DATA WAREHOUSE

Análise de dados (Big Data)

Data Warehouse

O que é Data Warehouse?

Um data warehouse (DW), ou armazém de dados é um banco de dados com dados históricos usados para análise e decisões das mais exóticas perguntas realizadas por executivos. Os dados contidos nos data warehouse são sumarizados, periódicos e descritivos. Com a manipulação desses dados os executivos podem tomar decisões baseadas em fatos e não em intuições e especulações. Os data warehouses são projetados para processamento on-line analítico (OLAP, *On-line Analytical Processing*) ao invés do processamento transacional on-line (OLTP, *On-line Transactional Processing)*.

Data Warehouse

O que é Data Warehouse?

Ferramentas OLAP para pesquisa inteligente de dados são chamadas de *data mining*. Delimitando a abrangência dos dados a uma área de negócio da empresa o data warehouse passa a se denominar *data mart*. É possível implementar um data warehouse com vários *data marts* distribuídos.

Data Warehouse

O que é Data Warehouse?

No mercado competitivo atual uma decisão errada pode decretar a morte de uma empresa. Decisões baseadas em dados fragmentados obtidos pelos sistemas de informações tradicionais não oferecem uma informação consistente, caso não exista uma forte integração entre eles. Um data warehouse concentra dados de diversos sistemas estruturados e outras bases de dados, em diferentes plataformas. Os dados antes de serem armazenados são filtrados, normalizados, reorganizados, sumarizados para constituírem uma base de dados confiável e íntegra. Muitas vezes uma informação está representada sob diversas formas, dependendo do sistema de informação. Por exemplo, um código de fornecedor pode ser diferente em dois ou mais bancos de dados.

Data Warehouse

O que é Data Warehouse?

Um data warehouse é projetado para garimpar informações escondidas nas montanhas de dados de uma empresa.

Os sistemas de informações são desenvolvidos e implementados visando o controle de um determinado processo na empresa. Em alguns casos, nem mesmo os analistas de sistemas conseguem ter a visão do todo. A maioria dos sistemas de informação é parametrizada, onde as pesquisas às informações são pré-definidas, não oferecendo flexibilidade ao usuário final (nem aos próprios analistas) para criar novas pesquisas de forma ágil e rápida.

Os data warehouses tem como premissa resolver essa questão, dando ao usuário final a flexibilidade necessária para pesquisas, mesmo para as mais exóticas. Foi dessa forma que a cadeia americana de supermercados Wal-Mart descobriu uma relação entre o consumo de fraldas descartáveis e o consumo de cervejas.

Data Warehouse

O que é Data Warehouse?

O banco de dados de um data warehouse deve ser projetado para processamento analítico on-line (OLAP), onde caracteriza-se pela ênfase na performance da recuperação das informações. Orientado à análise e processos de decisão pelos usuários finais através do uso de ferramentas especialmente desenvolvidas para o cruzamento multidimensional dos dados, os *data mining*. Essas ferramentas podem descobrir associações que nem mesmo o usuário imaginaria pesquisar. Os *data mining* são mais eficientes se usados em *data marts*, pois estes são orientados a determinados assuntos da empresa.

Os data warehouses devem permitir o download de informações para a utilização em outras ferramentas, tais como: planilhas eletrônicas e outros bancos de dados. Diferente dos bancos de dados orientados à transações on-line em tempo-real que trabalham centrados nas operações do dia-a-dia da empresa.

