ser utilizado em qualquer plataforma que execute X-Windows.

- Apertus Carleton Company

Passport – Automatiza o desenvolvimento e manutenção de DD em sistemas cliente/servidor ou ambiente de computador de grande porte. A capacidade do produto inicia na criação e carregamento do metadados, até todos os aspectos

do programa de desenvolvimento e manutenção do DD. Todos os processos são realizados a partir do metadados. No cerne do Passport existe uma lista de metadados. Este metadados pode ser implementado como uma estação independente, um computador de grande porte ou ambiente cliente/servidor. Passaport utiliza um ambiente gráfico no mapeamento das regras para "popular" o DD. Um programa de gerência permite gerenciar as regras já definidas. Permite o acesso a dados legados em seu ambiente original. Carrega os dados em todos os SGBD do mercado, incluindo DB2, Informix, Oracle, Red Brick, Sybase e Teradata.

- Forecross Corp

Convert Series – Uma família de ferramentas para migração que automatiza o processo de conversão, incluindo programas, definições de base de dados e os próprios dados. A ferramenta Convert IMS-DB converte aplicações de IDMS para SQL. O produto inclui Convert/IDMS-DB-to-DB2, Convert/IDMS-DB-to-Oracle, Convert/IDMS-DB-to-Sybase e Convert/IDMS-DB-to-Informix. Cada uma destas ferramentas opera em computadores de grande porte, OS/2 ou Unix e automatiza esquemas, dados e programas de conversão. A ferramenta Convert IMS/VSAM converte aplicações de VSAM para SQL. O produto inclui Convert/VSAM-DB-to-DB2, Convert/VSAM-DB-to-Oracle, Convert/VSAM-DB-to-Sybase e Convert/VSAM-DB-to-Informix. A série da Convert executa em computadores IBM de grande porte, OS/2 ou Unix.

3.5.3.2 De segunda geração, produtos baseados em acesso direto:

Têm como características: o processo de extração/transformação /carga ocorrerem no servidor; dados são extraídos diretamente dos arquivos origem e processados no servidor; os dados são transformados na memória e escritos diretamente na base destino; códigos são executados diretamente no servidor, nenhum código é compilado; existem várias funções, incluindo monitoração, programação, extração, limpeza, transformação, carga, indexação, agregação e metadados; uso eficiente de processamento paralelo, alto desempenho porque não existe gravação de arquivos intermediários; metadados é gerado automaticamente e mantido em repositório aberto.

As ferramentas comerciais mais representativas encontradas no mercado são as seguintes:

- Informatica Corporation

PowerMart Suite – Uma solução completa com ferramentas para criação do modelo de dados, um repositório para Metadados com carga automática das operações realizadas e uma ferramenta de extração/transformação/carga que retira os dados de bases de dados origem e coloca na base de dados destino.

- Sagent Technology

Data Mart Solution – Uma solução completa envolvendo ferramentas de modelagem, extração/transformação/carga, metadados e adicionalmente ferramentas de acesso e análise de dados.

Existem muitas outras ferramentas no mercado, mas orientadas a apenas um aspecto da extração de dados. As ferramentas apresentadas possuem uma maior abrangência de bases de dados e ambientes transacionais.

3.5.4 Ferramentas de acesso a dados

A ferramenta de acesso para usuário final tem como objetivo permitir o acesso ao DD com graus variados de recursos.

As ferramentas podem ser divididas em cinco grupos que são:

- Ferramentas de consulta – Ferramentas mais simples para consultas e geradores de relatórios básicos. Em geral, oferecem uma interface gráfica para geração de SQL, permitindo o uso de menus e botões para a especificação de elementos de dados, condições, critérios de agrupamento, sem que seja necessário aprender uma linguagem especializada para acesso ao banco. O processamento estatístico, neste caso, é limitado a médias, totais, desvios-padrão e algumas outras funções básicas de análise.

- Ferramentas de relatório – Ferramentas especializadas na emissão de relatórios, mas não atendem a usuários que precisem mais do que uma visão estática dos dados e que não pode mais ser manipulada. Ferramentas OLAP podem oferecer a este tipo de usuário maior capacidade de manipulação, permitindo analisar o porquê dos resultados obtidos. Estas ferramentas, muitas vezes, são baseadas em bancos de dados multidimensionais, o que significa que os dados precisam ser extraídos e carregados para as estruturas proprietárias do sistema, já que não há padrões abertos para o acesso de dados multidimensionais.

  Ex.: Seagate Crystal Reports.

- Ferramentas OLAP – Ferramentas que extraem um subconjunto de dados e disponibilizam aos usuários. Podem ser OLAP desktop, ROLAP e MOLAP. O primeiro é ferramentas de pequeno porte que rodam no

microcomputador cliente, no segundo caso temos uma ferramenta de três camadas que roda em servidor com uma base relacional e por último bases multidimensionais que apresentam grandes desempenhos mas utilizam bases multidimensionais proprietárias. O OLAP não é uma solução imediata. Configurar o programa de OLAP e ter acesso aos dados requer uma clara compreensão dos modelos de dados da empresa e das funções analíticas necessárias aos executivos e outros analistas de dados.

Ex: 1. Brio, Business Objects e Impromptu, Excel 2000.
2. DSS Agent, Decision Suite, MetaCube, SQL 7.
3. Oracle Discover, Pilot, Holos, SAS MDDB, SQL 7.

- Ferramentas de mineração – Ferramentas que através de algoritmos, como agrupamento, classificação e associação, permitem descobrir relações até então desconhecidas, e têm como característica pouca ou nenhuma intervenção do usuário na descoberta de conhecimento. Mineração de dados pode ser utilizado com os seguintes objetivos:

    - explanatório: explicar algum evento ou medida observada, tal como por que a venda de sorvetes caiu no Rio de Janeiro;

    - confirmatório: confirmar uma hipótese. Uma companhia de seguros, por exemplo, pode querer examinar os registros de seus clientes para determinar se famílias de duas rendas têm mais probabilidade de adquirir um plano de saúde do que famílias de uma renda;

    - exploratório: analisar os dados buscando relacionamentos novos e não previstos. Uma companhia de cartão de crédito pode analisar seus registros históricos para determinar que fatores estão associados a pessoas que representam risco para créditos.

Especialmente devido ao alto custo envolvido, estas ferramentas vinham sendo usadas, até o momento, quase que unicamente por grandes corporações e instituições governamentais. A maior parte das atividades de mineração de dados ficava restrita a especialistas, com empresas oferecendo seus serviços de análise, mas sem entregar aos clientes seus métodos e ferramentas. Com o grande aumento do volume de dados nas empresas e com o crescimento do uso de tecnologia de banco de dados, especialmente de Data Warehouse, as técnicas de mineração de dados assumiram papel importante no suporte aos processos de tomada de decisão e devem, aos poucos, ganhar mercado dentre empresas de menor porte. No entanto, no atual estado da arte destas ferramentas, ainda requerem um bom nível de conhecimentos – do domínio da aplicação, de estatística e da própria ferramenta.

