unioeste
Universidade Estadual do Oeste do Paraná

Datawarehouse e OLAP

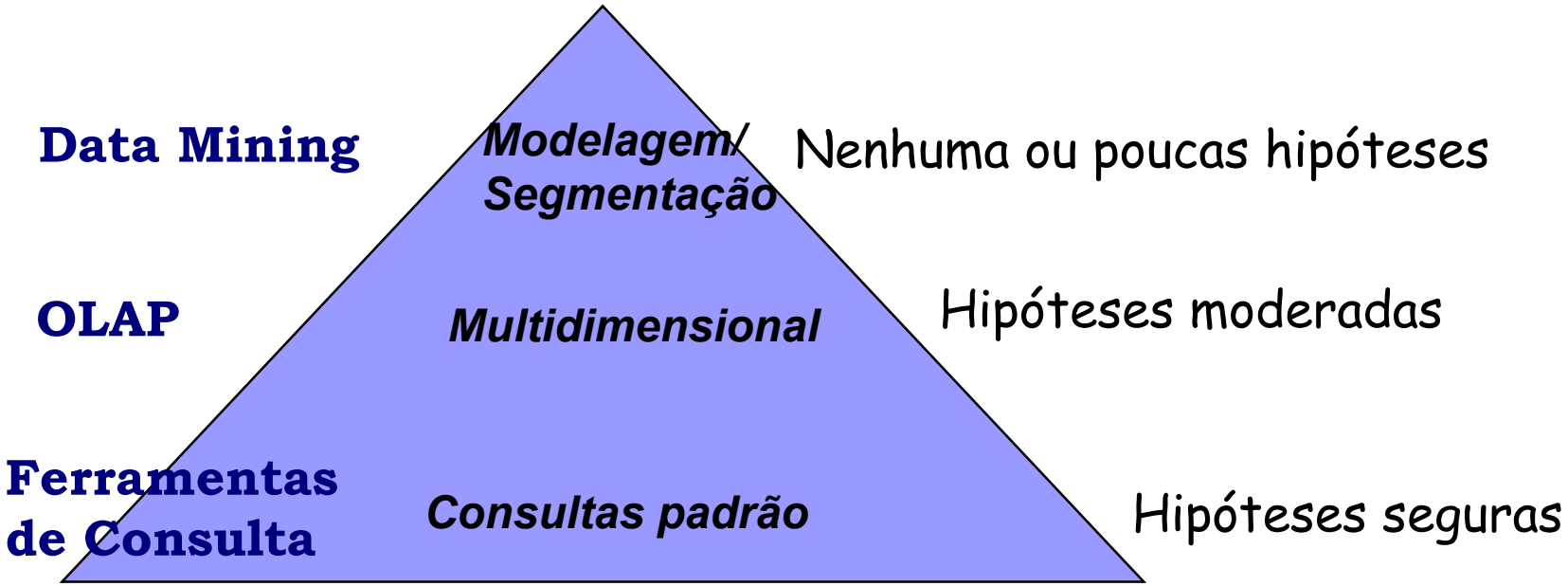
Clodis Boscarioli

Informática

Introdução

- ENTERPRISEWARE: Ferramentas que visam o aumento de produtividade de Grupos Funcionais dentro de uma empresa.

Ambientes analíticos



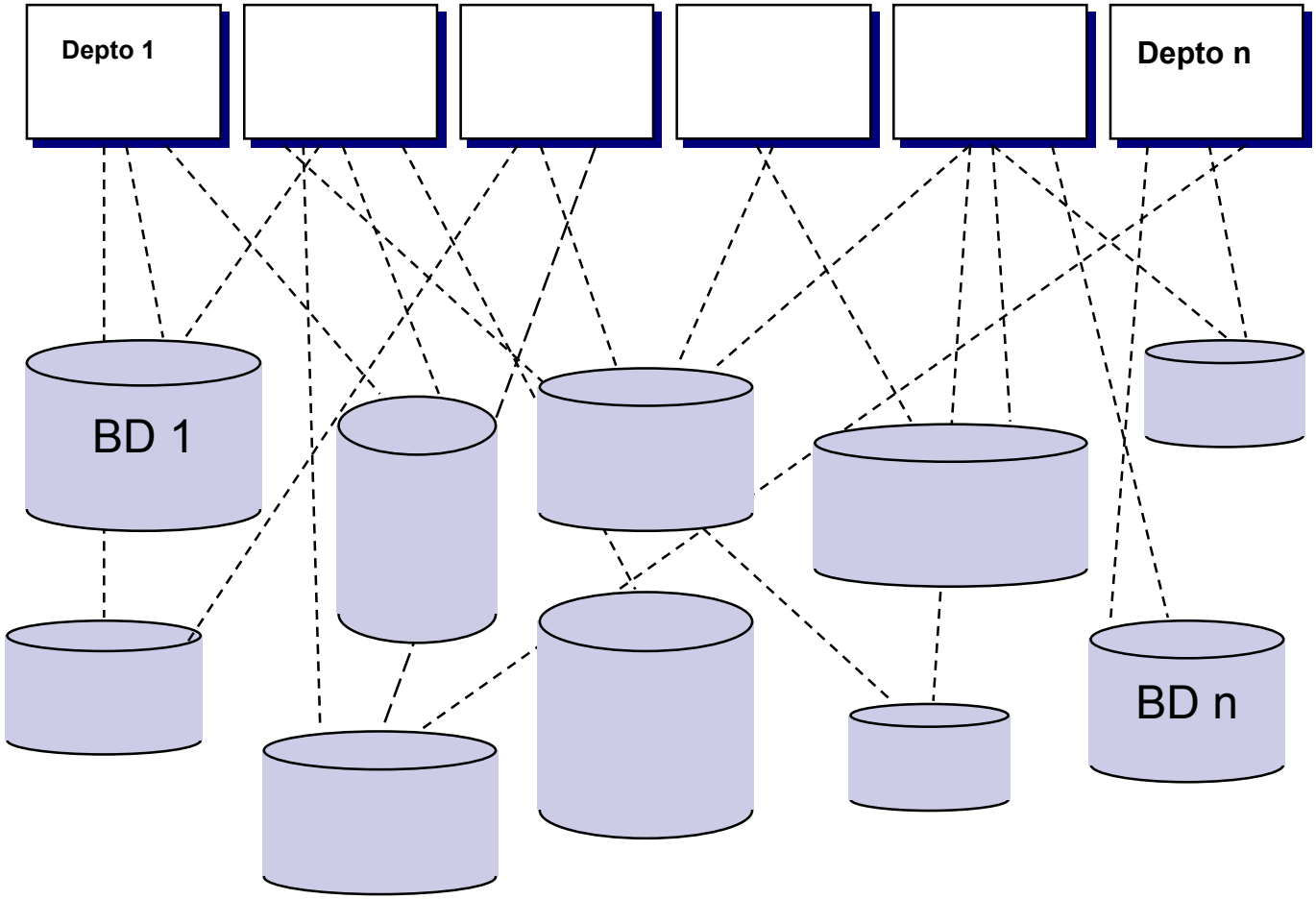
Conceitos Principais

Características	Transacional	Apoio à decisão
Conteúdo dos Dados	**Corrente**	**Histórico**
Natureza dos dados	**Dinâmica**	**Estática**
Atualização	**Contínua**	**Periódica**
Processamento	**Repetitivo**	**Analítico**
Tempo de resposta	**Segundos**	**Minutos**
Objetivo dos dados	**Funcional**	**Negócios**

Análises Idealizadas

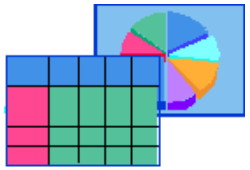
- Qual o desempenho dos nossos representantes em cada região?
- Para cada produto, qual o total de vendas no último ano?
- Como tem variado o índice de participação de cada produto em nossas vendas (*Product Share*) ao longo dos três últimos anos?
- Existe alguma relação entre o desempenho dos representantes e sua faixa de salário?

Realidade dos BD Corporativos



Um “bando” de dados

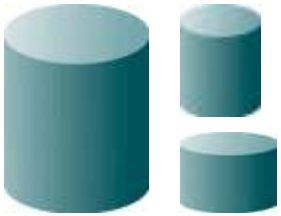
de todos tipos



provenientes de diversas fontes

oriundos de diversos meios

arquivados de diversos modos



DATA WAREHOUSE (DW) - Conceito

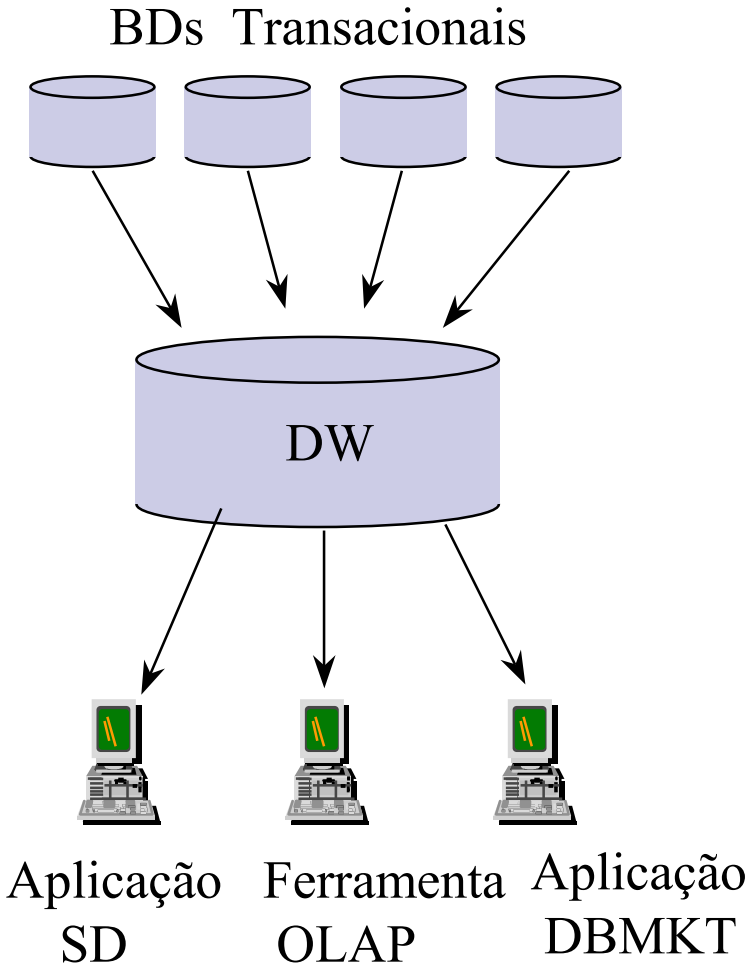
Armazém de Dados

É um amplo e flexível repositório de dados, que aglutina dados de fontes heterogêneas, projetado de modo a suportar o processo de tomada de decisão.

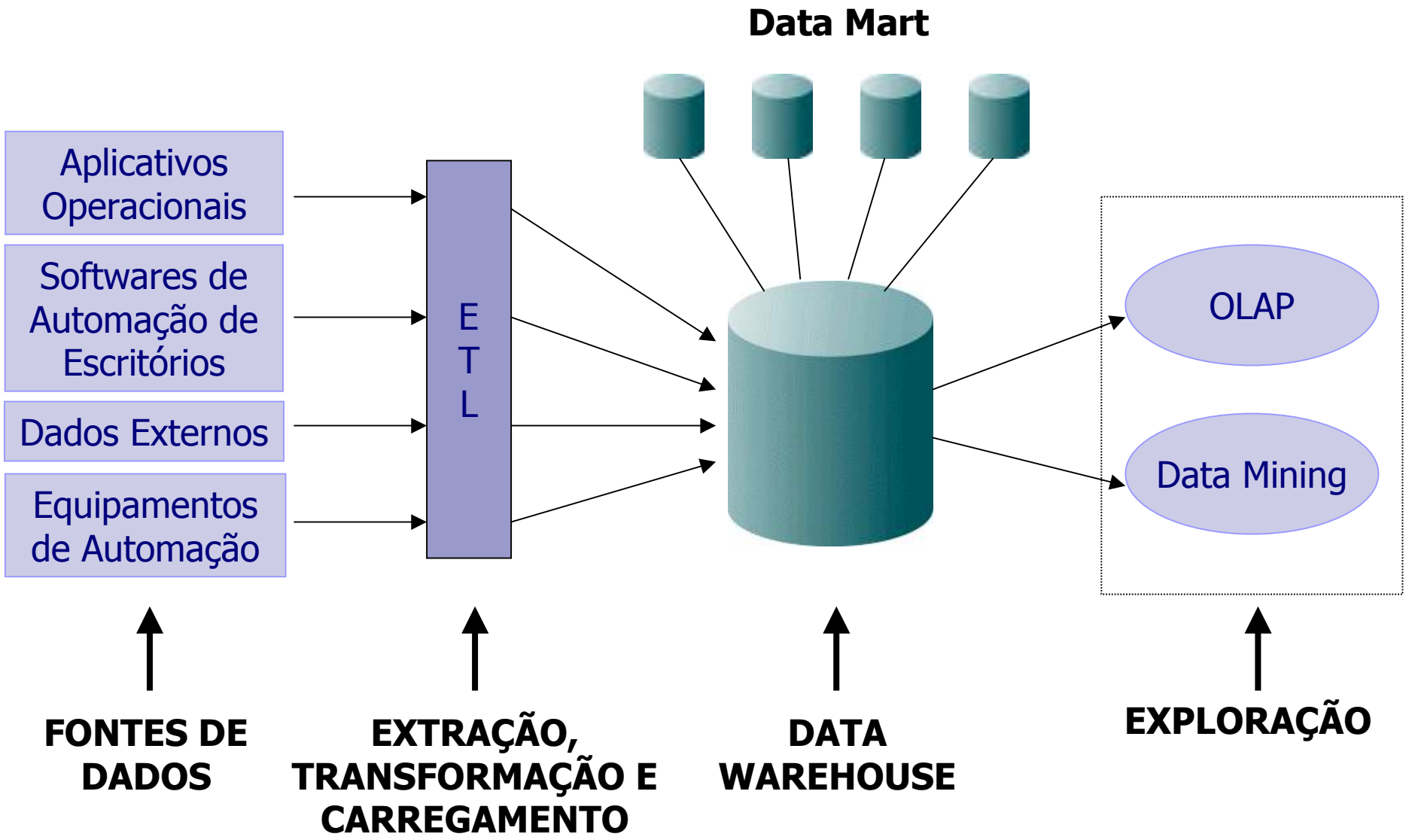
- **Ambiente separado**
- **Disponibilidade**
- **Integrado**
- **Retrato no tempo**
- **Orientado por assunto**
- **Fácil acesso**

Porque um Data Warehouse?