Data Warehouse

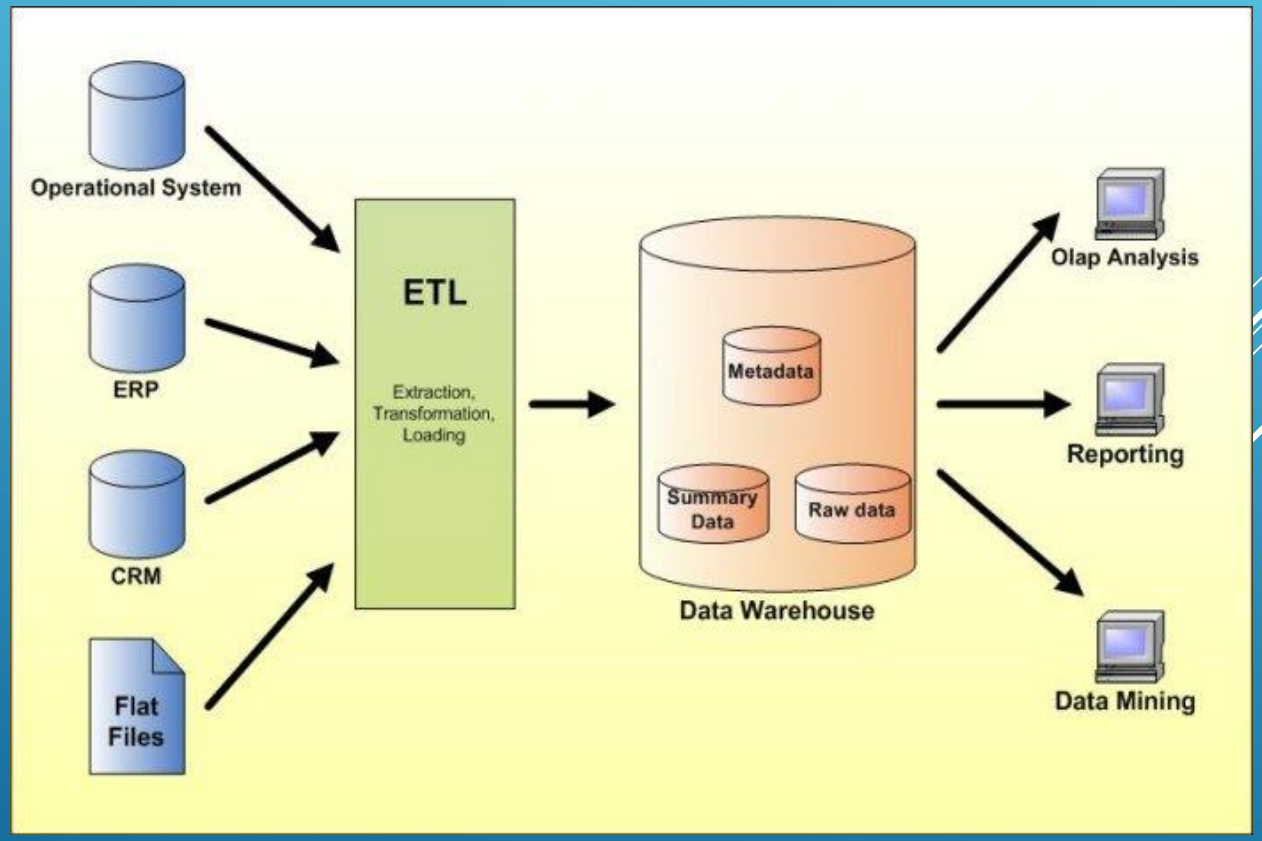
Benefícios do Data Warehouse:

- Mantém o histórico de dados, mesmo se os sistemas transacionais não os fizerem;
- Integra os dados de vários sistemas, permitindo uma visão consolidada de toda a operação, principalmente quando uma organização possui várias empresas com sistemas de informações diferentes e trabalha agressivamente em aquisições e fusões;
- Melhora a qualidade dos dados, criando uma padronização de códigos e descrições e identificando e corrigindo dados ruins;
- Apresenta as informações da organização de forma consistente;
- Fornece um único modelo de dados para toda a organização, independente da fonte;
- Reestrutura os dados de modo a satisfazer as necessidades dos usuários do negócio;
- Reestrutura os dados para melhorar o desempenho de consulta, mesmo para consultas analíticas complexas, sem afetar os sistemas em operação;
- Agrega valor às aplicações de negócio operacional, principalmente a gestão de relacionamento com clientes (CRM).

Data Warehouse

Arquitetura: A arquitetura construída sobre a base de dados gera relatórios para análise e estudo de ampla quantidade de informações obtidas. O data warehouse possibilita facilidade de análise para decisões estratégicas.

