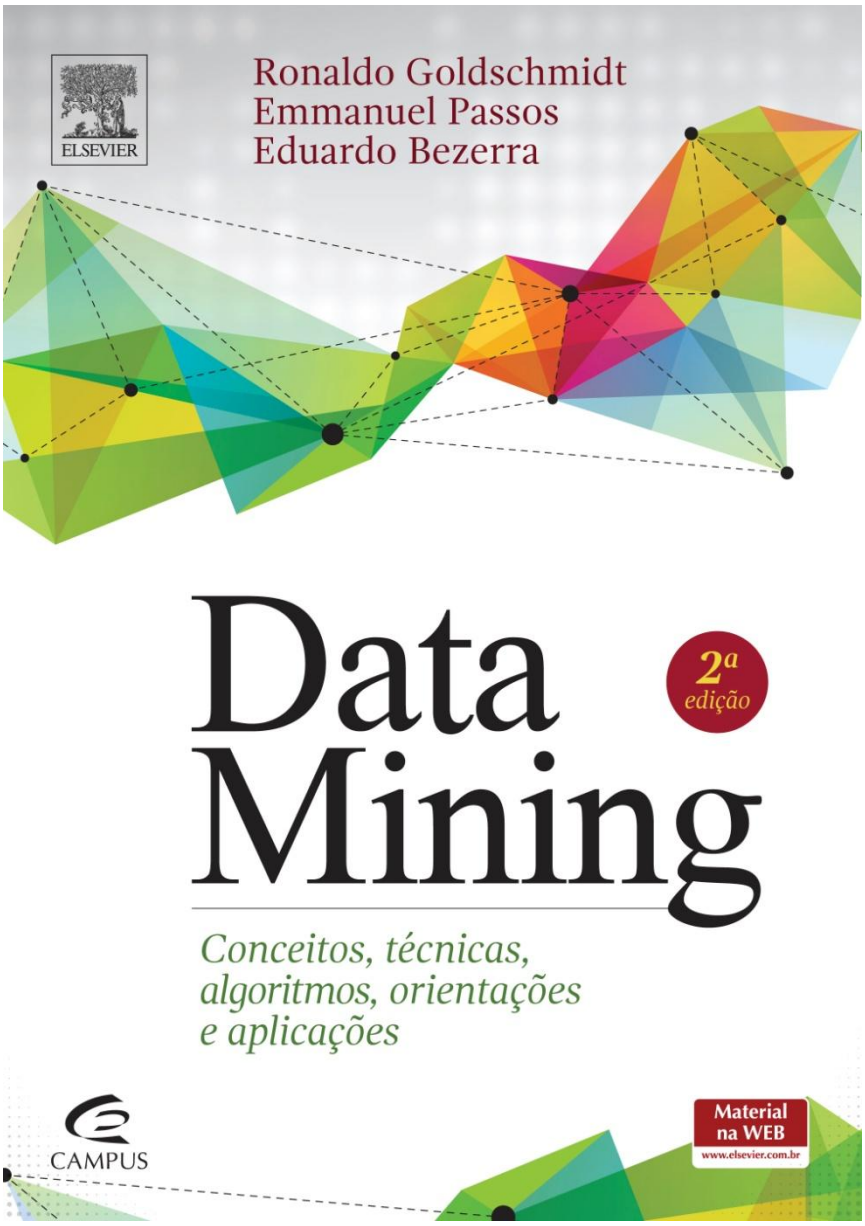


Complemento I - Noções Introdutórias em Data Warehouses

Esse documento é parte integrante do material fornecido pela WEB para a 2ª edição do livro ***Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações***.

Motivação e Conceitos Básicos

Com o advento da globalização, a competitividade entre as empresas no mundo dos negócios vem aumentando intensamente nos últimos anos. As empresas contemporâneas, cientes da necessidade de adaptação a este cenário, têm investido na captação, armazenamento, tratamento e aplicação da informação como diferencial estratégico e competitivo na condução dos negócios. Recursos da área da Tecnologia da Informação têm sido fundamentais neste processo. Em particular, muitos sistemas de informação vêm sendo desenvolvidos e utilizados em diversas aplicações.