- Integrar dados de múltiplas fontes
- Facilitar o processo de análise sem impacto para o ambiente de dados operacionais
- Obter informação de qualidade
- Atender diferentes tipos de usuários finais
- Flexibilidade e agilidade para atender novas análises



Ferramentas e Técnicas de BI



Elementos de um Data Warehouse

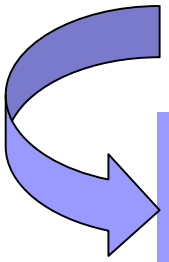
- Banco de Dados
- Ferramentas para Transformação e Integração de Dados
- Metadados
- Ferramentas de Acesso
- Data Marts
- Administração e Gerenciamento do Sistema de Data Warehouse

Componentes Potenciais de um DW

1. Repositório de Metadados
2. Ferramentas de Projeto CASE
3. Ferramentas de Extração, Transformação e Carga (ETL)
4. Ferramentas para Qualidade e Limpeza
5. Ferramentas para Replicação
6. Provedores de Interfaces de BD ODBC/OLE
7. Ferramentas de *Gateway* para BD Legados
8. Bancos de Dados Relacionais
9. (Bancos de Dados Não-Relacionais Legados)
10. Bancos de Dados Multidimensionais

Componentes Potenciais de um DW

11. Ferramentas OLAP
12. Ferramentas de Relatório e Consulta
13. Ferramentas de Data Mining
14. *Cross-Platform Batch Schedulers*
15. Ferramentas de Monitoramento e Controle
16. Pacotes de Aplicação para Data Warehouse



Todos estes componentes manipulam/geram metadados!

Projeto de Sistemas de DW

- Princípio:
  - Os dados que se deseja analisar estão disponíveis nos bancos operacionais da empresa.
  - Os bancos operacionais não são adequados para efetuar as operações analíticas.
- Estratégia:
  - Criar um novo sistema de banco de dados para armazenar as operações analíticas.
  - O sistema analítico é atualizado por rotinas automáticas executadas *off-line*, a partir de dados extraídos dos BDs operacionais.
  - As rotinas de transporte dos dados operacionais para o banco analítico realizam todas as consistências necessárias relativas à eliminação de dados desnecessários e ajuste da granularidade de tempo adotada para o banco analítico.
  - Os usuários podem realizar apenas operações de leitura sobre o banco analítico.

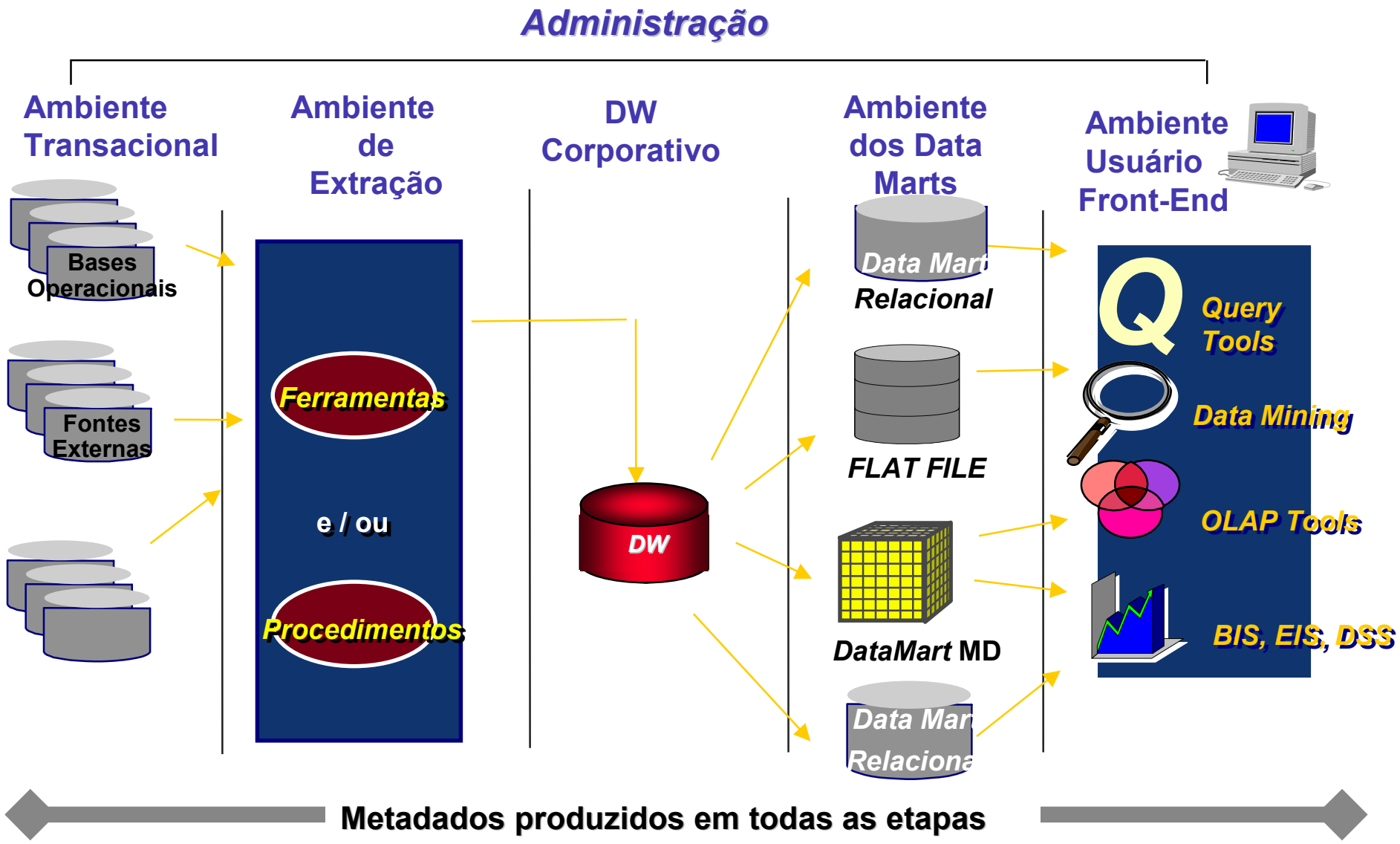
Administração e Gerenciamento do DW

- Características dos Sistemas de DW:
  - Tendem a ser 4 vezes maiores que os sistemas de banco de dados operacionais.
  - Não são sincronizados em tempo real com os dados operacionais.
- Funções ligadas ao gerenciamento do Sistema:
  - Gerenciamento de segurança e prioridades
  - Monitoramento das atualizações oriundas de fontes múltiplas
  - Verificação da qualidade dos dados
  - Gerenciamento e atualização dos metadados
  - Auditoria relativa ao uso do sistema de DW
  - Eliminação de dados obsoletos ou desnecessários
  - Replicação e distribuição de dados
  - *Backup* e recuperação

Os 9 Pontos de Decisão (Kimball)

1. Os processos e, portanto, a identidade das tabelas de fatos;
2. A granularidade (nível de detalhe) de cada tabela de fatos;
3. As dimensões de cada tabela de fatos;
4. Os fatos, incluindo fatos pré-calculados;
5. Os atributos da dimensão com descrições completas e terminologia apropriada;
6. Como rastrear dimensões de modificação lenta;
7. Os agregados, dimensões heterogêneas, minidimensões, modos de consulta e outras decisões de armazenamento físico;
8. A amplitude de tempo do histórico do banco de dados;
9. Os intervalos em que os dados são extraídos e carregos no DW.

Ambiente de Data Warehouse (Proposta Original)

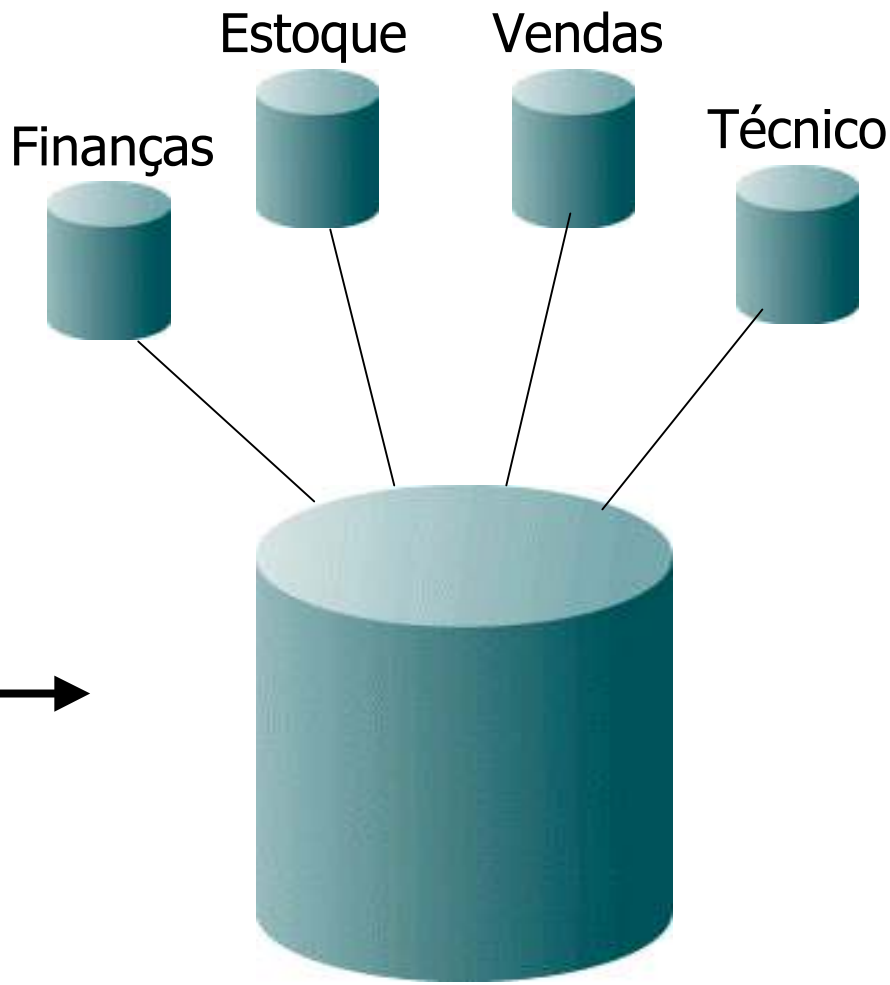


Data Marts

- Conjunto de dados não normalizados, sumarizados, relativos a uma área específica para análise de negócios.
- Podem ser independentes ou derivados de uma visão única concebida a partir do sistema de DW.

DW - Organização

DATA MART
Data warehouse departamental



DATA WAREHOUSE
Corporativo

Infra-estrutura Básica

Metadados

- Metadados são dados sobre os dados e são classificados em dois tipos:
  - Metadados Técnicos (*Operational Metadata*): Descreve como os sistemas operacionais são mapeados no sistema de datawarehouse.
    - Inclui informações sobre as fontes de dados, descrição das transformações, informações sobre as tabelas de destino, regras para extração dos dados, restrições de acesso, etc.
  - Metadados de Negócio: Descreve como o sistema de DW é mapeado com o modelo de dados de negócio dimensional do usuário, usado pelo seu sistema de apoio a decisão (DSS - *Decision Support System*).
    - Inclui informações sobre áreas de negócio, tipos de consulta, relatórios, etc.

Ferramentas para Transformação e Integração de Dados

- Compõe uma parte significativa do esforço (e do custo) na implantação de um DW.
- As principais dificuldades encontradas são:
  - Heterogeneidade dos bancos operacionais.
  - Heterogeneidade dos esquemas de dados (nomes e tipos diferentes para mesmos atributos).
- A extração e adequação dos dados oriundos dos bancos operacionais pode ser feita de duas formas:
  - através de rotinas escritas pelos programadores da empresa
  - através de ferramentas que automatizam a transferência dos dados.
- As principais funções a serem realizadas são:
  - Remover os dados indesejáveis dos bancos de dados analíticos.
  - Efetuar as conversões de nomes e tipos de dados.
  - Calcular sumários e dados derivados.
  - Estabelecer valores *default* para dados inexistentes.

