

JUNIOR CEZAR VIDOTTI

**PROJETO DE UM DATA WAREHOUSE:
ANÁLISE DE CUSTO/BENEFÍCIO**

CUIABÁ/MT
MARÇO DE 2001

PROJETO DE UM DATA WAREHOUSE: ANÁLISE DE CUSTO/BENEFÍCIO

JUNIOR CEZAR VIDOTTI

TRABALHO MONOGRÁFICO PARA CUMPRIMENTO DE CRÉDITOS DA DISCIPLINA DE PROJETO SUPERVISIONADO, DO CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO SOB A ORIENTAÇÃO DO PROF. DSC. EINSTEIN LEMOS DE AGUIAR.

CUIABÁ/MT
MARÇO DE 2001

VIDOTTI, JUNIOR CEZAR

VI, 35 pág. 29,7 cm (Depto. de Ciência da Computação, 2000.2)

Trabalho Monográfico - Universidade Federal do Mato Grosso, UFMT

1. Data Warehouse
2. Data Mart
3. DSS
4. Processamento Analítico

JUNIOR CEZAR VIDOTTI

**PROJETO DE UM DATA WAREHOUSE:
ANÁLISE DE CUSTO/BENEFÍCIO**

Prof. Dsc. Einstein Lemos de Aguiar

Prof. Evandro César Freiburger

Prof. Msc. Divino

Resumo do trabalho monográfico apresentada para o Curso de Bacharelado em Ciência da Computação / UFMT como parte do requisitos necessários para o cumprimento de créditos da disciplina de Projeto Supervisionado.

DATA WAREHOUSE: UMA ANÁLISE DE CUSTO/BENEFÍCIO

Junior Cezar Vidotti
Março de 2001

Orientador: Prof. Dsc. Einstein Lemos de Aguiar
Departamento de Ciência da Computação
Universidade Federal de Mato Grosso

Nossa sociedade passa atualmente por uma fase de extrema competição causada pela globalização. A vantagem da informação precisa e rápida extraída dos dados corriqueiros pode realmente fazer a diferença entre uma empresa e outra. O uso de um data warehouse e de ferramentas de análise de dados vêm ao encontro com os desejos dos executivos e gerentes de extrair o máximo de informação possível dos dados disponíveis.

A importância deste trabalho reside no fato das empresas necessitarem de informação rica e abundante e não saber como fazer para obtê-la, já que seus relatórios analíticos são insuficientes.

Mas qual é o custo deste benefício? E quais as justificativas precisas para estes custos?

Sendo assim, se torna imprescindível o estudo do custo/benefício da implantação de um data warehouse e do uso de ferramentas de processamento analítico, com o objetivo de tornar as empresas mais competitivas, de posse de informações que realmente façam a diferença.

AGRADECIMENTOS

Ao professor Einstein especialmente, por me orientar e prestar apoio, tanto psicologicamente quanto cientificamente na elaboração deste trabalho.

Ao meus pais, Ademar e Marileide, pelas pessoas maravilhosas que são.

À minha noiva, Fabiane, pela sua paciência e compreensão.

E a todos os que contribuíram, direta ou indiretamente, para a conclusão deste trabalho.

SUMÁRIO

INTRODUÇÃO.....	3
INTRODUÇÃO.....	3
CAPÍTULO I.	
HISTÓRICO, CONCEITOS E DEFINIÇÕES.....	5
I.1.HISTÓRICO DOS SISTEMAS DE INFORMAÇÃO.....	5
I.2.TIPOS DE PROCESSAMENTO DE DADOS.....	6
I.2.1.OLTP – AS APLICAÇÕES DO NEGÓCIO.....	6
I.2.2.OLAP – APLICAÇÕES SOBRE O NEGÓCIO.....	7
I.3.DATA WAREHOUSE.....	8
I.4.DATA MART.....	10
CAPÍTULO II.	
ASPECTOS DE PROJETO DE DATA WAREHOUSE.....	13
II.1.O FLUXO DE DADOS.....	13
II.2.OS METADADOS.....	14
II.3.O DATA WAREHOUSE E O MODELO DE DADOS.....	15
II.4.O DATA WAREHOUSE E O SGBD.....	16
II.5.A GRANULARIDADE.....	17
II.6.O PROCESSO DE ETL.....	18
II.7.TECNOLOGIAS DE APLICAÇÕES SOBRE O DATA WAREHOUSE E OS DATA MARTS.....	19
II.7.1.DSS.....	20
II.7.2.DATA MINING.....	20
II.7.3.CRM.....	21
CAPÍTULO III.	
CUSTOS DO DATA WAREHOUSE.....	22
III.1.CLASSIFICANDO OS CUSTOS.....	22
III.1.1.CUSTOS INICIAIS.....	22
III.1.2.CUSTOS PERIÓDICOS.....	24
III.2.FATORES DE VARIAÇÃO DO CUSTO.....	25

CAPÍTULO IV.

JUSTIFICAÇÃO DOS CUSTOS DO DATA WAREHOUSE.....	27
IV.1.ACESSO À INFORMAÇÃO.....	27
IV.1.1.FAZENDEIROS X EXPLORADORES.....	29
IV.2.DATA WAREHOUSE E DATA MARTS.....	29
IV.3.VANTAGENS NO NEGÓCIO.....	30
CONCLUSÃO.....	32
CONCLUSÃO.....	32
ANEXO I. FONTES DE INFORMAÇÃO NA INTERNET.....	34
ANEXO I. FONTES DE INFORMAÇÃO NA INTERNET.....	34
GLOSSÁRIO.....	35
GLOSSÁRIO.....	35
BIBLIOGRAFIA.....	36
BIBLIOGRAFIA.....	36

INTRODUÇÃO

Nossa sociedade passa atualmente por uma fase de extrema competição causada pela globalização. A vantagem da informação certa extraída de dados aparentemente inócuos pode realmente fazer a diferença entre uma empresa e outra.

Já foram vistos casos, no mundo dos negócios, de empresas milionárias irem a bancarrota em questão de dias (ou horas) por mudanças no panorama mundial. Algumas vezes isso poderia ser evitado realizando-se uma previsão momentos antes e determinar, se possível, soluções para evitar tais acontecimentos.

Uma empresa precisa de informações sintéticas e precisas sobre a realidade do negócio e as tendências futuras. O presente e o passado de uma empresa estão contidos em seus dados, gerados a todo instante pelas suas aplicações de negócio.

Mas, devido à diversidade de sistemas que estão em uso em uma empresa ao mesmo tempo, é muito difícil, custoso e demorado realizar uma análise integrada dos dados de todos esses sistemas. E, esse tipo de informação é requerido a todo instante por executivos e gerentes para que tomem decisões a níveis estratégicos e de controle de uma empresa.

Sendo difícil, custoso e demorado integrar, conciliar e reunir dados de várias fontes para criar este tipo de relatório, começou-se a pensar no conceito de data warehouse (armazém de dados, em inglês). Um data warehouse é um conjunto integrado de dados que extrai e reúne informações de todos os sistemas da empresa. A partir de um data warehouse, podemos construir sistemas de suporte à decisão, ferramentas de mineração de dados (descoberta de padrões entre os dados), entre outras aplicações específicas de análise de negócios.

O data warehouse é um dos passos iniciais na mudança da forma como as empresas tomam decisões e analisam informações. O mercado tornou-se mais competitivo e as empresas (componentes do mercado) devem acompanhar tal tendência. Uma possibilidade de aumentar a competitividade e diminuir custos é implantar um data warehouse e seus aplicativos para analisarem o negócios.

Antes de implantar um data warehouse, um executivo fará a seguinte pergunta: “Qual o custo deste tal de data warehouse e o quê que ele trará de bom para minha empresa?”

Sabendo que esta pergunta pode ser a causa mortis de qualquer projeto dentro de uma empresa propomos, como nosso objetivo, detectar quais os custos e benefícios que um data warehouse pode trazer para as empresas que necessitem da implantação de um.

Para que possamos atingir nosso objetivo, precisaremos rever alguns conceitos e definições sobre data warehouse e aspectos sobre o projeto do data warehouse para que possamos, então, atacar o problema proposto, realizar a análise do custo-benefício do data warehouse.

CAPÍTULO I.

HISTÓRICO, CONCEITOS E DEFINIÇÕES

Antes de iniciarmos o desenvolvimento do trabalho em si, forneceremos um panorama geral da evolução dos sistemas de informação, conceitos e definições acerca do data warehouse. Isso será importante para entendermos a importância do data warehouse para as empresas de hoje.

Começaremos estudando a história dos sistemas de informação, descreveremos os tipos de processamento decorrente da evolução desses sistemas e, então, explicaremos os conceitos de data warehouse e data marts.

I.1. Histórico dos sistemas de informação

Em 1946, com o início da guerra fria, foi criado o ENIAC, um grande computador capaz de realizar cerca de 200 operações por segundo movido por 18.000 válvulas, criado pela Universidade da Pensilvânia.

Cerca de uma década depois foi criado o transistor, capaz de executar praticamente a mesma tarefa das válvulas, porém com menor gasto de energia e com um tamanho muito menor que uma válvula.