A maioria dos sistemas de informação opera sobre bancos de dados chamados transacionais. Estes bancos de dados contêm informações detalhadas que permitem às empresas acompanhar e controlar processos operacionais. Por outro lado, existe uma demanda cada vez maior por sistemas de informação que auxiliem no processo de tomada de decisão. Gerentes e executivos necessitam de recursos computacionais que forneçam subsídios para apoio ao processo decisório, sobretudo nos níveis tático e estratégico das empresas.

Conceitualmente, um Data Warehouse é um conjunto de dados baseado em assuntos, integrado, não-volátil, variável em relação ao tempo, e destinado a auxiliar em decisões de negócios. A orientação a assunto, aliada ao aspecto de integração, permite reunir dados corporativos em um mesmo ambiente de forma a consolidar e apresentar informações sobre um determinado tema. Os dados são não voláteis, pois uma vez carregados no datawarehouse, estes não podem mais sofrer alterações. Cada conjunto de dados, ao ser carregado em um datawarehouse fica vinculado a um rótulo temporal que o identifica dentre os demais. Cada rótulo temporal fica associado, portanto, a uma visão instantânea e sumarizada dos dados operacionais que corresponde ao momento de carga do data warehouse. Desta forma, na medida em que o data warehouse vai sendo carregado com tais visões, pode-se realizar análises de tendências a partir dos dados.

Convém reforçar as diferenças entre data warehouses e bases de dados operacionais. Uma base de dados operacional é um banco de dados clássico que contém informações detalhadas a respeito do negócio em nível transacional.

Nas bases de dados tradicionais, normalmente, os dados encontram-se voltados para a representação de detalhes operacionais corporativos. Em data warehouses, os dados

encontram-se consolidados de forma a prestar informações para os níveis gerencial e estratégico das empresas. Em geral, os data warehouses devem disponibilizar dados sobre a história da empresa de forma a viabilizar consultas, descoberta de tendências e análises estratégicas a partir dos dados. O exemplo abaixo procura ilustrar, de forma simples, o exposto acima.

Considere uma base de dados transacional relacional contendo os detalhes de cada venda realizada por seus vendedores durante um mês. A cardinalidade desta relação representa o número de vendas realizadas durante o mês. Imagine, a título de exemplo, que tenham ocorrido 1000 vendas durante o mês.

Venda(Seqüencial, Data, Hora, Vendedor, Valor)

Um data warehouse sobre Venda, gerado a partir do exemplo acima poderia conter um resumo sobre as vendas realizadas:

Venda_Mensal(Mês, Ano, Vendedor, Total)

Convém perceber que nesta estrutura existiria apenas uma tupla referente a cada vendedor no mês, indicando o total vendido.