Dentre as técnicas utilizadas, encontram-se [BRAC 96][WEL 96][HER 96]:

- lógica baseada em casos: consiste na derivação de regras a partir de estudos de caso.

  Ex.: Remind, da Cognitive Systems.

- descoberta de regras: envolve a execução de algoritmos de análise de dados em grandes massas de dados na busca de padrões e correlações que possam subsidiar a formulação de regras. A busca pode ser dirigida (procurando por dados para apoiar uma determinada regra) ou não (permitindo que padrões de dados sugiram possíveis regras).

  Ex.: IDIS, da Information Discovery.
  KnowledgeSeeker, da Angoss Software.

- pontuação: utilizando um esquema de pontuação, dados históricos podem ser analisados e uma árvore de decisão construída baseada em um conjunto de valores;

- processamento de sinais: técnicas de processamento de sinais, tal como filtragem digital, podem identificar itens de observações com características similares.

  Ex.: Data Engine, da alemã MIT GmbH.

- fractais: uso de fractais para comprimir grandes bancos de dados sem perda de informação permite a análise sobre todo o universo do banco com um tempo de resposta surpreendente.

  Ex.: F-DBMS, da Cross/Z International.

- redes neurais: constitui-se em modelos de previsão baseados e, princípios similares àqueles do pensamento humano. Em uma rede de nós, cada nó recebe entrada e envia uma saída para nós subseqüentes baseado no que recebeu. A rede é "treinada" usando uma amostra de dados para determinar "pesos" apropriados para cada nó. Ela então produz valores específicos para dados subseqüentes.

  Ex.: Prism, da Nestor Inc.
  Neura/Ware, Neura/Ware Inc.

Na verdade, alguns dos complexos algoritmos usados em mineração de dados já existem há duas décadas. O governo americano tem usado sistemas de mineração de dados especializados usando redes neurais, lógica fuzzy e reconhecimento de padrões para investigar fraudes em impostos e outras áreas estratégicas.

As diversas técnicas de mineração de dados dão suporte a um conjunto de operações que diferem entre si pelo tipo de problema que são capazes de resolver. São elas: associações, padrões seqüenciais, séries temporais similares, classificação e regressão, e clusterização [BRA96].

- Associação

Associações são relacionamentos significativos entre itens de dados armazenados. O objetivo da operação é encontrar tendências, a partir de grande número de transações, que possam ser usadas para entender e explorar padrões de comportamento dos dados. Um exemplo seria o de varrer registros de terminais de ponto de venda e descobrir que tipos de itens são vendidos juntos, de forma a redefinir a disposição dos artigos na loja e sua promoção em campanhas publicitárias, permitindo explorar com maior eficácia essas associações.

- Padrões seqüenciais e séries temporais similares

Enquanto a associação encontra eventos que ocorrem juntos a partir de coleções lógicas, a operação de padrões seqüenciais encontra eventos relacionados que ocorrem ao longo de um período de tempo. Um exemplo deste tipo de operação seria a identificação de padrões de sintomas e doenças em pesquisas médicas.

Séries temporais similares podem ser usadas para identificar séries similares coletadas ao longo de um período de tempo. Como exemplo, pode-se considerar a identificação de empresas com padrão de crescimento similares, ações ou fundos de investimento com movimentos de preços parecidos.

- Classificação e regressão

Classificação e regressão usam dados existentes para criar modelos de comportamento de variáveis.

A operação de classificação cria automaticamente um modelo a partir de um conjunto inicial de registros. Esse conjunto serve de exemplo e é chamado de conjunto de treinamento. Os registros do conjunto de treinamento devem pertencer a um pequeno grupo de classes predefinidas. O modelo é composto de padrões, essencialmente generalizações em relação aos registros, os quais são usados para diferenciar as classes. Uma vez

obtido o modelo, este é usado para classificar automaticamente os demais registros.

O modo como as classes são criadas oferece vantagens em relação a métodos estatísticos. Os padrões podem ser produzidos a partir de um conjunto localizado de fenômenos, ao passo que métodos estatísticos devem agir sobre populações inteiras e de distribuição bem conhecida. Desta forma, é possível prever características de um pequeno percentual do conjunto de registros, o que não seria alcançado estatisticamente dada a inexpressividade dos registros sendo avaliados. Como exemplo, uma empresa de cartão de crédito poderia examinar algumas características de seus clientes e prever o nível de inadimplência associado. Tais características poderiam incluir renda, histórico de crédito, tipo e localização do emprego.

- Clustering

O agrupamento em clusters envolve segmentar a informação disponível em conjuntos definidos e homogêneos baseando-se em atributos específicos. O conceito de clustering já tem uma longa história em estatística, mas o que tem de novo em mineração de dados é o fato de poder também ser aplicada a itens não numéricos. Os resultados de uma operação de clusterização podem ser usados de duas diferentes maneiras: para produzir um sumário da base de dados ou como dados de entrada para outras técnicas, por exemplo, classificação, já que um cluster é um grupo menor e de mais fácil manuseio por parte de algoritmos de classificação.

Ex.: MineSet – Silicon Graphics.
Intelligent Miner – IBM.

- Ferramentas de visualização – Ferramenta de mineração que através de apresentação visual, normalmente em três dimensões, permite maior facilidade de análise.

As técnicas de visualização não são propriamente técnicas de mineração de dados, mas sim meios de analisar e observar os dados de uma determinada base de dados de forma gráfica.

A visualização fornece meios de obter sumários visuais dos dados de uma base de dados. No caso de técnicas de clusterização podem ser usadas ferramentas de visualização para determinar qual ou quais clusters criados são úteis ou interessantes para os métodos de mineração de dados.

As ferramentas de visualização podem ainda ser usadas como um mecanismo de compreensão da informação extraída por meio das técnicas de mineração de dados. Características difíceis de detectar pela simples

observação de linhas e colunas (com valores numéricos) podem se tornar óbvias se forem observadas graficamente.

Por meio de visualização podem ser realizadas técnicas interativas que permitam rápida e facilmente alterar o tipo de informação analisada, bem como o método usado (histogramas, gráficos de dispersão, etc.). Também é útil para a percepção de características que se aplicam a pequenos subconjuntos dos dados e que poderiam passar despercebidas se fossem utilizados meios estatísticos, pois estes consideram características genéricas.

Por meio destas técnicas podem ser encontrados características ou fenômenos pouco comuns ou interessantes sem que se esteja diretamente procurando por eles; também não é necessário saber que tipo de fenômeno deve ser analisado ou que questões específicas devem ser feitas, tal como acontece com métodos estatísticos, pois, em termos humanos, tais características se tornam explícitas quando os dados são representados graficamente [WEL 96].

Ex.: dbExpress - Computer Concepts Corp.
MineSet Rule Visualizer – Silicon Graphics.
Discovery – HYPERparallel.

4 Prova de conceito – empresa de telecomunicações

O objetivo deste capítulo é descrever as aplicações de DD em uma empresa de telecomunicações e apresentar a prova de conceito validando os conteúdos dos capítulos anteriormente apresentados. Esta prova de conceito será feita com dados oriundos da CRT na área de análise de desempenho de rede e qualidade de serviços.

