

DATA WAREHOUSE

Professor MSc Ly Freitas Filho

Site: www.lyfreitas.com

E-mail: ly@lyfreitas.com

Tendências: tecnologias

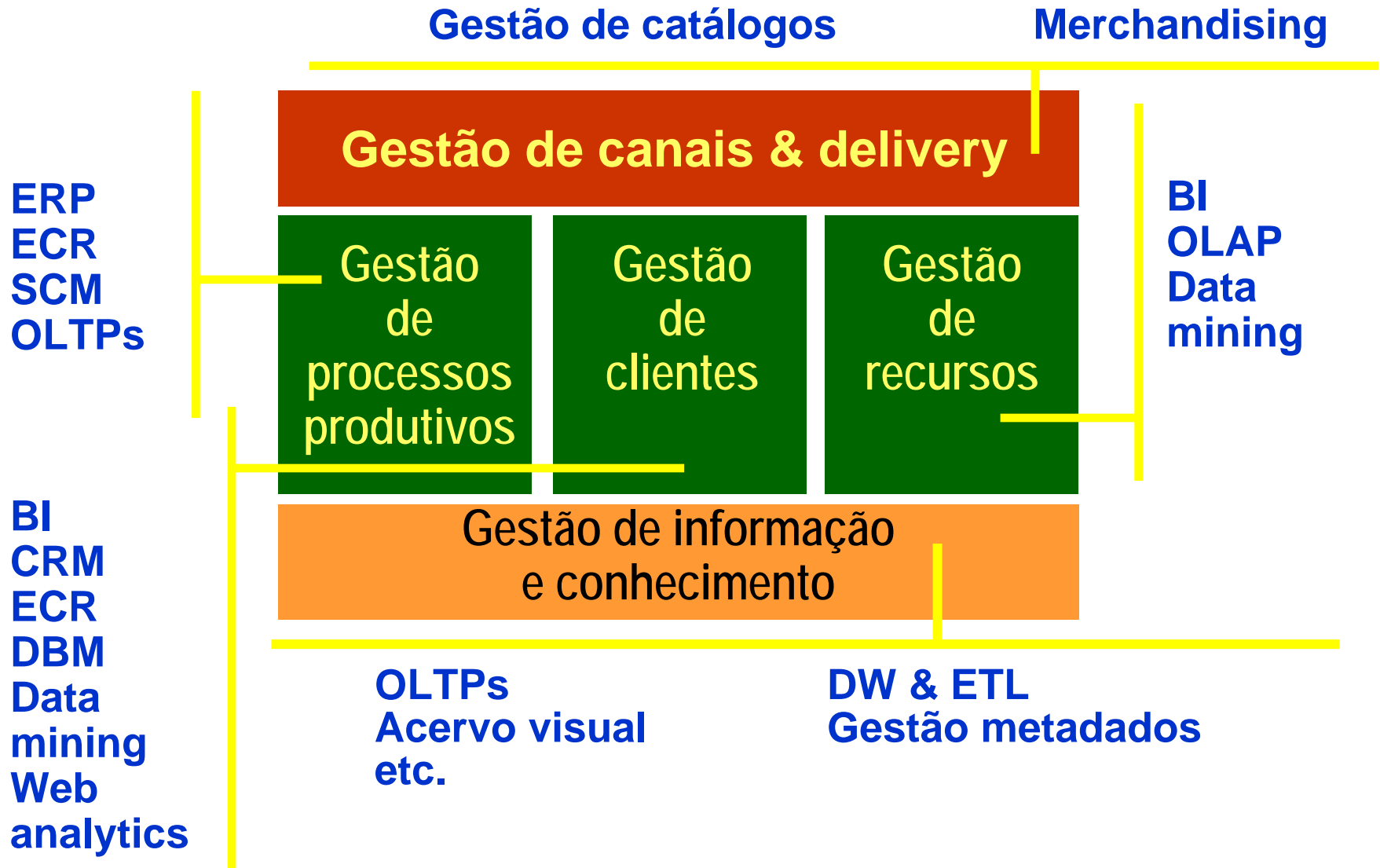
interactive media **wireless**
web analytics
gestão do conhecimento **business intelligence**
info-entertainment **web commerce**
web warehousing **content management**
virtual reality **gestão da cadeia de valor**
customer relationship management **technomarketing**
data mining **ensino à distância** **modelos preditivos**



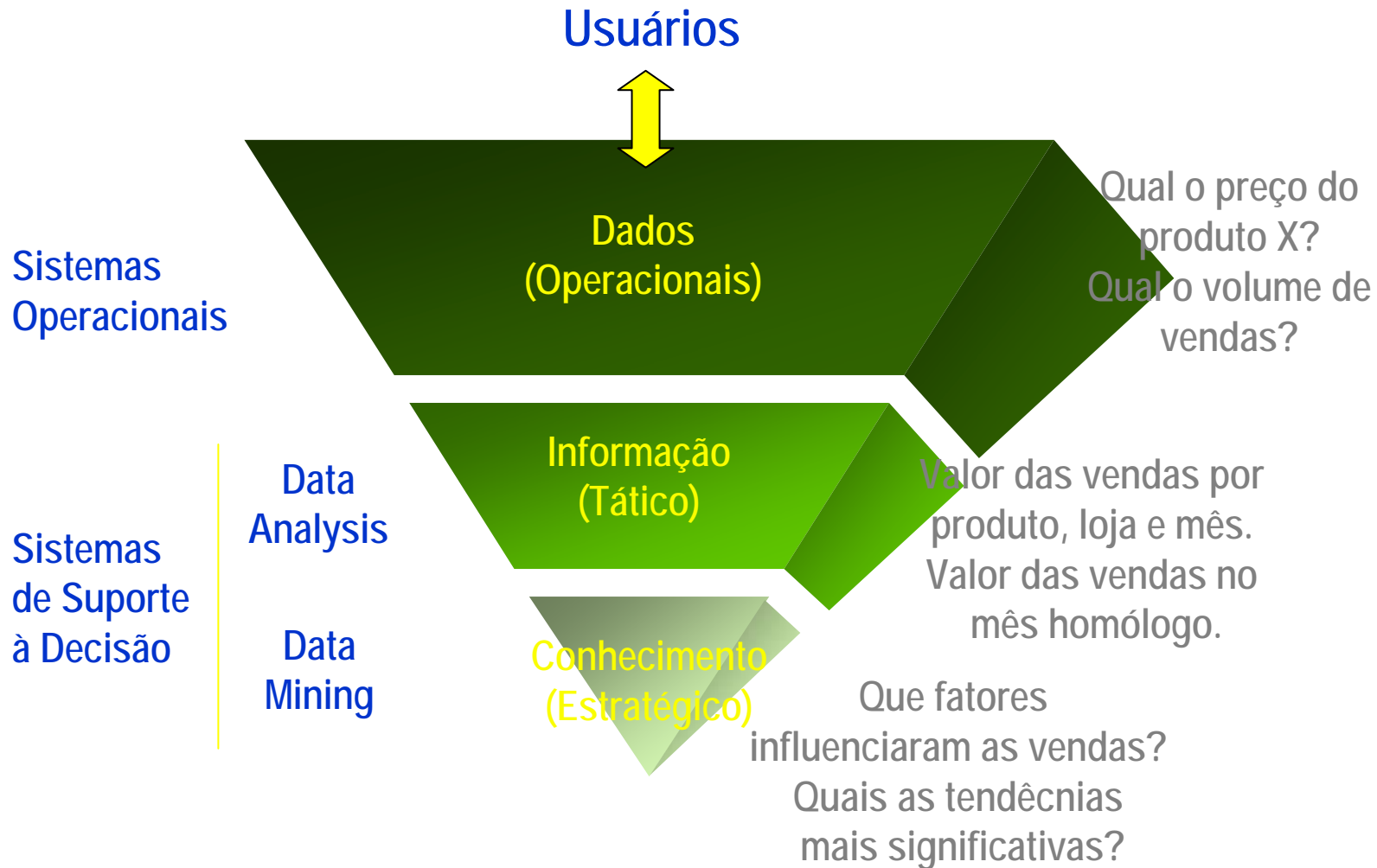
Business Intelligence: quadro de referência

