

UM PROJETO DE DATA WAREHOUSE

ANGELO LUIZ DE BORTOLI¹

angelodebortoli@onda.com.br

RESUMO

Com a chegada de novas ferramentas tecnológicas de análise de informação, os gerentes começaram a exigir dos Sistemas Transacionais respostas às suas solicitações. Como esses sistemas foram desenvolvidos para garantir a operação da Empresa, não estavam preparados para gerar e armazenar as informações estratégicas necessárias a um *Business Intelligence* eficiente.

Palavras – chave: Data Warehouse, Banco de dados, Informações, Armazenagem.

ABSTRACT

With the arrival of new technological tools of information analysis, the controlling had started to demand of the Systems Do business answers to its requests. As these systems had been developed to guarantee the operation of the Company, they were not prepared to generate and to store the necessary strategical information to a Business efficient Intelligence.

Key – words: Data Warehouse, Data base, Information, Storage.

INTRODUÇÃO

¹ Graduando em Sistemas de Informação pela Faculdade Mater Dei. Residente na Rua Xingu, nº 126. Centro. Pato Branco – PR. CEP: 85.501-230.

Inicialmente analisemos algumas definições, elaboradas por acadêmicos, autores e profissionais especializados em *Data Warehouse*, que podem nos dar uma primeira impressão sobre a Tecnologia.

Diz D. H. 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; trata-se de um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos.”

Segundo R. Kimball: “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*. Para entender o que é um *DW*, é importante fazer uma comparação com o conceito tradicional de banco de dados.”

Conforme Batini: “Um banco de dados é uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de uma empresa específica.” Os dados mantidos por uma empresa são chamados de ‘operacionais’ ou ‘primitivos’.

Dados operacionais referem-se aos dados do banco de dados, distinguindo-se de dados de entrada, dados de saída e outros tipos de dados.

Levando em consideração esta definição sobre dados operacionais, pode-se dizer que um *DW* é, na verdade, uma coleção de dados derivados dos dados operacionais para sistemas de suporte à decisão. Estes dados derivados são, muitas vezes, referidos como dados “gerenciais”, “informativos” ou “analíticos”.

Os bancos de dados transacionais, ou operacionais, armazenam as informações das transações diárias da empresa, são utilizados por todos os funcionários para registrar e executar operações pré-definidas, por isso seus dados podem sofrer constantes mudanças. Por não ocorrer redundância nos dados e as informações históricas não ficarem armazenadas por muito tempo, este tipo de BD não exige grande capacidade de armazenamento.

Já um *DW* armazena dados analíticos, destinados às necessidades da gerência no processo de tomada de decisões. Isto pode envolver consultas complexas que

necessitam acessar um grande número de registros, por isso é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. Um *DW* armazena informações históricas de muitos anos e por isso deve ter uma grande capacidade de processamento e armazenamento dos dados que se encontram de duas maneiras, detalhados e resumidos.

HISTÓRICO

Para se entender o avanço que culminou na chegada do conceito de *Data Warehouse* para a Tecnologia da Informação, é preciso lembrar como evoluíram os processos tecnológicos na área. O primeiro grande passo para os *Data Warehouses* foi dado em decorrência da revolução industrial e das grandes guerras mundiais.

No final dos anos 60, os computadores tornaram-se realmente indispensáveis a qualquer grande organização. Rodavam somente um aplicativo de cada vez, onde esses aplicativos eram executados sobre arquivos mestres. As aplicações eram caracterizadas por relatórios e programas, geralmente em COBOL. O uso de cartões perfurados era comum. Os arquivos mestres eram armazenados em arquivos de fitas magnéticas, que eram adequadas para o armazenamento de um grande volume de dados a baixo custo, mas apresentavam o inconveniente de terem que ser acessadas seqüencialmente.

Por volta de 1970, a época de uma nova tecnologia de armazenamento e acesso a dados, havia chegado: a introdução do armazenamento em disco, ou DASD. Surgiu um novo tipo de *software* conhecido como SGBD ou sistema de gerenciamento de banco de dados. Com o DASD e o SGBD surgiu a idéia de um “banco de dados”, também definido como uma única fonte de dados para todo o processamento. O banco de dados promoveu uma visão de uma organização “baseada em dados”, em que o computador poderia atuar como coordenador central para atividades de toda a empresa. Nesta visão, o banco de dados tornou-se um recurso corporativo básico.

Pela primeira vez as pessoas não estavam vendo os computadores apenas como misteriosos dispositivos de previsão. Em vez disso, os computadores eram vistos como uma verdadeira vantagem competitiva. A idéia dos sistemas de informação para os negócios começou a tomar forma. Em outras palavras, os computadores tornaram-se importantes máquinas de negócios, onde as empresas alcançaram mais eficiência.