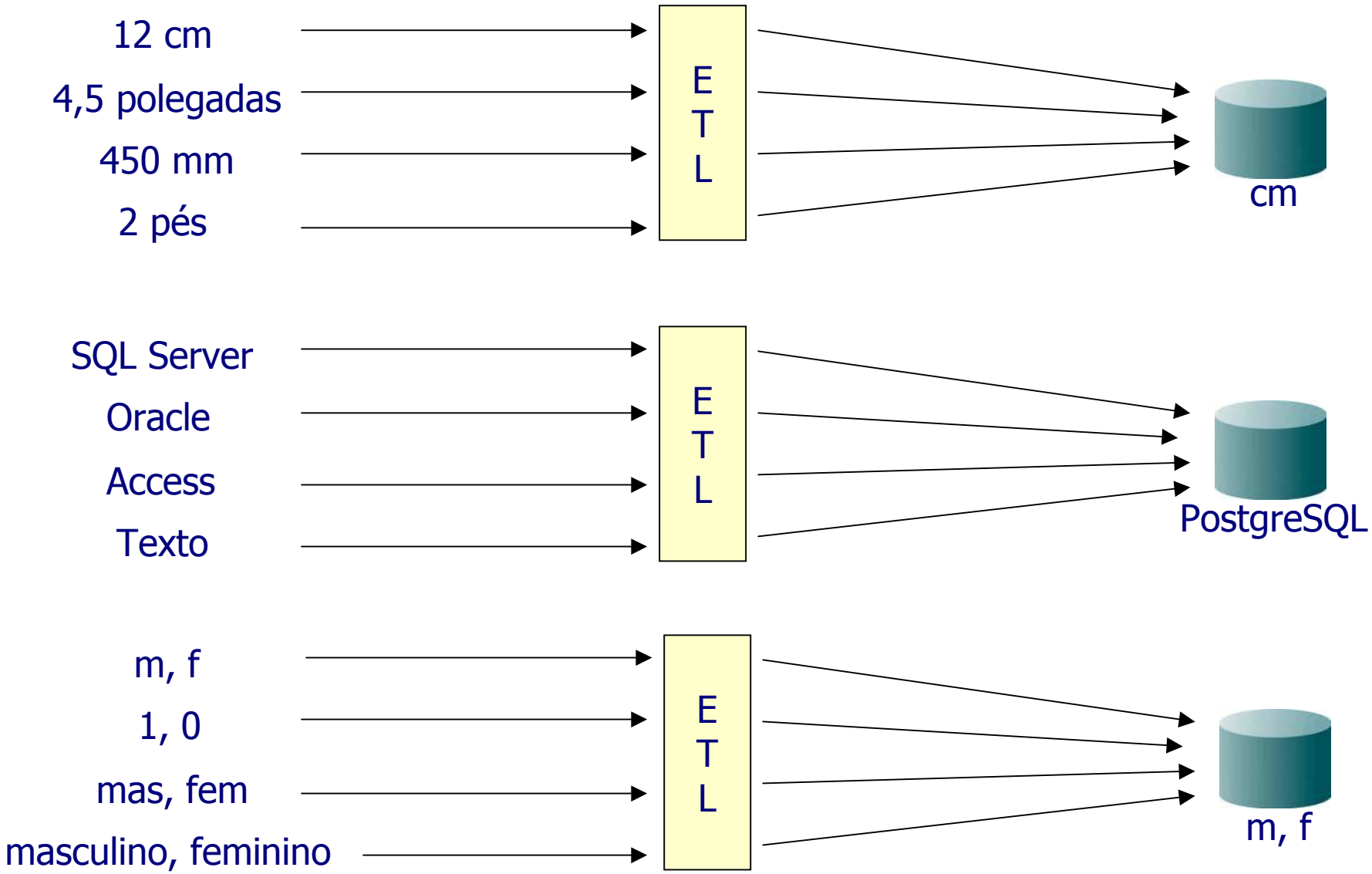
ETL – Extração

- Extração Seletiva: os dados são extraídos por meio de programas desenvolvidos especificamente para selecionar os dados a serem exportados;

- Manutenção por *logs* ou lotes: os dados são extraídos através dos registros automáticos (logs) ou de lotes de dados das transações efetuadas nos sistemas transacionais;

- Replicação Automática: os dados são extraídos através de um sincronismo automático entre dois bancos de dados;

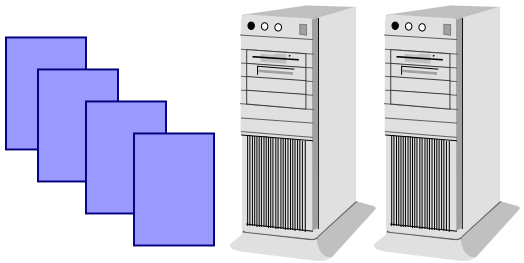
ETL – Transformação



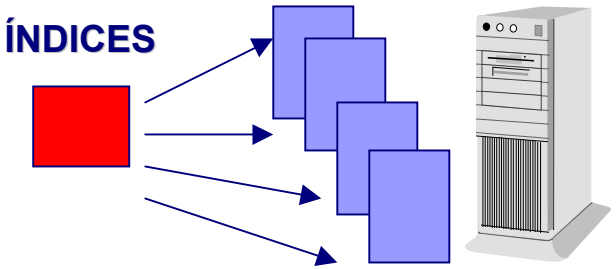
Banco de Dados

- As principais opções para o sistema de banco de dados do sistema de DW são:

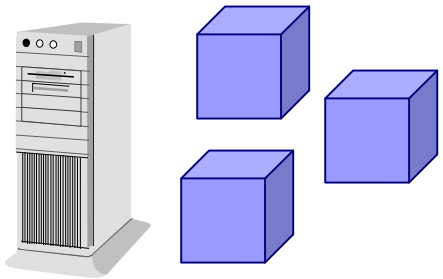
RELACIONAL COM HARDWARE ESPECIAL



RELACIONAL COM INDICES ESPECIAIS



MULTIDIMENSIONAL

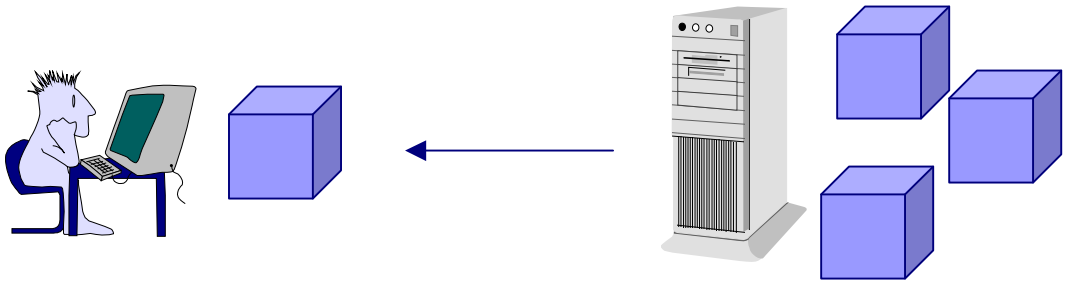


Projeto de DW em RDB

- Os dados de aplicações de DW são armazenados segundo o modelo em estrela:
  - Uma tabela de fatos com as métricas a serem avaliadas e as chamadas para as tabelas de dimensões.
  - Uma tabela para cada dimensão, contendo os níveis associados a cada dimensão.
- Por razões de desempenho, o modelo em estrela pode ser alterado segundo três estratégias principais:
  - Sumarização: Criação de tabelas de fatos redundantes, com dados já sumarizados (também chamadas de agregações).
  - Denormalização: Substituição dos relacionamentos da tabela de fatos pelos atributos da tabela de dimensões.
  - Particionamento: Fragmentação da tabela de fatos em tabelas menores (por exemplo, tabela de vendas do ano de 1999).

Banco de Dados Multidimensionais

- **MDD (Banco de Dados Multidimensionais)**
  - Armazenam informações em *arrays* de formato proprietário (os cubos), que correspondem às dimensões de negócio definidas pelos usuários.
  - Não são compatíveis diretamente com SQL. Eles são acessados por API's proprietárias desenvolvidas pelos fabricantes.
  - As consultas aos cubos são pré-processadas, aumentando muito o volume dos dados armazenados (em torno de 25 vezes).
  - Não permitem realizar relacionamentos entre os dados (*joins*).
  - Não suportam *update* incremental (os cubos precisam ser reconstruídos).



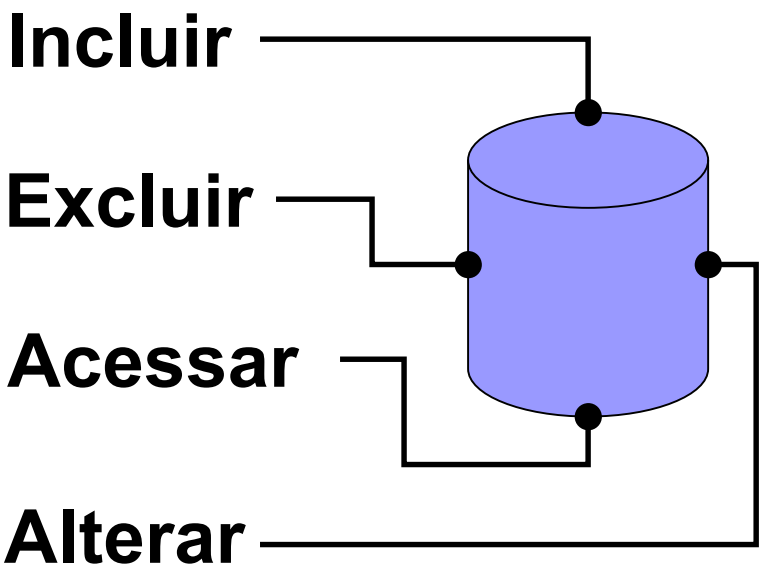
DADOS ARMAZENADOS EM CUBOS

Alternativas para Multidimensionalidade

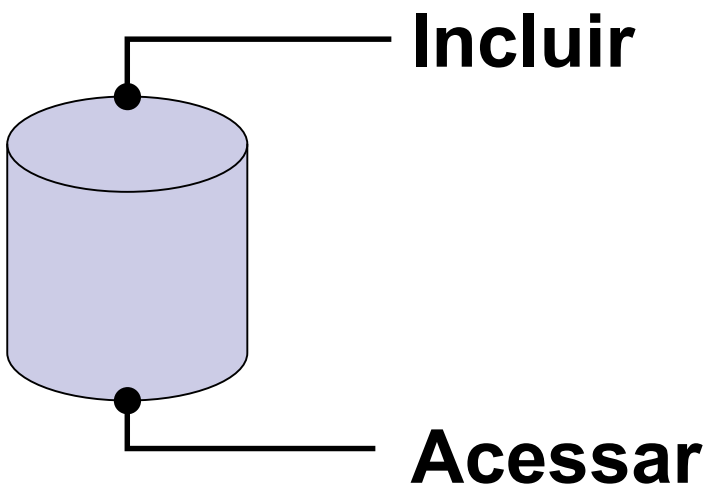
- MOLAP
  - MD Real
  - Armazena os dados em formato multidimensional
  - Não usa SQL como linguagem de acesso aos dados
- ROLAP
  - MD Virtual
  - Armazena os dados em formato relacional
  - Comandos SQL são gerados para acesso aos dados
- HOLAP
  - Híbrida
  - Mais usual atualmente