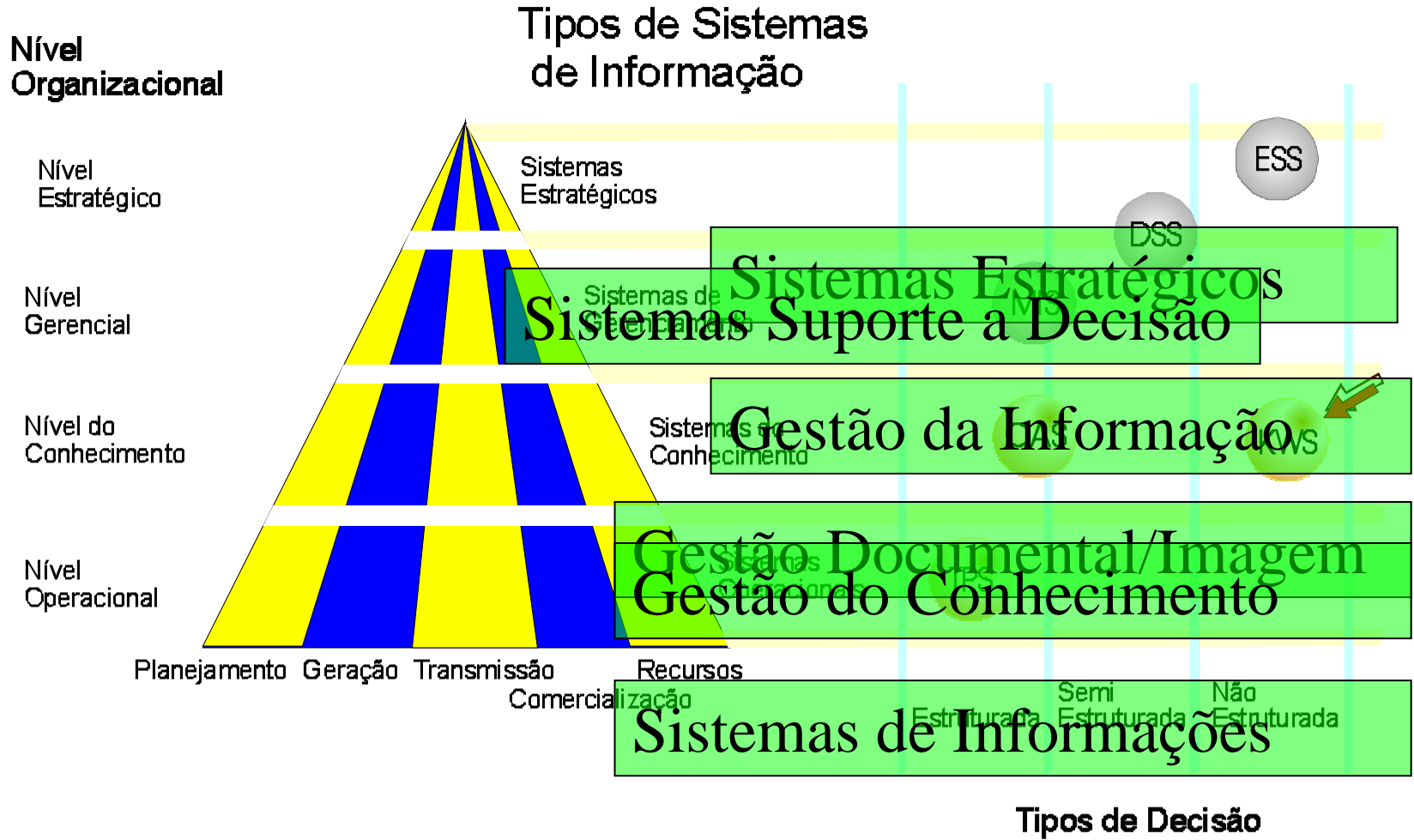


Business Intelligence: quadro de referência



Para o sucesso do negócio é necessário transformar os dados em informação e conhecimento

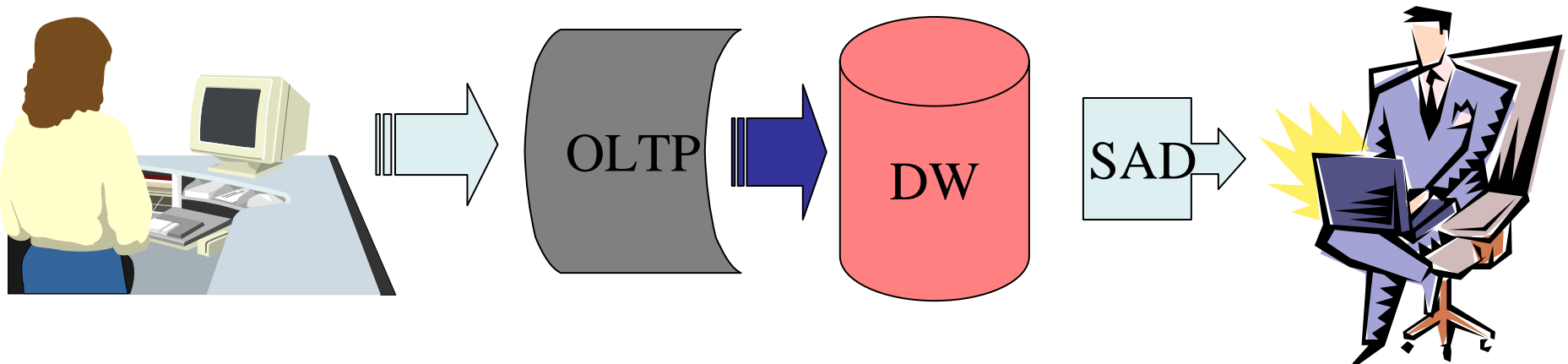




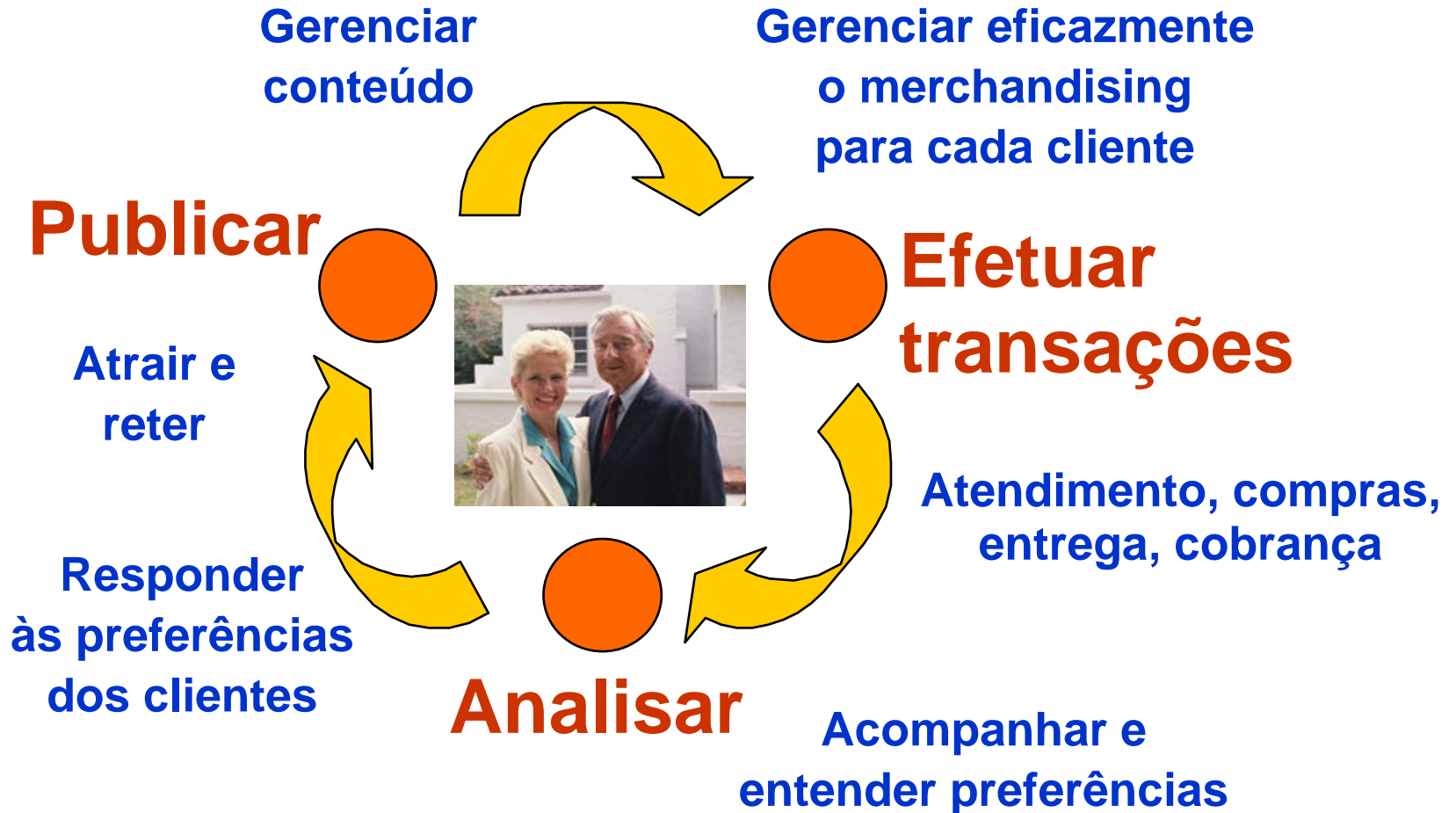
Ref. Management Informations Systems
K.C.Laudon; J.P.Laudon

Evolução dos Sistemas de Informação

- OLTP - Processo de transações On-Line: automatizar os processos, melhorar o desempenho e confiabilidade
- SAD - Sistemas de apoio a decisão: sistemas que ajudam decisores a tomar decisões em situações onde o julgamento humano é uma contribuição importante ao processo de resolução, mas existe uma limitação humana para processar informações



O Ciclo P-T-A

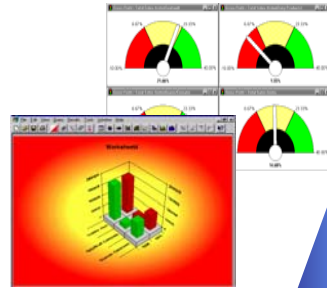


A arquitetura de infonegócios

Portal de acesso e distribuição

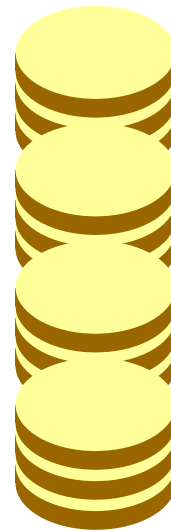


Análise e exploração



Ciclo PTA

Bases analíticas



Data Mart

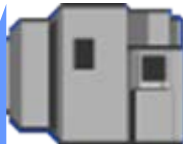
Extração e integração de dados



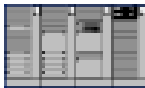
Data Warehouse ou ODS

Introdução/Negócios

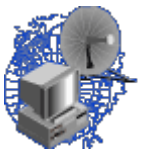
Fontes de dados



OLTP



Legado



Externo

As necessidades de informação estratégica e consolidada sempre existiram...