Os dados nesse sistema não são voláteis, não mudam, somente em caso de correções prévias, pois cada informação está disponível somente para leitura. A ferramenta mais utilizada para a geração de determinados dados é a Online Analytical Processing (OLAP).



Data Warehouse

Retrosperctiva

O conceito sobre o data warehouse surgiu nos anos 80, fase do amadurecimento dos sistemas de informação empresarial. Nessa época, o sistema OLTP não respondia à crescente demanda de análise via simples geração de relatórios.

O data warehouse passou a vigorar em grande uso nas grandes empresas, sendo uma ferramenta pertencente aoBusiness Intelligence, um mercado de inteligência e soluções por meio da informática.

Data Warehouse

Retrosperctiva

O sistema faz parte de núcleos de sistemas de informações gerenciais e apoio à decisão estratégica. As principais etapas desse sistema são o Armazenamento, a Modelagem Multidimensional e o Metadado.

- O Armazenamento ocorre a partir de um depósito único de rápido acesso para análises solicitadas, contendo dados de eventos passados de bancos transacionais. Busca-se a maior quantidade de informações possíveis para orientar novas decisões.
- A Modelagem Multidimensional ocorre pela utilização de normalização por parte dos sistemas de base de dados tradicionais, no objetivo de alcançar a consistência dos dados, aproveitamento de espaço para o armazenamento e mitigação de repetições de informação. A utilização de dados em formato normalizados amplia o desempenho das consultas.
- O Metadado é dado sobre dados, ou seja, dados sobre os sistemas que operam sobre determinados dados. A ferramenta essencial para o gerenciamento de um Data Warehouse é o repositório de metadados, devendo conter informações sobre origem de dados , regras, transformações e formatos. O metadado abrange : origem dos dados, fluxo de dados, formato dos dados, definições de negócio, etc.

Data Warehouse

Modelagem dimensional

A modelagem dimensional é uma metodologia que permite modelar logicamente dados para melhorar o desempenho de consultas e prover facilidade de utilização a partir de um conjunto de eventos básicos de medição. Os modelos dimensionais são compreensíveis, previsíveis, ampliáveis e resistentes ao ataque específico de grupos de usuários de negócio, por se manter fiel à simplicidade, ter uma perspectiva voltada para as necessidades analíticas da empresa, e especialmente ao seu formato simétrico, em que todas as dimensões normalmente são iguais pontos de entrada na tabela de fatos [KIMBALL, 2002]. Os modelos dimensionais são a base de muitos aprimoramentos de desempenho SGBD, inclusive agregações e métodos de indexação avançados.

Data Warehouse

Modelo Dimensional para Data Warehouse

O modelo dimensional para construção de banco de dados para Data Warehouse é uma forma de modelagem onde as informações se relacionam de forma que pode ser representada como um cubo. Sendo assim podemos fatiar este cubo e aprofundar em cada dimensão ou eixo para extrair mais detalhes sobre os processos internos que ocorrem na empresa que em um modelo relacional torna-se muito complicados de serem extraídos e muitas vezes até impossíveis de serem analisadas.

Data Warehouse

Modelo Dimensional para Data Warehouse

O modelo dimensional permite visualizar dados abstratos de forma simples e relacionar informações de diferentes setores da empresa de forma muito eficaz.

O que torna o Data Warehouse mais poderoso é que informações que se situam em vários sistemas, planilhas e arquivos espalhados por todos os setores da empresa, são reunidos em um banco de dados de forma dimensional, sendo assim tendo informações unificadas e padronizadas em um mesmo local.