Conceitos Principais

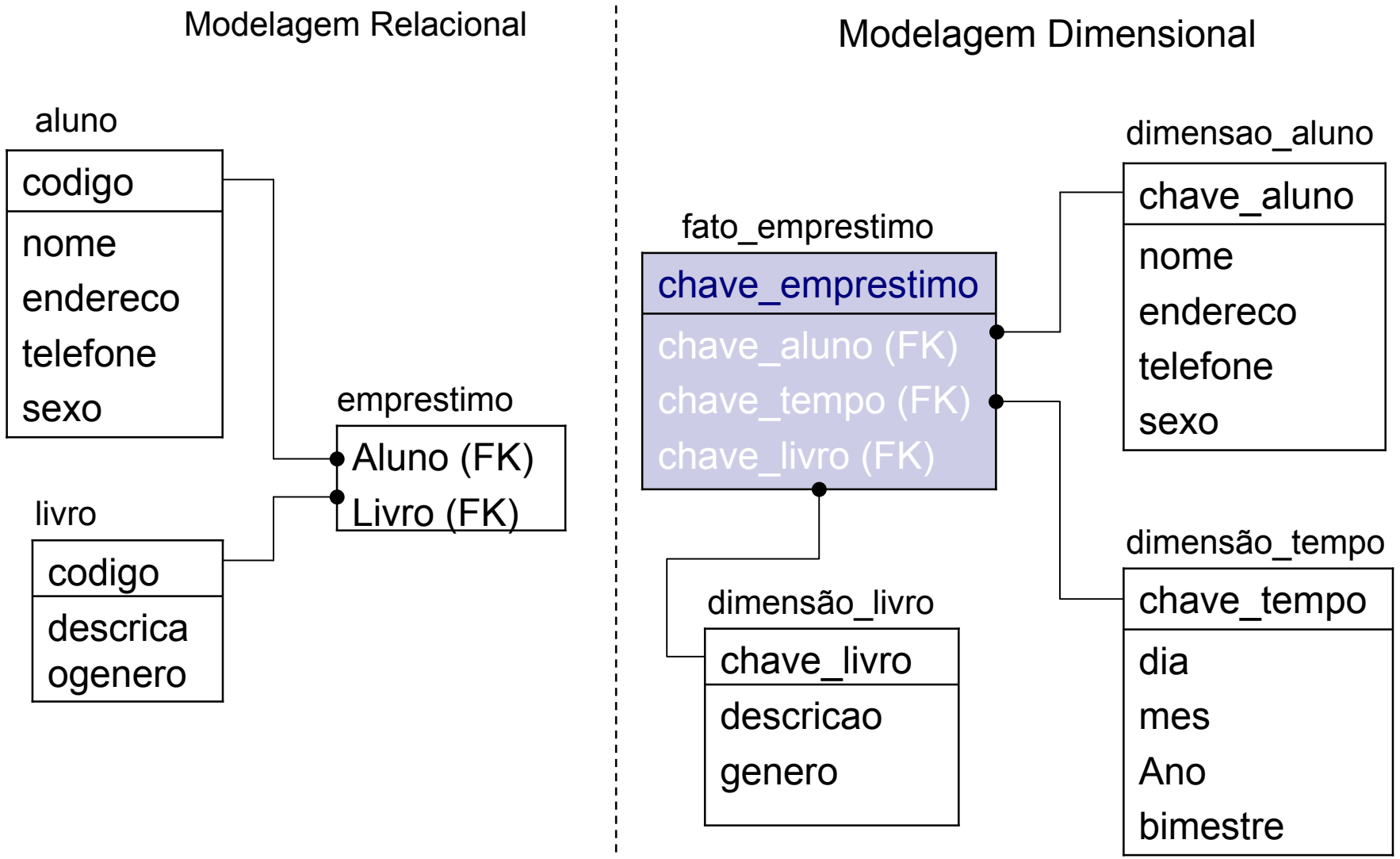
Banco de dados Transacional



Data Warehouse

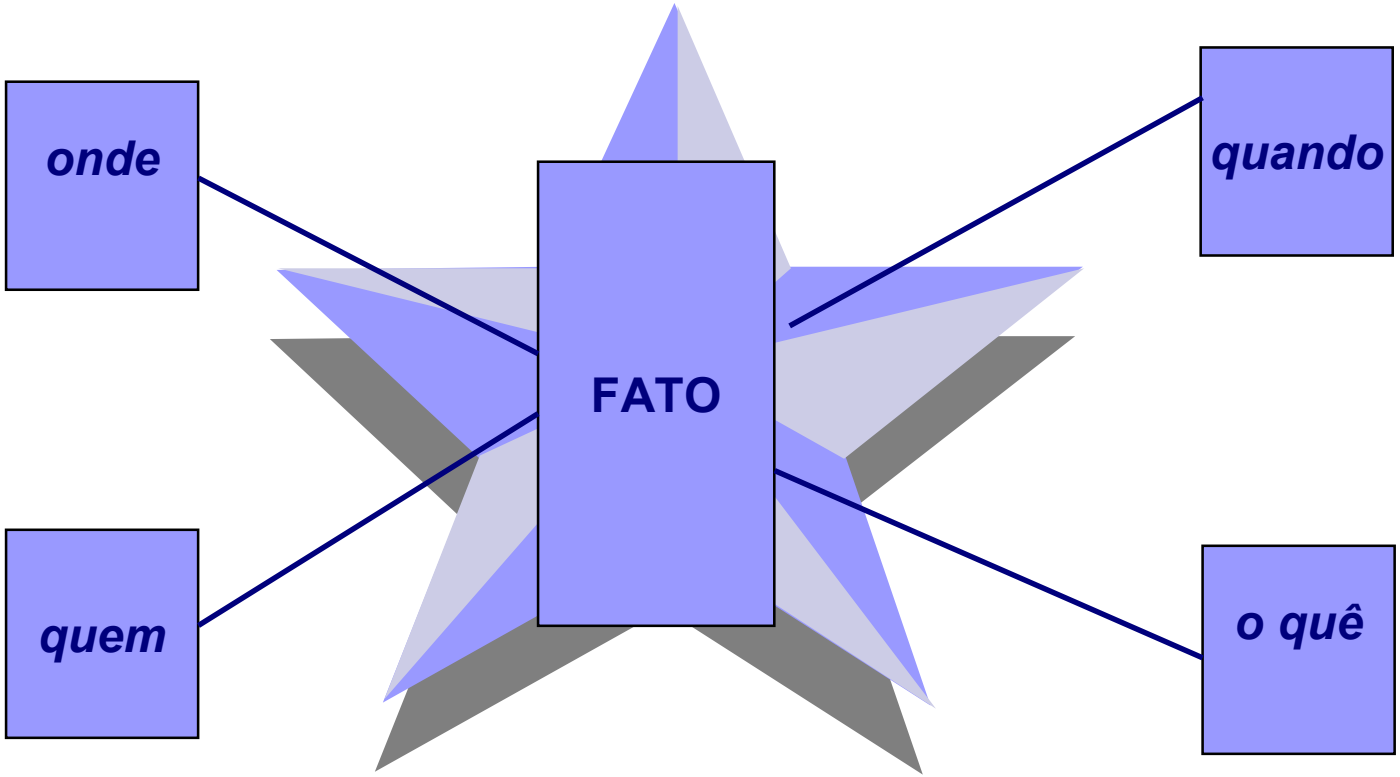


Diferenças na Modelagem

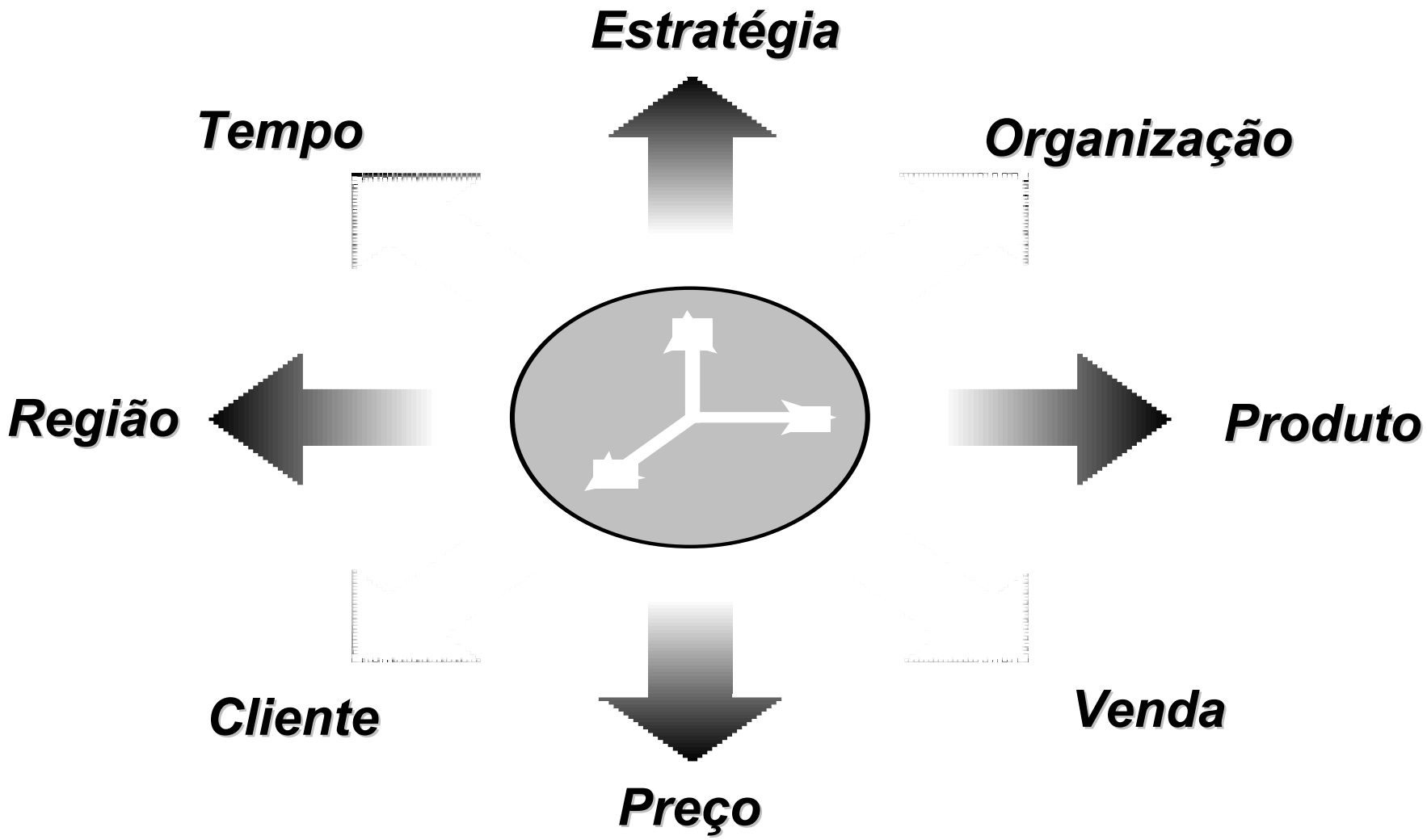


Modelo Dimensional → Esquema Estrela

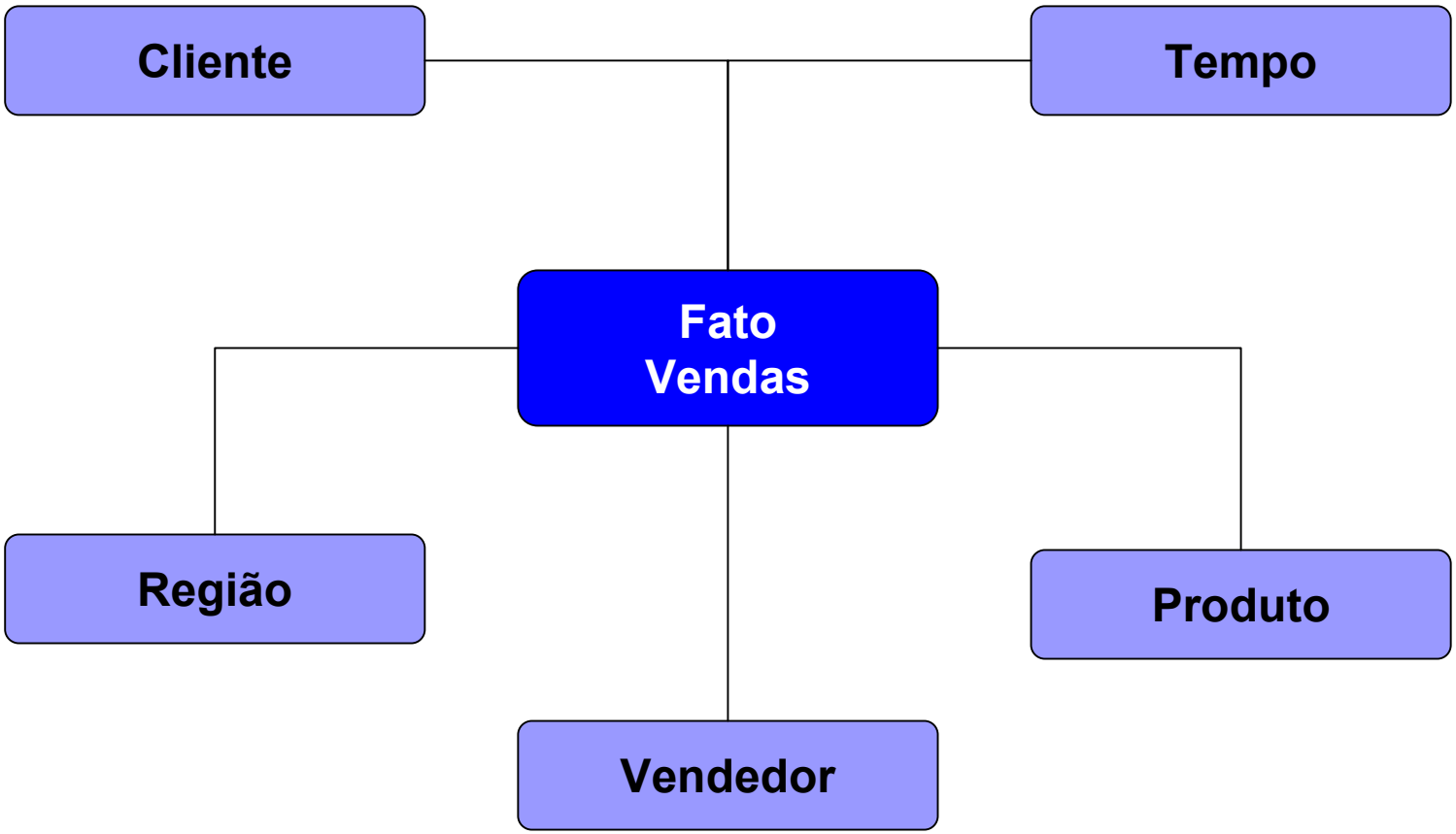
- Uma tabela de fatos cercada de tabelas de dimensões