- **Arquivos simples (poucos Mb)**
- **Linguagens Imperativas**
- **Análise dos Dados**
 - Pedida aos programadores
 - Equivalente a nova aplicação
 - Forma típica: impressões em papel
- **BDs Cliente/Servidor (muitos Gb)**
- **Ferramentas Específicas**
- **Análise dos Dados**
 - Diretamente pelos gestores
 - Forma típica: usando interfaces tipo “point-and-click”

1970

1980

1990

2000

- **BDs Centralizadas (muitos Mb)**
- **Linguagens Declarativas e Folhas de Cálculo**
- **Análise dos Dados**
 - Pedida a analistas e assessores
 - Usando “perguntas relacionais”
 - Forma típica: listas na tela ou folhas de cálculo

Anos 2000 o domínio do acesso Internet.

A importância da informação

- **SGBDs + Internet (muitos Tb)**
- **Ferramentas Específicas**
- **Análise dos Dados**
 - “Informação na ponta dos dedos”
 - Tecnologia “push”
 - Forma típica: Browser Web

“Ferramentas de interrogação e folhas de cálculo têm-se mostrado extremamente limitadas na forma como a informação pode ser agregada, apresentada e analisada” *E.F. Codd*

“A lacuna mais importante das bases de dados relacionais tem sido a incapacidade de consolidar, apresentar e analisar informação sobre múltiplas dimensões” *E.F. Codd*

“O maior desafio das empresas de tecnologias de informação é aprender a construir Bases de Informação e não Bases de Dados” *Peter Drucker*

“Informação sobre dinheiro está a tornar-se mais importante que o dinheiro propriamente dito.” *John Reed, President of Citicorp/Citibank*

Data Warehouse

- **É um conjunto de dados íntegros, integrados e históricos, não voláteis, organizados por assunto que servirão de base aos sistemas de suporte à decisão – SSD ou sistemas de apoio à decisão - SAD.**

Data Warehouse

- a fonte de consulta de um empreendimento (Kimball et al, 1998)
- coleção de dados **orientada a assunto, integrada, não volátil e variável em relação ao tempo**, que tem por objetivo dar **apoio aos processos de tomada de decisão** (Inmon, 1997)

Data Warehouse

- uma base de dados analítica que dá apoio a processos decisórios + **recursos de acesso intuitivos** (Poe et al, 1998)
- um **processo**, e não um produto, para a **montagem e administração de dados provenientes de várias fontes** com o propósito de obter uma visão simples e detalhada de parte de todo o negócio (Gardner, 1998)

Quando organizar os dados?

- Grande volume de dados, dificuldade no acesso
- Resultados do mesmo negócio apresentados com valores diferentes por áreas diferentes
- Dificuldade em localizar os dados relevantes ao negócio
- Pouca confiabilidade nos dados apresentados.
- Tempo de resposta muito ruim, quando se tenta pesquisar uma informação no banco de dados.

Um Data Warehouse é uma arquitetura de sistemas com um processo complexo de construção

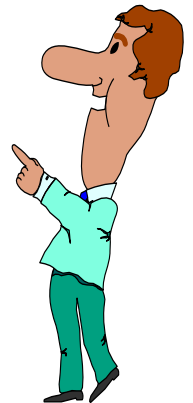
- um Data Warehouse é uma **ARQUITETURA...**
não é um produto ou tecnologia
- um Data Warehouse **CONSTRÓI-SE...**
não se compra
- um Data Warehouse é um processo **COMPLEXO...**
não um simples projeto



“Primeiro surgiu a arquitetura, a seguir a metodologia depois (e apenas depois) surgiram as ferramentas”

Data Warehouse a informação estratégica e consolidada do seu negócio

- **Permite a análise consolidada dos dados da organização. Estrutura a informação de forma multidimensional e hierárquica orientada aos conceitos de negócio**
- **Flexibilidade na construção de análises, permitindo navegação nos dados e rápidas mudanças de perspectiva**
- **Interface avançada com os utilizadores. Ferramentas de acesso da nova geração com capacidade de disponibilização de informação via Web, Wap e Voz**



Foco no negócio: uma das diferenças entre Sistemas Operacionais e Sistemas de Suporte à Decisão

	Sist. Operacionais	Data Warehouse
Fontes	internas	internas + externas
Organização	aplicação (processo)	tema (negócio)
Natureza	val. correntes	val. históricos
Otimização	normalização	redundância
Dimensão BD	Mb a Gb	Gb a Tb
Tipo Utilização	burocrática/repetitiva	analítica/exploratórias
Tempos Resposta	instantâneos	minutos, horas
Previsão Carga	possível	difícil
Atualização	atômica, alta freq.	blocos, baixa freq.

No cerne desse novo ambiente "projetado" está a percepção de que há fundamentalmente duas espécies de dados:

Dados Primitivos e
Dados Derivados.

Dados Primitivos

São dados detalhados utilizados na condução das operações cotidianas da Organização.

Dados Derivados

São dados resumidos ou calculados de forma a atender às necessidades da área estratégica da Organização.

Data Warehouse X Data Mart

- **Data Warehouse** – contém todas as informações da companhia, vindas de múltiplas fontes de dados operacionais, dispostas de forma integrada e consolidada
- **Data Marts** – contém um subconjunto dos dados corporativos para atender um departamento ou uma unidade de negócio.

Datawarehouse X Datamart

Datawarehouse



Datamart

Datawarehouse X Datamart

Qual fazer primeiro????

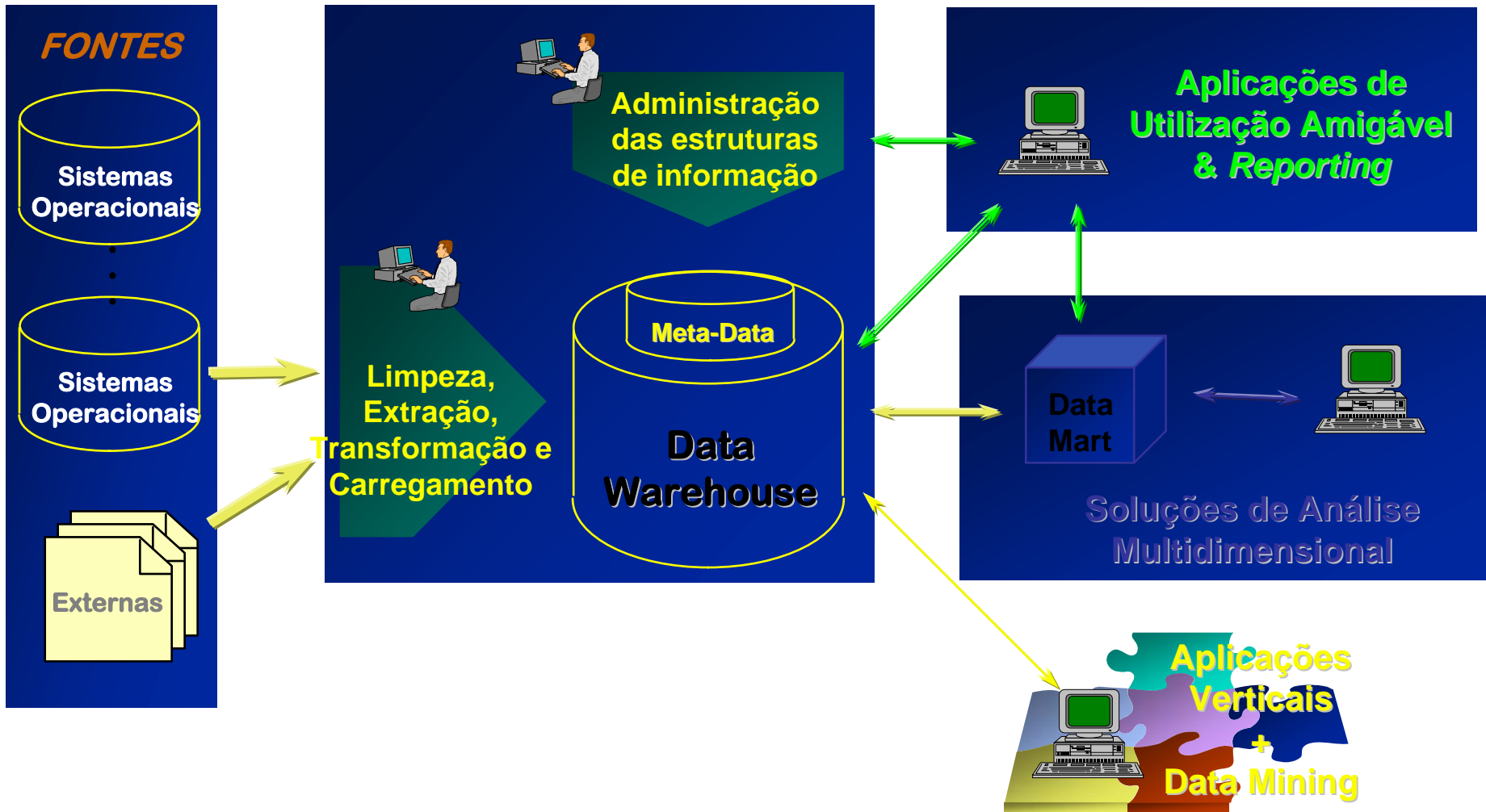
Data Mart (DM)