Na década de 60 foram produzidos computadores menores e mais baratos baseados em transistores, tornando viável o uso da computação no meio comercial. O meio de armazenamento eram fitas magnéticas, podendo armazenar grandes quantidades de dados a baixo custo, apesar do inconveniente da lenta leitura seqüencial – os dados precisavam ser procurados na fita magnética, o que dificultava a recuperação de um dado no meio ou final da fita.

Na década de 70 foi desenvolvida a tecnologia DASD (Direct Access Storage Device, ou dispositivo de armazenamento de acesso direto), tecnologia a qual originaram os discos rígidos. Como esta nova tecnologia possuía leitura aleatória, possibilitou-se o desenvolvimento dos DBMS (Data Base Management System). Um DBMS (ou Sistema Gerenciador de Banco de Dados, SGBD, em português) é um programa que controla a gravação e leitura de dados em um sistema. É possível, através dos SGBDs, centralizar os dados para facilitar o processamento e tornar a gravação/leitura dos dados no meio físico transparente para as aplicações. Criou-se, assim, o conceito de “banco de dados”.

Banco de dados era definido, assim, como uma única fonte de dados para todo o processamento. Através disso, o computador adquiriu uma nova forma de atuação dentro da empresa, passando a suprir a empresa com dados de uma forma nunca vista antes. A partir daí o computador passou a não ser apenas um instrumento de cálculo e previsão, e sim uma ferramenta de análise e operação do negócio e uma forte vantagem competitiva. Pode-se considerar que a “era da informação” estava prestes a nascer.

Nos anos 80 e 90, com a criação dos banco de dados relacionais, das linguagens de 4ª. geração a idéia do computador como ferramenta de negócio indispensável consolidou-se.

Mas foi com a invenção do computador pessoal, e como conseqüência o barateamento dos semi-condutores, que os sistemas de informação chegaram à mesa dos funcionários da maior parte das empresas.

1.2. Tipos de processamento de dados

A evolução dos computadores também acarretou uma grande mudança no enfoque dos sistemas da empresa. Passou-se a dividir os sistemas em dois grandes ramos:

- OLTP – On Line Transactional Processing
- OLAP – On Line Analytical Processing

1.2.1. OLTP – As aplicações do negócio

As aplicações baseadas em OLTP (processamento transacional) são as aplicações do negócio. As aplicações do negócio são os sistemas que a empresa utiliza em seu dia-a-dia a nível operacional. Tais aplicações oferecem suporte às regras de negócio que a empresa utiliza, sendo freqüentemente chamados de sistemas de produção ou aplicações operacionais.

Os sistemas de produção trabalham com dados chamados dados operacionais ou transacionais. Os dados operacionais são estruturados de acordo com a regra de negócio da empresa. Esses sistemas estão preparados para suprir as necessidades de operação no dia-a-dia da empresa, mas não suportam a análise integrada dos dados.

1.2.2. OLAP – Aplicações sobre o negócio

As aplicações baseadas em processamento analítico (OLAP) são comumente chamadas de aplicações sobre o negócio. As aplicações sobre o negócio analisam as informações obtidas através das aplicações operacionais, ajudando executivos e altos gerentes a interpretar mudanças na realidade de negócio e, assim, mudar as regras de negócio da empresa.

Os dados usados por essas aplicações são chamados de dados analíticos. Os dados analíticos são nada mais que dados operacionais otimizados para a análise, e não para transações.

Não é certo interpretar que não se possa extrair relatórios e análise de aplicação operacionais. Isto é possível, mas demanda grande esforço para reunir, integrar e apresentar cada relatório. Já nos sistemas de suporte a decisão esse trabalho se resume a pesquisar dados já reunidos e integrados em uma única base de dados.

Quadro 1.

A mudança de enfoque dos dados	
<ul style="list-style-type: none"> • DADOS OPERACIONAIS • Baseados em aplicações • Detalhados recentes • Comunidade funcional • Performance • Acesso a um registro por vez • Transações • Necessidade de alta disponibilidade • Alta probabilidade de acesso 	<ul style="list-style-type: none"> • DADOS ANALÍTICOS • Baseados em assuntos ou negócios • Detalhados históricos, resumidos e refinados • Comunidade gerencial • Versatilidade • Acesso a um conjunto de registros por vez • Análise • Necessidade de baixa disponibilidade <p style="text-align: center;">Baixa ou média probabilidade de acesso</p>

Fonte: INMON (1992), modificado.

I.3. Data warehouse

“Um data warehouse é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais”.

W.H. INMON (1992: 33)

De acordo com R. KINBALL (1996: 310), data warehouse é “uma cópia dos dados transacionais estruturado para relatórios e análise”. Isto é, data warehouse nada mais é que os dados operacionais, usados no dia a dia da empresa pelas aplicações do negócio, integrados e transformados de uma forma que possamos extrair informações integradas sobre a corporação e seus processos.

O data warehouse é um repositório de dados preparados para análise prontos para serem acessados. É importante ressaltar que o data warehouse é somente a estrutura na qual as ferramentas de processamento analítico, e não tais ferramentas em si.

A primeira característica marcante de um data warehouse é a orientação ao assunto. Os banco de dados operacionais baseiam-se nas aplicações da empresa, sendo projetado para agir da melhor forma para um melhor desempenho na aplicação operacional. Já o data warehouse atua sob outro paradigma. Os dados devem estar organizados de acordo com os principais assuntos da empresa, como clientes e pedidos.

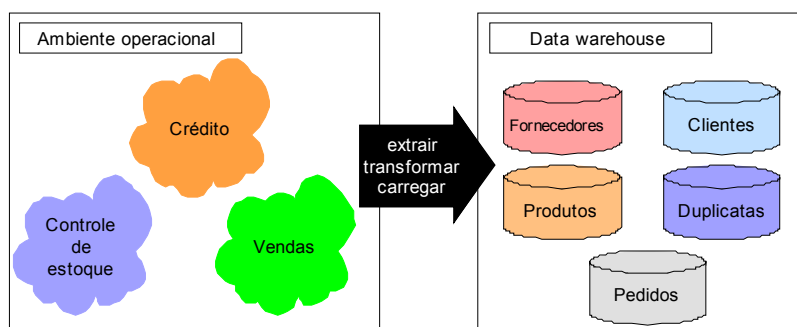


Figura 1. A orientação ao assunto é um fator fundamental do data warehouse.

Outro aspecto de um data warehouse é a forte integração de dados. Nas diversas fontes de dados operacionais, talvez projetados por pessoas diferentes ou em épocas diferentes, os mesmos dados podem aparecer de formas diferentes. Praticamente não há uma padronização e convenção dos dados, medidas e formatos de dados no ambiente operacional, tratando-se de aplicações diferentes. No data warehouse, é adotado um padrão para dados, medidas e formatos. Ou seja, os dados estão integrados, sob uma única forma.

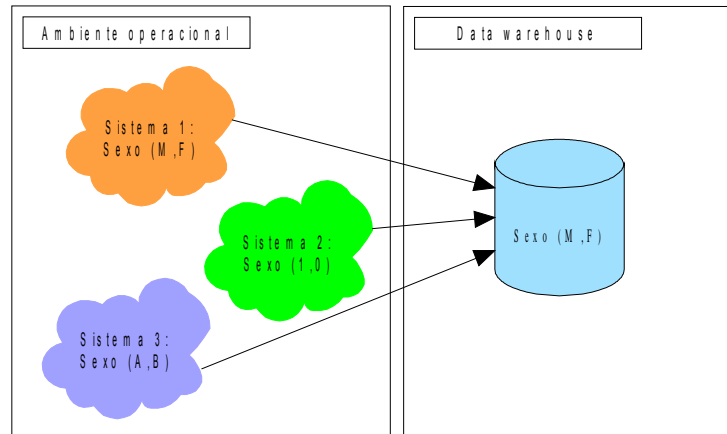


Figura 2.A integração dos dados para a uniformidade e facilidade de acesso.

Pode-se também dizer que os dados no data warehouse não são voláteis. No ambiente operacional, os dados são acessados por muitas pessoas ao mesmo tempo. Quando duas ou mais pessoas manuseiam o mesmo dado, podem acontecer falhas e inconsistências no sistema. O dado recuperado há 1 minuto atrás já pode não estar atualizado, pois alguém pode tê-lo alterado ou apagado. Quando uma pessoa tenta acessar um dado que outra está gravando pode acarretar um travamento do sistema de ambos usuários. Isso acontece por que os dados operacionais podem ser inseridos, atualizados, apagados e acessados. Para evitar essas falhas, as aplicações operacionais trabalham com sistemas de transação de dados, o qual não entraremos em detalhes. Tal proteção consome certa quantidade de recursos do sistema. No data warehouse, não há necessidade desse tipo de proteção, pois o risco de acontecer esse tipo de problema é mínimo. Os dados no data warehouse são inseridos uma única vez e, a partir daí, somente há leituras nesses dados.

Outra característica interessante no data warehouse é a variância temporal dos dados. O banco de dados operacional é uma foto dos dados no presente momento. Quando um dado é modificado, o banco de dados não reflete mais a realidade da hora em que o dado foi criado. Não se pode saber qual era a posição de certo dado há dois anos atrás, por exemplo. Como dissemos, no data warehouse há somente inserções e consultas. E como é feita a atualização dos dados a partir das fontes operacionais?