DW - Dimensões



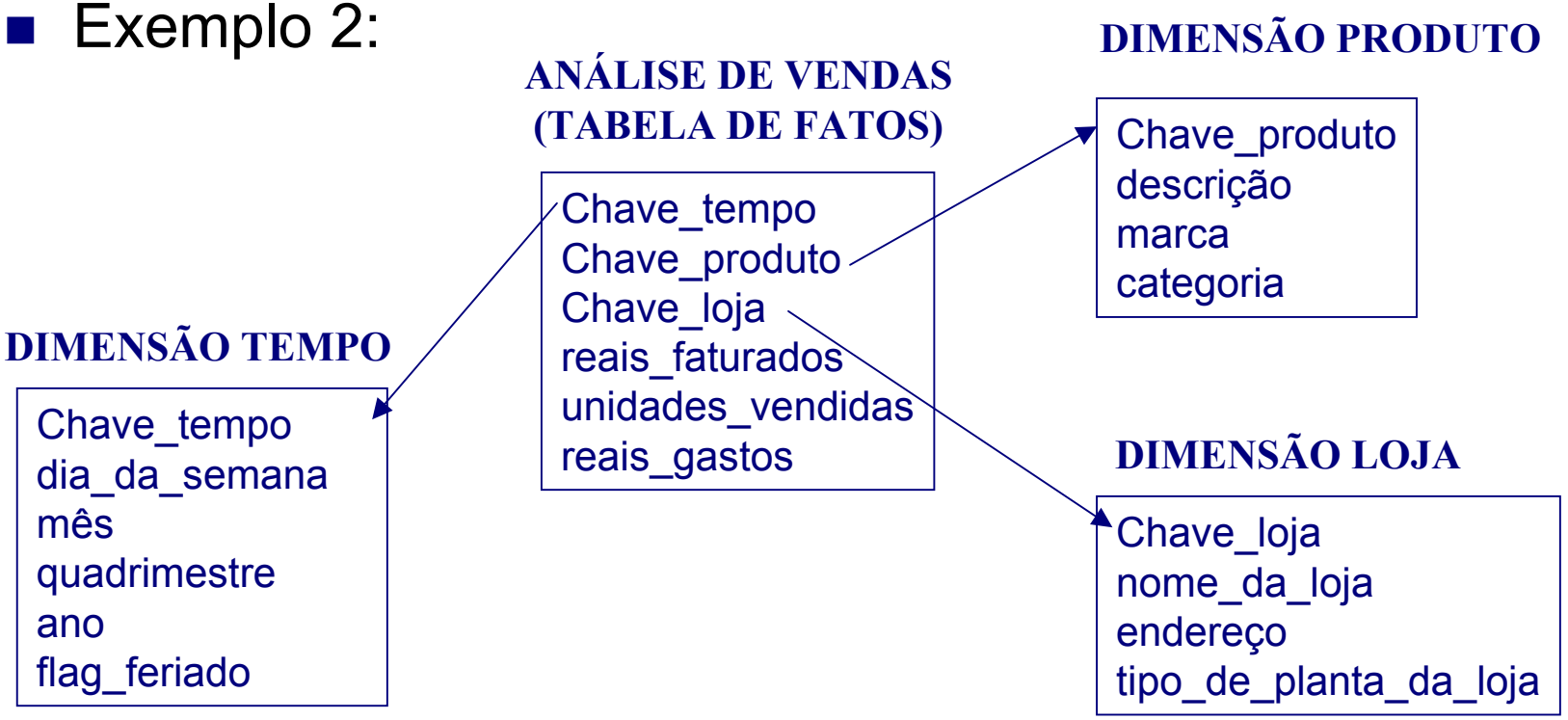
Modelo Dimensional → Esquema Estrela



Modelo Dimensional → Esquema Estrela

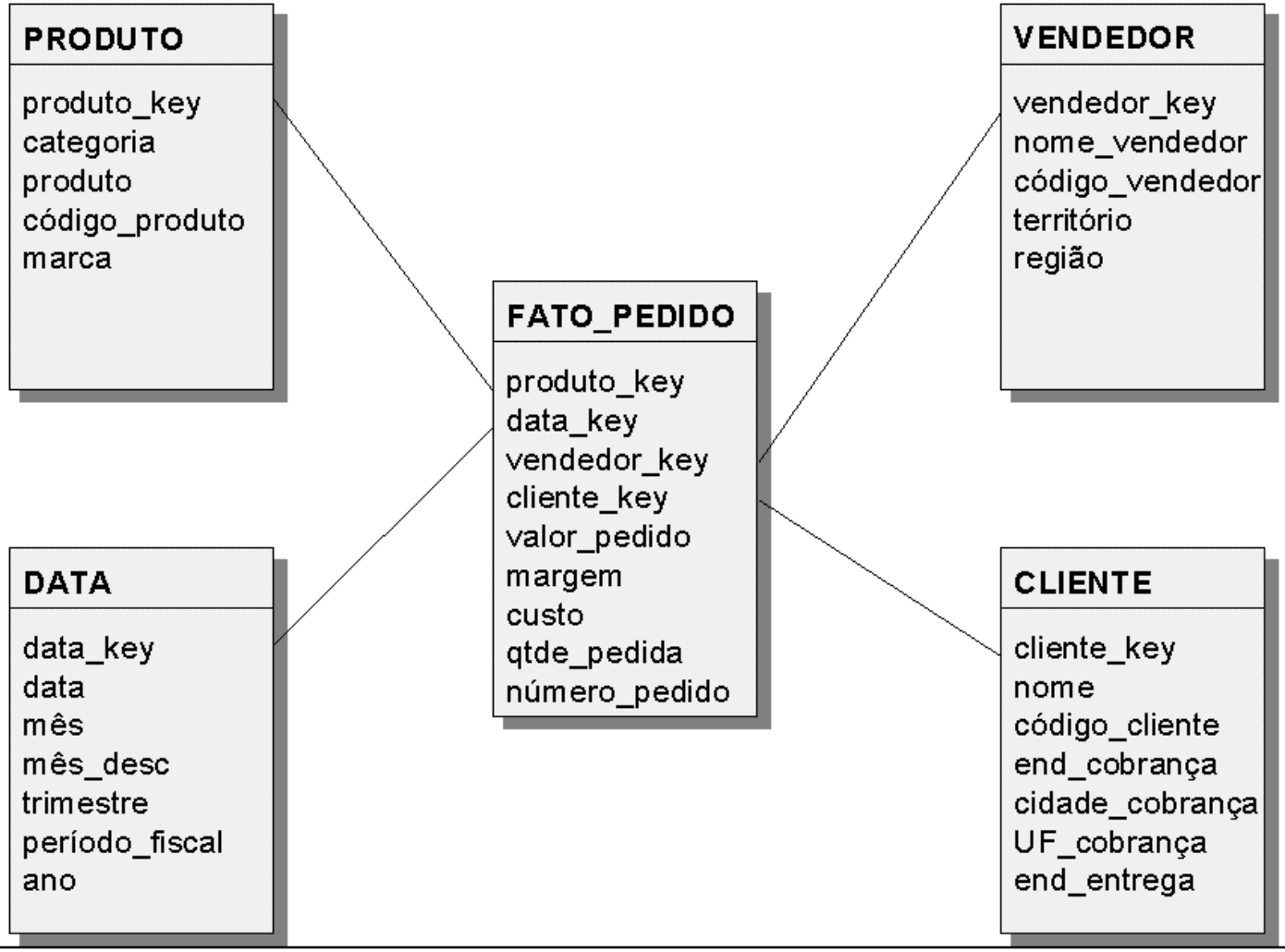
- O projeto de um banco de dados dimensional é do tipo *top-down,* isto é, ele é projetado a partir do tipo de análise que se quer efetuar.

- Exemplo 2:



Modelo Dimensional → Esquema Estrela

- Exemplo 3:



Junção lógica entre tabelas Fato e Dimensão

Tabela Fato

TEMPO KEY	PRODUTO KEY	MERCADO KEY	Valor	Quantidade
980715	101	4030	348,00	140
980716	101	4030	287,00	114
980716	101	4010	443,00	170
980717	101	2010	580,00	232
980718	101	2010	686,00	274

Dimensão Mercado

MERCADO KEY	Mercado Descrição	Região	Estado	Cidade	Loja	Nível
1010	Região Sul	1				4
1020	Região Sudeste	2				4
2010	SP	2	10			3
2020	RJ	2	20			3
3010	São Paulo	2	10	101		2
3020	Campinas	2	10	102		2
4010	Loja ABC	2	10	101	1001	1
4020	Loja DEF	2	10	102	1002	1
4030	Loja GHI	2	10	102	1003	1

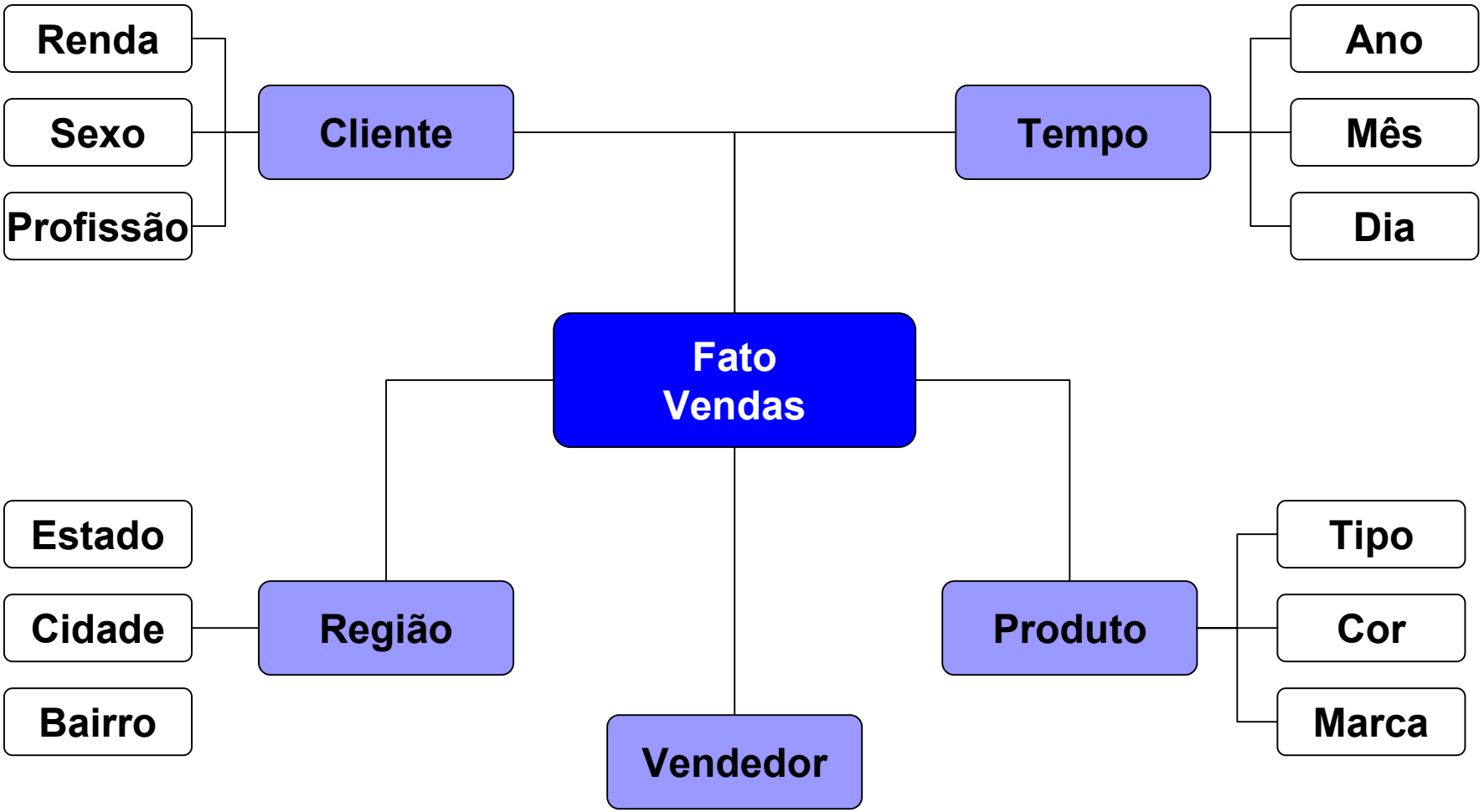
Atributos típicos da Dimensão Tempo

Tempo key	Descrição	Resolução ou Nível	Sequência na Resolução	Indicador Corrente	Ano	Semestre	Mês	Dia
...	...	...	...	...	...	...	...	...
101	Janeiro 1999	Mês	501	S	1999	011999	011999	
102	Fevereiro 1999	Mês	502	S	1999	011999	021999	
103	Março 1999	Mês	503	S	1999	011999	031999	
104	Abril 1999	Mês	504	S	1999	011999	041999	
105	Maio 1999	Mês	505	S	1999	011999	051999	
106	Junho 1999	Mês	506	S	1999	011999	061999	
113	Janeiro 2000	Mês	513	S	2000	012000	012000	
114	Fevereiro 2000	Mês	514	S	2000	012000	022000	
115	Março 2000	Mês	515	S	2000	012000	032000	
116	Abril 2000	Mês	516	S	2000	012000	042000	
21	1 Sem. 1999	Semestre	101	S	1999	011999		
22	2 Sem. 1999	Semestre	102	S	1999	021999		
23	1 Sem. 2000	Semestre	103	S	2000	012000		
24	2 Sem. 2000	Semestre	104	S	2000	022000		
3	1999	Ano	3	S	1999			
4	2000	Ano	4	S	2000			
5	2001	Ano	5	N	2001			
6	2002	Ano	6	N	2002			
...	...	...	...	...	...	...	...	...