- Data Warehouse de pequena capacidade usado para atender a uma unidade específica de negócios
 - projeto piloto
 - atender necessidades imediatas de um Processo
 - restrições (custo, tempo, conhecimento tecnológico)
 - desempenho
 - aprendizagem, aceitação

Data Warehouse (DW)

- Data Warehouse (corporativo)
 - integração de seus data marts
 - requer um planejamento global que norteie o desenvolvimento de DMs individuais
 - integração em sistemas operacionais

A arquitetura de referência de um Data Warehouse: processos de ETC, Metadata, Data Mart e Reporting.



Granularidade

É o nível de detalhe ou de resumo contido nas unidades de dados existentes no DW

É a unidade de medida mínima de um modelo de DW .

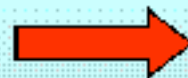
É a combinação de uma linha da tabela de fatos, associada a uma linha de uma ou mais dimensões .

Agregação

São registros sumarizados logicamente redundantes com os dados Granulares do DW

Finalidades: (melhorar o tempo de reposta as consultas; reduzir o tempo de processamento; reduzir espaço de armazenamento)

VENDAS - LOJA XX (R\$)					
FILIAL	DIA				
	13/09	14/09	15/09	16/09	17/09
1	3000	2500	2000	3000	5000
2	1000	1500	1200	1800	2500
3	4000	3500	3000	3400	6000



VENDAS - LOJA XX (R\$)	
FILIAL	SEMANA
1	15500
2	8000
3	19900

Metadados

O metadado representa a definição dos dados contidos no DW, é através dele, que o usuário fica sabendo como as entidades estão representadas, de onde surgem, como foram transformadas e como podem ser utilizadas.

O metadado corresponde a um catálogo e dependendo de sua estrutura poderá conter várias informações.

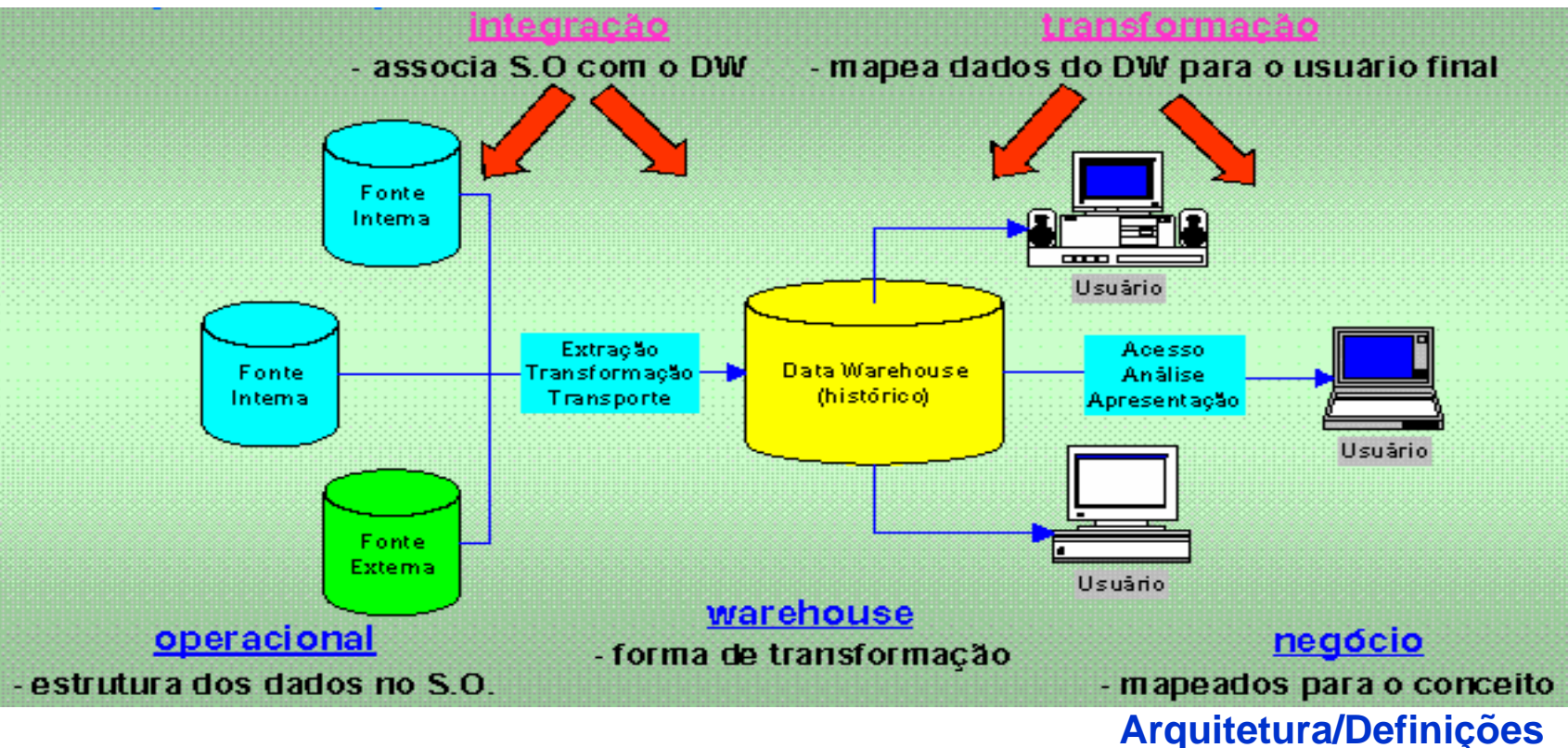
Metadados

No ambiente de DW, os metadados armazenam informações sobre todo ciclo de vida:

- De onde o dado veio?
- Como foi calculado?
- Quando foi realizado o processo de ETL?
- Estatísticas de utilização.
- Mudanças na política de negócios.
- e muito mais...