Existe uma nova dimensão (ou chave) presente nas tabelas de dados do data warehouse. É a dimensão tempo. Ao inserir dados no data warehouse, esses são adicionados e a chave tempo recebe o valor do momento da atualização. Ao atualizar os dados a partir das fontes operacionais, os dados (para o data warehouse) recebem o mesmo tratamento: são novamente inseridos no data warehouse com a chave tempo do momento da atualização. Então, podemos dizer que o data warehouse é

variante no tempo. Isto é, o DW é o álbum de fotografias dos bancos de dados operacionais em todas as suas modificações. Podemos saber, exatamente, qual o estado de um determinado dado há 5 anos atrás.

A IBM (International Business Machines, uma grande empresa que atua em vários ramos de tecnologia) geralmente usa o termo “information warehouse” para designar o data warehouse. É, por vezes, uma definição muito melhor para um data warehouse. Os dados são material bruto gerados por produtores de dados (as aplicações operacionais) e são usadas pelo processadores de dados (ferramentas de análise) para criar informação. Informação são os dados trabalhados que geraram algum tipo de conhecimento ou significado. O data warehouse mantém, em muitos casos, os resumos e extratos armazenados fisicamente. Isto é, ele mantém fisicamente informações prontas para serem acessadas. Essas informações são consumidas pelos gerentes e executivos que fazem uso de um data warehouse.

O data warehouse é, muitas vezes, relacionado às necessidades dos executivos em obter informações para a tomada de decisões estratégicas. Na verdade, não é somente isso. O data warehouse é uma ferramenta que extrai dados brutos de toda a corporação, transforma-os em informação palpável através de softwares de mineração de dados e tem a capacidade de disponibilizar dados para toda a empresa. Há aplicações de mais diversos tipos, tanto ferramentas estratégicas quanto ferramentas operacionais, rodando sobre a plataforma do data warehouse.

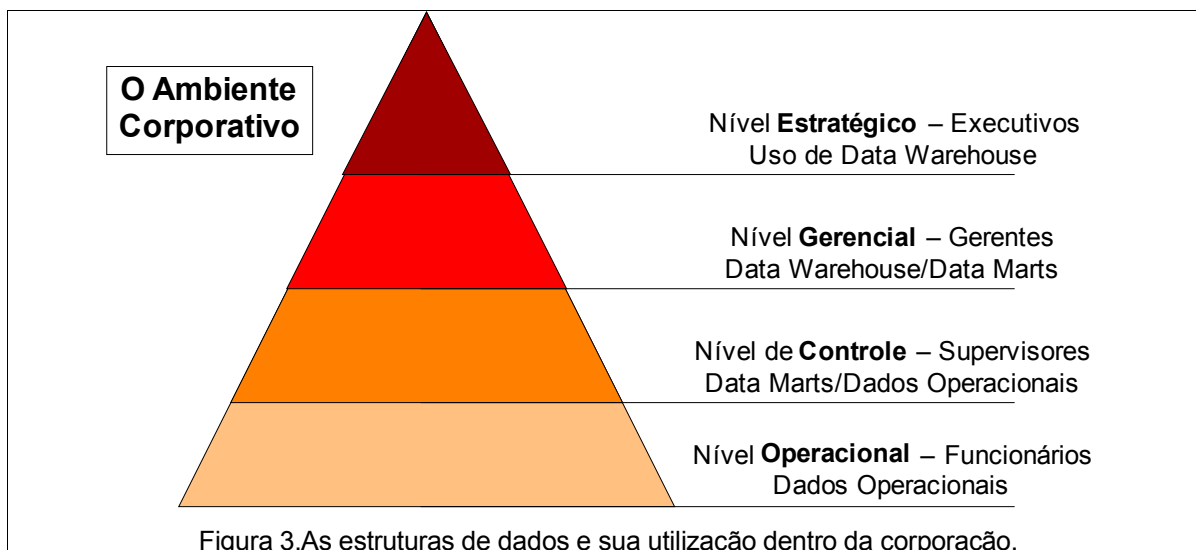
1.4. Data mart

De acordo com INMON (2000a), “um data mart é uma coleção de assuntos organizados para o suporte de decisões baseado nas necessidades de um departamento”.

Um data mart é um repositório de dados extraído das fontes operacionais ou de um data warehouse que é projetado para servir as necessidades particulares de um grupo específico. A ênfase do data mart é atender as necessidades específicas em termos de análise, conteúdo, apresentação e facilidade de uso do grupo alvo de usuário. O data mart é criado para fazer com que os dados analíticos pareçam mais familiares aos usuários finais.

Na prática, podemos adotar o data warehouse e data mart trabalhando juntos. O data warehouse busca atender as necessidades estratégicas da organização, pois reunir informações de toda a empresa. Já os data marts atendem as necessidades gerenciais da empresa mais a nível operacional. Podemos dizer que o data warehouse

é uma estrutura corporativa, ao passo que um data mart é uma estrutura departamental.



Há dois tipos básicos de data marts: dependentes e independentes do data warehouse. Se o data mart atualiza seus dados através do data warehouse, ele é considerado dependente. Se o data mart faz isto através dos dados operacionais, é um data mart independente.

O data mart geralmente contém poucos dados históricos, mas isso pode variar de acordo com a necessidade de cada departamento. Sem os dados históricos os data marts podem ser livremente indexados, ao contrário do data warehouse, que contém grandes quantidades de dados. O data warehouse contém dados granulares, isto é, altamente detalhados, enquanto o data mart possui dados agregados e sumarizados, em sua maioria.

Mas um fato interessante é que vários produtos no mercado implantam o data warehouse e, a partir daí, cria visões específicas para cada grupo de usuário (departamentos), podendo tal “visão dos dados” ser chamada de data mart. São os chamados data marts virtuais.

Quadro 2.

Data warehouse	Data mart
Corporativo	Departamental
Altamente granular	Sumarizado e agregado
Dados históricos	Dados recentes
Grandes volumes de dados	Volumes de dados limitados
Versatilidade	Especificidade

Um comparativo entre data warehouse e data mart.

Neste capítulo, estudamos a história e evolução dos sistemas de informação (chegada dos computadores às empresas), tipos de processamento de dados (OLTP e OLAP) e conceitos de data warehouse e data mart. No próximo capítulo estudaremos aspectos acerca do projeto e da tecnologia que envolvem o data warehouse, a nível lógico e físico.

CAPÍTULO II.

ASPECTOS DE PROJETO DE DATA WAREHOUSE

O data warehouse, como já foi visto, está inserido no contexto da análise dos dados e no armazenamento de informações analíticas tratadas extraídas do ambiente operacional. Mas como os dados fluem do ambiente operacional para o data warehouse e como esses dados são usados? Qual a tecnologia empregada nesses processos?

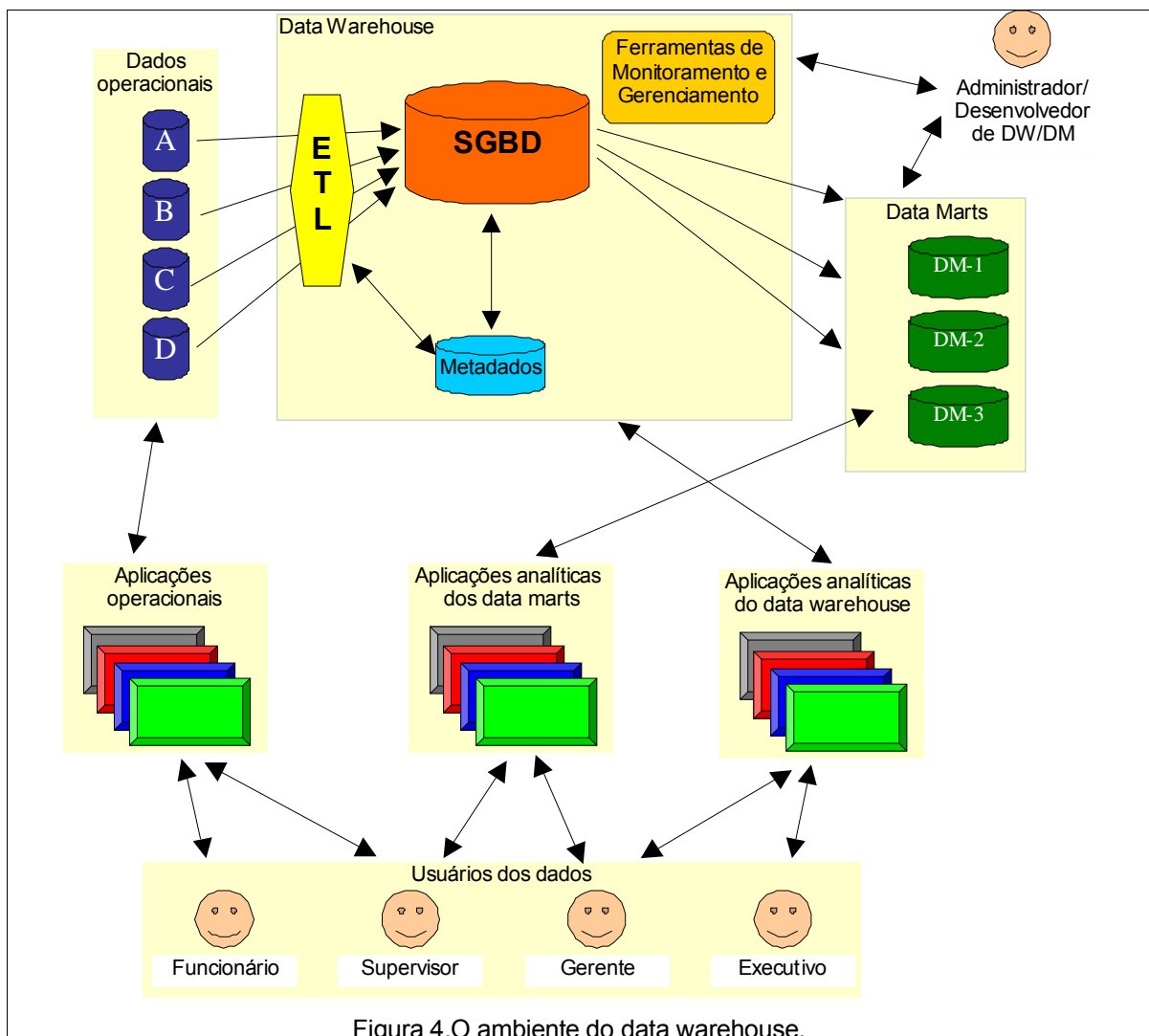
Neste capítulo veremos as características ambientais do data warehouse, como a relação do data warehouse com o SGBD (Sistema Gerenciador de Banco de Dados) e com o modelo de dados, a importância da granularidade e as tecnologias empregadas no processamento dos dados contidos no data warehouse.