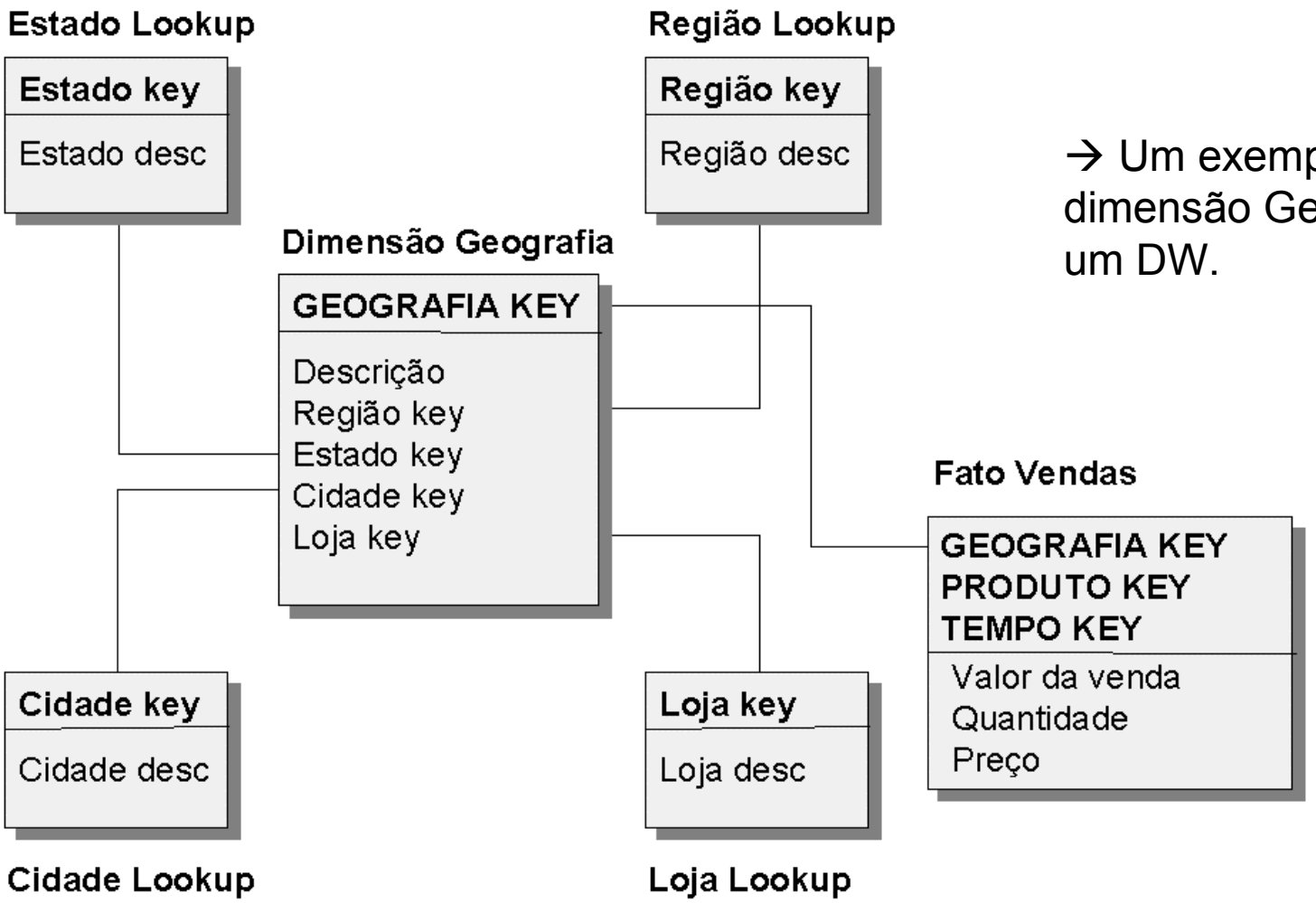
Modelo Dimensional → Esquema *Snowflake*

- O esquema *Snowflake* pode ser considerado um *Star* normalizado, pois emprega uma combinação de normalização da base de dados, para manter a integridade e reduzir os dados armazenados de forma redundante, com uma desnormalização para obter melhor desempenho.
- Neste esquema as dimensões são normalizadas em subdimensões, e cada nível da hierarquia fica em uma subdimensão. Por esta razão, não há necessidade de utilizar o indicador de nível que existe nos esquemas do tipo *Star*.
- A tabela principal da dimensão tem uma chave para cada nível hierárquico representado na subdimensão e não mais uma única chave, como no *Star.*

Modelo Dimensional → Esquema *Snowflake*



Modelo Dimensional → Esquema *Snowflake*



→ Um exemplo para a dimensão Geografia de um DW.

OLAP (*Online Analytical Processing*)

- Conjunto de processos para criação, gerência e manipulação de dados multimensionais para análise e visualização, visando maior compreensão dos dados pelos usuários finais.

- É usual a expressão “ferramenta” OLAP, referindo-se aos sistemas com estas funcionalidades e que são, juntamente com o SGBD, a base de um DW.

- Facilidade para fazer análises, definir agregações e cruzamentos, permitindo visualizar os dados em múltiplos níveis de hierarquias e diferentes perspectivas.

Agregações das Informações

- Apesar dos dados no DW serem armazenados segundo a granularidade definida, muitas das consultas realizadas necessitam, além das informações detalhadas, de informações sumariadas ao longo das dimensões.
- A informação armazenada no nível de detalhe é importante, porém o acesso à informação em níveis sumariados permite aos analistas de negócio terem uma visão global do modelo de negócios analisado.
- Estas consultas, partindo de uma base onde existem apenas os dados de nível básico, ou seja, do nível mais detalhado, se for necessário sumariar os dados no momento da execução, todo o processo de análise será sobrecarregado.

Agregações das Informações

- Um determinado conjunto de vários agregados pré-computados faz-se necessário para acelerar cada uma das consultas, sendo que o efeito sobre o desempenho é considerável, obtendo reduções drásticas no tempo de processamento, motivo pelo qual é um recurso bastante eficiente para controlar o desempenho do DW.
- Exemplos:
  - Agregado unidirecional: totais de *categoria* por *loja* por *dia*;
  - Agregado unidirecional: totais de *cidade* por *item de produto* por *dia*;
  - Agregado unidirecional: totais mensais por *item de produto* por *loja*;
  - Agregado bidirecional: totais de *categoria* por totais de *cidade* por *dia*;
  - Agregado bidirecional: totais de *categoria* por totais mensais por *loja*;
  - Agregado bidirecional: totais de *cidades* por totais mensais por *item de produto*;
  - Agregado tridirecional: totais de *categoria* por totais de *cidade* por totais mensais.

OLTP *versus* OLAP

OLTP

- Mais frequente
- Mais previsível
- Pequena quantidade de dados por consulta
- Consulta a dados básicos
- Dados correntes
- Poucas derivações complexas

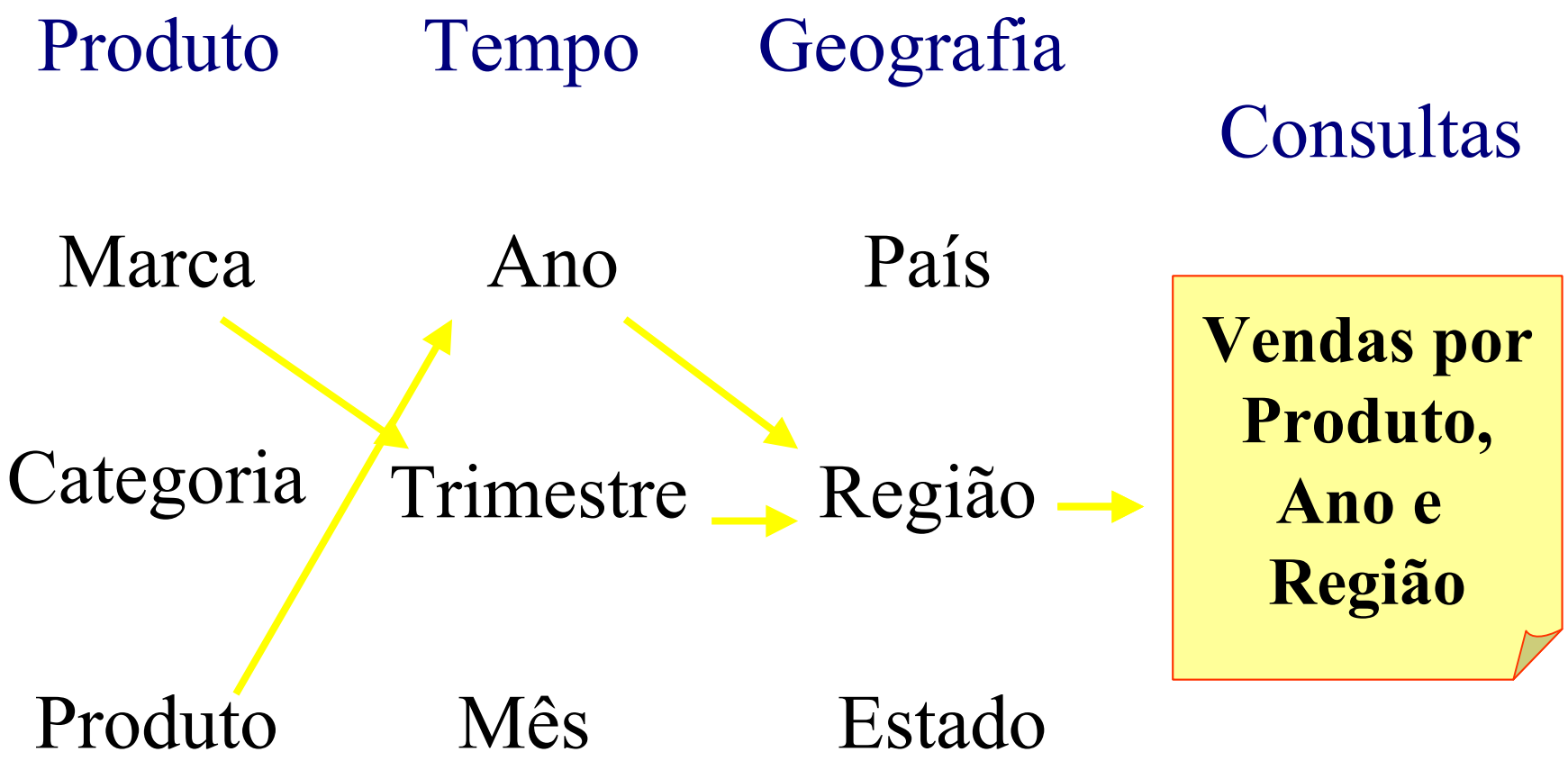
OLAP

- Menos freqüente
- Menos previsível
- Grande quantidade de dados por consulta
- Consulta a dados derivados
- Dados correntes, passados e projeções
- Muitas derivações complexas

SQL versus OLAP

- Desvantagens do SQL:
  - Consultas relacionadas a problemas reais relativamente simples são traduzidas em consultas SQL complexas, envolvendo diversas varreduras, agregações, junções e classificações de tabelas.
  - A linguagem SQL é relativamente pobre no suporte de funções matemáticas para manipular dados históricos (Por exemplo, cálculo da flutuação da média dos últimos três meses).
- Desvantagens do OLAP:
  - Quando o número de dimensões aumenta, o número de células aumenta exponencialmente.

Hierarquias e Agregados



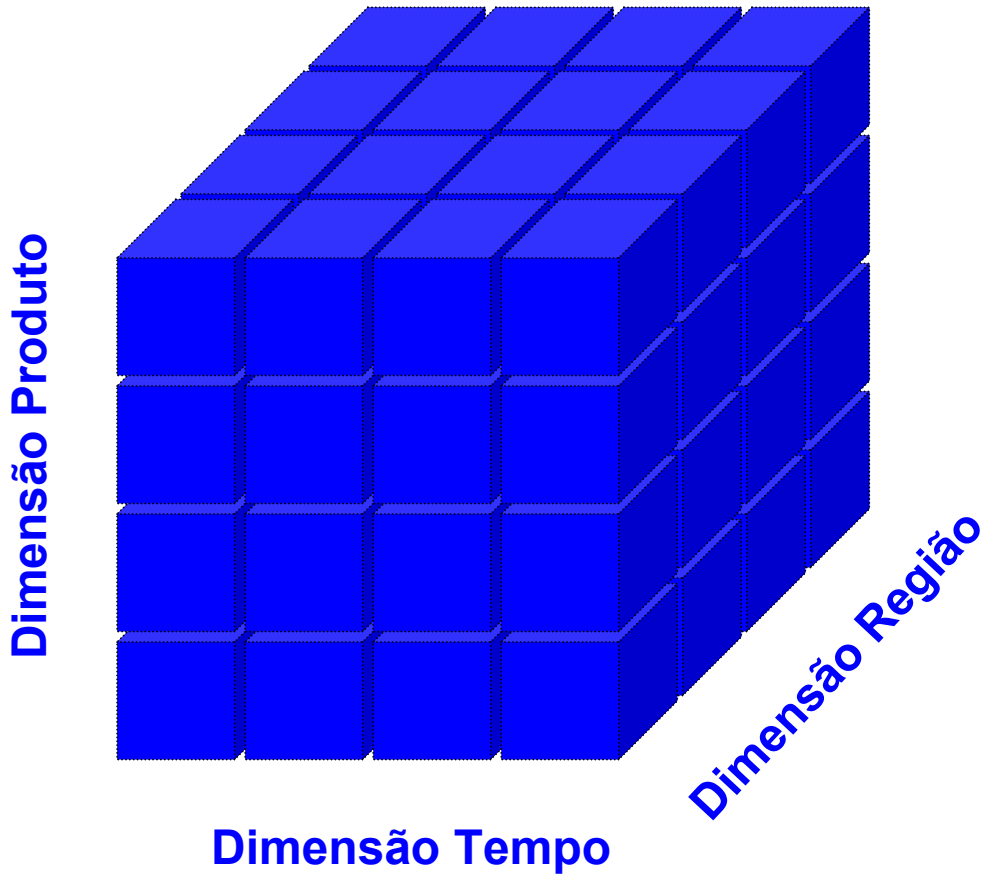
Operações OLAP Usuais

- **Navegação pelas hierarquias e seus elementos:** permite selecionar as perspectivas sob as quais se deseja visualizar as variáveis ou medidas;
- **Cruzamentos:** permitem sumariar fatos por diferentes combinações das dimensões;
- **Drill down:** navegação ao longo das dimensões na direção de maior detalhe;
- **Roll up (Drill up):** navegação ao longo das dimensões na direção de menor detalhe;