Data Warehouse

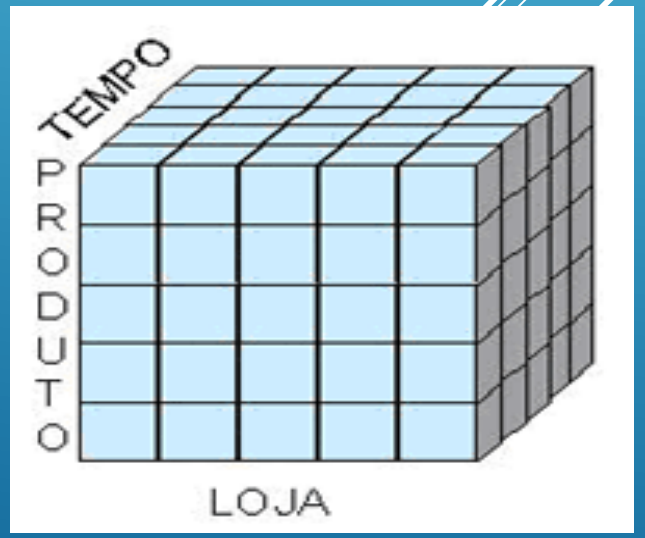
Cubo

Vejamos o caso de uma empresa que possui várias lojas filiais e que deseja acompanhar o desempenho de suas vendas ao longo do tempo. Um desenhista de Data Warehouse visualiza estas informações de uma forma como um cubo que pode ser descrito com três dimensões principais que são:

- Tempo
- Loja
- Produto

Na intersecção destas três dimensões está a quantidade de produtos que foi vendido.

Neste modelo cada cubo menor, ou seja, a intersecção entre as dimensões ou eixos representa uma quantidade de um produto que foi vendido em uma determinada loja em uma data especifica.



Data Warehouse

Detalhando o Cubo anterior

Mas se quisermos saber e controlar também se os produtos que foram vendidos participavam de uma promoção teríamos que ter mais uma dimensão chamada **PROMOÇÃO**, e se quisermos controlar em cada momento as equipes de **marketing** que atuaram em cima das promoções e das lojas devemos ter mais outra dimensão, e se quisermos controlar os **clientes** que compraram os produtos teríamos que ter uma dimensão Clientes, sendo assim teríamos um modelo com seis dimensões. Tantas dimensões não é possíveis desenhar graficamente, mas seguem o mesmo conceito de cubo, pois é possível navegar, aprofundar-se, detalhar e acompanhar os desempenhos destas dimensões ao longo do tempo. Um modelo dimensional pode ter quantas dimensões forem necessária.

Data Warehouse

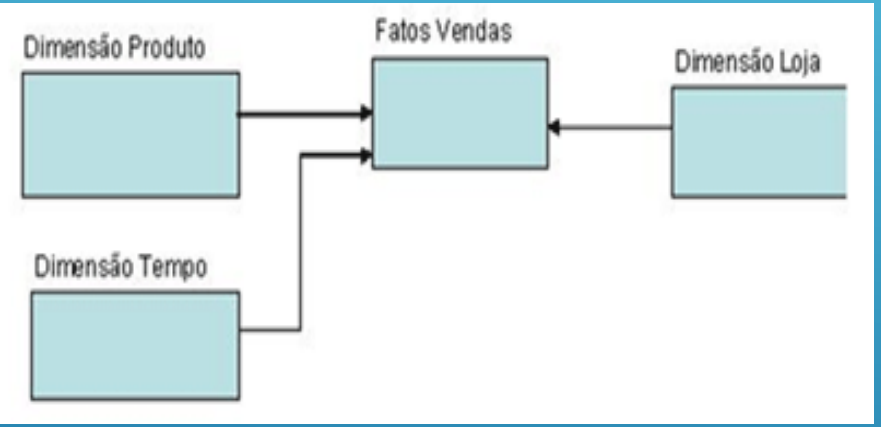
Modelo Dimensional para Data Warehouse

Um modelo de dados dimensional é extremamente simples, isto facilita para os usuários deste banco de dados identificarem onde estão localizadas as informações e permite que os softwares naveguem por estes bancos de dados com eficiência. Um outro fator importante para a modelagem dimensional é a velocidade de acesso a uma informação, com modelos simples sem muitas tabelas para relacionar, é muito rápido para extrair as informações necessárias.

Data Warehouse

Modelo Dimensional para Data Warehouse

Um modelo dimensional conta basicamente com uma tabela de fatos central e tabelas dimensionais ligadas diretamente a elas.



Os Fatos e Dimensões são tabelas do banco de dados, só que no modelo dimensional adquirem nomes de Fatos e Dimensões de acordo com a função da tabela.