II.1. O fluxo de dados

Na figura 4, vemos os elementos e o fluxo dos dados em um ambiente de data warehousing. Dados, como dito anteriormente, são material bruto fornecidos pelos produtores de dados. Através de uma ação dos funcionários (nível operacional), as aplicações operacionais criam dados dos mais diversos tipos, como pedidos, clientes, produtos, etc. Tais dados não servem para propósitos de tomada de decisões estratégicas e gerenciais. Os dados, para serem efetivamente usados para fins analíticos, devem ser transformados e carregados no data warehouse. Durante estas transformações (realizadas pelas ferramentas ETL – Extrair, Transformar e Carregar), são criados resumos e agregamentos entre esses dados que os transformam em informações de baixo nível. Informações são os dados trabalhados que geraram algum tipo de conhecimento ou significado. Através de um resumo de dados, estes passam a

ter algum significado para os gerentes e executivos. Isto é, os dados passam a ser agora informações, embora em um nível ainda baixo.

Quando essas informações de baixo nível são processadas por aplicações OLAP, elas são transformadas em informações de mais valor. Essas informações é que vão suprir as necessidades dos executivos e gerentes, relativas à estratégia e gerenciamento do negócio. A estratégia é a tomada de decisões a nível corporativa, enquanto o gerenciamento é a nível departamental. Idem para data warehouses (nível estratégico) e data marts (nível gerencial).



II.2. Os metadados

Uma parte importante do data warehouse são os metadados. Eles são os chamados dados sobre os dados. São o índice do conteúdo de um data warehouse. É um mapeamento dos dados do modo como foram extraído das fontes operacionais e como estão inseridos dentro do DW. Os metadados definem e descrevem os dados de negócio e os tipos de dados. São definições de tabelas e seus atributos, domínios e

relações. A forma como os metadados estão organizados são essenciais para que os usuários façam um bom uso do data warehouse. Eles são geralmente armazenados em repositórios separados ao alcance dos usuários.

As ferramentas de monitoramento detectam gargalos de desempenho, compõem estatística de acesso e são capazes de responder perguntas como quem, quando e de que modo determinado dado foi acessado. Através das ferramentas de monitoramento são detectados quais dados não foram acessados e, assim, desocupar o espaço desses dados, passando-os para um dispositivo de fita magnética.

II.3. O data warehouse e o modelo de dados

No ambiente do data warehouse INMON (1997: 81) destaca três modelos de dados: corporativo (genérico), operacional (aplicações operacionais) e analítico (data warehouse). O modelo corporativo de dados contém primitivas básicas sobre o tipo de dados necessários para a corporação. Os outros dois modelos (operacional e analítico), são derivações do modelo corporativo.

O modelo operacional é um modelo parecido com o corporativo, mas com fatores relativos a performance destacados, como uma adequação ao processo operacional. Este modelo trabalha junto o modelo de orientação do processo e está num modelo clássico de dados, em forma altamente normalizadas¹.

O modelo analítico trabalha num modelo de normalização chamado star join. Esse modelo apresenta menores níveis de normalização. Os dados são divididos em fatos e dimensões. Os fatos são os acontecimentos, como um pedido de vendas, por exemplo. As dimensões são as variantes relacionadas aos fatos, como produtos, tempo, fornecedor, clientes, etc. Os fatos são, em geral, dados numéricos, fáceis de serem analisados e tidos em grande quantidade. As dimensões representam dados textuais, são difíceis de analisar e são relativamente poucos (em relação aos fatos).

Uma comparação do desempenho dos modelos clássico e star join, realizada pelo site DWBrasil (2000), apontou que o modelo star join é 30% mais rápido que o modelo clássico quando se trata de análise de dados. Foi usada uma tabela fato com cem mil registros e o aplicativo de data mining Business Objects 5.0 com base de dados Access. O modelo clássico apresentou uma média de 79 segundos para realizar cada consulta, enquanto o star join demorou 57 segundos na média.

Considerando-se que um grande data warehouse pode apresentar tabelas fato superiores a milhões de registros e levando-se em conta que essa diferença cresce

¹ Veja o verbete normalização no glossário para uma explicação mais detalhada sobre essa forma de dados.

exponencialmente, de acordo com os realizadores da pesquisa, podemos crer que tal resultado representa uma grande economia na recuperação de dados. E essa economia de processamento e tempo é devida somente a uma característica básica de um projeto bem feito de um data warehouse e de seu modelo dados.

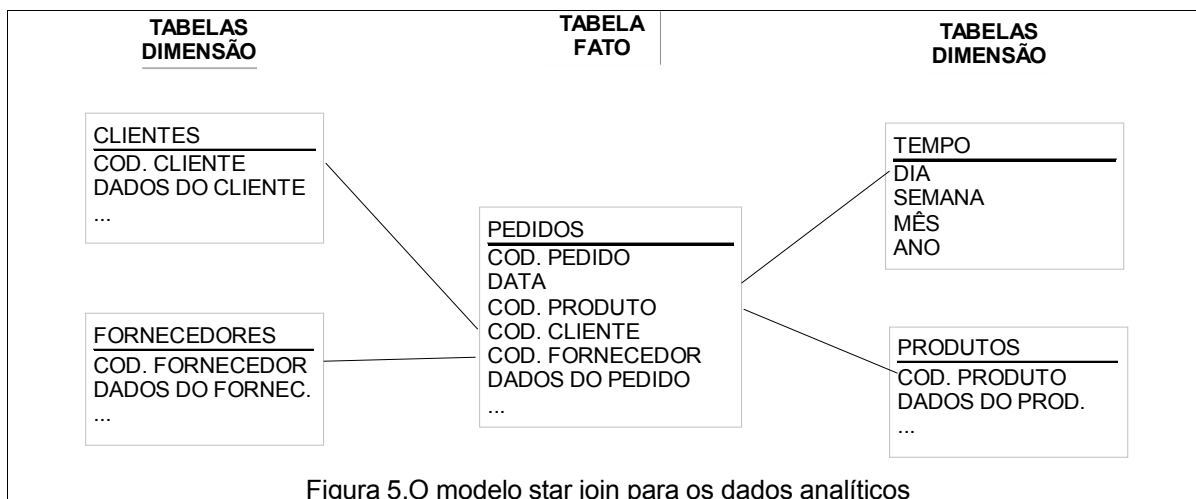


Figura 5.O modelo star join para os dados analíticos

II.4. O data warehouse e o SGBD

Há, hoje em dia, sistemas de gerenciamento de banco de dados específicos para o trabalho com data warehouse. Este tipo de SGBD é otimizado para o processamento analítico realizado no ambiente de data warehousing. INMON (1997; 173) destaca vários pontos necessários para que um SGBD supra as necessidades de um data warehouse.

Conforme vimos no capítulo anterior, o processamento transacional deve ser suportado por um SGBD que contenha proteção contra dead-locks (travamentos). No SGBD para DW não há necessidade de tratamento pesado desses recursos, evitando o overhead (sobre-carregamento) do sistema.

O SGBD usado no data warehouse deve ter facilidade para manusear grandes quantidades de informações. O data warehouse contém muito mais dados que o ambiente operacional, pois ele contém grande quantidade de dados históricos detalhados e também resumos desses dados, enquanto as fontes de dados trabalham com dados recentes somente em seu nível atômico.

É interessante também notar que os SGBD-DW podem possuir um sistema mais robusto e bem aproveitado de índices de dados devido ao fato de somente serem inseridos e acessados os dados. Num SGBD normal, com as inserções, atualizações e deleções de dados, os índices tendem a se deteriorar e exigir um sistema de controle independente.

É muito importante também notar que no DW há muitos meios físicos de armazenamento de dados (memória, disco rígido, fitas magnéticas e outros dispositivos de armazenamento ótico e/ou magnético de massa). O SGBD deve ser capaz de gerenciar os vários tipos de mídia, o que se torna complexo existindo grandes quantidades de dados.