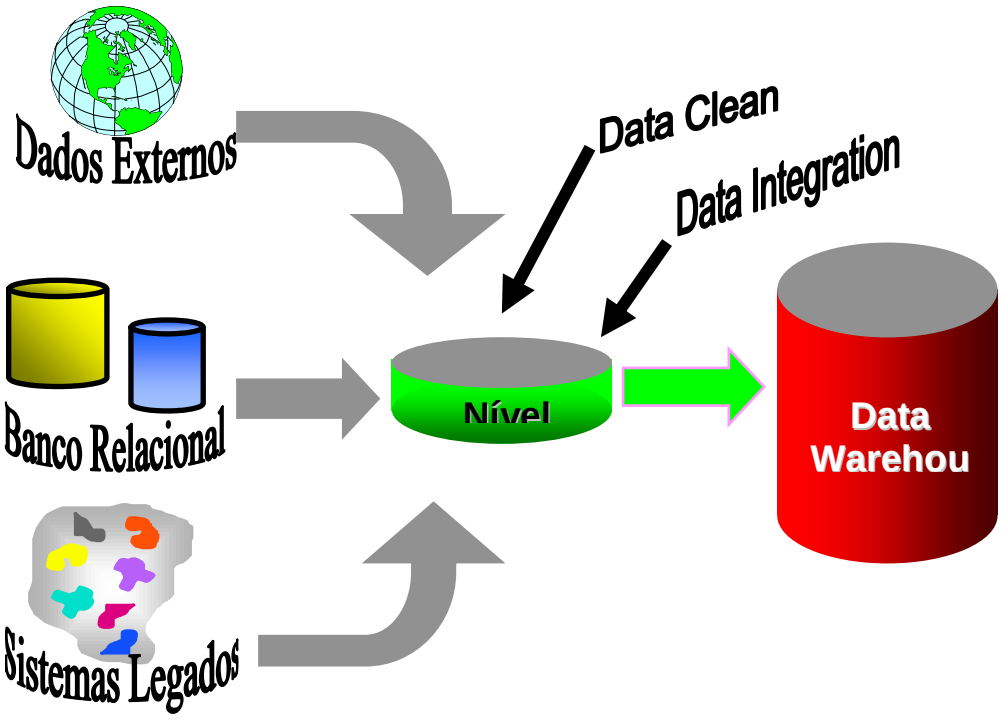
O nível de consolidação de informação varia em função da necessidade de cada aplicação e deve ser definido durante o processo de modelagem e construção do data warehouse. A granularidade de uma informação corresponde ao grau de consolidação de informação envolvido. Por exemplo, a relação *Venda* acima apresenta uma maior granularidade do que a relação *Venda_Mensal*. O detalhamento dos dados na primeira relação é maior do que na segunda.

Sistemas de Informação que utilizam bases de dados transacionais são muitas vezes denominados de aplicações OLTP (*On-Line Transactional Processing*). Por outro lado, Sistemas de Informação que acessam datawarehouses são usualmente chamados de aplicações OLAP (*On-Line Analytical Processing*). Em geral, estes últimos permitem visualização e navegação pelos dados sob diversas perspectivas e níveis de detalhe.

A seguir estão relacionados alguns exemplos de aplicações em datawarehouses:

- Pesquisa de fraudes;

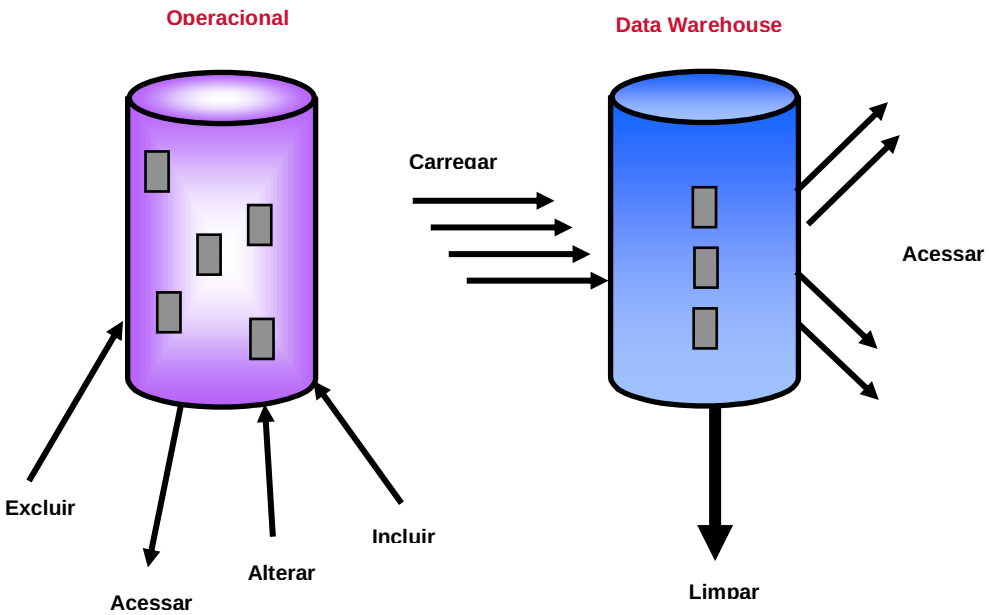
- Análise de crédito;
- Análise de sazonalidade da produção;
- Análise de risco;
- Integração de Informações de Clientes;
- Rentabilidade de Clientes e Produtos;
- Análises de Resultados de Vendas;
- Análises de Ações de Marketing.