Nas décadas de 70 e 80, grandes aperfeiçoamentos tecnológicos resultaram em novos sistemas de informação que custavam bem menos e eram bem mais poderosos. Com o surgimento dos bancos de dados relacionais a informatização nas Empresas já acontecia a passos largos: as pessoas mais influentes e poderosas tinham acesso aos microcomputadores e a sua facilidade de uso aumentou muito. Com o processamento de transações *online* de alta performance, surgiram os sistemas de reservas aéreas em nível mundial, sistemas bancários globais e cartões de créditos internacionais.

A chegada de novas tecnologias, como os PC's e as linguagens de 4ª geração, permitiu-se que o usuário final assumisse um papel mais ativo, controlando diretamente os sistemas e os dados, fora do domínio do clássico processamento de dados. Com essa evolução, as empresas começaram a perceber que poderiam analisar de forma otimizada seus dados, ou seja, descobriram que poderiam incrementar seus recursos de *Business Intelligence*.

Essa descoberta muda o enfoque que até então fora atribuído ao conjunto de informações – sistemas. Nasce um novo conceito para a tecnologia da informação, onde os sistemas informatizados passaram a pertencer a dois grupos:

- Sistemas que tratam o negócio: dão suporte ao dia a dia do negócio da empresa, garantem a operação da empresa, e são chamados de **SISTEMAS TRANSACIONAIS**;
- Sistemas que analisam o negócio: sistemas que ajudam a interpretar o que ocorreu e a decidir sobre estratégias futuras para a empresa – compreendem os **SISTEMAS DE SUPORTE A DECISÃO**.

Com a chegada de novas ferramentas tecnológicas de análise de informação, os gerentes começaram a exigir dos Sistemas Transacionais respostas às suas solicitações. Como esses sistemas foram desenvolvidos para garantir a operação da Empresa, não estavam preparados para gerar e armazenar as informações estratégicas necessárias a um *Business Intelligence* eficiente.

Em atendimento às solicitações dos gestores em relação à deficiência da análise de informação nos sistemas legados, surgiu no mercado os chamados Programas Extratores. Esses programas extraem informações dos Sistemas Transacionais com o intuito de trabalhá-las em outros ambientes. Muitas vezes essas extrações ocorriam em arquivos intermediários, onde as informações sofriam novos tratamentos. Isso provocava uma falha na integridade das informações acarretando, muitas vezes, uma falta de credibilidade dos dados, uma queda da produtividade e a informação sendo publicada com valores diferentes. Além disso, pelo fato de que os Sistemas Transacionais geravam um grande volume de dados e pela diversidade dos sistemas implantados nas empresas as pesquisas – relatórios – realizadas eram produzidas muito lentamente. Nos tempos do Clipper e do Cobol fazer um relatório desse nível significava perder muitas horas sobre o computador, pois se fazia necessário que fossem extraídos os dados de vários sistemas, muitas vezes esses não conversavam entre si.

Apesar dessas razões, é importante salientar que é possível a prática de *Business Intelligence* com os sistemas operacionais da empresa, e com outras fontes de dados, como planilhas eletrônicas e dados em papel, mas esse procedimento implica em grande possibilidade de equívocos, já que esses dados são oriundos de várias fontes independentes, e não possuem entre si relação de integridade. Outro fator importante que prejudicava as decisões foi a falta de registro dos fatos históricos nos Sistemas Transacionais, pois estes trabalhavam com uma situação instantânea dos negócios.

Para resolver este problema, começou-se a estudar uma forma de se armazenar a informação contida nos sistemas transacionais numa base de dados central, para que houvesse integração total dos dados da empresa. Além disso, era

necessário manter o histórico das informações e fazer com que ela fosse disposta dimensionalmente, ou seja, o analista de negócios poderia visualizar um mesmo fato através de diversas dimensões diferentes. O nome dado a essa modalidade de Sistema de Apoio à Decisão foi o *Data Warehouse*, ou em português, Armazém de Dados.

Com o surgimento do *DATA WAREHOUSE* são necessários novos métodos de estruturação de dados, tanto para armazenamento quanto para a recuperação de informações. Cabe ressaltar que as perspectivas e técnicas necessárias para projetar o *DATA WAREHOUSE* são profundamente diferente dos SISTEMAS TRANSACIONAIS. Os usuários, o conteúdo dos dados, a estrutura dos dados, o *hardware* e o *software*, a administração, o gerenciamento dos sistemas, o ritmo diário, as solicitações, as respostas e o volume de informações são diferentes.