Operações OLAP Usuais

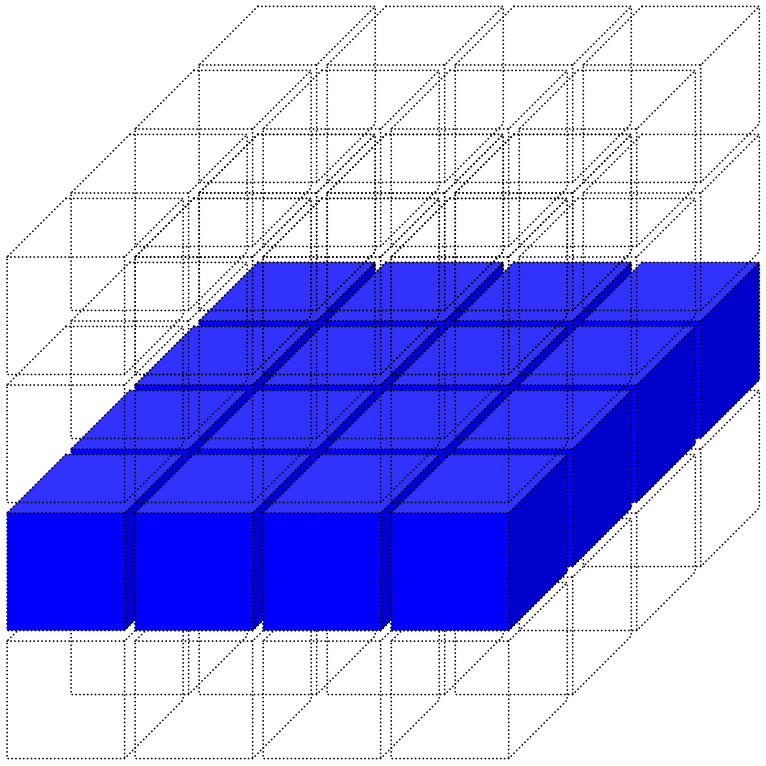
- **Rotação:** capacidade de inverter colunas e linhas; Navegação ao longo da dimensões na direção de maior detalhe;
- **Slice and Dice**: Caminha através de um dado específico. Seleção definindo um subcubo;
  - *(Ex: vendas onde cidade = 'Porto Alegre' e data = '1/15/90')*
- **Cálculo e ranking.**
  - **(***Ex: top 3% das cidades por média de rendimentos)*