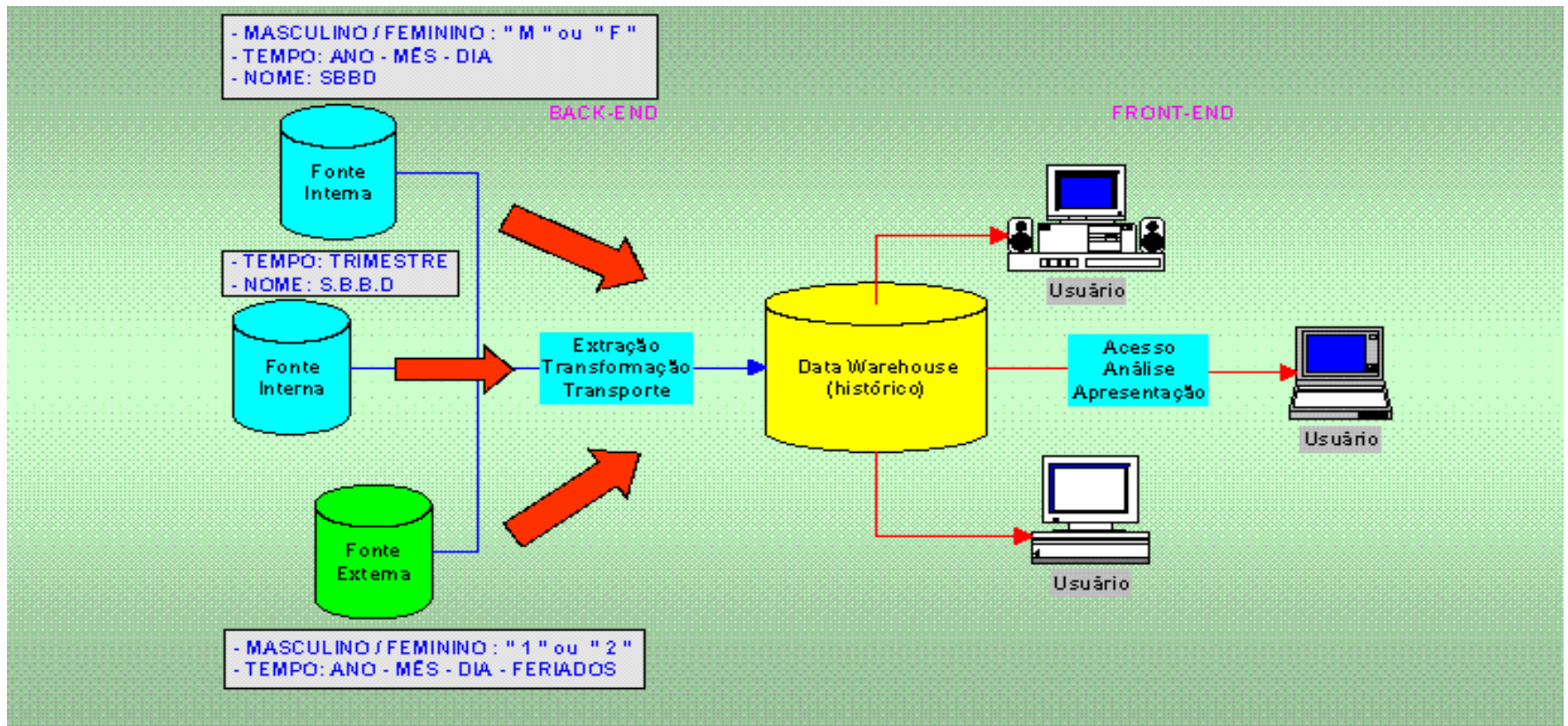
Metadados

Dados sobre dados”. Provêm informações sobre a estrutura de dados e as relações entre estas dentro ou entre bancos de dados. São também informações mantidas a cerca do DW em lugar das providas pelo DW



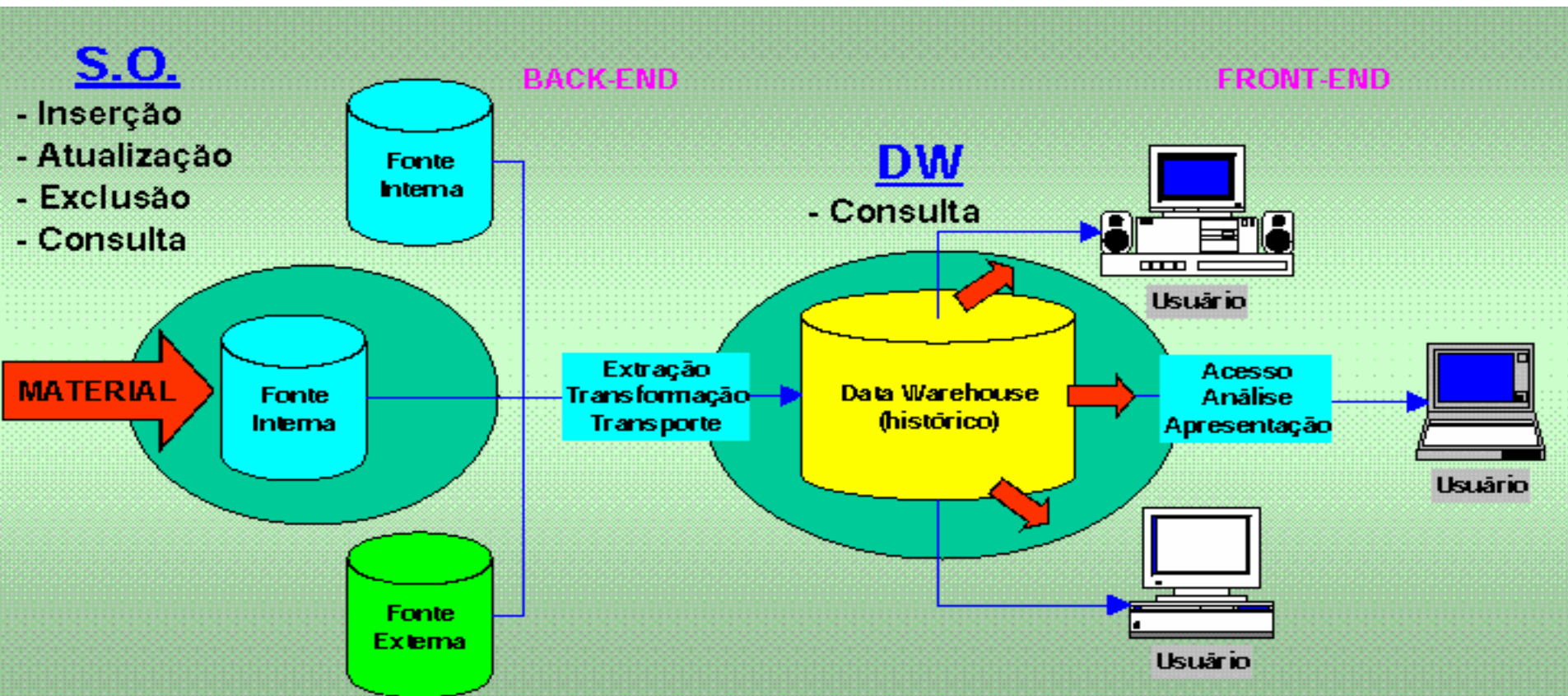
Integrado

Os dados fonte de sistemas OLTP são modificados e convertidos para um estado uniforme de modo a permitir a carga no DW.



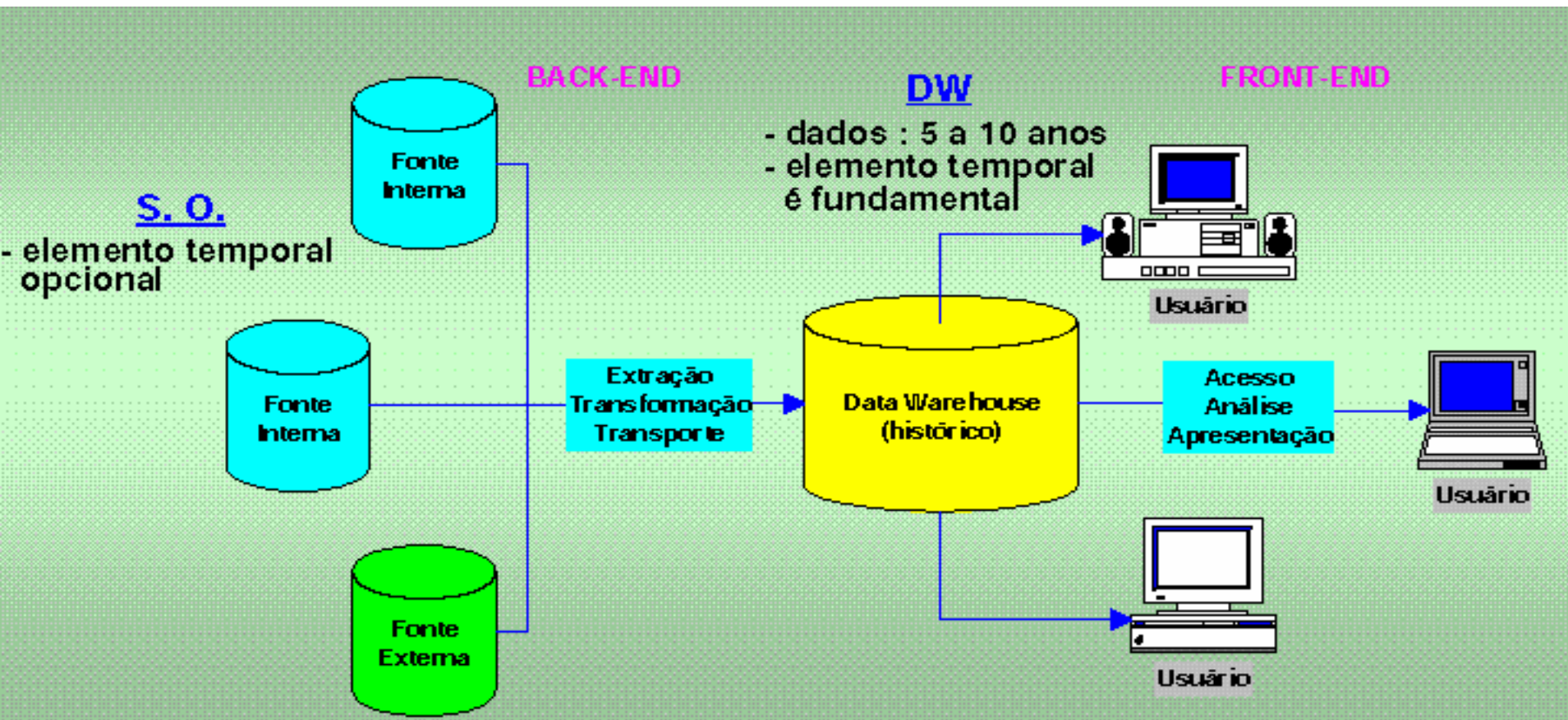
Não Volátil

Os dados após serem extraídos, transformados e transportados para o DW estão disponíveis aos usuários somente para consulta