Em geral, a passagem de dados de um ambiente operacional legado para um data warehouse não é tão simples quanto uma mera extração e carga de dados. Muitas vezes há a necessidade de transformação e consolidação dos dados. Existem ferramentas que permitem o chamado processo de ETC (Extração, Transformação e Carga) dos dados, muito úteis para auxiliar no processo de criação de data warehouses.O processo de extração dos dados oriundos de diversas fontes de dados operacionais, transformação e carga em datawarehouses normalmente prevê um estágio intermediário de preparação dos dados, sincronização e integração dos dados. Durante este estágio, os dados permanecem em uma área de armazenamento intermediária entre as bases de dados operacionais e o datawarehouse. Ações de limpeza e integração dos dados são realizadas neste momento, conforme ilustra a figura abaixo.

Em bases de dados operacionais, em contra-partida aos datawarehouses, são permitidas atualizações constantes sobre os dados. O conteúdo de um datawarehouse somente sofre alterações no momento de carga de novo conjunto de dados, associado a um novo rótulo temporal.

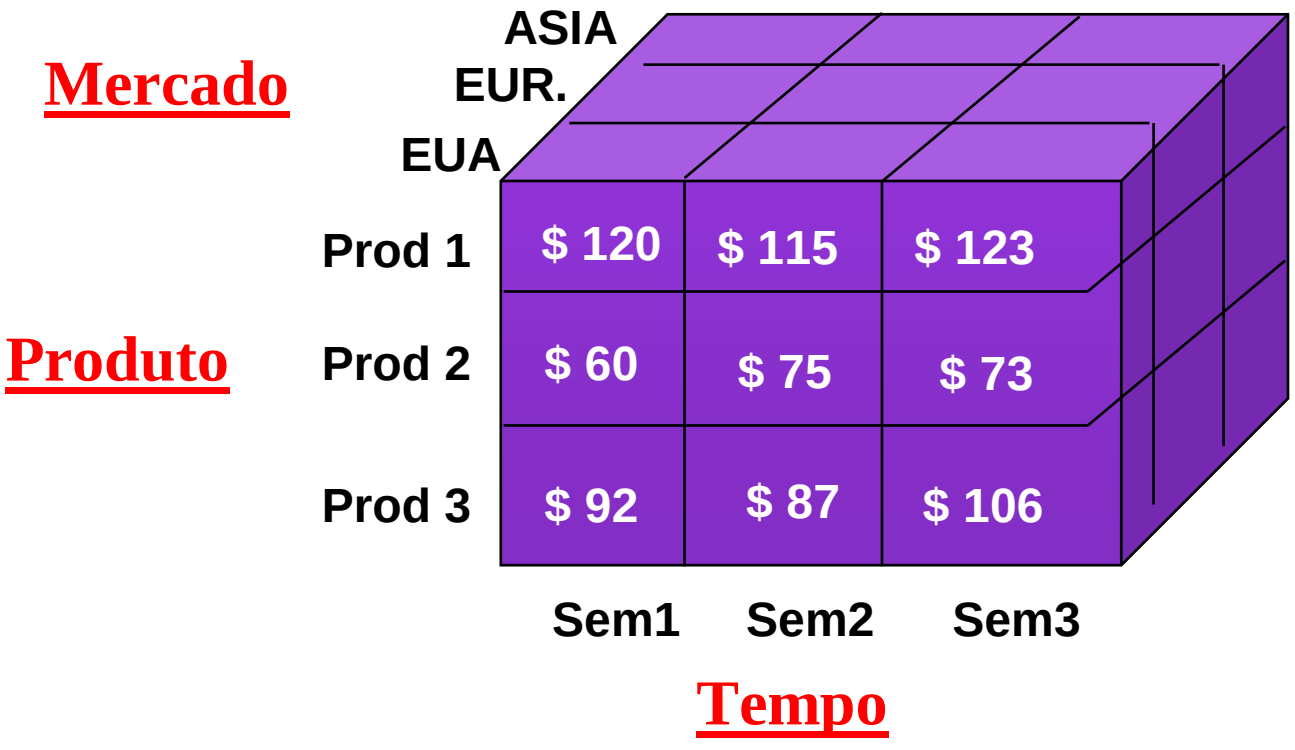
Como a construção e a manutenção de data warehouses envolvem um contínuo pré-processamento dos dados (limpeza, transformações, integração, dentre outras ações), estes repositórios são fontes potenciais de informação para serem submetidas ao processo de descoberta de conhecimento em bases de dados.