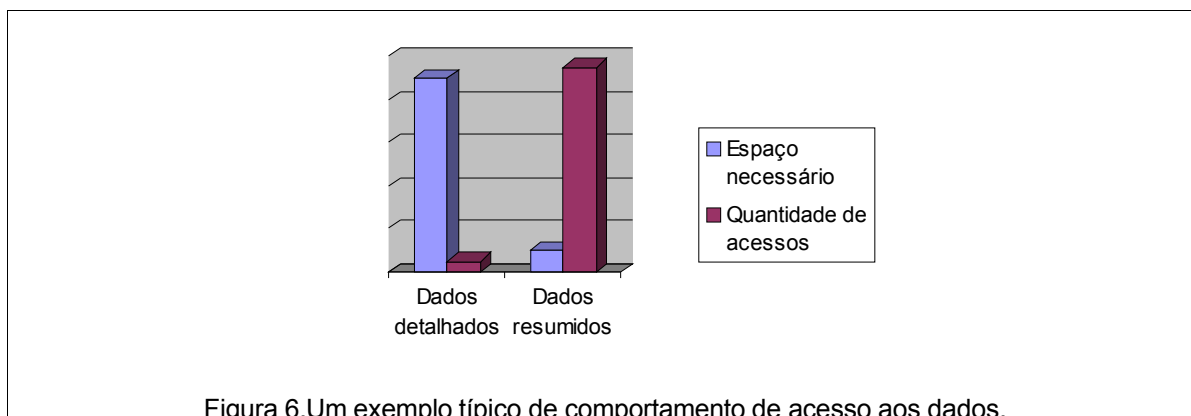
II.5. A granularidade

Os níveis de granularidade são parte fundamental de um data warehouse. A granularidade é o nível de detalhe contido nas unidades de dados existentes no data warehouse. Os dados dentro do data warehouse podem estar disponíveis de várias formas: granulares, levemente resumidos, altamente resumidos e assim por diante. Quando os dados são processados no ambiente operacional, eles realizam operações nos dados em seu menor nível de granularidade. Os detalhes são utilizados em seu nível máximo. No ambiente analítico, o data warehouse, os dados detalhados, depois de resumidos, são usados em pouquíssimos casos, se os níveis de granularidade forem bem projetados.

A questão da granularidade é importante por que pode diminuir consideravelmente o tempo de acesso aos dados que realmente interessam (as informações já resumidas) e diminuem a quantidade de discos rápidos e caros necessários para armazenar uma grande quantidade de dados históricos.

O exemplo de vários níveis de granularidade nos dados são as vendas de uma rede de varejo. Todos os dias, vendedores realizam muitas vendas para clientes. Cada venda representa o nível atômico de dados (o menor nível de granularidade). Ao entrar no data warehouse, esses dados a respeito das vendas são carregadas no seu nível atômico e, então, sofrem um processo de resumo dos dados. Esses dados são resumidos em total das vendas feitas por um vendedor no dia, total das vendas para um determinado cliente na semana, total das vendas da empresa no dia, na semana, no mês e assim por diante. Os dados resumidos representam muito menos gasto físico de espaço, mas simbolizam muito mais informações que o nível mais detalhado de dados, além de serem dados mais fáceis de se recuperar.

Quanto aos dados detalhados históricos, isto é, não recentes, podem ser colocados em meios de armazenamento de massa mais baratos, como fitas magnéticas, quando as ferramentas de monitoramento do data warehouse apontarem que eles são usados em muitos poucos casos. Os resumos desses dados continuarão acessíveis a partir dos meios de armazenamento rápido.



II.6. O processo de ETL

O processo de ETL (Extraction, Transformation and Load) é a parte do data warehouse responsável pela carga de dados e sua atualização a partir dos dados operacionais. Ele se resume em localizar onde estão os dados a serem extraídos, extraí-los, transforma-los para o formato de dados do DW, integrá-lo conforme especificações dos metadados e, finalmente, lançá-los dentro do data warehouse. Os metadados do DW são baseados no modelo corporativo de dados.

Segundo INMON (1997: 115), pode parecer um processo simples, dando a impressão que uma programação manual de uma simples ferramenta de extração dará o serviço como feito. Mas não é isso o que acontece geralmente. Ao começar a realizar o serviço braçal de extração, os desenvolvedores do data warehouse se dão conta que esta tarefa é muito mais complexa do que parecia no início.

INMON cita alguns problemas que configuram o ETL como uma tarefa árdua e complexa, como:

- Há uma mudança na tecnologia de dados – é necessário ler os dados no SGBD herdado e grava-los em um SGBD mais novo e atual, capaz de suportar o data warehouse. Isso incorre numa necessidade de ferramentas para a comunicação entre esses dois tipos de base (middleware).
- Os dados precisam ser convertidos:
 - Datas, medidas e padrões precisam ter uma única face. Exemplo: as datas em um banco de dados operacional estão no formato AA/MM/DD e em outro BD estão MM/DD/AA. No data warehouse foi adotado o padrão DD/MM/AAAA, então todos os dados precisam ser reformatados para entrar no DW. Essa tarefa é extremamente mecânica e árdua, e poderia ser realizada por conversores automáticos, por exemplo, ao invés de programada manualmente.

- O formato lógico dos dados, como formatos de texto EBCDIC e ASCII, devem ser convertidos também
- As chaves dos dados precisam ser reestruturadas. Em casos mais simples é somente uma questão de adicionar a dimensão tempo, em outros casos as chaves nos diversos sistemas herdados são diferentes e precisam ser convergidas para um único formato.
- A seleção dos dados que irão entrar no DW é extremamente custosa, podendo haver várias fontes para os mesmos dados
- É necessário eliminar a redundância de dados. Um mesmo cliente é cadastrado em várias aplicações diferentes, por exemplo. Ao entrar no DW, esse cliente deverá aparecer somente uma vez.
- O relacionamento entre os dados em programas herdados não são documentados ou são colocados em uma lógica extremamente otimizada para tais programas. INMON coloca este problema como um dos piores a serem resolvidos, pois desvendar os dados atrelados dessa forma é um passo muito demorado e custoso.
- A grande variedade de fontes de dados. Em consequência, há igual número de formatos e relacionamentos de dados a serem trabalhados.
- Correção de dados. Dados não confiáveis de fontes operacionais devem ser corrigidos para não haver dúvidas quanto à veracidade das informações do data warehouse. Um simples algoritmo de correção pode resolver, mas talvez sejam necessárias ferramentas de inteligência artificial para tratar tais dados.

É necessário um cuidado especial com as ferramentas de ETL. No caso de pequenos data warehouses, com poucas fontes de dados e poucas tabelas, pode-se considerar a possibilidade de programação manual da extração. Mas com um data warehouse um pouco maior é praticamente obrigatório o uso de ferramentas automatizadas para a ETL dos dados.

II.7. Tecnologias de aplicações sobre o data warehouse e os data marts

O datawarehouse é, acima de tudo, uma grande banco de dados corporativo com dados preparados para análise. Tal banco de dados possui alguns dados “prontos para consumo”, como os dados em diferentes níveis de sumarização e agregação. Mas com os dados disponíveis através do data warehouse, pode ser feito muito mais. Ferramentas estão amplamente disponíveis no mercado e um bom engenheiro de informação pode muito bem desenvolver soluções personalizadas para cada empresa.

Essas ferramentas, ditas ferramentas de processamento analítico (OLAP), são um dos grandes benefícios em se possuir um data warehouse bem construído.

Vejamos algumas dessas ferramentas que trabalham o data warehouse.

II.7.1. DSS

De acordo com POWER (1997: 1), “um sistema de suporte a decisão é um sistema de computador interativo cujo objetivo é ajudar gerentes a tomar decisões.”

Segundo GREENFIELD (2000: 1), um sistema de suporte à decisão é uma ferramenta especificamente idealizada para permitir que os usuários finais de negócios possam realizar análises geradas por computador dos dados em si.

Um DSS tem 4 características básicas (Turban):

- Incorporam modelos e dados.
- Ajudam executivos e gerentes a tomar decisões a partir de tarefas semi-estruturadas ou ainda não estruturadas.
- Eles suportam, ou preferencial substituem, o julgamento gerencial.
- O objetivo do DSS é aumentar a eficiência das decisões, e não aumentar a eficiência com que as decisões são tomadas.

Os DSS usualmente suportam análises “what if” (“e se”). Tais análise podem fazer previsões se o usuário fizer mudanças nas variáveis que influem em processo.

Vamos dar um exemplo: “E se aumentássemos nossos preços em 5%, o que isso iria influir em nossas vendas daqui a 2 meses?”. O DSS formularia uma resposta satisfatória se ele pesquisar o histórico de vendas e analisar como foram as vendas após uma subida de preços.

II.7.2. Data Mining

Ferramentas de data mining, ou mineração de dados, são comuns no ambiente de data warehouse. Data mining é uma classe de aplicações analíticas que procuram por padrões escondidos nos dados do data warehouse. As ferramentas realizam uma filtragem de grande quantidade de dados usando muitas técnicas, tais como raciocínio baseado em casos, visualização de dados, pesquisa e análise fuzzy e redes neurais.

Com ferramentas de data mining, podemos descobrir, por exemplo, qual o perfil do cliente devedor, que produtos que são vendidos juntos, qual o tipo do paciente desenvolve um tipo específico de doença, padrões comportamentais de poços de petróleo e muitas outras coisas. A mineração de dados se encaixam tanto em pesquisas comerciais quanto em pesquisas de cunha científico.