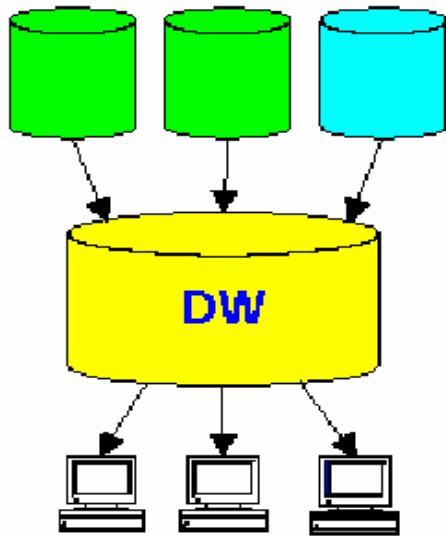


Variável em Relação ao Tempo

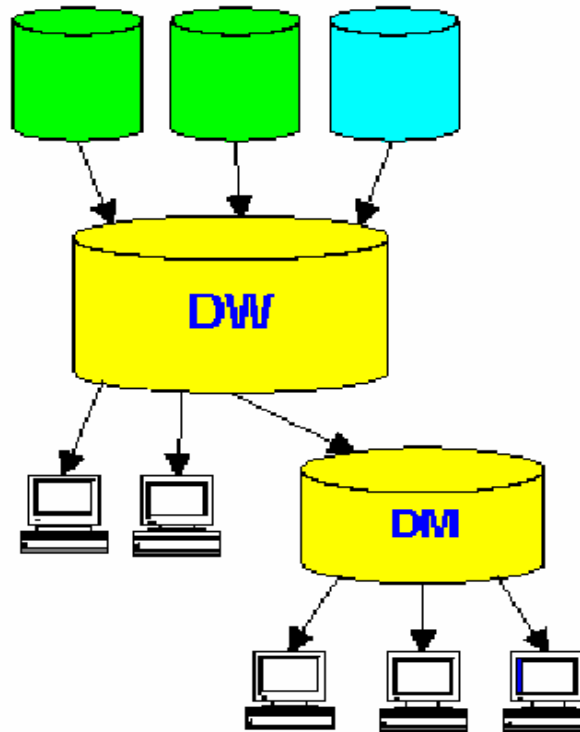
Os DW devem armazenar dados por um período de tempo.
O elemento tempo é fundamental



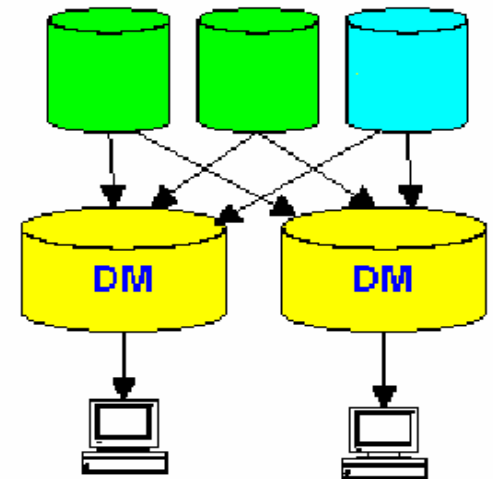
Topologias



Topologia Centralizada

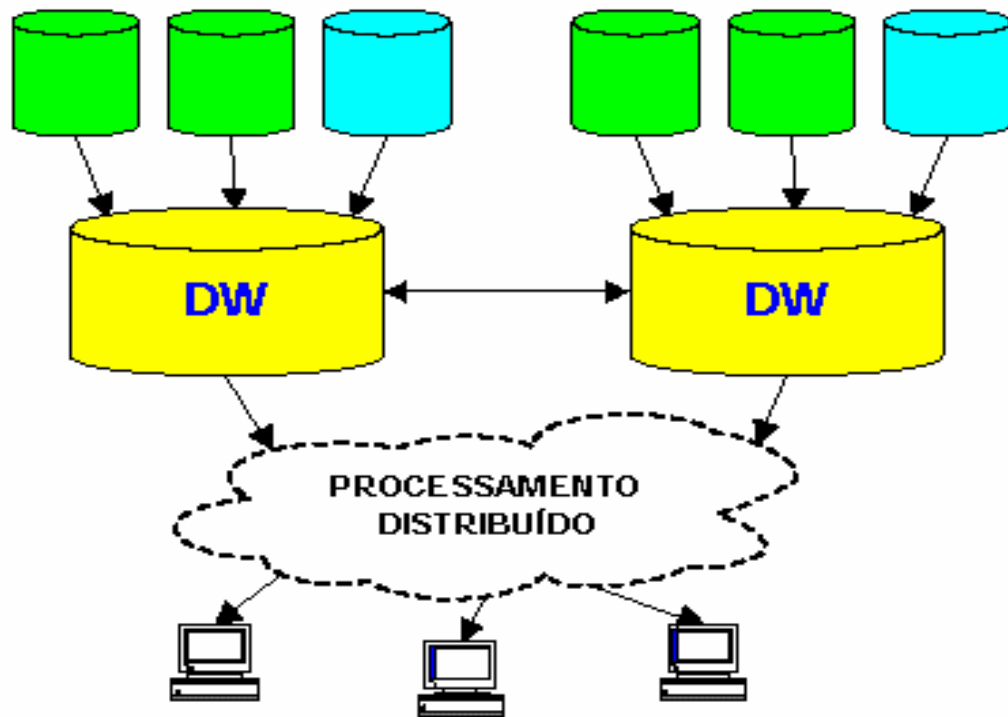


**Topologia DW e DM
(DM dependente)**

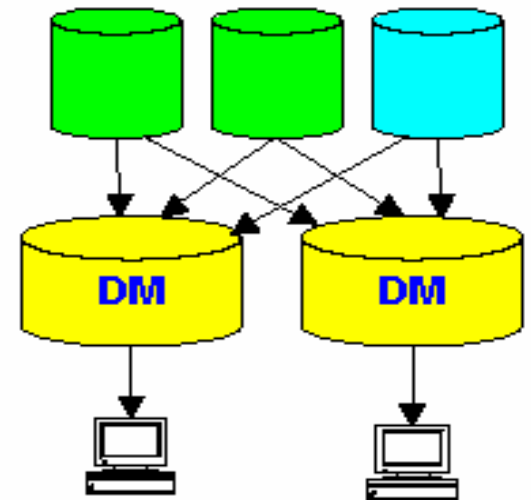


**Topologia DM
(DM independente)**

Topologias



Topologia DW Distribuído



**Topologia DW
(Desenvolvimento Estratégico)
Arquitetura/Topologias**

Sistema Fonte

Um sistema operacional de registros cuja função é capturar as transações de negócios, as vezes são chamados de sistemas legados .

Importância dos Dados Corporativos

Com a globalização, as corporações estão cada vez mais necessitando de informações confiáveis em um tempo hábil para tomada de decisões.

A implantação de um sistema de suporte à decisão passa a ser um diferencial em uma corporação, pois oferece condições para que os níveis gerenciais definam os rumos da companhia com base em dados consistentes.

Data Staging Area

Área de transição dos dados (dados estagiários) e definição dos processos para limpeza, transporte, combinação, integração, melhoramento e preparação dos dados para uso no Data Warehouse

