

SUMÁRIO

1. INTRODUÇÃO.....	2
2. O QUE É DATA WAREHOUSE?.....	2
3. O QUE DATA WAREHOUSE NÃO É.....	4
4. IMPORTANTE SABER SOBRE DATA WAREHOUSE.....	5
4.1 Armazenamento.....	5
4.2 Modelagem.....	6
4.3 Metadado.....	6
4.4 Data Marts.....	8
4.5 Extração de Dados.....	8
5. OS PROCESSOS DE DATA WAREHOUSE.....	9
6. CONCLUSÃO.....	10
7. REFERÊNCIAS.....	10

1. INTRODUÇÃO

Todos nós sabemos que os bancos de dados são de vital importância para as empresas e também estamos cientes de que sempre foi difícil analisar os dados neles existentes. Hoje em dia, as grandes empresas detêm um volume enorme de dados e esses estão em diversos sistemas diferentes espalhados por ela. Assim, não conseguíamos buscar informações que permitissem tomarmos decisões embasadas num histórico dos dados. Por um outro lado, em cima desse histórico podemos identificar tendências e posicionar a empresa estrategicamente para ser mais competitiva e conseqüentemente maximizar os lucros diminuindo o índice de erros na tomada de decisão.

Por fim, introduziu-se um novo conceito no mercado, o Data Warehouse (DW). Este consiste em organizar os dados corporativos de maneira integrada, com uma única versão da verdade, histórico, variável com o tempo e gerando uma única fonte de dados, que será usada para abastecer os Data Marts (DM). Isso permite aos gerentes e diretores das empresas tomarem decisões embasadas em fatos concretos e não em intuições, cruzando informações de diversas fontes. Isso agiliza a tomada de decisão e diminui os erros. Tudo isso num banco de dados paralelo aos sistemas operacionais da empresa.

Segundo a (Aspect International Consulting, 1997), cerca de 88% dos diretores admitem que dedicam quase 75% do tempo às tomadas de decisão apoiadas em análises subjetivas, menosprezando o fato de que por volta de 100% deles tem acesso a computadores. Atualmente esse número deve ter diminuído, porque existem muitos Data Warehouses sendo utilizados.

2. O QUE É DATA WAREHOUSE?

Um Data Warehouse (ou armazém de dados, ou depósito de dados no Brasil) é um sistema de computação utilizado para armazenar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada. O Data Warehouse é:

Orientado a Assunto: A primeira característica de um Data Warehouse é que ele está orientado ao redor do principal assunto da organização. O percurso do dado orientado ao assunto está em contraste com a mais clássica das aplicações orientadas por processos/funções ao redor dos quais os sistemas operacionais mais antigos estão organizados.

Integrado: Facilmente o mais importante aspecto do ambiente de Data Warehouse é que dados criados dentro de um ambiente de Data Warehouse são integrados. SEMPRE. COM NENHUMA EXCEÇÃO. A integração mostra-se em muitas diferentes maneiras: na convenção consistente de nomes, na forma consistente das variáveis, na estrutura consistente de códigos, nos atributos físicos consistente dos dados, e assim por diante.

Não Volátil: sempre inserido, nunca excluído.

Variante no Tempo: posições históricas das atividades no tempo.

O data warehouse possibilita a análise de grandes volumes de dados coletados dos sistemas transacionais (OLTP). São as chamadas séries históricas que possibilitam uma melhor análise de eventos passados, oferecendo suporte às tomadas de decisões presentes e a previsão de eventos futuros. Por definição, os dados em um data warehouse não são voláteis, ou seja, eles não mudam, salvo quando é necessário fazer correções de dados previamente carregados. Os dados estão disponíveis somente para leitura e não podem ser alterados.

A ferramenta mais popular para exploração de um data warehouse é a Online Analytical Processing OLAP ou Processo Analítico em Tempo Real, mas muitas outras podem ser usadas.