II.7.3. CRM

CRM significa Customer Relationship Management, ou Gerenciamento de Relações com o Cliente. Aplicações de CRM usam o data warehouse para criar um histórico integrado sobre os clientes e poder oferecer um atendimento personalizado para estes.

Assim, podemos saber facilmente quantos clientes foram adicionados ao sistema, o rendimento com cada cliente, agrupamento de clientes (de acordo com inadimplência, lucro, etc.), produtos que atraem clientes específicos, pontos de contato do cliente com a empresa e assim por diante.

CAPÍTULO III.

CUSTOS DO DATA WAREHOUSE

O data warehouse, como qualquer outra estrutura, possui suas próprias características e custos. Sabendo-se quais os gastos de se construir e operar um data warehouse, podemos antecipar e gerenciar estes custos. Dessa forma, a carga desse investimento na instituição pode se fazer de forma organizada e prevista, aumentando com isso a satisfação da organização com relação ao data warehouse.

III.1. Classificando os custos

A perspectiva de INMON (2000c: 3) sobre a classificação dos custos decorrentes da implantação do data warehouse é a divisão, quanto ao tempo do investimento, entre custo inicial e custo periódicos. Também pode ser dividido, quanto ao tipo de investimento, em custos capitais e custos operacionais. Os custos capitais são decorrentes do gasto de dinheiro referente a compra de equipamentos, serviços e estrutura. Os custos operacionais é o gasto com pessoal, não necessariamente gastos financeiros, mas sim gastos de tempo para o desenvolvimento de soluções. Este tempo é precioso e não é somente gastando-se mais dinheiro com pessoal que o tempo de desenvolvimento do data warehouse irá diminuir.

III.1.1. Custos iniciais

O custo inicial é o investimento realizado na implantação do data warehouse. Segundo INMON (2000), ele pode ser classificado como:

Quadro 3.

Hardware	Software	Pessoal
<ul style="list-style-type: none"> • Processamento • Armazenamento • Estrutura de rede 	<ul style="list-style-type: none"> • SGBD • Ferramentas de gerenciamento e monitoramento do sistema • Criação e povoamento dos metadados • Criação ou aquisição dos programas de ETL (Extraction, Transform and Load) 	<ul style="list-style-type: none"> • Consultoria em DW • Administrador de data warehouse (banco de dados) • Administrador de rede • Analistas de sistema • Programadores Treinamento dos usuários

Os elementos formadores do custo inicial de um data warehouse.

Os gastos com hardware são investimento em estrutura física de suporte ao data warehouse, como computadores, discos rígidos, dispositivos de fitas de dados, estrutura de rede. Segundo especialistas, o maior gasto nesse quesito é a parte de armazenamento, pois um grande data warehouse pode chegar a ocupar vários terabytes de dados. Então para diminuir esse gasto, são instalados sistemas de armazenamento em massa, como fitas magnéticas e armazenamento ótico, decaindo o custo muitas vezes. Deve-se realizar uma previsão da tamanho inicial do data warehouse e uma estimativa de crescimento do data warehouse para que seja adquirido o hardware necessário para conter o crescimento.

O investimento em software é o custo de aquisição de um SGBD, ferramentas de gerenciamento e monitoramento do sistema, da criação de programas de ETL e a criação e povoamento dos metadados.

A criação e povoamento da estrutura dos metadados é um passo importantíssimo, segundos os especialistas em DW, para um projeto bem sucedido. Os metadados devem estar amplamente disponíveis para os usuários finais do data warehouse (principalmente os exploradores) para que façam uso do data warehouse.

Há uma grande variedade de programas de ETL disponíveis para os usuários de data warehouse. Deve-se configurar estes programas para buscar os dados operacionais, integra-los e transforma-los para que alimentem o data warehouse.

Os gastos com pessoal são o fator mais importante do data warehouse. Inicialmente, deve-se contratar uma consultoria para uma avaliação das necessidades e proposição das melhores alternativas para a criação do data warehouse. A partir daí, contrata-se ou realoca-se profissionais de administração do data warehouse e redes, analistas e programadores.

O melhor investimento que se pode fazer para que um data warehouse (ou qualquer outro projeto de informática) é o treinamento dos usuários. Usuários bem

treinados são essenciais para o sucesso de qualquer projeto. São os usuários que irão proporcionar o retorno à todo e qualquer investimento em data warehouse, portanto iremos estudá-los melhor mais a frente (Cap. III – Fazendeiros x exploradores).

Uma solução alternativa e viável é a terceirização dos serviços de pessoal. Há muitas empresas que disponibilizam profissionais para a construção e manutenção do data warehouse, além de oferecerem treinamento para os usuários do data warehouse.

III.1.2. Custos periódicos

Os custos periódicos são os gastos derivados da manutenção, atualização e uso do data warehouse. Podemos dividi-los em:

Quadro 4.

Hardware	Software	Pessoal
<ul style="list-style-type: none"> • Manutenção • Acréscimos 	<ul style="list-style-type: none"> • Manutenção do SGBD • Manutenção das ferramentas de ETL ou criação de utilitários de processamento de log • Manutenção dos metadados • Aquisição de novas ferramentas 	<ul style="list-style-type: none"> • Consultoria em DSS • Administrador de data warehouse (banco de dados) • Administrador de rede • Analistas de sistema • Programadores <p>Treinamento dos usuários</p>

Os elementos formadores dos custos periódicos do data warehouse.

Os gastos periódicos com hardware estão relacionados com manutenção do equipamento existente, devido ao desgaste natural das peças, e com acréscimo de dispositivos de processamento/armazenamento de acordo com o crescimento do data warehouse.

Os gastos com software estão em sua maioria relacionados com o gasto de pessoal, como a manutenção do software e da estrutura de metadados. Eventualmente, pode haver aquisição de novas ferramentas para a manutenção e uso do data warehouse.

Os gastos com pessoal diminuem em relação ao custo inicial, pois a mão de obra pesada usada no dia a dia do data warehouse é menor. Os custos com treinamento ainda devem continuar sendo uma constante no orçamento do data warehouse. Esta parte, como já dito, pode ser terceirizada.

III.2. Fatores de variação do custo

Quanto cada parte de investimento mencionado no item anterior irá custar ao bolso do investidor? Podemos dizer que não há parâmetro para afirmarmos com certeza quanto o data warehouse irá custar.

O que podemos afirmar é que há uma série de muitos fatores que influenciam diretamente no custo do data warehouse. Assim, não podemos, sem o levantamento de todos estes fatores, que variam de organização para organização, dizer quanto o data warehouse irá custar, nem inicialmente, nem periodicamente.

Os fatores que influem no custo são:

- A quantidade de dados gerados pelo dia a dia operacional – determina quanto crescimento o equipamento e software do data warehouse precisará conter para trabalhar com folga pelos próximos dois anos.
- O histórico que deve ser armazenado no data warehouse – a quantidade de dados que poderá ser armazenada em dispositivos de massa, como fitas magnéticas, que custa barato, mas são lentas e se tornam um obstáculo para um processamento rotineiro dos dados.
- O nível de detalhamento dos dados necessário – quanto maior o detalhe contido nos dados, mais espaço de armazenamento é necessário. O nível de detalhe atômico é o usado nas aplicações operacionais.
- O nível do usuário final – quanto de treinamento os usuários necessitam para se adaptar ao data warehouse.
- Se o mercado está num mercado competitivo ou não – que tipo de análise deve ser feita pelos programas que rodam sobre o data warehouse.
- A velocidade com que o data warehouse precisa ser construído – quanto mais rápido, mais profissionais e, como consequência, maior deve ser o investimento.
- Se o data warehouse será construído manualmente ou através de ferramentas automáticas – o processo manual demanda menos gasto com software mas muito mais gastos com pessoal, tanto no custo inicial quanto no custo periódico.
- O número de data marts a ser construído – quantos departamentos necessitam de sua própria visão dos dados analíticos, tendo a necessidade de construir seus próprios data marts.

- A quantidade de sumários e dados sumarizados a ser mantido – quanto mais sumários, menor a quantidade de acesso aos dados detalhados, que podem estar armazenados em dispositivos de armazenamento de massa.

A todos esses fatores pode-se também somar:

- A escolha do SGBD correto, que contenha o maior numero de ferramentas de ETL e administração nativas.
- A escolha do hardware correto, de acordo com as necessidades de processamento e armazenamento.
- A contratação de bons profissionais para realizar cada tipo de serviço.
- Erros na implantação devido a “economias” na hora de fazer o projeto do data warehouse.

Devido a esses fatores, podemos assumir que não uma empresa “típica” em que possamos basear um custo fixo. Também devemos lembrar que estamos trabalhando o custo do data warehouse, e não de todo o sistema DSS. A esses fatores podemos adicionar os custos das ferramentas de análise e sistemas de informação que irão trabalhar sobre o data warehouse.

CAPÍTULO IV.

JUSTIFICAÇÃO DOS CUSTOS DO DATA WAREHOUSE

Podemos dizer que o data warehouse é um investimento alto. Mas o que ele pode nos oferecer em troca de tamanho custo? O retorno desse investimento vem a curto, médio ou longo prazo? Qual a melhor argumentação para se justificar a implantação de um data warehouse?