Data Warehouse

Modelo Dimensional para Data Warehouse

Uma tabela de Fatos, em nosso exemplo "Fatos Vendas" contém medições sobre o negócio como a quantidade de produtos que foi vendido, contém o valor da venda e o valor unitário do produto vendido. Além destas informações de fatos, esta tabela contém chaves para as tabelas de dimensões. Uma tabela de fatos é extremamente grande referente à quantidade de registros que contém, neste exemplo ela armazena todas as vendas de cada produto feitas em cada loja todos os dias. É comum uma tabela de fatos alcançar alguns Gibabytes logo nos primeiros meses de uso do Data Warehouse.

Data Warehouse

Modelo Dimensional para Data Warehouse

As tabelas de Dimensões contém descrições textuais sobre cada um elementos que fazem parte do processo, no exemplo que citamos temos três dimensões (Tempo, Loja e Produto) as tabelas dimensionais contém vários atributos que descrevem em detalhes todas as características que possam definir e serem úteis para futuras pesquisas no Data Warehouse.

Data Warehouse

Modelo Dimensional para Data Warehouse

A dimensão Produto deve ter descrições curtas e detalhadas sobre o produto, deve também ter o tamanho, peso, categoria, cor, departamento, marca, tipo da embalagem, etc. Ou seja todos os atributos que podem definir o produto e que possam ser utilizados para futuras pesquisas e analises que ajudarão o empresário a tomar decisões sobre seu negócio.

A dimensão Loja deve conter informações sobre as lojas que fazem parte do complexo do negócio, dentre estas informações deve ter descrições como endereço, CEP, região, cidade, bairro, telefone, gerente, etc.

A dimensão Tempo deve ter detalhes sobre o calendário para que facilite pesquisas estratégicas, então a dimensão tempo não deve ter somente a data em que o produto foi vendido, mas deve conter informações como dia no mês, dia na semana, número do dia na semana, mês, número do mês no ano, ano, número da semana no ano, número de semanas corridas, número de meses corridos trimestre, período fiscal, indicador de feriado, indicador de fim de semana, indicador de último dia do mês, etc.

Data Warehouse

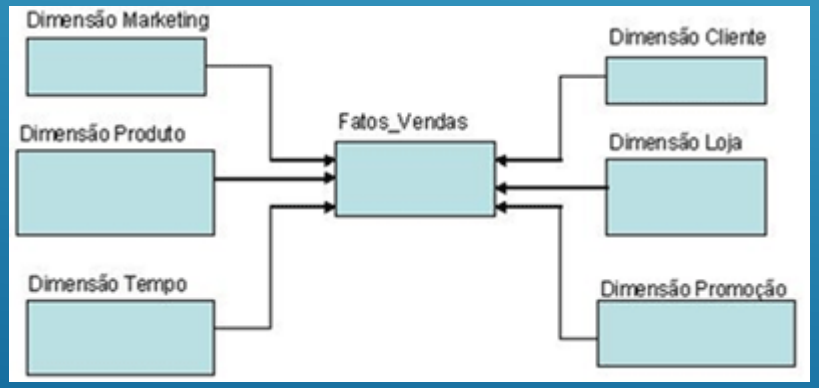
Modelo Dimensional para Data Warehouse - Tipos de Modelos Dimensionais

- O Modelo Estrela (Star Schema)
- O Modelo Floco de Neve (Snow Flake)

Data Warehouse

Tipos de Modelos Dimensionais – Estrela (star Schema)

No modelo estrela todas as tabelas relacionam-se diretamente com a tabela de fatos, sendo assim as tabelas dimensionais devem conter todas as descrições que são necessária para defini uma classe como Produto, Tempo ou Loja nela mesma, ou seja, as tabelas de dimensões não são normalizadas no modelo estrela, então campos como Categoria, Departamento, Marca contém suas descrições repetidas em cada registro, assim aumentando o tamanho das tabelas de dimensão por repetirem estas descrições de forma textual em todos os registros.