Os data warehouse surgiram como conceito acadêmico na década de 80. Com o amadurecimento dos sistemas de informação empresariais, as necessidades de análise dos dados cresceram paralelamente. Os sistemas OLTP não conseguiam cumprir a tarefa de análise com a simples geração de relatórios. Nesse contexto, a implementação do data warehouse passou a se tornar realidade nas grandes corporações. O mercado de ferramentas de data warehouse, que faz parte do mercado de Business Intelligence, cresceu então, e ferramentas melhores e mais sofisticadas foram desenvolvidas para apoiar a estrutura do data warehouse e sua utilização.

Atualmente, por sua capacidade de sumarizar e analisar grandes volumes de dados, o data warehouse é o núcleo dos sistemas de informações gerenciais e apoio à decisão das principais soluções de business intelligence do mercado.

Segundo Inmon, *Data Warehouse é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão.*

Kimball define assim: *é um conjunto de ferramentas e técnicas de projeto, que quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos permitirá que planejem e construam um data warehouse.*

3. O QUE DATA WAREHOUSE NÃO É

Produto: O Data Warehouse não é um produto e não pode ser comprado como um software de banco de dados. O sistema de Data Warehouse é similar ao desenvolvimento de um ERP, ou seja, ele exige análise do negócio, exige o entendimento do que se quer retirar das informações. Apesar de existirem produtos que fornecem uma gama de ferramentas para efetuar o Cleansing dos dados, a modelagem do banco e da apresentação dos dados, nada disso pode ser feito sem um elevado grau de análise e desenvolvimento.

A linguagem: O sistema de Data Warehouse não pode ser aprendido ou codificado como uma linguagem. Devido ao grande número de componentes e de etapas, um sistema de Data Warehouse suporta diversas linguagens e programações desde a extração dos dados até a apresentação dos mesmos.

Projeto: O sistema de Data Warehouse pode ser pensado mais como um processo. Ele também pode ser pensado como uma série de projetos menores que convergem para a criação de um único sistema de corporativo de Data Warehouse. Devido a natureza evolutiva do DW, é mais fácil aceitá-lo como um processo que está sempre em crescimento do que em um projeto com início-meio-fim, o que definitivamente ele parece mas não é.

Modelagem: O sistema de Data Warehouse não é somente um modelo de banco de dados e não é constituído por mais de um modelo. Existe o processo todo do sistema de BI/DW que compreende todos os procedimentos de ETL, Cleansing e apresentação das informações ao usuário final.

Cópia do sistema OLTP: Alguns acreditam que o sistema de Data Warehouse é somente uma cópia do sistema transacional existente na empresa. Assim como somente um modelo de dados não faz um sistema de BI/DW, uma cópia de um sistema transacional o faz menos ainda. Existem ferramentas que conseguem extrair dados dos sistemas transacionais existentes e criar relatórios a partir das informações coletadas, mas mesmo eles estão montando um pequeno conjunto de metadados e armazenando a informação em algum local.

4. IMPORTANTE SABER SOBRE DATA WAREHOUSE

* Um dos maiores problemas no desenvolvimento do DW é a compreensão dos dados, onde as dimensões devem ser definidas conforme a necessidade de visualização do usuário, ou seja, é tentador pensar que a criação do DW consiste em apenas extrair dados operacionais e inseri-los no Data Warehouse.

* O valor de DW não está em colecionar dados e sim saber gerenciar aqueles dados sendo transformados em informações úteis.

* Considerando complexa a construção de um DW, faz-se necessário um amplo estudo para geração de uma metodologia a fim de se obter sucesso no empreendimento.

Além disso, é necessário saber a respeito de algumas questões que representam verdadeiro desafio na implementação de um Data Warehouse:

* Integração de dados e metadados de várias fontes.

* Qualidade dos dados: limpeza e refinamentos.

* Sumarização e agregação de dados.

* Sincronização das fontes com o Datawarehouse para assegurar a atualização.

* Problemas de desempenho relacionados ao compartilhamento do mesmo ambiente computacional para abrigar as bases de dados corporativas operacionais e o Data Warehouse.