Entender essa tecnologia com certeza ajudará os empresários a descobrir novas tendências e caminhos para competir numa economia globalizada, onde a concorrência é acirrada, trazendo melhores produtos ou serviços para o mercado com maior rapidez sem aumento dos custos.

CARACTERÍSTICAS

Segundo Inmon, um *DW* deve ser orientado por assuntos, integrado, variável no tempo e não volátil. Essas são as principais características de um *DW* as quais iremos descrever em maiores detalhes o que quer dizer cada uma delas logo abaixo.

1. Orientação por Assunto

Trata-se de uma característica marcante de um *DW*, pois toda modelagem será voltada em torno dos principais assuntos da empresa. Enquanto todos os sistemas transacionais estão voltados para processos e aplicações específicas, os *DW's* objetivam assuntos.

2. Integração

Esta característica talvez seja a mais importante do *DW*. É através dela que iremos padronizar uma representação única para os dados de todos os sistemas que formarão a base de dados do *DW*. Por isso, grande parte do trabalho na construção de um *DW* está na análise dos sistemas transacionais e dos dados que eles contêm. Esses dados geralmente encontram-se armazenados em vários padrões de codificação, isso se deve aos inúmeros sistemas existentes nas empresas, e que eles tenham sido codificados por diferentes analistas. Isso quer dizer que os mesmos dados podem estar em formatos diferentes.

3. Variação no Tempo

Segundo W. H. Inmon, “os *Data Warehouses* são variáveis em relação ao tempo”. Isso quer dizer que em um *DW* é normal mantermos um horizonte de tempo bem superior ao dos sistemas transacionais, enquanto no OLTP mantemos um histórico curto dos dados; no *DW* guardamos esses dados num período maior. Isso é bastante lógico porque num sistema transacional a finalidade é de fornecer as informações no momento exato, já no *Data Warehouse*, o principal objetivo é analisar o comportamento das mesmas durante um período de tempo maior. Fundamentados nessa variação, os gerentes tomam as decisões baseados em fatos e não em intuições. Seguindo a mesma linha de raciocínio é válido dizer que os dados nos sistemas transacionais estão sendo atualizados constantemente, cuja exatidão é válida somente para o momento de acesso. Os dados existentes num *DW* são como fotografias que refletem os mesmos num determinado momento do tempo. Essas fotografias são chamadas de *snapshots*.

A dimensão tempo, sempre estará presente em qualquer fato de um *DW*, isso ocorre porque, como falamos anteriormente, sempre os dados refletirão num determinado momento de tempo, e obrigatoriamente deverá conter uma chave de tempo para expressar a data em que os dados foram extraídos. Portanto podemos dizer que os dados armazenados corretamente no *DW* não serão mais atualizados tendo-se assim uma imagem fiel da época em que foram gerados.

Assim como os dados é importante frisar que os metadados, também possuem elementos temporais, porque mantêm um histórico das mudanças nas regras de negócio da empresa. Os metadados são responsáveis pelas informações referentes ao caminho do dado dentro do DW.

4. Não Volatilidade

No *DW* existem somente duas operações, a carga inicial e as consultas dos *front-ends* aos dados. Isso pode ser afirmado porque a maneira como os dados são carregados e tratados é completamente diferente dos sistemas transacionais. Enquanto nesses sistemas temos vários controles e *updates* de registros, no *DW* temos somente *inserts* e *selects* de dados. Por exemplo, num sistema de contabilidade podemos fazer alterações nos registros. Já no *DW*, o que acontece é somente ler os dados na origem e gravá-los no destino, ou seja, no banco modelado multidimensional.

Deve-se considerar que os dados sempre passam por filtros antes de serem inseridos no *DW*. Com isso muitos deles jamais saem do ambiente transacional, e outros são tão resumidos que não se encontram fora do *DW*. “Em outras palavras, a maior parte dos dados é física e radicalmente alterada quando passam a fazer parte do *DW*. Do ponto de vista de integração, não são mais os mesmos dados do ambiente operacional. À luz destes fatores, a redundância de dados entre os dois ambientes raramente ocorre, resultando em menos de 1 por cento de duplicações”, essa definição dada por Inmon é muito válida.

5. Localização

Os dados podem estar fisicamente armazenados de três formas:

- Centralizado: Num único local centralizando o banco de dados em um *DW* integrado, procurando maximizar o poder de processamento e agilizando a busca dos dados. Esse tipo de armazenagem é bastante utilizada, porém há o inconveniente do investimento em *hardware* para comportar a base de dados

muito volumosa, e o poderio de processamento elevado para atender satisfatoriamente as consultas simultâneas de muitos usuários.