4.1 Utilização de "Call Detail Record" (CDR) ou Bilhetes em aplicações para empresas de telecomunicações [TEC 95] [MAT 97]

Uma das indústrias que mais experimentam a concorrência acirrada nos últimos tempos é a de telecomunicações. A área de marketing tem que trabalhar para manter os clientes. A área de SI tem que gerenciar proativamente as mudanças de carga de tráfico na rede devido a novos serviços, que têm crescido em taxas inesperadas como Internet e telefonia fixa sem fio (*wireless*). E a área de finanças tem que garantir os resultados positivos com preços cada vez mais baixos, devido à intensa concorrência nos mais variados segmentos de serviço.

Mas esta empresa já tem uma ferramenta para realizar as análises necessárias, as chamadas "Call Detail Record" (CDR) ou simplesmente bilhete. Utilizando DD, OLAP e técnicas de mineração, as empresas podem obter conhecimentos úteis e difíceis de obter para se tornarem orientadas ao cliente e auxiliar os gerentes a realizarem decisões mais corretas e melhor substanciadas sobre marketing, cobrança e gerência de rede de telecomunicações.

Podemos comparar o bilhete com registros de transações bancárias ou registros de reservas de companhias aéreas. O bilhete é gerado no início de toda interação do cliente com a empresa e alterado no caminho que percorre a ligação. No caso do Brasil, até o momento, só são criados bilhetes de chamadas interurbanas, as chamadas locais são somente registradas pelo número de pulsos ocorridos, sem informação de destino ou rota utilizada.

Mesmo assim, o número de bilhetes gerados mensalmente é bastante grande – na CRT são gerados entre 120 a 150 milhões de bilhetes com tamanho de 80 bytes cada um.

Existem várias possibilidades de análise baseadas nos bilhetes. A seguir, algumas destas aplicações:

Uma delas é a análise de problemas de origem/destino e qualidade do serviço. O de qualidade já é exigido pela ANATEL para verificar o desempenho das empresas. E é neste caso onde será realizada a prova de conceito.

Outra aplicação seria na detecção de fraudes, através de uso de duas técnicas, sendo o primeiro passo o uso de mineração de dados para criar perfil de clientes, incluindo tempo e distância, padrões de chamadas, limites de crédito e destino das chamadas. Com estes dados a empresa compara os dados reais de chamada com o perfil do usuário.

- A primeira técnica é detecção de fraude em tempo real. Neste caso a empresa monitora a duração da chamada durante a sua execução antes da geração do bilhete. Se a duração for maior que o usual para aquele usuário, é imediatamente enviada para a área de fraudações e se possível contatado o cliente para abortar o chamado e/ou bloquear o número do telefone.
- A segunda técnica é de detecção de fraude de assinatura. Neste caso a empresa monitora o comportamento das chamadas. Se for fora do padrão usual daquele usuário, é imediatamente enviada para a área de fraudações e é contatado o cliente para verificar este comportamento.

A aplicação mais popular entre as consultorias e fornecedores de solução atualmente é a dirigida para área de marketing denominada como "Customer Relationship Management" ou Gerência de Relacionamento com Clientes, que atua com todos os contatos que o cliente tem com a empresa.

- Através do uso de bilhetes agregados dos valores de cobrança (extraídos normalmente do Sistema de Faturamento) e utilizando várias técnicas de análise, determina os clientes e mercados mais rentáveis, os menos rentáveis, criam segmentações e controla resultados de campanhas. Os clientes preferenciais podem então ser atingidos por propaganda e promoções, por exemplo, desconto por atingir determinado volume de chamadas. E os clientes não rentáveis podem ter sua tarifa aumentada, se a ANATEL permitir, para torná-los também rentáveis. Bilhetes faturados também podem ser utilizados para vendas cruzadas de serviços.

Ainda outra aplicação é a análise de *churn,* termo em inglês que designa a mudança de operadora de serviços (mais orientada para telefonia celular, onde existe menor grau de fidelização).

- Utilizando-se bilhetes faturados, consegue-se determinar antecipadamente a mudança de hábitos de um cliente e realizar uma ação para que esta conduta não o leve a cancelar o serviço.

Estas são algumas das possibilidades de aplicação de um DD em telecomunicações, mas segundo [TEC 95] pode-se resumir como sendo nas áreas de marketing e de negócio as maiores aplicações, como segue:

TABELA 3 – Aplicações de DD para telecomunicações

Marketing	**Outras áreas**
Gerência de Relacionamento de Clientes	Análise de uso de serviços
Criação de perfil de clientes	Análise de abastecimento
Retorno de clientes perdidos	Análise de falhas de serviços
Análise de promoções	Análise de desempenho de rede e planejamento
Análise de penetração de mercado	Análise de capacidade
Análise de canais de venda	Análise de orçamento
Rentabilidade de produtos	Análise de estoque
Análise de faturamento	Relatórios para órgãos reguladores
Análise consolidada de serviços	Análise financeira
Análise de domicílio	
Análise de perda de clientes	
Análise de outros produtos como cartão de crédito.	

4.2 Prova de conceito

Esta prova de conceito vem atender a uma dificuldade real da CRT, na área de medições e registros (área responsável por indicadores de desempenho da rede e da qualidade dos serviços), de verificar os bilhetes originais que participaram no cálculo dos indicadores de desempenho.

4.2.1 O problema

Os indicadores são calculados baseados nos bilhetes obtidos por um sistema de coleta de bilhetes gerados nas centrais bilhetadoras, espalhadas pelo Rio Grande do Sul, e que posteriormente são enviados para um Sistema de Gerência de Desempenho de Tráfego, na forma de vários arquivos por central, para o cálculo dos indicadores. Numa etapa seguinte estes bilhetes com os indicadores associados são disponibilizados via ftp para carga na área usuária para análise três dias após a coleta.

O trabalho de verificação é feito através de carga dos dados, vários arquivos txt, para uma base de dados em MS-Access na forma de uma tabela única, onde as consultas são feitas diretamente sobre a tabela, indexando as

colunas referentes aos indicadores de forma a agrupar os bilhetes por cada indicador analisado. Desta forma não existe qualquer tipo de comparação entre os indicadores ou outras alternativas como realizar filtros por central bilhetadora ou rota de encaminhamento, etc. Além de demandar um esforço muito grande pela área usuária para carregar os dados no MS-Access e também para o analista do negócio para realizar indexações variadas sobre um sw e uma máquina inadequada para tal, principalmente pelo volume gerado por estes dados.

Os dados utilizados para gerarem os indicadores são relativos a aos períodos críticos, 9h às 11h, 14h às 16h e 20h às 22h, num dia determinado no mês onde são coletados todos os bilhetes gerados, num total aproximado de 6 milhões de registros com tamanho de 150 bytes.

4.2.2 A solução

O primeiro passo foi verificar com o usuário as necessidades apresentadas no item anterior.