Poderemos medir a satisfação de um empresa com seu data warehouse à media que se tem um maior retorno por um menor investimento. O retorno, como veremos, se dará em formas de vantagens no negócio, aumentando a competitividade e o lucro da empresa detentora do data warehouse. O investimento em um data warehouse não compensa se este custo não for pago gradualmente com o uso dele. Então, para que o data warehouse pague-se mais facilmente, é preciso do equilíbrio entre um data warehouse barato e de alto benefício para a empresa.

IV.1. Acesso à informação

Dos vários benefícios que o data warehouse traz ao cotidiano de uma empresa, o mais simples de se ver é a facilidade de extração de relatórios integrados e sua diminuição de custo.

Ao extrair relatórios sem um data warehouse, uma empresa que possui várias aplicações operacionais precisa:

1. Localizar em quais sistemas estão os dados.
2. Recuperar os dados.
 - Extrair dados de cada um dos sistemas operacionais. Eles podem estar baseados em ADABAS, Oracle, DB2, etc.

3. Integrar/transformar os dados.
 - Padronizar os dados.
 - Reconciliar os dados.
 - Modificar as estruturas de dados.
4. Reunir os dados.
5. Construir o relatório.

A força de desenvolvimento envolvida em produzir um relatório é muito custosa. E não podemos reaproveitar relatório já prontos, pois as necessidades de dados mudam para cada relatório. Podemos afirmar que o custo de extração de cada relatório pode variar de milhares a milhões de dólares e demorar vários meses para ficar pronto.

Com o data warehouse, o processo de extração dos relatórios se simplificam, pois os passos de 1 a 4 no modelo anterior são feitos uma única vez e após isso é só construir cada relatório. Então, relatórios que eram feitos em meses passam a ser confeccionados em dias ou horas.

Considerando a o número de fontes de dados e b o número de relatórios, o custo total de extração sem o DW é de $a \times b$, já que deverão ser minerados todas as fontes para cada relatório. O custo com o data warehouse cai para $a + b$, partindo-se do princípio que os dados já estão no data warehouse e que é só construir cada relatório.

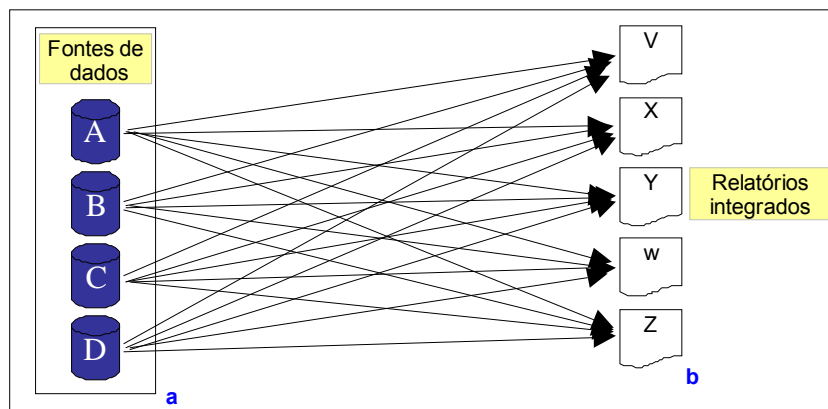


Figura 7.O custo com relatórios antes do data warehouse ($a \times b$).

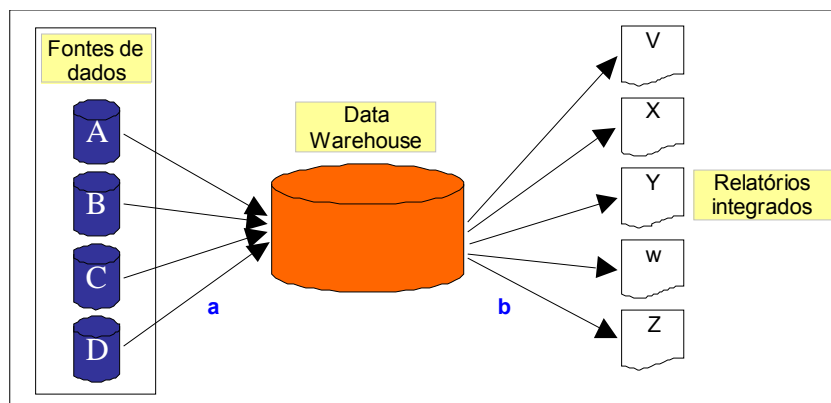


Figura 8.O custo com relatórios após o data warehouse (a + b).

Nesse diferencial de custos não está incluso o tempo gasto para a extração de cada relatório. Num ambiente empresarial, tempo é muito importante ao se tomar uma decisão.

IV.1.1. Fazendeiros x Exploradores

INMON divide os usuários do data warehouse em dois grupos distintos: o fazendeiro e o explorador. O fazendeiro está presente no dia-a-dia da empresa, usando os caminhos habituais para acesso aos dados e retornando informações de pouco valor em um ritmo constante de extração. O explorador faz pesquisas ocasionais no data warehouse e raramente ele acha alguma informação que possa ser útil mas, muitas vezes, tal informação é de grande valor.

Então, por quem justificar o custo do data warehouse, pelo usuário diário que retorna informações baratas ou pelo explorador que pode encontrar uma verdadeira preciosidade em forma de informação?

Logicamente, devemos optar pelo fazendeiro como nosso justificador de custos. O retorno do investimento através do uso diário e rotineiro do data warehouse é constante e não oferece riscos. Em contrapartida, o explorador é inconstante e é muito arriscado fazer compromissos de que as informações poderão ser usadas.

O fazendeiro é imprescindível, mas o explorador também é muito importante. Podemos adotar o modelo exploratório para descobrir novas informações associadas aos nossos dados óbvios.

IV.2.Data warehouse e data marts

Outra abordagem dos benefícios do data warehouse é a construção do data warehouse como “ponte” para a produção de data marts. Data marts, como vimos no capítulo 1, é uma estrutura departamental do data warehouse.

Podem-se construir data marts sem um data warehouse, mas o custo de toda a estrutura independente de data marts a longo prazo irá crescer muito em relação a um data warehouse envolto com data marts.

Vejamos por quê: para a construção de um data mart, iremos proceder da mesma forma como na construção de um data warehouse: extraindo, transformando e integrando dados pertinentes à área departamental do data mart. Então, o custo inicial da construção de data marts independentes é $a \times b$ (número de fontes) \times b (número de data marts). Da mesma forma como acontece com os relatórios, os data marts consomem apenas $a + b$, pois os dados são extraídos para o data warehouse uma única vez e a partir daí que vão nutrir os data marts.

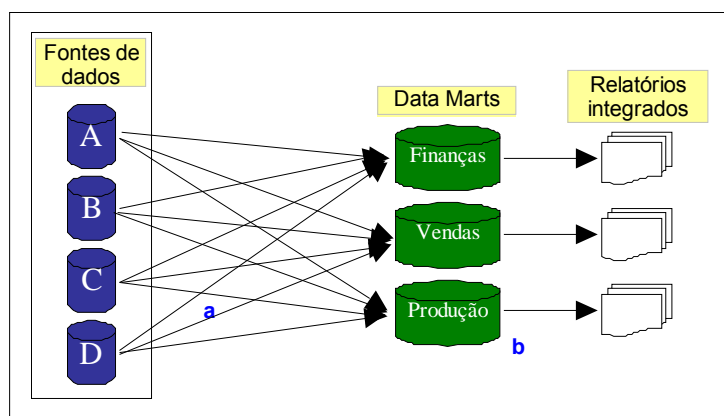


Figura 9. Data marts independentes. Custo $a \times b$.

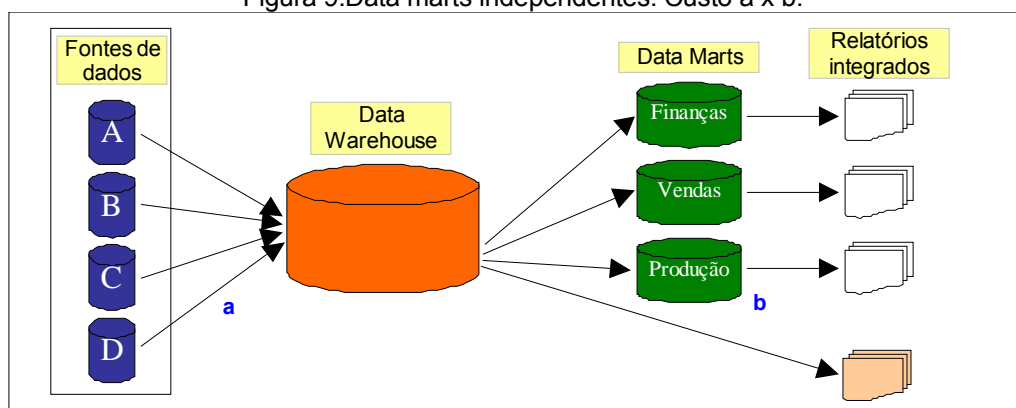


Figura 10. Data marts construídos a partir do data warehouse. Custo $a + b$.

IV.3. Vantagens no negócio

A justificação do custo do data warehouse baseada em vantagens no negócio pode ser interessante, embora tal benefício não possa ser quantificado, pois embora possamos medir o crescimento da empresa antes e depois do data warehouse vários outros fatores podem ter acarretado tais mudanças na realidade dos negócios.