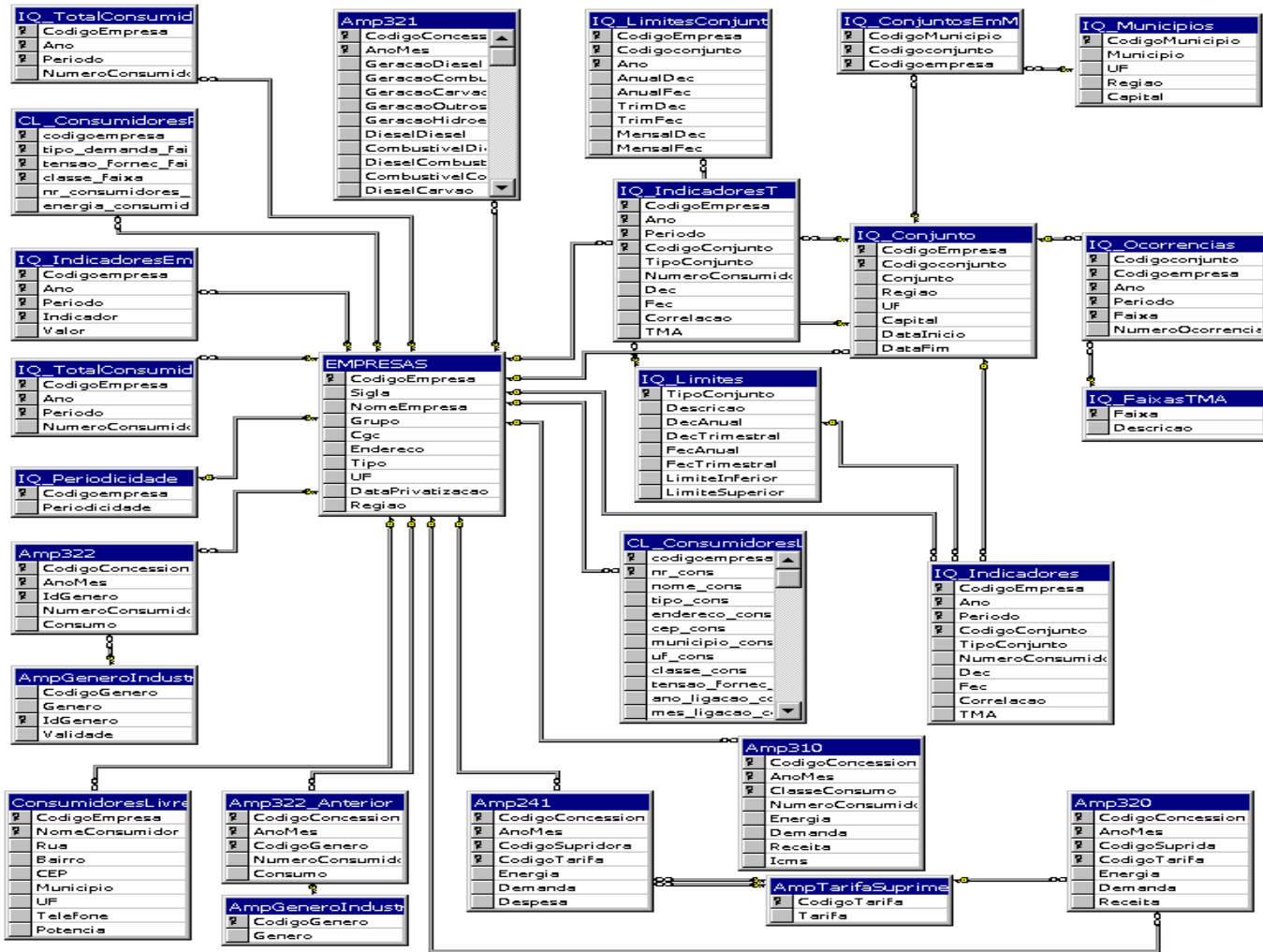
Presentation Server

Máquina física alvo no qual os dados do Data Warehouse estão organizados e armazenados para consulta direta pelos usuários finais, servidores de relatórios e outras aplicações.

Modelo Dimensional

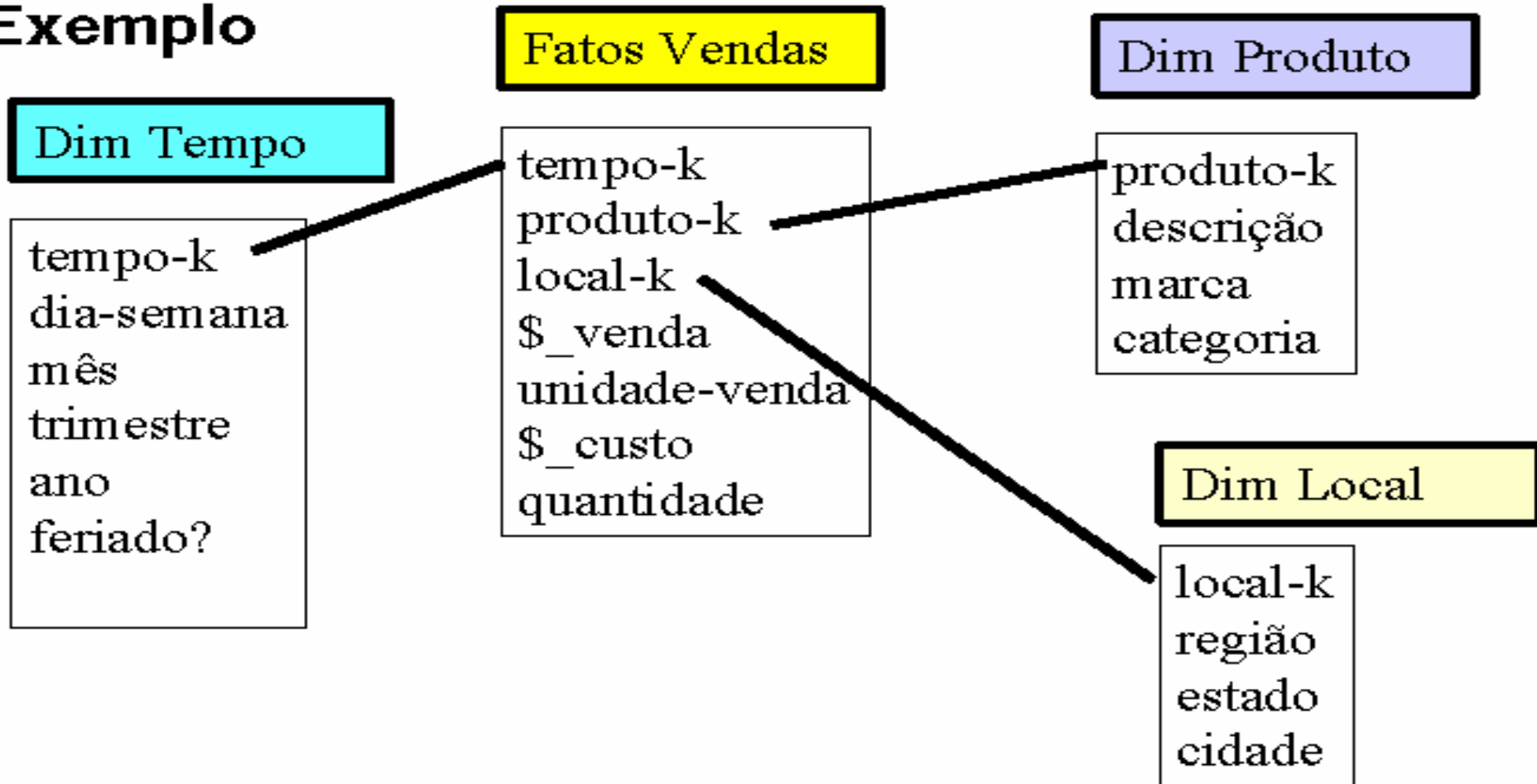
Uma metodologia específica para modelar dados, uma alternativa ao modelo ER, contém a mesma informação que o modelo ER, mas o pacote de dados está em um formato simétrico cujo objetivo é facilitar a consulta, melhorar a performance e flexível a mudanças.