Exemplos no Cubo de Dados

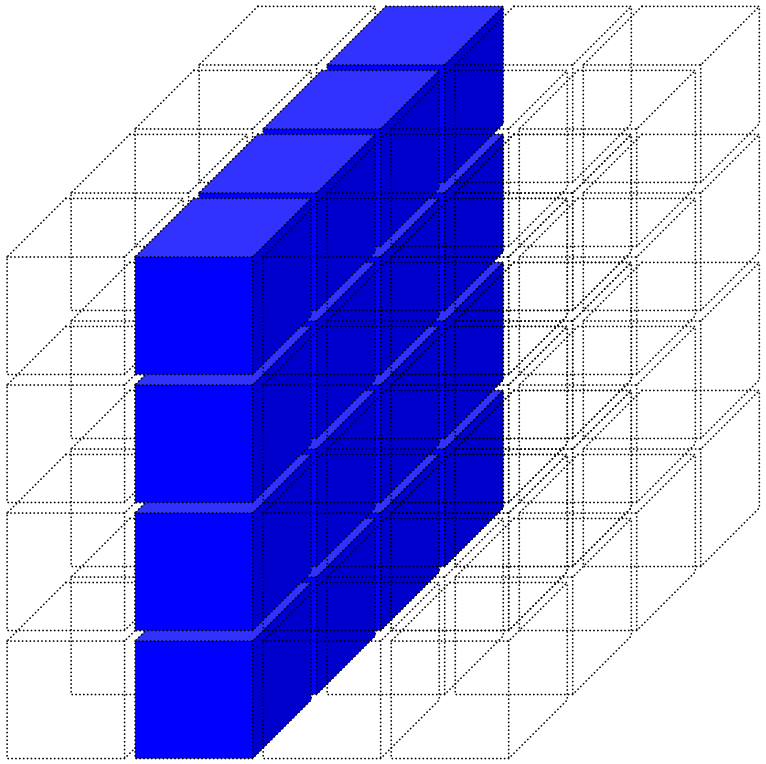


Slice and Dice

Visão Produto

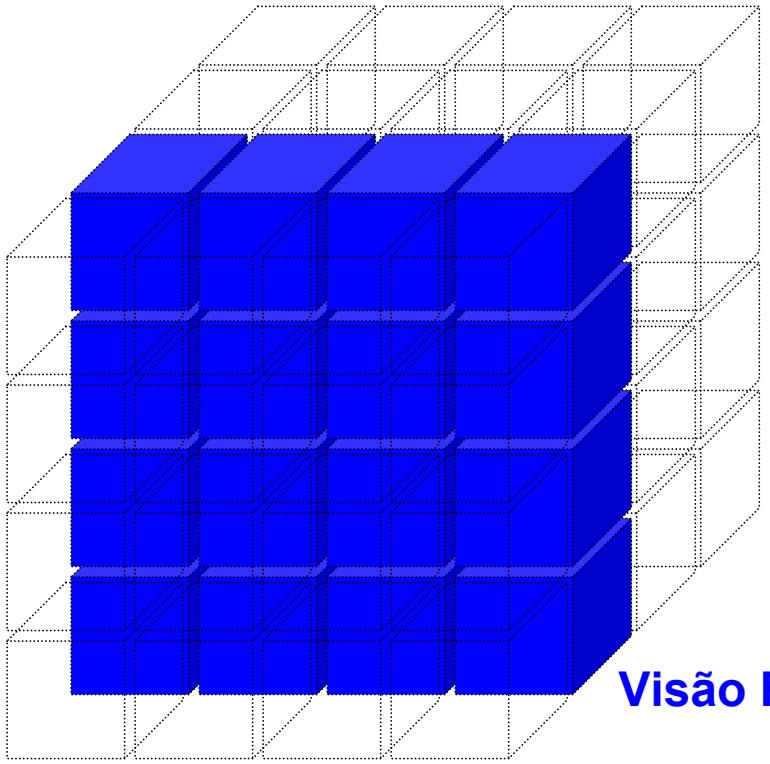


Slice and Dice



Visão Tempo

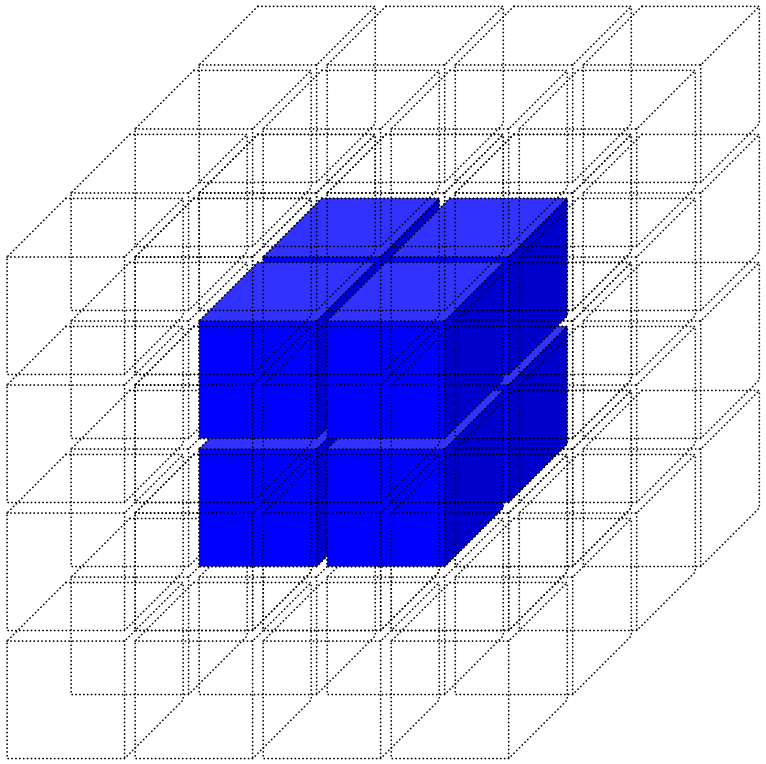
Slice and Dice

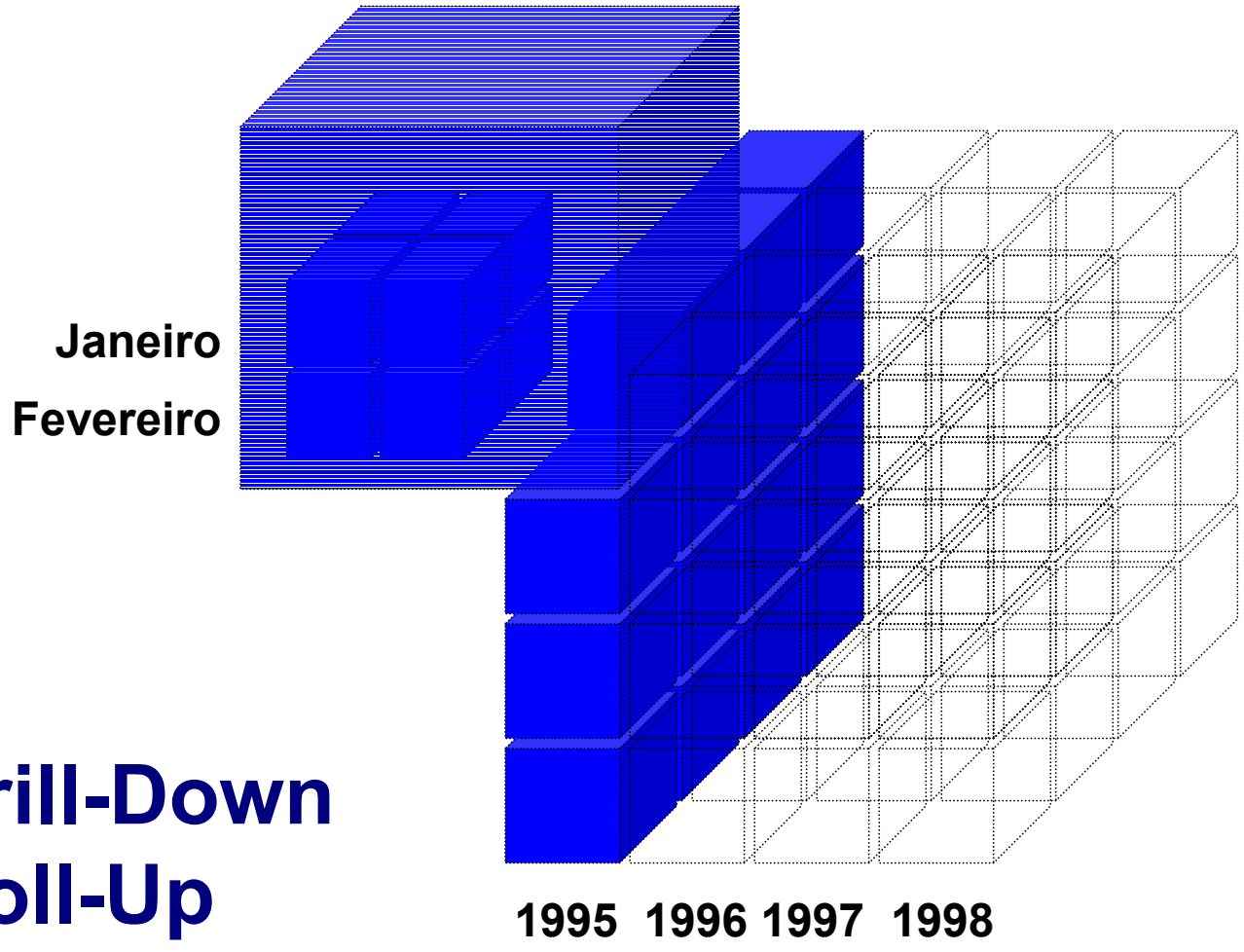


Visão Região

Slice and Dice

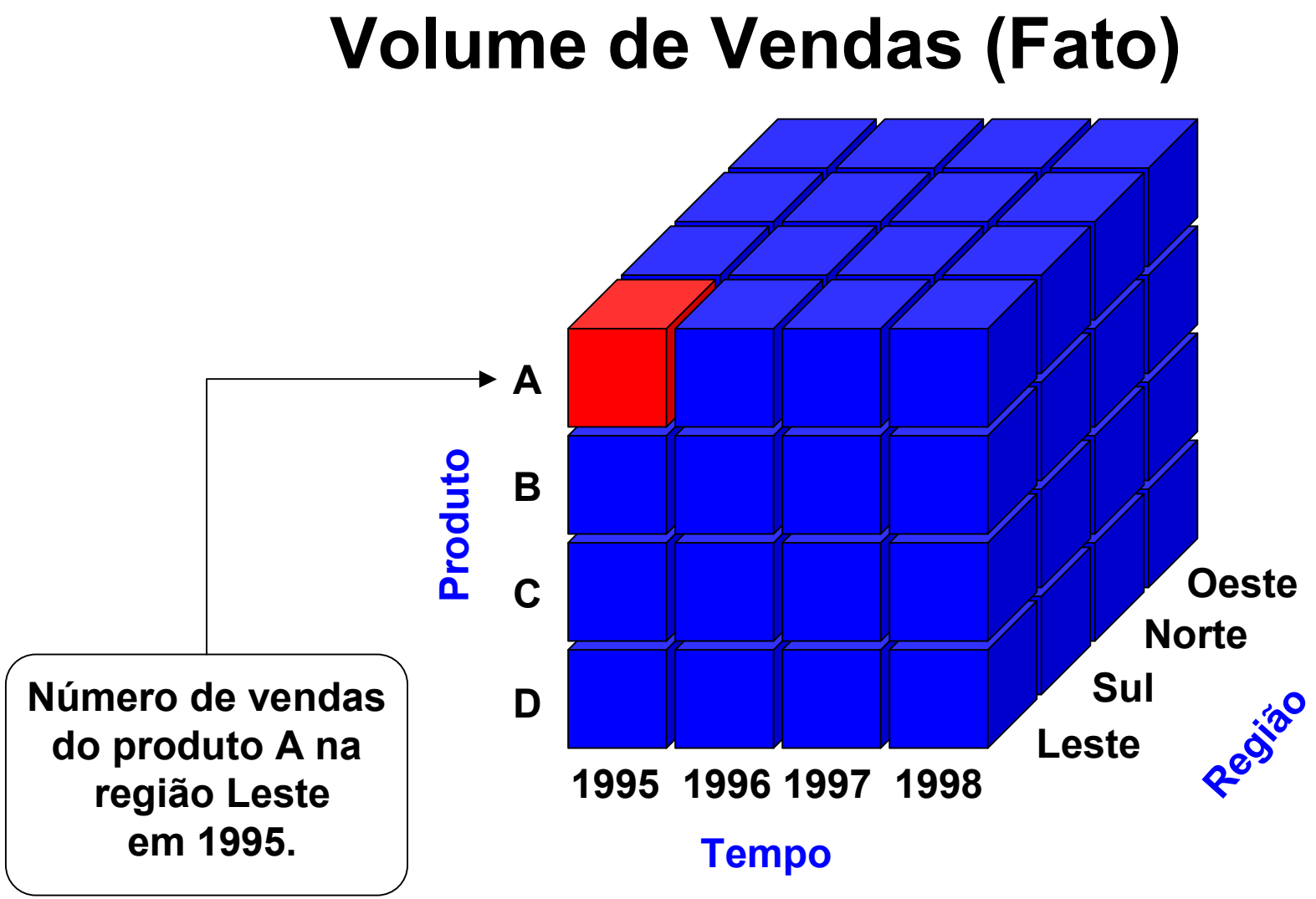
Visão ad-hoc





Drill-Down
Roll-Up

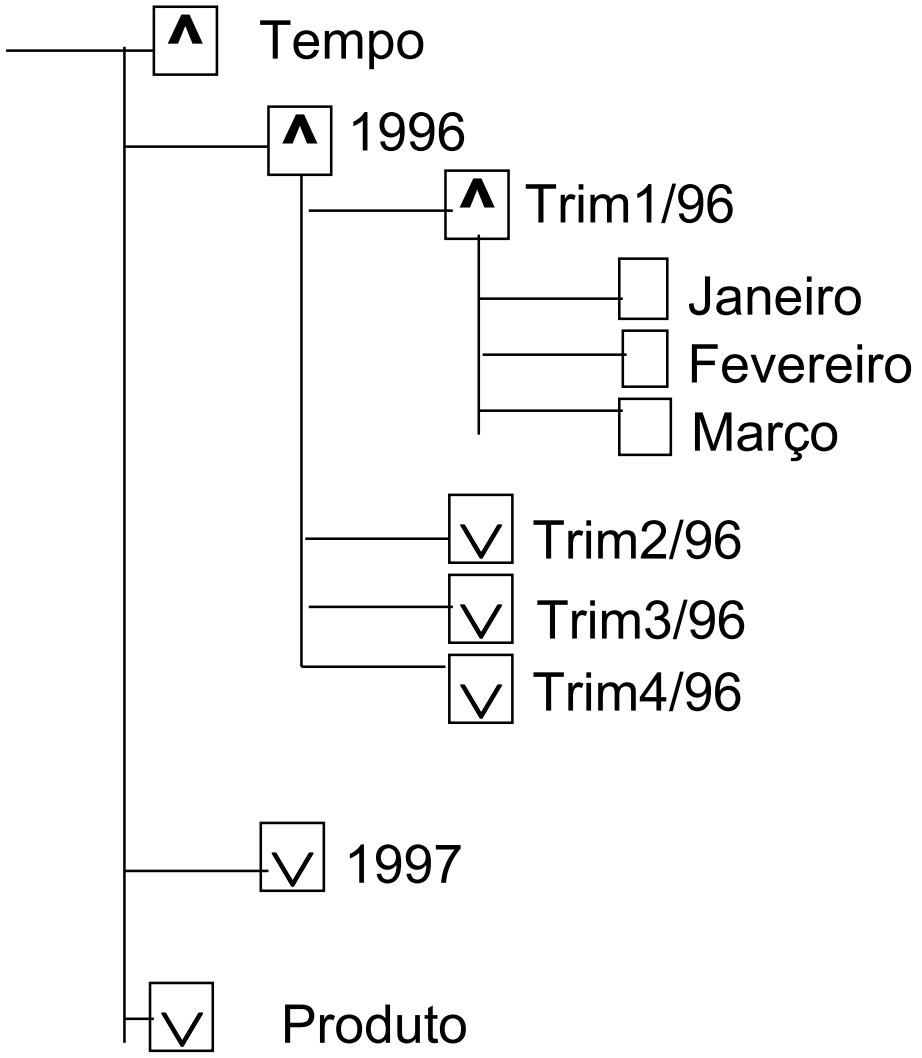
Analisando o Cubo



Dimensões vistas em ferramentas OLAP

Dimensão Tempo

Chave_Tempo
Mes Trimestre Ano



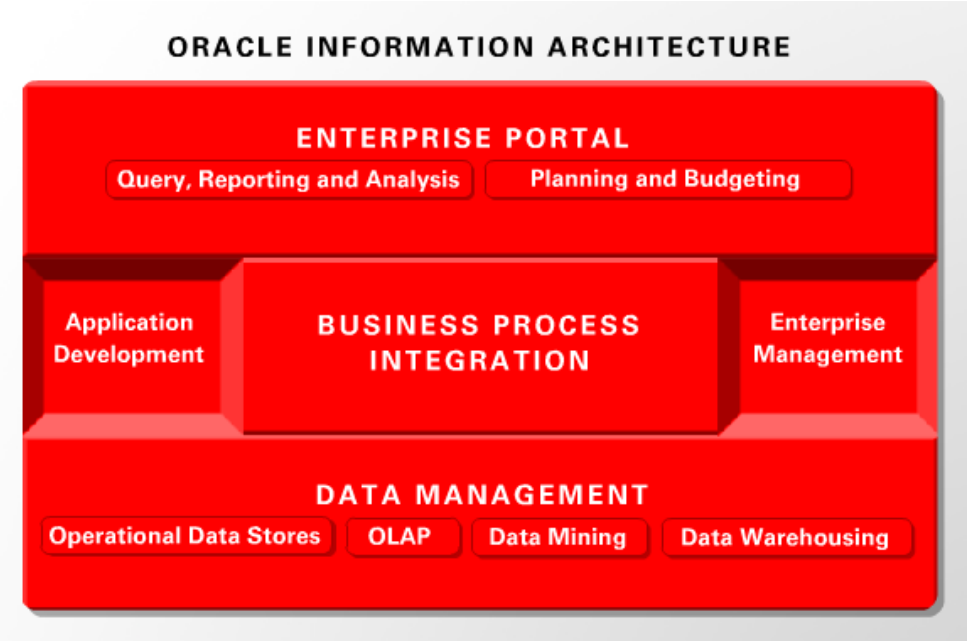
Ferramentas de OLAP

- DynamiCube 3.0
  - http://www.datadynamics.com
  - Exemplos no site.
- Maestro
  - http://www.hperinf.com.br
  - Hyper Consultoria em Informática LTDA
  - Ferramenta ROLAP, cujo SQL gerado faz acesso, via ODBC, a BDs relacionais como Oracle, SyBase, DB2, etc ou até mesmo para ambientes menores, Access, FoxPro, DBase.

Ferramentas para SAD

Oracle

- *Oracle Warehouse Builder*
- *Oracle Partitioning*
- *Oracle Data Mining*
- *Oracle OLAP*

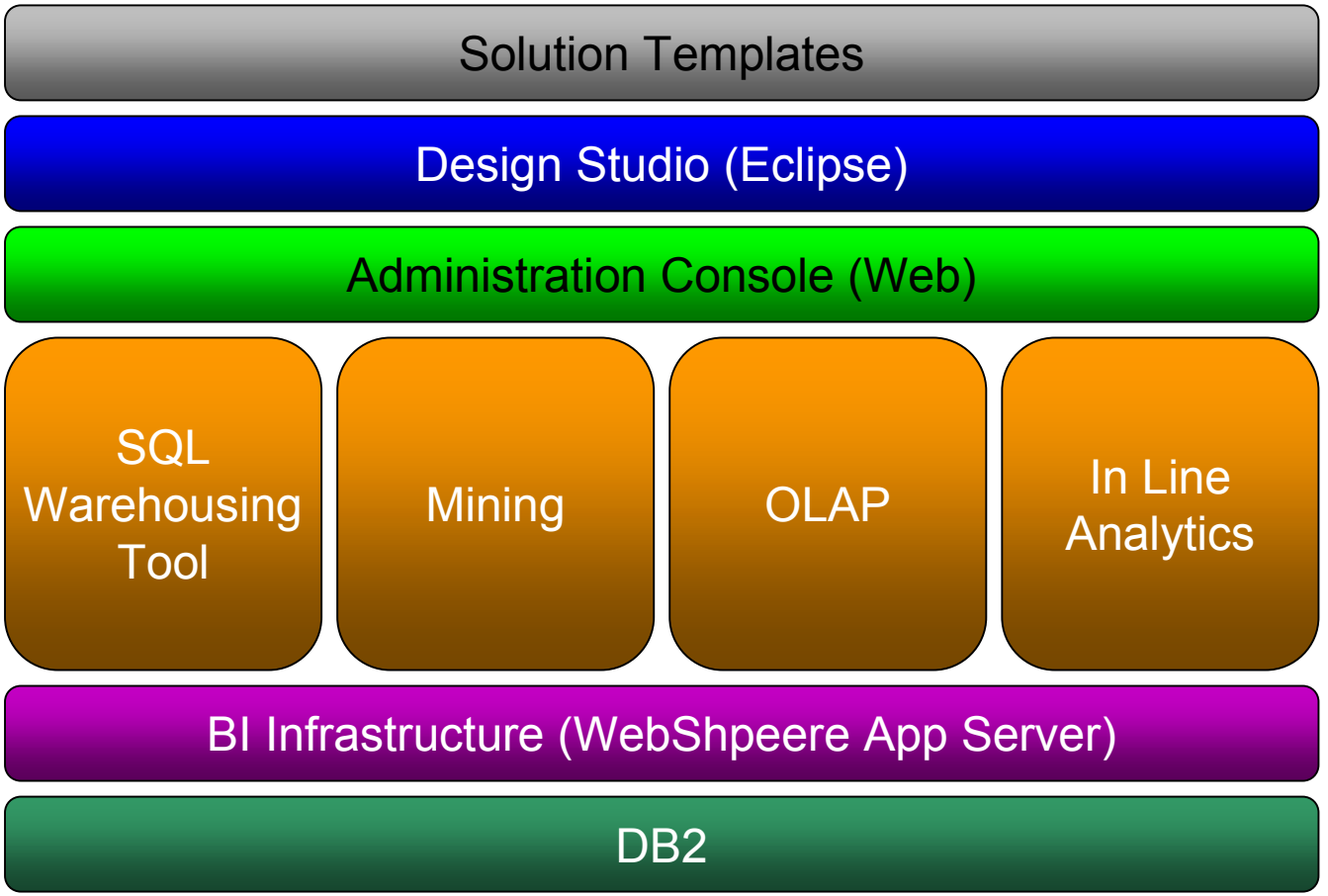


Microsoft

- *SQL Server Business Intelligence (BI) Development Studio.*
  - **Integration Services (SSIS)**
  - **Analysis Services (SSAS)**
  - **Reporting Services**
  - **Data-mining**

Ferramentas

DB2 Data Warehouse Edition for Linux, Unix and Windows



Referências Bibliográficas

- **Introdução a Banco de Dados (Apostila, Cap. 10). Prof. João Eduardo Ferreira (IME/USP)**
- **Notas de aula da Prof. Maria Luiza M.Campos (DCC/IM/UFRJ)**
- **Notas de aula do Prof. Edgard Jamhour (PPGIA/PUCPR)**
- **Eric Thomsen. OLAP – Construindo Sistemas de Informações Mutidimensionais. Editora Campus. Rio de Janeiro, 2002.**
- **Ralph Kimball. Data Warehouse Toolkit. Editora Makron Books. São Paulo, 1998.**
- **Laudon & Laudon. Gerenciamento de Sistemas de Informação. 3ª Edição. Editora LTC. Rio de Janeiro, 2001.**
- **Sistemas de Banco de Dados. (Cap. 28) Ramez Elmarsri e Sham Navathe. 4ª Edição. Ed. Pearson, 2005.**
- **Sites oficiais dos fornecedores das tecnologias.**