Ao analisarmos as necessidades apresentadas, vemos como solução criar um Data Mart incremental que possibilite a reutilização de suas definições e estrutura para atender a demandas futuras que virão a seguir. Neste caso, fica evidente a utilização de agregações e sumarização para aumentar a velocidade das respostas à análise pretendida.

O hardware e o software foram escolhidos tendo em vista a disponibilidade da CRT e a adequação ao projeto.

Os softwares escolhidos foram MS SQL 7 Server e suas ferramentas para criação do Data Mart e do cubo OLAP e MS Excel 2000 /Intranet para acesso aos dados. A versão 7 do MS SQL incorpora várias ferramentas integradas ao SGBD dirigidas ao mercado de DD, como descritos no item 2.5.1.1 , além de melhorias significativas nos SGBD.

O hardware disponibilizado foi um servidor ACER Altos 9000 com um processador Pentium Pro 200 MHz, 256 Mbytes de memória RAM, e raid de disco com 14 gigabytes.

A escolha pelo software Microsoft para esta solução leva em conta também a viabilização de implantação da solução por um custo baixo, ferramentas de fácil uso, ferramenta de acesso já conhecida pelo usuário e já disponível na empresa e principalmente a adequação ao escopo pretendido.

4.2.3 Metodologia

1. Identificar a origem da solicitação, um patrocinador, riscos e taxa de retorno.
   - A área de medições e registro, com o patrocínio do gerente da área e superintendente, baixo risco e com retorno alto mensurável por homem/hora utilizado na solução atual.
2. Avaliar as necessidades dos usuários e identificar as funcionalidades desejadas.
   - Dispor de uma ferramenta que faça a carga de dados oriunda do Sistema de Gerência de Desempenho de Tráfego, faça um tratamento nos dados e disponibilize para análise.
   - Dispor de uma ferramenta que permite analisar, rapidamente, os bilhetes que determinaram os indicadores de desempenho, permitindo filtros por diversos itens disponíveis no bilhete, bem como agregações por prefixos.
3. Projetar uma arquitetura corporativa de Depósito de Dados de longo prazo no papel.
   - Arquitetura proposta atende às necessidades da área de tráfego.
4. Definir necessidades funcionais para o projeto inicial.
   - Determinar os bilhetes que originaram os indicadores.
   - Permitir agregações por prefixo.
   - Permitir filtros utilizando todos os dados dos bilhetes.
   - Permitir o acesso via Intranet.
   - Pesquisar e selecionar componentes de DD e ferramentas.
   - Hardware – Acer Altos 9000, 256 Mbytes Ram, 14 gigabytes de disco rígido.
   - Software
     - Sistema operacional – MS Windows NT Server e MS Windows 95 cliente.
     - SGBD, OLAP, Ferramenta de extração, transformação e carga, metadados, Gerência - MS-SQL Server 7.
     - Ferramenta de Acesso - MS Excel 2000.
5. Projetar a base de dados destino.
   - Script de criação da Tabela destino, veja Anexo 1.

6. Construir o mapeamento de dados, extração, transformação e regras de limpeza.

Veja VBScript de extração, transformação e carga no Anexo 2.

7. Construir agregações, sumarizações, particionamento e distribuição.
8. Construir projeto piloto para um Data Mart incremental utilizando um exato subset da arquitetura corporativa desenvolvida no item 3.
9. Construir um Data Mart incremental adicional.
    - Está fora do escopo pretendido, mas é facilmente alcançável com as etapas anteriores já realizadas. Neste item somente seria utilizado um volume de dados muito superior e com tempos de execução também muito maiores.
10. Amplie para um DD central utilizando a arquitetura corporativa.
    - Fora do escopo deste projeto.
11. Mantenha e administre o DD.
    - Deverá ser feito iterativamente como sugerido pelo método espiral apresentado após a realização do item 10 desta metodologia.

5 Dificuldades na integração automática de dados de bases transacionais

O objetivo deste item é apresentar uma solução que identifica, utilizando inteligência artificial, atributos semelhantes em bases de dados heterogêneas, mas principalmente ressaltar a dificuldade encontrada na automatização deste processo.

A imensidão dos dados força a utilização de técnicas para automatizar os processos de extração dos dados, os quais requerem domínio adicional de conhecimento se é para ser feito inteligentemente. Um sistema para extração de dados, portanto, deve ser capaz de utilizar apropriadamente o domínio do conhecimento em conjunção com a aplicação de algoritmos para automatizar o processo de integração de dados de bases heterogêneas.

Um exemplo de como identificar atributos semelhantes em bases de dados diferentes é apresentado a seguir, com o objetivo de mostrar a dificuldade do processo de extração de dados e o alto grau de complexidade que se pode atingir na tentativa de automatizar este processo [LI 94].

Um procedimento de integração semântica automática utiliza: os nomes de atributos, os valores dos atributos e domínio, a especificação dos campos. A seguir, são detalhados os procedimentos para estas três partes.

5.1 Comparando o nome dos atributos

Sistemas têm sido desenvolvidos para automatizar a integração de base de dados. [HAY 90] analisou o problema de equivalência de atributos com o sistema MUVIS. MUVIS é um sistema baseado em conhecimento para integração de visões. Ele assiste o projetista de base de dados na representação de visões de usuários e integra estas visões em uma visão conceptual global. MUVIS determina o grau de similaridade e dissimilaridade entre dois objetos durante uma fase de pré-integração.

A similaridade e dissimilaridade no MUVIS são baseados na comparação de nome de campos dos atributos. A equivalência dos atributos é determinada comparando os aspectos de cada um e computando um valor de peso para similaridade e dissimilaridade. Uma recomendação é produzida de como efetuar a integração.

A maioria das ferramentas desenvolvidas para assistir os projetistas em estabelecer a correspondência de objetos através da comparação de nomes dos atributos funciona muito bem com homônimos, e os usuários determinam as falsas. No entanto, objetos diferentes podem ter diferentes sinônimos que não são facilmente determinados na inspeção. Com isso, o problema é transferido para construção do dicionário de sinônimos. E até

mesmo um dicionário de sinônimos tem suas limitações, porque é difícil um projetista de base de dados definir um nome de campo utilizando somente palavras que podem ser encontradas num dicionário ou abreviações sem ambigüidades, em alguns casos é difícil utilizar uma palavra única ao invés de uma frase. Atualmente modelos semânticos não permitem capturar o estado do mundo real completamente e interpretações do mundo real mudam com o passar do tempo.

5.2 Comparando valores e domínios utilizando conteúdo dos dados

Outro enfoque [LAR 89] na determinação de equivalência de atributos é através da comparação do domínio dos atributos. Neste processo os relacionamentos e os conjuntos de entidades podem ser integrados baseados no domínio de seus relacionamentos. A determinação destes relacionamentos demanda muito tempo e é tedioso. Se cada esquema tem 100 tipos de entidades e uma média de cinco atributos por tipo de entidade, então 250.000 pares de atributos devem ser considerados. Outro problema é a falta de tolerância a falha. Pequenas quantidades de dados incorretos podem levar a uma conclusão errada dos relacionamentos de domínios.