Um dos maiores desafios na construção e na manutenção de data warehouses reside na busca pela qualidade dos dados. Diversos são os tipos de problemas existentes nas bases de dados operacionais e que devem ser tratados no projeto de data warehouses. Alguns deles estão citados a seguir:

- Ausência de informação
- Valores inválidos
- Ausência de integridade referencial
- Violações de regras de negócios
- Cálculos inválidos
- Formatos não padronizados
- Duplicação de informação e inconsistência
- Falhas na modelagem das bases de dados operacionais

Um Data Mart é uma porção física ou lógica de um Data Warehouse para atender a uma área específica da empresa. Trata-se de um subconjunto do data warehouse. Muitas vezes



data marts são criados de forma a oferecer simplicidade, menor custo e agilidade ao processo de construção e manutenção de data warehouses.Uma estratégia comum na construção de data warehouses envolve a construção paulatina destes por meio de data marts.O cubo de dados (ou hipercubo de dados) é um recurso que permite o cruzamento e a visualização dos dados em aplicações OLAP. A figura abaixo contém um exemplo de cubo de dados com 3 dimensões (local, produto e período), onde cada célula mostra o total vendido de um determinado produto em uma determinada semana em um determinado local.

Os metadados assumem um papel de grande relevância na manutenção e na expansão de data warehouses. Devem conter informações sobre diversos aspectos, dentre os quais, podem ser destacados:

- A fonte original dos dados e como ter acesso
- Os responsáveis pela informação original
- Como são criados os relatórios e para que servem
- As consultas disponíveis para acesso a determinadas informações
- Como as definições de negócio e terminologia mudaram ao longo dos anos
- Quais premissas de negócio foram assumidas na modelagem do data warehouse

Modelagem Multidimensional

A modelagem multidimensional é uma forma de Modelagem de Dados voltada para concepção e visualização de conjuntos de medidas que descrevem aspectos comuns de um determinado assunto. É utilizada especialmente para sumarizar e reestruturar dados, apresentando-os em visões que suportem a análise dos valores envolvidos.

Enquanto a modelagem de dados tradicional assegura o cumprimento de restrições e evita redundância de informação, a modelagem multidimensional facilita a realização de consultas por usuários não técnicos, acelerando o desempenho destas consultas e admitindo redundância de informação.

O exemplo abaixo ilustra uma modelagem multidimensional onde, a partir das informações de ação, bolsa e mês, expressa-se a lucratividade da ação:

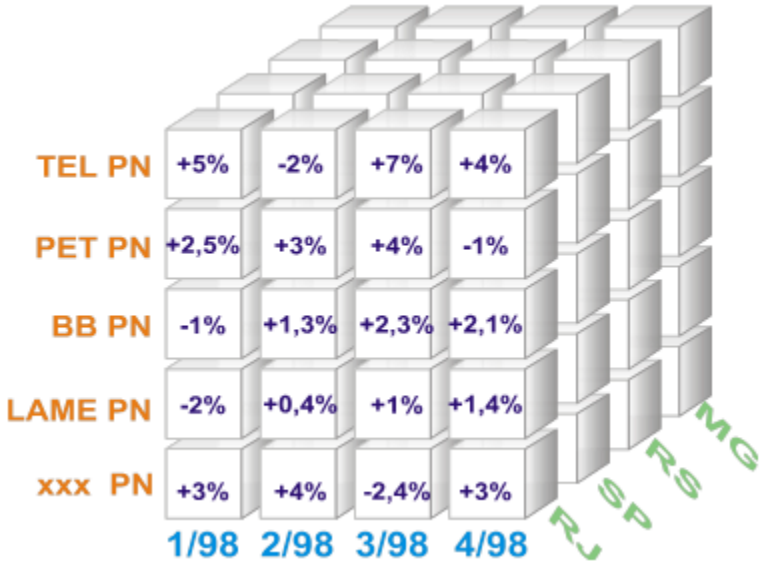
Ação	Bolsa	Mês	Lucratividade
Tel PN	Rio de Janeiro	Janeiro	+5%
Tel PN	Rio de Janeiro	Fevereiro	-2%
Tel PN	Rio de Janeiro	Março	+7%
Tel PN	São Paulo	Janeiro	+4%
Tel PN	São Paulo	Fevereiro	-1%
Tel PN	São Paulo	Março	+4%
Pet PN	São Paulo	Janeiro	+2,5%
BB PN	Rio de Janeiro	Janeiro	-1%

O cubo de dados possui recursos adequados para visualização das informações modeladas em um formato multidimensional. Um modelo multidimensional possui três componentes básicos:

1. Fatos - Um fato é uma coleção de itens de dados, composta de dados de medida e de contexto. Representa um item, ou uma transação ou um evento associado ao tema da modelagem. Exemplo: uma tupla da relação acima.

2. Dimensões – Uma dimensão é um tipo de informação que participa da definição de um fato. No exemplo: ação, local, mês. As dimensões determinam o contexto do assunto. Normalmente são descritivas ou classificatórias. Em geral, as perguntas "O que? Quem? Onde? Quando?" ajudam a identificar as dimensões de um assunto.
3. Medidas – Uma medida é um atributo ou variável numérica que representa um fato. Exemplos: valor da ação, número de evasões escolares, quantidade de produtos vendidos, valor total de venda, etc.

A figura abaixo apresenta o cubo de dados (com apenas 3 dimensões) do exemplo mencionado:

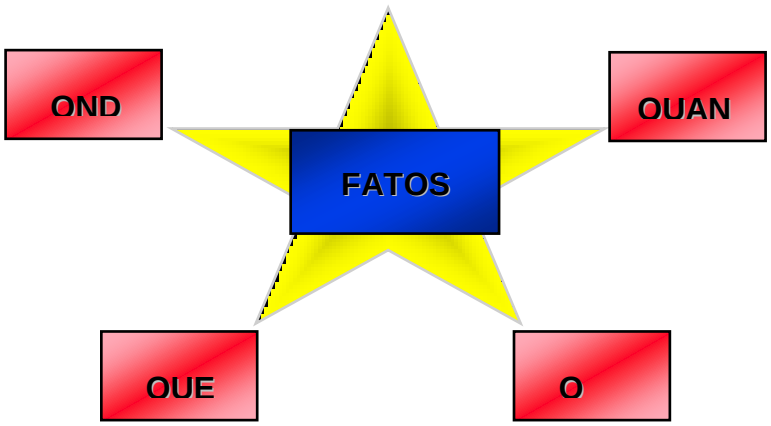


Existem diversos operadores OLAP que permitem acessar os dados em modelos multidimensionais. A seguir encontram-se indicados alguns deles:

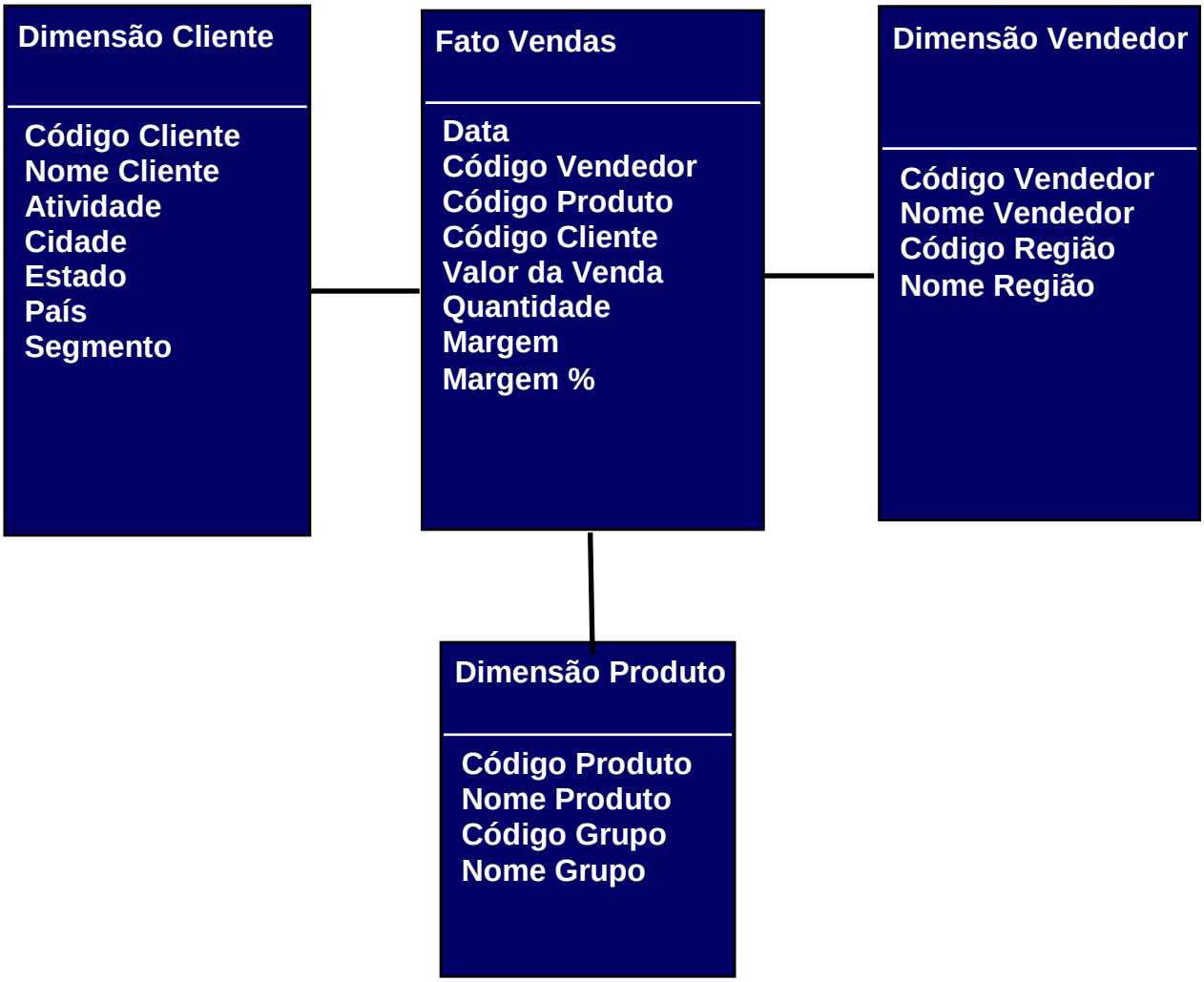
- Drill up/down – Utilizado para aumentar ou reduzir o nível de detalhe da informação acessada. Exemplo: Vendas por país ←→ Vendas por UF.
- Slicing – Utilizado para selecionar as dimensões a serem consideradas na consulta. Exemplo: Visualizar as vendas, separadas por país e por mês.
- Dicing – Utilizado para limitar o conjunto de valores a ser mostrado, fixando-se algumas dimensões. Exemplo: Vendas no estado de Minas, de um determinado produto em um determinado ano.
- Pivoting – Utilizado para inverter as dimensões entre linhas e colunas. Exemplo: Ao visualizar vendas por produto e por estado, aplicar o operador para visualizar as vendas por estado e por produto.