4.1 ARMAZENAMENTO

Um *Data Warehouse* pode armazenar grandes quantidades de informação, às vezes divididas em unidades lógicas menores que são chamadas de Data Marts. O esquema de dados mais utilizado é o "Star Schema" (Esquema Estrela), também conhecido como Modelagem Multidimensional. Apesar de bastante utilizado, não existe um padrão na indústria de software para o armazenamento de dados. Existem, na verdade, algumas controvérsias sobre qual a melhor maneira para estruturar os dados em um *Data Warehouse*. Geralmente, o *Data Warehouse* não armazena informações sobre os processos correntes de uma única atividade de negócio, mas sim cruzamentos e consolidações de várias unidades de negócios de uma empresa.

4.2 MODELAGEM

Os sistemas de base de dados tradicionais utilizam a normalização, no formato de

dados para garantir consistência dos dados e uma minimização do espaço de armazenamento necessário. Entretanto, frequentemente as transações e consultas em bases de dados normalizadas são lentas. Um *Data Warehouse* utiliza dados em formato mais de-normalizados. Isto aumenta a performance das consultas e, como benefício adicional, o processo torna-se mais intuitivo para os utilizadores comuns.

4.3 METADADO

O conceito **Metadado** é considerado como sendo os "dados sobre dados", isto é, os dados sobre os sistemas que operam com estes dados. Um repositório de metadados é uma ferramenta essencial para o gerenciamento de um *Data Warehouse* no momento de converter dados em informações para o negócio. Entre outras coisas, um repositório de metadados bem construído deve conter informações sobre a origem dos dados, regras de transformação, nomes e *alias*, formatos de dados, etc. Ou seja, esse "dicionário" deve conter muito mais do que as descrições de colunas e tabelas: deve conter informações que adicionem valor aos dados.

Tipo de Informação considerada Metadado

Os metadados são utilizados normalmente como um dicionário de informações e, sendo assim, devem incluir:

Origem dos Dados – Todo elemento de dado precisa ter identificado, sua origem ou o processo que o gera. Esta identificação é muito importante no caso de se necessitar saber informações sobre a fonte geradora do dado. Esta informação deve ser única, ou seja, cada dado deve ter uma e somente uma fonte de origem.

Fluxo de Dados – Todo elemento de dado precisa ter identificado os fluxos nos quais sofre transformações. É importante saber que dados servem de base para que processos.

Formato dos Dados – Todo elemento de dados deve ter identificado seu tamanho e tipo de dado.

Nomes e Alias – Todo elemento de dados deve ser identificado por um nome. Este nome pode ser da Área de Negócios ou um nome técnico. No caso de serem usados alias para os nomes, pode-se ter os dois. Devem existir padrões para criação de nomes e alias (ex.: convenções para abreviações), evitando assim ambigüidades.

Definições de Negócio – Estas definições são as informações mais importantes contidas nos metadados. Cada elemento de dado deve ser suportado por uma definição do mesmo no contexto da Área de Negócio. O método de manutenção destas informações também deve ser muito consistente, de forma que o usuário possa obter facilmente definições para as informações desejadas. Nestas definições devem ser evitadas referências a outros metadados que necessitem de uma segunda pesquisa para melhor entendimento.

Regras de Transformação – São consideradas como sendo as Regras de Negócio codificadas. Estas regras são geradas no momento da extração, limpeza e agrupamento dos dados dos Sistemas Operacionais. Cada regra de transformação codificada deve estar associada a um elemento de Metadado. Se mais de uma aplicação contiver a mesma regra de transformação, deverá ser garantido que estas sejam idênticas.

Atualização de Dados – O histórico das atualizações normalmente é mantido pelo próprio banco de dados, mas definir um elemento de metadado, indicando as datas de atualização dos dados, pode ajudar o usuário no momento de verificar a atualidade dos dados e a consistência da dimensão tempo do *Data Warehouse*.

Requisitos de Teste – Identifica os critérios de julgamento de cada elemento de dado. Valores possíveis e intervalos de atuação. Deve conter também padrões para procedimentos de teste destes dados.

Indicadores de Qualidade de Dados – Podem ser criados índices de qualidade baseados na origem do dado, número de processamentos feito sobre este dado, valores atômicos X valores sumariados, nível de utilização do dado, etc.