5.3 Comparando especificações de atributos

Este enfoque [NAV 86] utiliza as especificações de campos para determinar a similaridade e dissimilaridade de um par de atributos. Assume-se que para uma dada base de dados, diferentes projetistas tendem a ter esquemas similares e restrições de projeto semelhante, porque eles devem ter a mesma tecnologia e conhecimento no projeto de uma base dados. Então informação sobre os atributos como comprimento, tipo, e restrições podem ser utilizados como discriminadores para determinar a semelhança que dois atributos equivalentes apresentam. Resultados experimentais mostram que características de esquema são muito efetivas com discriminadores. Esta técnica pode ser utilizada com outros enfoques, como um primeiro passo para eliminar atributos incompatíveis. No entanto, esta técnica necessita de uma base teórica para desenvolvimento de heurísticas para graus de similaridade e dissimilaridade. Outra fraqueza é que os esquemas podem não estar disponíveis.

5.4 Semântica de base de dados

Neste item serão apresentadas as informações utilizadas como discriminadores. Note que não são somente estas informações possíveis, mas são aquelas mais adequadas e disponíveis. Este método tem uma vantagem: a facilidade da descoberta da utilidade dos discriminadores escolhidos é feita

automaticamente; mas não prejudica a utilização de discriminadores adicionais que não dão uma boa base para comparação de atributos. Para entrada do classificador, a informação é mapeada para um vetor de valores com intervalo entre 0 e 1, onde cada item do vetor representa um discriminador. A escolha de discriminadores pode ser feita somente para cada um dos DBMS em questão, e a técnica apresentada permite que se use discriminadores ao invés de outros enfoques apresentados. Uma parte da informação que não se utiliza é o nome do atributo. Isto já foi muito estudado e é complementar a este trabalho. A integração de comparação de nomes com este método é uma área de estudo futuro.

5.5 Especificações de campo

As características de uma especificação de campo ao nível de esquema são: tipo, comprimento e tipo de dados suplementares como formatos, e a existências de restrições (chave primária, chaves estrangeiras, chaves candidatas, valores e intervalo de restrições, proibição de valores nulos, e restrições de acesso). Não é difícil de extrair estas informações de uma base de dados. Muitas bases de dados relacionais armazenam estas informações numa tabela, permitindo consultas SQL para extrair informações. Informações de categoria como tipo de dados necessitam um tratamento especial. Por exemplo, se convertemos tipos de dados para o intervalo [0,1] e atribuímos os valores 1, 0,5 e 0 para tipos de dados data, numérico e caractere respectivamente, então estamos dizendo que uma data está mais próxima de um campo numérico que de caractere. O classificador determina se isto é verdadeiro. Na verdade, entrada de categorias é convertida em um vetor de valores binários (1,0,0 para tipos data, 0,1,0 para tipos numéricos, e 0,0,1 para tipos caractere).

Em alguns casos (como dados de arquivos legados) nós não teremos dados do esquema de forma acessível. Neste caso, temos que realizar um processo manual, ferramentas comerciais como DBStar podem automaticamente extrair informações de esquema de arquivos.

5.6 Conteúdo dos dados

Os conteúdos dos dados de diferentes atributos tendem a ser diferentes mesmo que seu esquema como tipo de dado e comprimento seja o mesmo, porque o seu padrão de dados, distribuição de valor, agrupamento ou outras características são diferentes. Por exemplo, “SSN”e “Balancete” podem ser projetados como campos numéricos de 9 dígitos; não podem ser distinguidos baseados somente em seu esquema e restrições de projeto. No entanto, seus conteúdos de dados são diferentes e portanto seus padrões de dados, distribuição de valores e médias são diferentes. Portanto, examinando o

conteúdo dos dados, a técnica pode corrigir ou aumentar a eficácia dos resultados.

Note que isto não é o mesmo que análise de domínios. Análise de domínio compara o conteúdo completo de cada par de atributos. Esta técnica realiza uma análise de cada atributo para obter um conjunto de características que descrevem os dados. Estas características são divididas em dois tipos: Caractere e Numérico. Tipos que não se enquadram nestes são raro o bastante para que as outras informações sejam suficientes para discriminá-los de outros tipos.

5.7 Padrões de dados para campos tipo caractere

1. A razão do número de caracteres numéricos e o total do número de caracteres. Por exemplo, a razão na placa de carro para estados diferentes serão diferentes. Esta razão para Last_Name ou First_Name deve ser zero. Mas para o campo Stud_Id cujo tipo de dados é designado como caracter (999-99-9999), esta razão é 9/11. Para o campo endereço, esta razão deve ser menor.

2. Razão entre caracteres brancos e total de caracteres: Um campo Last_Name ou First_Name conterá poucos espaços em branco. Um endereço normalmente conterá alguns espaços em branco.

3. Estatísticas no tamanho: Além do simples tamanho do campo, comparamos a média, variância, e coeficiente de variância da parte utilizada ao tamanho máximo. Um campo de ID tipicamente utilizará todo campo mas um nome com certeza variará bem mais.

5.8 Padrões de dados para campos numéricos

Para campos numéricos, podemos utilizar análise estatística dos dados como um discriminador. A estatística utilizada é:

1. Média: A média é uma das características para os dados cujos tipos de dado são números. Por exemplo, as contas de poupança e contas correntes normalmente terão médias diferentes. O peso médio de navios e carros deve ser diferente.

2. Variância: A variância é a medida de variabilidade de um conjunto de valores. Ele coloca um peso maior em observações que estão mais afastadas do padrão, porque ele eleva ao quadrado o desvio-padrão.

3. Coeficiente de Variação (CV): este é a raiz quadrada da variância dividido pela média.

5.9 Método da integração semântica

Os discriminadores provêem uma boa quantidade de informações sobre as características dos atributos. No entanto, é difícil determinar quais os discriminadores que auxiliarão e quais serão apenas ruído.

Redes neurais emergiram como uma técnica poderosa no reconhecimento de padrões. Redes neurais podem aprender as similaridades entre dados diretamente de instâncias de dados, e empiricamente inferir soluções a partir de dados sem conhecimento prévio de regularidades. A vantagem de utilizar redes neurais para determinação de equivalência de atributos sobre métodos com regras fixas é [LI 94]:

1. Redes neurais podem realizar tarefas como classificação e generalização sem ser alimentadas com regras, desde que tenham sido previamente treinadas, e não programadas.

2. Os pesos designados podem ser ajustados dinamicamente de acordo com os dados de entrada.

3. Redes neurais podem generalizar por causa de sua habilidade de responder corretamente a dados não utilizados no treinamento.

Primeiro, a informação disponível de uma base individual é utilizada como dado de entrada para um algoritmo para mapa auto-organizado para categorizar os atributos. Segundo, a resposta do classificador é utilizada para treinar os dados para o aprendizado das categorias e algoritmo de reconhecimento. O algoritmo treinado de reconhecimento então determina a similaridade entre pares de atributos de bases de dados diferentes.