- Distribuídos: são *Data Marts*, armazenados por áreas de interesse. Essa pode ser uma saída interessante para quem precisa de bastante performance, pois isso não sobrecarrega um único servidor, e as consultas serão sempre atendidas em tempo satisfatório.
- Por níveis de detalhes: processo em que as unidades de dados são mantidas no *DW*. Pode-se armazenar dados altamente resumidos num servidor, dados resumidos noutra nível de detalhe intermediário no segundo servidor e os dados mais detalhados – atômicos – num terceiro servidor. Para mudar de nível é necessário que ocorra um dos seguintes eventos: os dados são sintetizados, arquivados ou eliminados.

6. Credibilidade dos Dados

A credibilidade dos dados é o muito importante para o sucesso de qualquer projeto. Discrepâncias simples de todo tipo podem causar sérios problemas quando se quer extrair dados para suportar decisões estratégicas para o negócio das empresas. Dados não dignos de confiança podem resultar em relatório inúteis, que não têm importância alguma. "Se você tem dados de má qualidade e os disponibiliza em um *DW*, o seu resultado final será um suporte à decisão de baixo nível com altos riscos para o seu negócio", afirma Robert Craig, analista do Hurwitz Group.

"Não é apenas a escolha da ferramenta certa que influi na qualidade dos dados", afirma Richard Rist, vice-presidente *Data Warehousing Institute*. Segundo ele, conjuntos de coleções de dados, processos de entrada, metadados e informações sobre a origem dos dados, são importantíssimos. Outras questões como a manutenção e atualização dos dados e as diferenças entre dados para bancos transacionais e para uso em *Data Warehousing* também são cruciais para o sucesso dos projetos.

7. Granularidade

Granularidade nada mais é do que o nível de detalhe ou de resumo dos dados existentes num *DW*. Quanto maior for o nível de detalhes, menor será o nível de granularidade. O nível de granularidade afeta diretamente o volume de dados armazenados no *DW*, e ao mesmo tempo o tipo de consulta que pode ser respondida.

8. Metadados

Os Metadados são um dos tópicos mais interessantes e, de certa forma, confusos do ambiente do *Data Warehouse*. Interessantes por serem os dados de controle de um projeto de *DW*. Confusos por não terem uma definição muito clara para a maioria das pessoas. Metadados são dados que fazem referência a outros dados.

Todas as fases de um projeto de *Data Warehouse*, desde a modelagem até a visualização da informação, geram metadados. Neles estarão contidos informações como atributos das tabelas, agregadas utilizadas, cálculos necessários, descrições, periodicidade das cargas, histórico de mudanças etc.

Segundo Inmon, os metadados mantêm informações sobre “o que está e onde”, no *DW*. Tipicamente os aspectos que sobre os quais os metadados mantêm informações são:

- A estrutura dos dados, segundo a visão do programador;
- A estrutura dos dados, segundo a visão dos analistas de SAD;
- A fonte de dados que alimenta o *DW*;
- A transformação sofrida pelos dados no momento de sua migração para o *DW*;
- O modelo de dados;
- O relacionamento entre o modelo de dados e o *DW*;
- O histórico das extrações de dados.

9. Processos de Carga

Esta etapa é uma das fases mais críticas de um *Data Warehouse*, pois envolve a fase de extração dos dados dos sistemas transacionais ou de outras fontes, como planilhas, arquivos ou textos. A fase de Filtragem consiste basicamente em garantir a integridade dos dados e, por fim, a fase de Carga dos Dados no *Data Warehouse*.

Quando os dados são movidos de sistemas transacionais para o ambiente de *Data Warehouse*, parece que nada além de simples extrações de dados de um local para o outro está ocorrendo. Em virtude desta enganosa simplicidade, muitas vezes as empresas acabam perdendo tempo e dinheiro por ter que refazer toda esta parte de extração.

O processo de carga dos dados passa por três etapas: extração, filtragem e a carga propriamente dita.

A extração de dados do ambiente operacional para o ambiente de *data warehouse* demanda uma mudança na tecnologia. Pois muitas vezes os dados são transferidos de um banco de dados hierárquico, tal como o ADABAS, para uma nova tecnologia de SGBD para *Data Warehouse*.

A seleção de dados do ambiente operacional pode ser muito complexa, pois muitas vezes é necessário selecionar vários campos de um sistema operacional para compor um único campo no *data warehouse*. Os dados são reformatados. Podem existir várias fontes de dados diferentes para compor uma informação.