- Data Surfing – Executar uma mesma análise em outro conjunto de dados. Exemplo: Ao visualizar as vendas no Brasil, aplicar o operador para realizar a mesma consulta na Inglaterra.

Existem diversas formas de modelagem física de um data warehouse. Uma das mais populares é a esquema estrela. Neste esquema, uma relação central de fatos é cercada por relações que correspondem às dimensões do problema. As dimensões no esquema estrela são usualmente denominadas pontos cardeais, conforme mostra a figura abaixo:

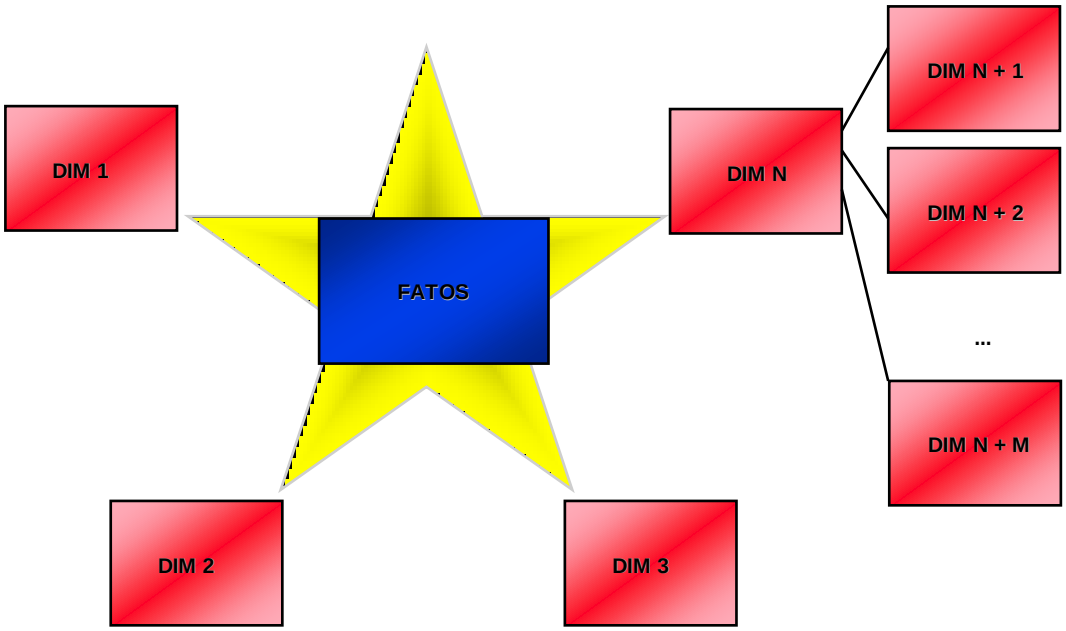


Abaixo encontra-se um exemplo de data warehouse modelado segundo o esquema estrela:



Os fatos estão na tabela Vendas. As dimensões estão representadas pelas tabelas Produto, Cliente e Vendedor. As medidas neste exemplo são: valor da venda, quantidade, margem e margem%.

O esquema estrela pode ser estendido de forma a compor o esquema "Flocos de Neve" (SnowFlake). Neste esquema, as dimensões podem ser associadas a novas dimensões, conforme ilustra a figura abaixo.



Para Saber Mais

Para maiores detalhes sobre data warehouses, sugerimos as referências a seguir.

Machado, Felipe. ***Tecnologia e Projeto de Data Warehouse***, 6ª ed, Rio de Janeiro: Érica, 2013.

Inmon William H.; Strauss, Derek; Neushloss, Genia. ***DW 2.0: The Architecture for the Next Generation of Data Warehousing: The Architecture for the Next Generation of Data Warehousing***, Morgan Kaufmann, 2010.