Procedimento de integração semântica

A Figura 21 mostra o diagrama de integração semântica. A única ação humana se dá para designar o limiar e conferir os resultados.

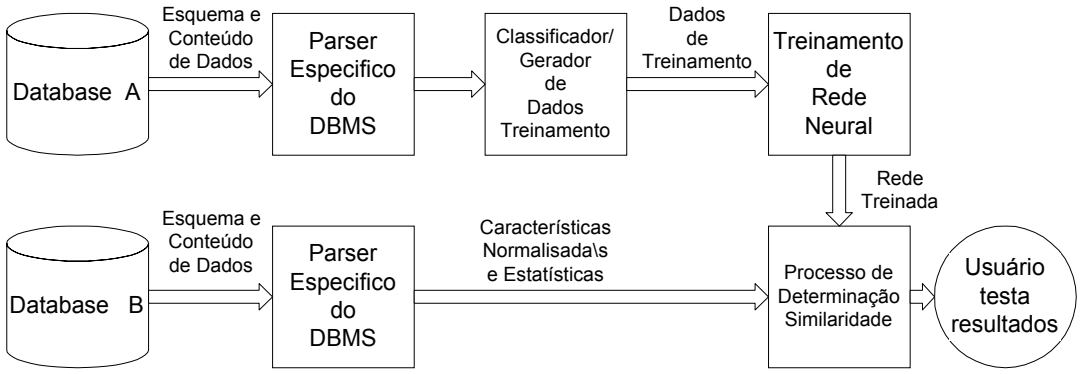

FIGURA 21 – Procedimento de integração semântica

Os passos necessários para o procedimento de integração semântica são:

Passo 1: Utilização de utilitários específicos de cada DBMS para obter informações de cada BD a ser integrado. O sistema transforma esta informação em um formato comum. O resultado destes utilitários deve incluir a informação dos esquemas, estatísticas dos valores dos dados, e tipos de características disponíveis.

Passo 2: O sistema cria um mapa auto-organizado com N nós na camada de entrada. Utiliza informação da base de dados A como entrada para o mapa auto-organizado recém-criado. Esta rede classifica os atributos na base de dados A em M categorias. M não é predeterminado; o número de categorias (grupo) criadas depende do valor do limiar estabelecido pelo treinador do sistema. M deve ser o número de atributos distintos na BD, isto é, os atributos que não têm chave estrangeira de outras BD. O resultado desta etapa é o número de categorias (M) e o peso dos grupos (M vetores N pesos). O peso dos grupos centros então são marcados como dados de treinamento para treinar a rede criada no passo 3.

Passo 3: O sistema cria uma rede de três camadas de aprendizado por propagação para trás com N nós como camada de entrada, M nós como camada de saída e (N+M)/2 nós na camada escondida. Durante o treinamento, a rede muda o peso de forma que cada nó na camada de saída representa um grupo.

Passo 4: A entrada para a rede treinada no passo 3 são as informações dos atributos de outra BD (base de dados B). A rede então dá a similaridade entre os atributos da BD B e cada categoria da BD A.

Passo 5: Os usuários do sistema verificam e confirmam a saída da rede treinada.

Neste exemplo de sistema de identificação de atributos semelhantes em bases diferentes se faz uso de redes neurais para analisar dados relativos aos atributos que se quer analisar e obter graus de similaridade para, finalmente, um analista determinar se são ou não semelhantes.

O objetivo deste item no trabalho é mostrar que o processo de extração é o mais complexo de se realizar, especialmente quando se lida com base de sistemas legados.

6 Conclusão

Os modelos de dados tradicionais funcionam efetivamente no mundo de sistemas de operação, sendo adotados na maioria de aplicações com uso de BD relacionais. Devido ao sucesso deste modelo, os primeiros sistemas de suporte a decisão surgiram utilizando este mesmo modelo e aplicados diretamente sobre as bases transacionais. Mas com o crescimento destas bases o desempenho ficou muito pobre, os sistemas difíceis de usar e manter, e impactando no desempenho dos sistemas transacionais.

A metodologia proposta utiliza a arquitetura de Data Marts incrementais para o desenvolvimento de DD. Com isto temos o melhor de dois mundos, inicia-se pequeno com custo baixos e retorno em curto prazo, mas não se perde a visão corporativa dos dados, permitindo a união posterior de Data Marts criados em um DD centralizado.

Um ponto crucial na adoção desta arquitetura é a escolha de plataforma de software e hardware que tenha a devida escalabilidade, isto é, que permita iniciar o projeto em uma plataforma, por exemplo, com equipamentos Intel/NT e cresça até servidores Unix multiprocessados com processamento paralelo, sem a necessidade de troca de SGBD, alteração nas aplicações ou trabalhos exaustivos na migração de plataforma. Outro ponto também importante é a arquitetura corporativa definida, descrevendo de forma clara e única todas as entidades do DD, a forma como é utilizada, fórmulas e períodos de tempo que permitam um entendimento claro do processo realizado, disponibilizado através do uso de metadados.

O uso de modelagem dimensional nos projetos de DD visa essencialmente a melhora do desempenho das consultas, e isto se consegue utilizando um número reduzido de tabelas (dimensões e fatos), altamente denormalizadas, que evitam ter que realizar um grande número de uniões durante uma consulta. Este modelo tem outra vantagem, ao apresentar as informações num formato que tem paralelo com o mundo de negócios, facilitando o uso do usuário final. Além disso, a estrutura simples deste modelo facilita a manutenção e expansão do DD.

A contribuição deste trabalho é de apresentar uma metodologia que minimize as causas de insucesso em implementações de DD, elaborada a partir de análises de várias arquiteturas e metodologias, sendo uma metodologia alternativa aqui apresentada. Mas existem vários itens desta metodologia que podem gerar trabalhos adicionais com grande importância para o resultado de uma implantação de DD. Um deles seria a avaliação do modelo de dados dimensional e a influência de agregações e sumarizações no desempenho de consultas, outro seria a análise de ferramentas de extração, de forma a diminuir ao máximo a intervenção via programação nesta tarefa, e por último seria a análise das soluções de SGBD no mercado para DD e a

identificação para quais situações estes SGBD apresentariam o melhor resultado.

A prova de conceito realizada está dirigida a atender à necessidade específica da área de análise de desempenho, e as ferramentas Microsoft não possuem a escalabilidade necessária para atingirmos um DD centralizado como preconiza este trabalho. Mas o objetivo principal deste projeto é utilizar conceitos apresentados neste trabalho e utilizar uma ferramenta de mercado – a disponibilizada para este piloto foi o MS-SQL7 e suas ferramentas integradas –, mas esta ferramenta, por suas limitações, só seria recomendável para a área usuária envolvida, já que só se interessa por um volume pequeno de dados e um histórico de duas semanas para suas análises. Já a área de marketing, pelo volume de dados e o longo prazo de histórico de que necessita inviabilizaria sua utilização, principalmente pela falta de escalabilidade. Para se ter uma idéia, se fôssemos criar o DD/DM utilizando todos os bilhetes gerados em um mês, estaríamos com 120 a 150 milhões de registros com 200 bytes cada, ou seja, aproximadamente 22 gigabytes/mês. Para uma solução de Gerência de Atendimento de Clientes teríamos que trabalhar com este volume de dados acrescidos de dados externos.