Quando há vários arquivos de entrada, a escolha das chaves deve ser feita antes que os arquivos sejam intercalados. Isso significa que, se diferentes estruturas de chaves são usados nos diferentes arquivos de entrada, então deve-se optar por apenas uma dessas estruturas. Os arquivos devem ser gerados obedecendo a mesma ordem das colunas estipuladas no ambiente de *data warehouse*.

Valores padrões devem ser fornecidos. Às vezes pode existir um campo no *data warehouse* que não possui fonte de dados, então a solução é definir um valor padrão para estes campos. *Data warehouse* espelha as informações históricas necessárias, enquanto o ambiente operacional focaliza as informações correntes.

Após a definição de como deverão ficar os dados no *data warehouse*, há a necessidade de filtragem dos dados para colocá-los no padrão definido.

O momento de carga é a parte de Integridade dos dados, onde se faz necessário checar os campos que são chaves estrangeiras com suas respectivas tabelas para certificar-se de que os dados existentes na tabela da chave estrangeira estão de acordo com a tabela da chave primária.

A carga incremental normalmente é feita para tabelas fatos, e a carga por cima dos dados é feita em tabelas dimensões, onde o analista terá que deletar os dados existentes e incluí-los novamente. Apesar de existirem ferramentas de Carga como o DTS (*Data Transformation Service*), ainda tem-se a necessidade de criar rotinas de carga para atender determinadas situações que poderão ocorrer.

10. Metodologia de Levantamento

Apesar de serem displicentemente ignoradas em muitos *Data Warehouses*, as metodologias de levantamento de dados gerenciais são indispensáveis ao sucesso de um Sistema de Apoio à Decisão que pretende atender às necessidades do usuário de negócio. Quando se fala em *DW*, muitos profissionais da área de TI pensam logo em construir rotinas de extração de dados dos sistemas legados para posterior carga num modelo dimensional, em detrimento de um entendimento das necessidades dos entendedores de negócio.

Para tal entendimento, foram criadas metodologias de levantamento de dados Gerenciais, como a JAD e o DMD, que são baseadas em reuniões de trabalho, onde os participantes, orientados por um profissional com prática nesta etapa, extraem conhecimentos sobre o negócio.

Necessariamente, ao aplicar a metodologia, a única preocupação é com termos e questões gerenciais. Alguns profissionais aplicam esta etapa já pensando na base de dados, com suas dimensões e fatos, causando assim confusão na cabeça dos entendedores do negócio e uma maior possibilidade de falhas na modelagem posterior. Em alguns casos, quando aplicada a todos departamentos de uma empresa, a metodologia provoca fenômenos interessantes como a descoberta de

processos e análises redundantes, fazendo com que a própria corporação seja otimizada.

Portanto, fica clara a necessidade de se implementar uma metodologia de levantamento de dados gerenciais, antes de se iniciar a implantação física de um *Data Warehouse*. Além de gerar como produto um SAD bem estruturado e modelado, esse procedimento também pode ser muito benéfico para a saúde organizacional da empresa.

CONCLUSÃO

Com base nestes conceitos podemos concluir que o *DW* é um conjunto de técnicas e bancos de dados integrados, projetados para suportar as funções dos Sistemas de Apoio à Decisão, onde cada unidade de dados está relacionada a um determinado assunto, ou fato. Esses bancos de dados é que darão subsídio de informações aos gerentes e diretores de empresas, para analisarem tendências históricas dos seus clientes e com isso melhorarem os processos que aumentem a satisfação e fidelidade dos mesmos. No *DW* os dados podem ser retirados de múltiplos sistemas de computação normalmente utilizados há vários anos e que continuam em operação, como também podem ser de fontes externas da empresa. *Data Warehouses* são construídos para que tais dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais. Os dados de um *DW* podem ser compostos por um ou mais sistemas distintos e sempre estarão separados de qualquer outro sistema transacional, ou seja, deve existir um local físico onde os dados desses sistemas serão armazenados.

REFERÊNCIAS

DATA WAREHOUSE. Disponível em <http://www.datawarehouse.inf.br>. Acesso em 21 de março de 2004.

DATA WAREHOUSE. Disponível em <http://www.assuncao.eti.br/luisdwh.htm>. Acesso em 01 de abril de 2004.

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

_____; WELCH, J. D.; GLASSEY, K. L. **Gerenciando Data Warehouse: técnicas práticas para monitorar operações e performances, administrar dados e ferramentas, gerenciar alterações e crescimento**. 1ª ed.- São Paulo: Ed. Makron Books, 1999.