Triggers Automáticos – Podem existir processos automáticos associados aos metadados definidos. Estes processos ou *triggers* devem estar definidos de forma que possam ser consultados por usuário e desenvolvedores, para que os mesmos não venham a criar situações conflitantes entre as regras definidas nestes processos.

Responsabilidade sobre Informações – Deve ser identificado o responsável por cada elemento de dados do *Data Warehouse* e também o responsável pela entrada de metadados.

Acesso e Segurança – Os metadados devem conter informação suficiente para que sejam determinados os perfis de acesso aos dados. Deve-se poder identificar que usuários podem ler, atualizar, excluir ou inserir dados na base. Deve haver, também, informações sobre quem gerencia estes perfis de acesso e como se fazer contato com o Administrador da Base de Dados.

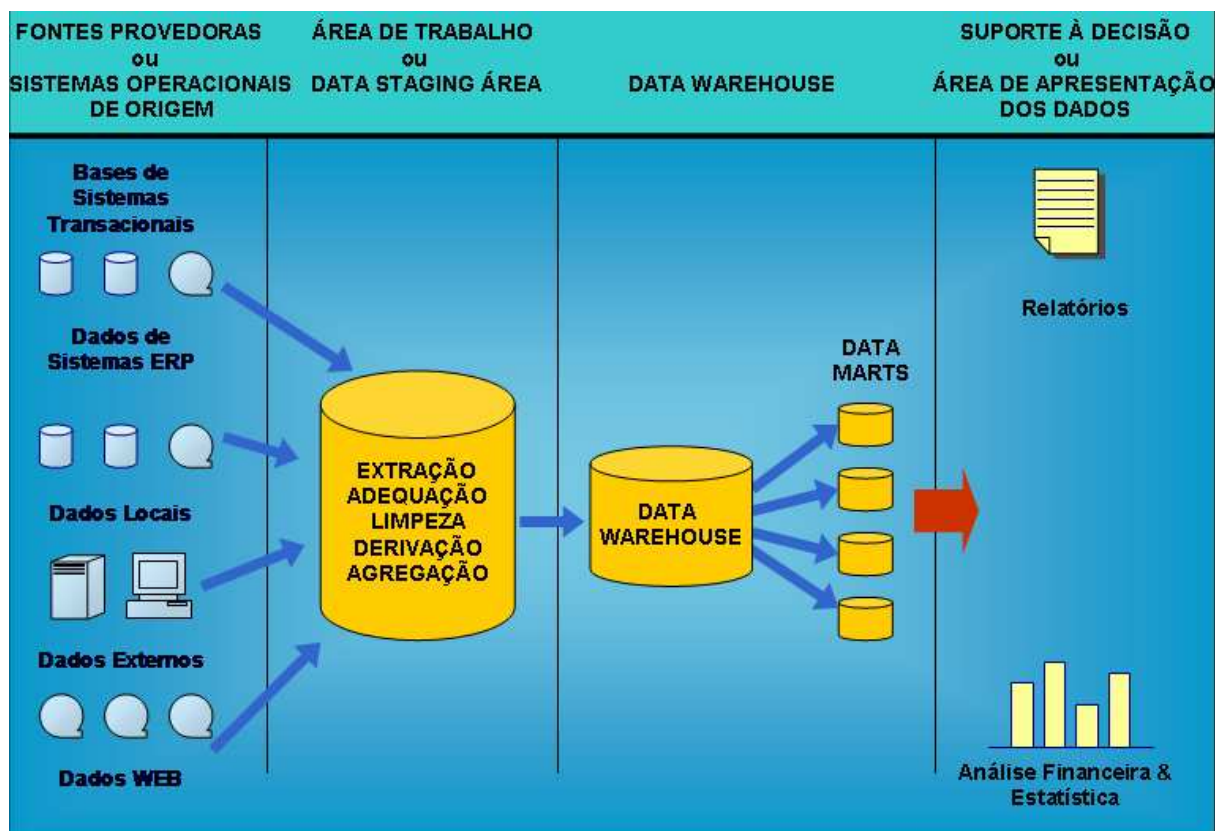
4.4 DATA MARTS

O *Data Warehouse* é normalmente acessado através de *Data Marts*, que são pontos específicos de acesso à subconjuntos do *Data Warehouse*. Os *Data Marts* são construídos para responder prováveis perguntas de um tipo específico de usuário. Por exemplo: um *Data Mart* financeiro poderia armazenar informações consolidadas dia-a-dia para um usuário gerencial e em periodicidades maiores (semana, mês, ano) para um usuário no nível da diretoria. Um *Data Mart* pode ser composto por um ou mais cubos de dados. Hoje em dia, os conceitos de *Data warehouse* e *Data Mart* fazem parte de um conceito muito maior chamado de *Corporate Performance Management*.

4.5 EXTRAÇÃO DE DADOS

Os dados introduzidos num *Data Warehouse* geralmente passam por uma área conhecida como área de *stage*. O *stage* de dados ocorre quando existem processos periódicos de leitura de dados de fontes como sistemas OLTP. Os dados podem passar então por um processo de qualidade, denormalização e gravação dos dados no *Data Warehouse*. Esse processo geralmente é realizado por ferramentas ETL.

5. OS PROCESSOS DE DATA WAREHOUSE



[Adaptado de [SunExpert Magazine](#), Outubro 1998.]

Sistemas operacionais de origem – São os sistemas operacionais de registro ou sistemas transacionais que capturam as transações da empresa. Os sistemas de origem devem ser considerados como externos ao *data warehouse* porque se presume que se tenha pouco ou nenhum controle sobre o conteúdo e o formato dos dados nesses sistemas. Os sistemas de origem também são chamados **Sistemas Legados** ou **OLTP**;