Um data warehouse bem construído é capaz de assegurar e aumentar a participação da organização no mercado, aumentar o lucro e diminuir custos. Esse

fenômeno se dá por quê a empresa munida de um data warehouse pode tomar decisões estratégicas mais rapidamente do que as empresas que não o possuem. As informações fluem através de toda a empresa através do data warehouse, acusando se há uma queda no mercado e podem ser tomadas decisões a tempo antes que as perdas sejam maiores. Os lucros podem aumentar, pois a empresa deixará de perder ao realizar negócios menos arriscados, baseando-se no histórico integrado dos negócios anteriores que o data warehouse disponibiliza. A diminuição de custos ficou evidente ao estudamos a produção de relatórios antes e depois do data warehouse.

Partindo-se desses princípios, pode-se deduzir que o data warehouse realize alterações nos sensores costumeiros do crescimento da empresa, tais como número de clientes, valor do faturamento, valor do lucro e assim por diante.

Tais vantagens são comprovadas por empresas que já implantaram o DW, consolidando o data warehouse como ferramenta redutora de custos, agilizadora de processos e medidora de qualidade e crescimento das empresas.

CONCLUSÃO

Após intensa pesquisa bibliográfica sobre a literatura específica e sites especializados em data warehouse, podemos tirar várias conclusões.

Não é possível dizer o custo específico para um data warehouse sem um minucioso projeto e levantamento das necessidades específicas da organização sobre o data warehouse. É necessário realizar uma pesquisa para saber as variáveis que a empresa apresenta, tais como volume de dados, tempo de resposta, tempo de construção do data warehouse, nível de instrução do usuário final, entre outros fatores mencionados no decorrer do trabalho. Cada um dos fatores altera o custo do data warehouse e faz com que este se torne impossível de se prever.

Aos fatores de variação de custo, podemos acrescentar a abordagem na construção do data warehouse. A empresa, ao iniciar um projeto de data warehouse, deve estar consciente de que a abordagem ao problema é um fator fundamental para a determinação do custo e a satisfação com o data warehouse. Um exemplo de abordagem errônea é a tentativa (e frustração) de construção do data warehouse em uma única etapa, pois um data warehouse construído iterativamente facilita o feedback com os usuários e correção do data warehouse.

Os benefícios da implantação de um data warehouse são evidentes, de acordo com o que foi apresentado durante o trabalho. É possível ter benefícios em pouco mais de um mês após o início da construção do data warehouse. Em contrapartida, os benefícios não estão bem delineados para a empresa até o término da construção do data warehouse, segundo INMON. A corporação começará a realizar uma mudança interna para incorporar o data warehouse à sua rotina de trabalho e com isso irá descobrir mais usos – benefícios – do data warehouse.

Uma das dificuldades de realização do trabalho foi a baixa disponibilidade de material sobre custos do data warehouse. Há uma lacuna neste trabalho sobre os tipos de ferramentas de construção e manutenção do data warehouse e seus custos. Uma proposta para as próximas pesquisas nesta área é um estudo mais aprofundado sobre este tema. Há um grande conjunto de ferramentas de data warehousing e cada uma possui suas próprias vantagens e custos, e com certeza surgirão muitas outras.

O estudo para a realização deste trabalho foi muito proveitoso, pois o data warehouse e suas aplicações são atualmente o estado-da-arte no que se refere a sistemas de informações.

ANEXO I. FONTES DE INFORMAÇÃO NA INTERNET

Bill Inmon Website – www.billinmon.com

Página de William H. Inmon. Contém artigos, white papers, bibliografia entre outras informações sobre data warehousing e tecnologia da informação

DSS Resources – www.dssresources.com

Página de D. J. Power. Recursos e muitas informações sobre DSS e seus agregados, como data warehouse, data mining, OLAP, etc.

ReserchIndex: Computer Science Directory – <http://citeseer.nj.nec.com/directory.html>

Base de dados on-line de trabalho científicos na área da ciência da computação. Possui vários modo de visualização de conteúdo e um sistema de cruzamento de referências bibliográficas que facilitam a pesquisa.

The Data Warehouse Information Center – <http://www.dwinfocenter.org/>

Ponto de partida para a busca de informações sobre data warehouse. Contém muitas informações e links para outras páginas sobre o assunto.

What Is – www.whatis.com

Base de dados com definições sobre termos tecnológicos.

SearchDatabase – www.searchdatabase.com

Base de dados sobre assuntos relacionados com base de dados

SearchDataWarehousing – www.searchdatawarehousing.com

Base de dados de procura de informações sobre data warehousing.

GLOSSÁRIO

- Análise de terminal – Trabalho de analisar e mostrar dados uma vez que tais dados estão no data warehouse. Portanto, a análise de terminal é a atividade analítica após os dados serem armazenados no DW. Alguns componentes da análise de terminal são: descoberta do problema, a preparação, gerenciamento e interpretação do relatório, a análise do conjunto-resultado, e a visualização do resultado.
- Dados em nível atômico – dados em seu menor nível de granularidade.
- Granularidade – O nível de detalhe contido em uma unidade de dados. Quanto menos detalhes dos dados, mais baixo é o nível de granularidade. Quanto mais detalhes, mais alto é o nível de granularidade.
- Sistema Gerenciador de Banco de Dados (SGBD) – Programa que gerencia uma coleção de dados organizados de uma maneira específica. Conhecido em inglês como DBMS, ou Data Base Management System.
- Sistema Gerenciador de Banco de Dados Relacional (SGBD-R) – Programa que gerencia uma coleção de dados organizados como um conjunto de tabelas formalmente descritas em que o dado pode ser acessado e remontado de diversas maneiras sem a necessidade de reorganizar as tabelas de dados. É usada a linguagem SQL para praticamente todos os SGBD-R.
- Sumarização – Método para diminuir a granularidade dos dados. Os dados são agrupados por categorias ou tempo para que fiquem mais compactos e fiquem mais fáceis de analisar.
- Dead-lock – Travamento do sistema devido à concorrência de vários usuários ao mesmo recurso. Em bases de dados operacionais é utilizado o sistema de transações para evitar tal condição. Em data warehouse não há a necessidade desse sistema pois as operações se resumem a inserir e consultar. Já nas bases de dados operacionais, é necessário um recurso adicional de recuperação de dead-locks que consomem recursos do sistema.
- ETL (Extract, Transform and Load) – Extrair, Transformar e Carregar. Processo de transformação dos dados transacionais para os dados analíticos.
- Leitura seqüencial – Forma de leitura de dados que lê todos os dados de uma fita magnética até encontrar a informação procurada. O outro método usado é a leitura aleatória, que “salta” a leitura até o ponto exato onde o dado está gravado.
- Linguagens de 4a. geração (L4G) – Linguagens de programação projetada para permitir ao usuário final acesso irrestrito aos dados.
- Normalização – Método para a redução de uma estrutura complexa de dados para um estrutura mais simples e estável. Em geral, o processo o remove as estruturas redundantes de dados, chaves e relacionamentos de um modelo conceitual de dados. (Power, 2000)
- Semi-condutores – Elementos básicos internos de um computador, chamados também de chips ou circuitos integrados.

BIBLIOGRAFIA

- [1] CAMPOS, Maria Luiza; ROCHA Filho, Arnaldo V., Tutorial de Data Warehouse, <http://genesis.nce.ufrj.br/dataware/tutorial/home.html>, NCE-UFRJ, Brasil.
- [2] GREENFIELD, Larry, A Definition of Decision Support, LGI Systems Inc., 22/06/00, <http://www.dwinfocenter.org/decsup.html>
- [3] GUPTA, Vivek R., An Introduction to Data Warehousing, <http://system-services.com/dwintro.asp>, System Services Corp.
- [4] INMON, William H., Como Construir o Data Warehouse – 2ª. Ed., Editora Campus, Brasil, 1997.
- [5] INMON, William H., Data Mart Does Not Equal Data Warehouse, DMReview.com, 07/2000.
<http://www.datawarehouse.com/iknowledge/articles/article.cfm>
- [6] INMON, William H., The Data Warehouse Budget, <http://www.billinmon.com/library/whiteprs/???.pdf>, EUA, 2000.
- [7] INMON, William H., The Data Warehouse Environment: Quantifying Cost Justification and Return on Investment, <http://www.billinmon.com/library/whiteprs/MsCJ.pdf>, EUA, 11/2000.
- [8] INMON, William H., What is a Data Warehouse, <http://www.billinmon.com/library/whiteprs/earlywp/ttdw.pdf>, EUA, 2000.
- [9] KINBALL, Ralph, The Data Warehouse Toolkit, Wiley, EUA, 02/1996
- [10] ORR, Ken, Data Warehousing Technology, <http://www.kenorinst.com/dwpaper.html>, 2000.
- [11] Power, D. J., Decision Support Systems Hyperbook. Cedar Falls, IA: DSSResources.COM, pre-publication PDF version, 2000, acessado em 23/02/2001 na URL <http://dssresources.com/dssbook/>.
- [12] POWER, D.J., What is a Decision Support System, acessado em 20/02/2001 na URL <http://www.dssresources.com/papers/whatisadss/index.html>
- [13] DWBRASIL, Snow x Star, acessado em 02/03/2001 na URL http://www.dwbrasil.com.br/html/snow_star.html