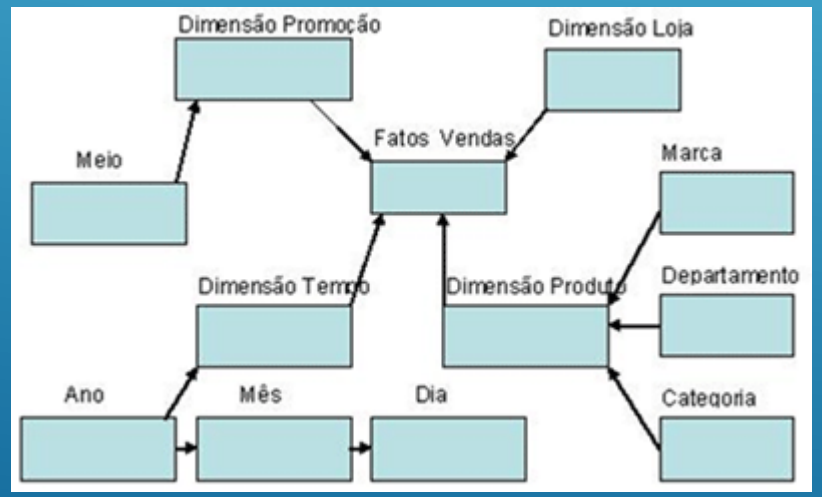


O esquema estrela é uma estrutura simples, com poucas tabelas e ligações(relacionamentos) bem definidas (POE, KLAUER, BROBST, 1998), assemelha-se ao modelo de negocio, o que facilita a leitura e entendimento, não só pelos analistas, como por usuários finais não familiarizados com estruturas de banco de dados. Permite a criação de um banco de dados que facilita a execução de consultas complexas, podendo ser realizadas de modo eficiente e intuitivo pelo usuário

Data Warehouse

Tipos de Modelos Dimensionais – Floco de neve (Snow Flake)

No modelo Floco as tabelas dimensionais relacionam-se com a tabela de fatos, mas algumas dimensões relacionam-se apenas entre elas, isto ocorre para fins de normalização das tabelas dimensionais, visando diminuir o espaço ocupado por estas tabelas, então informações como Categoria, Departamento e Marca tornaram-se tabelas de dimensões auxiliares.



No modelo Floco existem tabelas de dimensões auxiliares que normalizam as tabelas de dimensões principais. Na figura anterior estas tabelas são (Ano, Mês e Dia) que normalizam a Dimensão Tempo, (Categoria, Departamento e Marca) que normalizam a Dimensão Produto e a tabela Meio que normaliza a Dimensão Promoção.

Data Warehouse

Considerações sobre ambos Modelos Dimensionais

O Modelo Floco (Snow Flake) reduz o espaço de armazenamento dos dados dimensionais mas acrescenta várias tabelas ao modelo, deixando-o mais complexo, tornando mais difícil a navegação pelos softwares que utilizarão o banco de dados. Um outro fator é que mais tabelas serão utilizadas para executar uma consulta, então mais JOINS de instrução SQL serão feitos, tornando o acesso aos dados mais lento do que no modelo estrela.

O Modelo Estrela (Star Schema) é mais simples e mais fácil de navegação pelos softwares, porém desperdiça espaço repetindo as mesmas descrições ao longo de toda a tabela, porém análises feitas mostram que o ganho de espaço normalizando este esquema resulta em um ganho menor que 1% do espaço total no banco de dados, sendo assim existem outros fatores mais importantes para serem avaliados para redução do espaço em disco como a adição de agregados e alteração na granularidade dos dados, estes temas serão abordados em colunas posteriormente.

Data Warehouse

Considerações sobre ambos Modelos Dimensionais

Portanto, o que é recomendado, é utilizar um modelo estrela, pois fornece um acesso mas rápido aos dados e é mais fácil de se navegar. Criar tabelas auxiliares para dimensões, somente para dimensões especificas quando for estritamente necessário ou quando demonstrar um beneficio que justifique a perda de desempenho nas consultas, que pode não ser tão grande dependendo da forma que estas tabelas são construídas e a quantidade de registros que elas contiverem.

Data Warehouse

Fontes:
http://pt.wikipedia.org/wiki/Data_Warehouse
http://conteudo.imasters.com.br/1446/datawarehouse.ppt
http://imasters.com.br/artigo/3836/gerencia-de-ti/modelo-dimensional-para-data-warehouse/