Algumas questões representam verdadeiros desafios na implantação de um *Data Warehouse* e gostaria de realçá-los:

- Integração de dados e metadados de várias fontes e fornecedores de soluções;
- Qualidade dos dados: limpeza e refinamentos;
- Sumarização e agregação de dados;
- Sincronização das fontes com o *Data Warehouse* para assegurar a atualidade deste;
- Problemas de desempenho relacionados ao compartilhamento do mesmo ambiente computacional e recursos humanos para abrigar e atender os BDs corporativos operacionais e o *Data Warehouse*;
- Justificar a necessidade de investimento em plataformas de arquitetura escalável e paralela.

Anexo 1 Script de criação da tabela-Destino

```
/* Microsoft SQL Server - Scripting                */
/* Server: ANDROMEDA                               */
/* Database: Bilhetes1                             */
/* Creation Date 25/01/2000 16:05:18               */

/****** Object:Table [dbo].[BilhetesTSO2] Script Date: 25/01/2000 16:04:13 ******/
if exists (select * from sysobjects where id = object_id(N'[dbo].[BilhetesTSO2]')
and OBJECTPROPERTY(id, N'IsUserTable') = 1)
drop table [dbo].[BilhetesTSO2]
GO

/******Object:Table [dbo].[BilhetesTSO2]  Script Date: 25/01/2000 16:04:19 ******/
CREATE TABLE [dbo].[BilhetesTSO2] (
        [Bilhetador] [varchar] (7) NULL ,
        [Assinante A] [numeric](28, 0) NULL ,
        [Parte Tarifada] [varchar] (20) NULL ,
        [cat] [varchar] (3) NULL ,
        [Assinante B] [numeric](28, 0) NULL ,
        [FDS] [varchar] (3) NULL ,
        [Duracao] [varchar] (7) NULL ,
        [Data_Hora] [datetime] NULL ,
        [CO] [varchar] (2) NULL ,
        [RecNumber] [varchar] (11) NULL ,
        [Rota_Entrada] [varchar] (5) NULL ,
        [Juntor_Entrada] [varchar] (5) NULL ,
        [Rota_Saida] [varchar] (5) NULL ,
        [Juntor_Saida] [varchar] (5) NULL ,
        [Call_Type] [varchar] (4) NULL ,
        [Call_Status] [varchar] (3) NULL ,
        [Assinante_A_Discado] [varchar] (15) NULL ,
        [Assinante_B_Discado] [varchar] (15) NULL ,
        [Prestadora] [varchar] (7) NULL ,
        [Indicador] [varchar] (15) NULL ,
        [Indicador FCN7] [varchar] (5) NULL ,
        [Indicador T6] [varchar] (5) NULL ,
        [ELemento_Rede_N1B] [int] NULL ,
```

```
        [Elemento_Rede_N2B] [int] NULL ,
        [Elemento_Rede_N3B] [int] NULL ,
        [OK] [int] NULL ,
        [NR] [int] NULL ,
        [OU7] [int] NULL ,
        [CO2] [int] NULL ,
        [CO1] [int] NULL ,
        [OU5] [int] NULL ,
        [OU3] [int] NULL ,
        [OU8] [int] NULL ,
        [CO3] [int] NULL ,
        [LO] [int] NULL ,
        [DSC] [int] NULL ,
        [CO0] [int] NULL ,
        [Bilhete] [varchar] (90) NULL
) ON [PRIMARY]
GO
```

Anexo 2 Script de transformação do arquivo-texto.

```
'**********************************************************************
' Visual Basic Transformation Script
' Copy each source column to the
' destination column
'**********************************************************************

Function Main()
        DTSDestination("Bilhetador") = DTSSource("Col001")
        DTSDestination("Assinante A") = DTSSource("Col002")
        DTSDestination("Parte Tarifada") = DTSSource("Col003")
        DTSDestination("cat") = DTSSource("Col004")
        DTSDestination("Assinante B") = DTSSource("Col005")
        DTSDestination("FDS") = DTSSource("Col006")
        DTSDestination("Duracao") = DTSSource("Col007")
        DTSDestination("Data_Hora") = DTSSource("Col008")
        DTSDestination("CO") = DTSSource("Col009")
        DTSDestination("RecNumber") = DTSSource("Col010")
        DTSDestination("Rota_Entrada") = DTSSource("Col011")
        DTSDestination("Juntor_Entrada") = DTSSource("Col012")
        DTSDestination("Rota_Saida") = DTSSource("Col013")
        DTSDestination("Juntor_Saida") = DTSSource("Col014")
        DTSDestination("Call_Type") = DTSSource("Col015")
        DTSDestination("Call_Status") = DTSSource("Col016")
        DTSDestination("Assinante_A_Discado") = DTSSource("Col002")
        DTSDestination("Assinante_B_Discado") = DTSSource("Col017")
        DTSDestination("Prestadora") = DTSSource("Col018")
        DTSDestination("Indicador") = DTSSource("Col019")
        DTSDestination("Indicador FCN7") = DTSSource("Col020")
        DTSDestination("Indicador T6") = DTSSource("Col021")
        DTSDestination("ELemento_Rede_N1") = DTSSource("Col022")
        DTSDestination("Elemento_Rede_N2") = DTSSource("Col023")
        DTSDestination("Elemento_Rede_N3") = DTSSource("Col024")
        DTSDestination("Bilhete") = DTSSource("Col025")

        If DTSSource("Col019") = "ok" then
                DTSDestination("OK")=1
```

```
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0

elseif DTSSource("Col019") = "nr" then
        DTSDestination("OK")=0
        DTSDestination("NR")=1
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
        elseif DTSSource("Col019") = "ou7" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=1
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
```

```
elseif  DTSSource("Col019") = "co2" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=1
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
elseif  DTSSource("Col019") = "co1" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=1
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
        elseif  DTSSource("Col019") = "ou5" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=1
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
```

```
        DTSDestination("CO0")=0
elseif  DTSSource("Col019") = "ou3" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=1
        DTSDestination("OU8")=0
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
elseif  DTSSource("Col019") = "ou8" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=1
        DTSDestination("CO3")=0
        DTSDestination("LO")=0
        DTSDestination("DSC")=0
        DTSDestination("CO0")=0
elseif  DTSSource("Col019") = "co3" then
        DTSDestination("OK")=0
        DTSDestination("NR")=0
        DTSDestination("OU7")=0
        DTSDestination("CO2")=0
        DTSDestination("CO1")=0
        DTSDestination("OU5")=0
        DTSDestination("OU3")=0
        DTSDestination("OU8")=0
        DTSDestination("CO3")=1
        DTSDestination("LO")=0
```

```
            DTSDestination("DSC")=0
            DTSDestination("CO0")=0
        elseif  DTSSource("Col019") = "lo" then
            DTSDestination("OK")=0
            DTSDestination("NR")=0
            DTSDestination("OU7")=0
            DTSDestination("CO2")=0
            DTSDestination("CO1")=0
            DTSDestination("OU5")=0
            DTSDestination("OU3")=0
            DTSDestination("OU8")=0
            DTSDestination("CO3")=0
            DTSDestination("LO")=1
            DTSDestination("DSC")=0
            DTSDestination("CO0")=0
        elseif  DTSSource("Col019") = "dsc" then
            DTSDestination("OK")=0
            DTSDestination("NR")=0
            DTSDestination("OU7")=0
            DTSDestination("CO2")=0
            DTSDestination("CO1")=0
            DTSDestination("OU5")=0
            DTSDestination("OU3")=0
            DTSDestination("OU8")=0
            DTSDestination("CO3")=0
            DTSDestination("LO")=0
            DTSDestination("DSC")=1
            DTSDestination("CO0")=0
        elseif  DTSSource("Col019") = "co0" then
            DTSDestination("OK")=0
            DTSDestination("NR")=0
            DTSDestination("OU7")=0
            DTSDestination("CO2")=0
            DTSDestination("CO1")=0
            DTSDestination("OU5")=0
            DTSDestination("OU3")=0
            DTSDestination("OU8")=0
            DTSDestination("CO3")=0
```

```
            DTSDestination("LO")=0
            DTSDestination("DSC")=0
            DTSDestination("CO0")=1
        end if

        Main = DTSTransformStat_OK
End Function
```