A *data staging* área – É tanto uma área de armazenamento como um conjunto de processos, e normalmente denomina-se ETL (*Extract – Transformation - Load*).

Data Warehouse e Data Mart – A área de apresentação dos dados é o local em que os dados ficam organizados, armazenados e tornam-se disponíveis para serem consultados diretamente pelos usuários, por criadores de relatórios e por outras aplicações de análise. Essa área é tudo o que a comunidade de negócio vê e acessa através das ferramentas de acesso a dados (DB2, ESSBASE, etc). Um *data mart* trata de problema departamental ou local e é definido como um subconjunto

altamente agregado de dados, normalmente escolhido para responder a uma questão de negócio específica ao invés da corporação inteira;

Ferramenta de acesso a dados – O último componente principal do ambiente de *data warehouse* é a ferramenta de acesso a dados. Por definição, toda ferramenta de acesso a dados consulta os dados na área de apresentação do DW.

6. CONCLUSÃO

Através dessas novas tecnologias como o Data Warehouse, permitirá aos administradores descobrir novas maneiras de diferenciar sua empresa numa economia globalizada, deixando-os mais seguros para definirem as metas e adotarem diferentes estratégias em sua organização, conseguindo assim visualizarem antes de seus concorrentes novos mercados e oportunidades atuando de maneiras diferentes conforme o perfil de seus consumidores.

7. REFERÊNCIAS

FACTDATA. In: Business Intelligence & Data Warehouse. 2008. Disponível em: <http://factdata.com/index.php?option=com_search&searchword=data>. Acesso em: 19 abr. 2008.

WIKIPÉDIA. In: Data Warehouse. 2008. Disponível em: http://pt.wikipedia.org/wiki/Data_Warehouse>. Acesso em 26 abr. 2008.

FIC. In: Faculdades Integradas de Caratinga. 2008. Disponível em: <www.ficmg.edu.br/professores/glauber_costa/materias/bd_ii/arquivos/auladw.ppt>. Acesso em 26 abr. 2008.

PARANA. In: Edições. 2008. Disponível em: <<http://www.pr.gov.br/batebyte/edicoes/1997/bb62/warehouse.htm>>. Acesso em: 20 abr. 2008.

DW. In: Data Warehouse. 2008. Disponível em: <<http://www.datawarehouse.inf.br/dw.htm>>. Acesso em: 20 abr. 2008.