Bibliografia

[BAR 96] BARQUIM, R.; EDELSTEIN, H. **Planning and Design the Data Warehouse.** Upper Saddle River: Prentice Hall, 1997.

[BRA 96] BRACKET, M. **The Data Warehouse Challenge-Taming the Chaos.** New York:John-Wiley & Sons, 1996.

[BRAC 96] BRACHMAN, R. J. et al. Mining Business Databases. **Communications of the ACM**, Special Issue on Data Mining, New York, v. 39, n. 11, p.42-48, Nov. 1996.

[DAT 77] DATE, C. J. **An Introduction to Database Systems**. Reading: Addison-Wesley, 1977.

[DAT 99] DATA WAREHOUSE INSTITUTE. **10 Most Common Errors.** Disponível em : <http://www.datawarehousing.com>. Acesso em: 10 jun. 1999.

[FLA 97] FLANAGAN, T.; SAFDIE, E. **Data Warehouse Technical Guide.** Disponível em : <http://www.sybase.com/products/dataware/techguide.html>. Acesso em : 20 mar. 1997.

[HAC 97] HACKNEY, D. **Understanding and Implementing Successful Data Marts.** Reading: Addison-Wesley, 1997.

[HAM 95] HAMMER, J. et al. **The Stanford Data Warehousing Project.** Disponível em: <ftp://pub/widom/1995/warehouse-research.ps>.Acesso em : 30 maio 1997.

[HAR 96] HARJINDER, G.; RAO, P. C**. The official Guide to Data Warehousing.** Indianapolis: Que Corporation, 1996.

[HER 96] HERRMANN, S. **Estudo de Mineração de Dados:** trabalho individual**.** Porto Alegre: CPGCC da UFRGS, 1996.

[HAY 90] HAYNE, S.; RAM, S. Multi-user view integration system (MUVIS). In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 6., 1990. **Proceedings …** Los Angeles: IEEE Computer Society, 1990. p.402-409.

[INM 94] INMON, W.; HACKATHORN, R. **Using the Data Warehouse.** New York: John-Wiley & Sons, 1994.

[INM 96] INMON, W. **Building the Data Warehouse.** New York: John-Wiley & Sons, 1996.

[KIM 96] KIMBALL, R. **The Data Warehouse Toolkit.** New York: John-Wiley & Sons, 1996.

[KIM 98] KIMBALL, R. **The Lifecycle Toolkit**. New York: John-Wiley & Sons, 1998.

[LI 94] LI, W.; CLIFTON, C. Semantic Integration in Heterogeneous Database Using Neural Networks. In: VLDB CONFERENCE, 20.,1994. **Proceedings…** Santiago de Chile: Morgan Kaufmann, 1994. p.1-12.

[LAR 89] LARSON, J. A.; SHAMKANT, N. B.; ELMASRI, R. A theory of attribute equivalence in database with application to schema integration. **IEEE Transactions on Software Engineering**, New Jersey, v.15, p. 449-463, Apr. 1989.

[MAR 98] MARTIN, J**. James Martin Data Warehousing Process.** Disponível em: <http://www.jamesmartin.com>.Acesso em: 15 out. 1998.

[MAT 93] MATHEUS, C.; CHAN, P. K.; PIATTETSKY-SHAPIRO, G. System for Knowledge Discovery in Databases. **IEEE Transactions on Knowledge and Data Engineering**, New Jersey, v. 5, Dec. 1993.

[MAT 97] MATTISON, R. M. **Data Warehousing and data mining for telecommunications.** Boston: Artech House, 1997. 270p.

[MIC 99] MICROSOFT. **Business Intelligence.** Disponível em : <http:// www.microsoft.com/brasil/sql/business.htm>. Acesso em: 10 out. 1999.

[MOR 96] MORIARTY, T.; GREENWOOD, R. P. Data's Quest from Source to Query. **Database Programming & Design**, San Mateo,v. 9, n. 10, p. 55-61, Oct. 1996.

[NAV 86] NAVATHE, S. et al. P. Integrating user views in database design. **Computers**, New York,v. 19, n. 01, p. 50-62, Jan. 1986.

[NCR 99] NCR DO BRASIL. Como Construir um Data Warehouse. **Workshop NCR Data Warehouse**, São Paulo, v. 1, n. 6, p. 1-23, Mar. 1999.

[ORA 98] ORACLE. **Oracle Data Mart Suite Cookbook-Release 2**. Nashua, 1998.

[SIM 98] SIMON, A. **90 days to the Data Mart**. Reading: John Wiley & Sons, 1998.

[TEC 95] TECHNOLOGY RESEARCH INSTITUTE. **Data Warehousing and Decision Support Systems in Telecommunications**. Disponível em: <http://www.technology-research.com >. Acesso em: 25 out. 1995.

[WAT 95] WATTERSON, K. Ferramentas para Mineração de Dados. **Revista Byte**, São Paulo, v. 20, n. 10, p. 81-88, out. 1995.

[WEL 96] WELDON, J. L. Data Mining and Visualization. **Database Programming and Design**, San Mateo, v. 9, n. 5, p. 29-35, May 1996.

[WID 95] WIDOM, J. **Research Problems in Data Warehousing**. Disponível em: <http://dbpubs.stanford.edu/pub/1995-24>. Acesso em: 07 jul. 2000.

[ZYT 91] ZYTKOW, J. M.; BAKER, J. **Knowledge Discovery in Databases.** Cambridge: AAI/MIT, 1991. p. 31-45.