Modelo Relacional



Dados Corporativos

Exemplo



Perguntar

selecionar dimensões e atributos

selecionar medidas

Ano Região Vendas

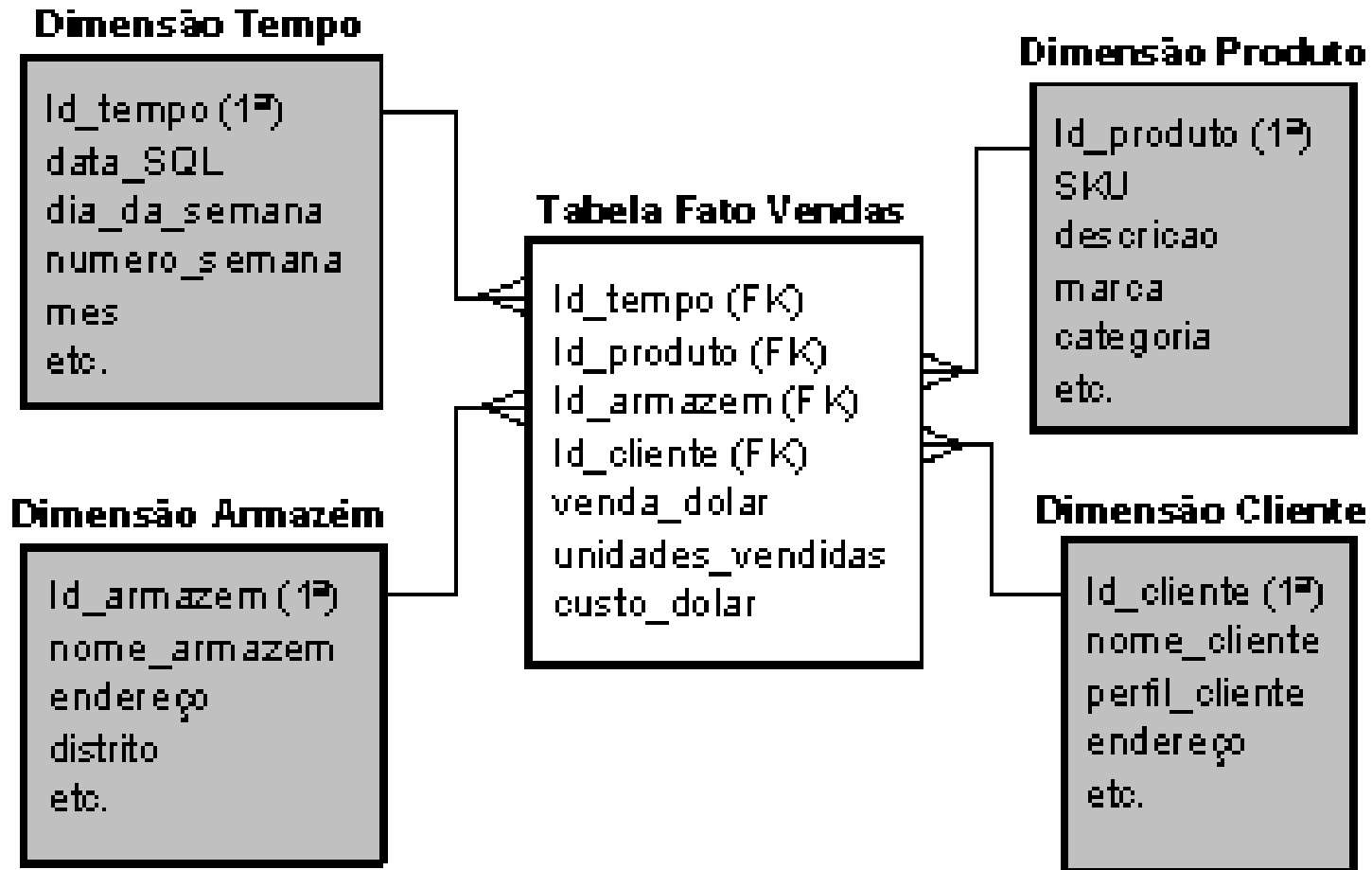
1997 Central 700

1997 Sul 1000

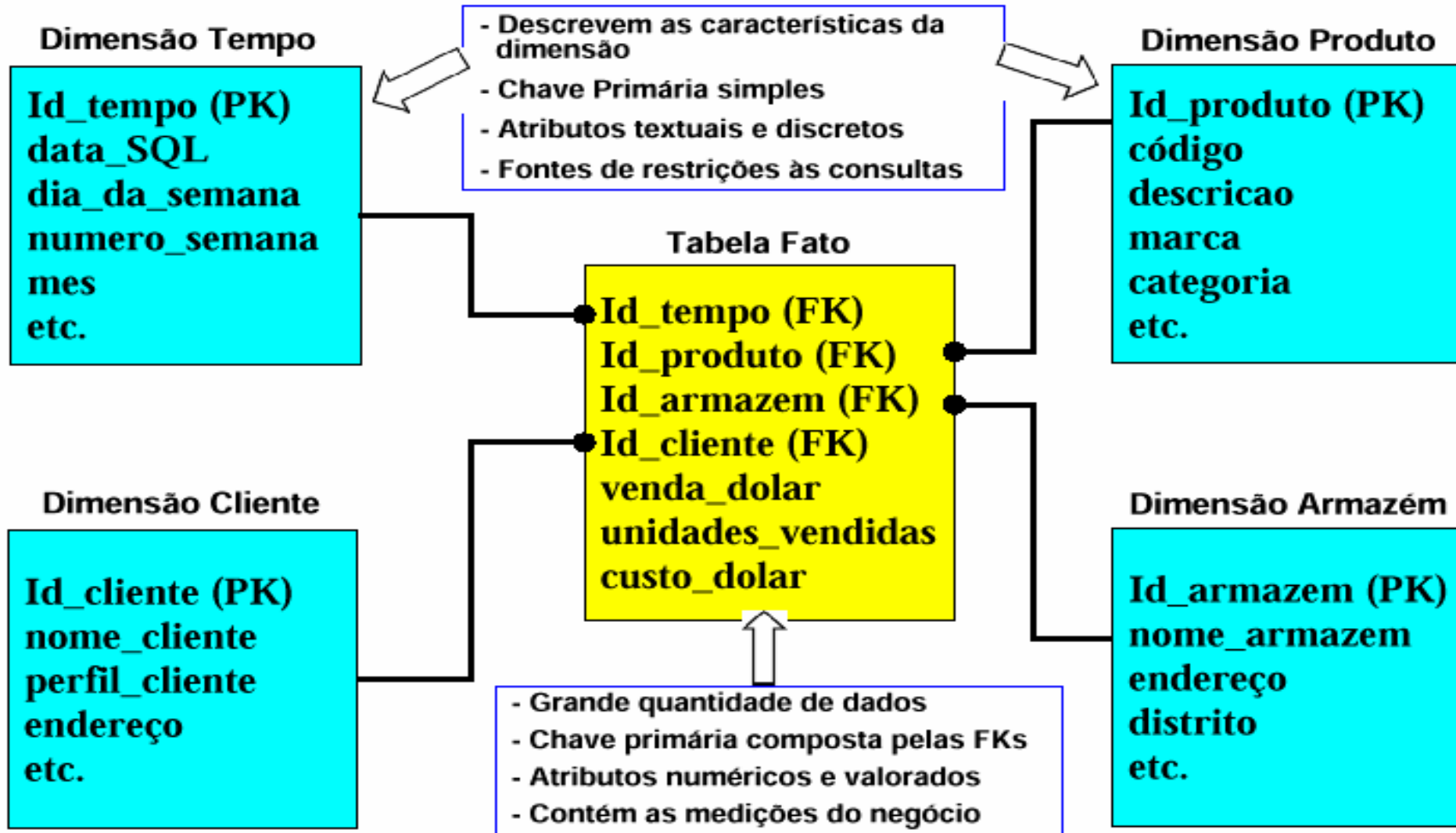
restringir valor de dimensões

```
select ano, região, sum(vendas)
from fatos, tempo, local
where junção de fatos com local e tempo
and ano = 1997
group by ano, regioao
```

Esquema Estrela



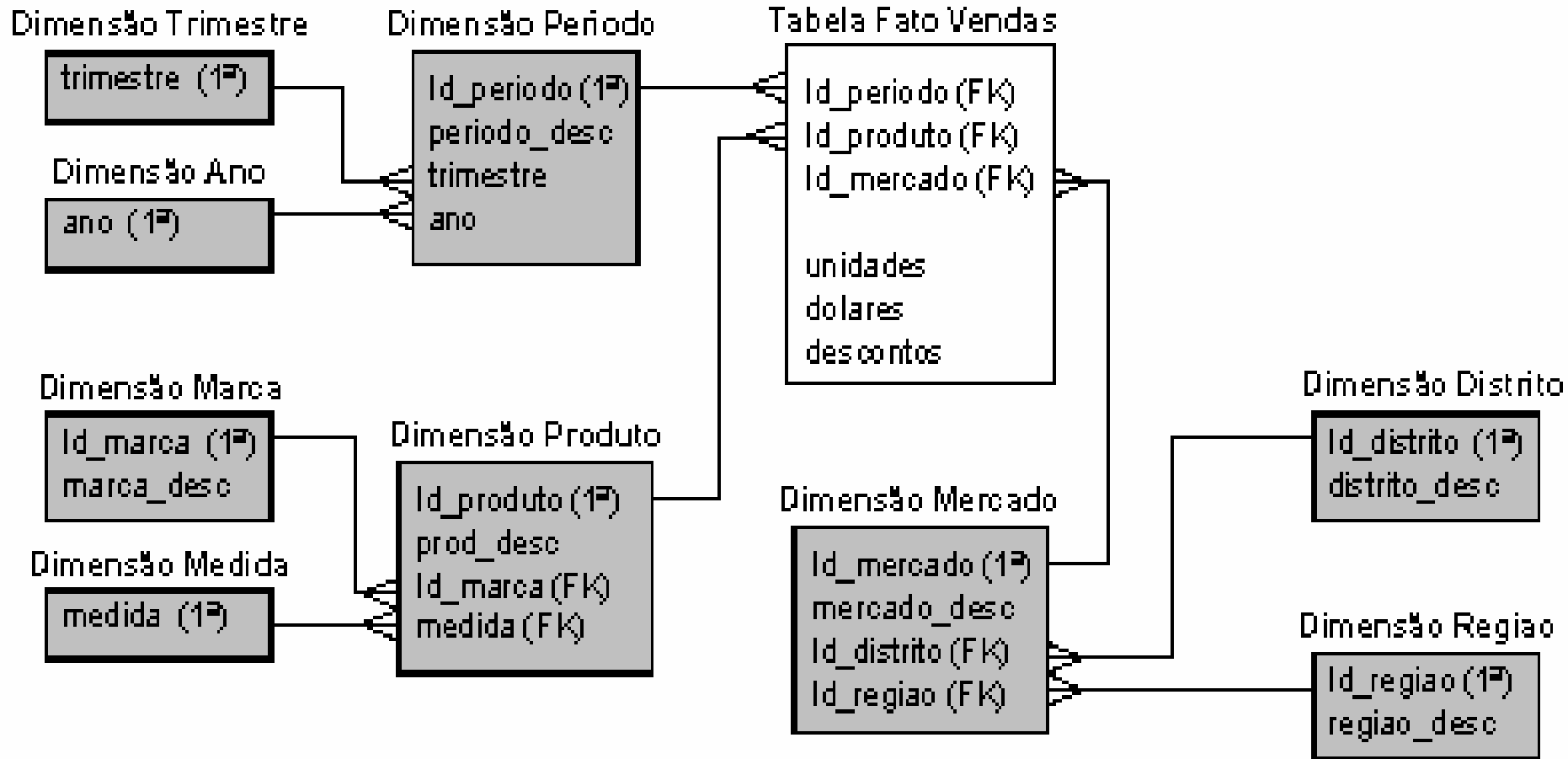
Modelagem Dimensional



Esquema Floco de Neve

- Desdobra-se as tabelas de dimensões removendo alguns campos para tabelas separadas conectando as mesmas com a tabela original através de chaves artificiais
- Geralmente não é recomendado num ambiente de DW
- *Snowflacking* - esquema onde aplica-se a normalização
- O excesso de chaves baixa a eficiência da consulta

Esquema Flocos de Neve



Comparar e Apresentar

Cálculos simples no conjunto de resultados

Diferença entre colunas



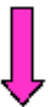
Ano	Mês	Soma (Vendas)	Soma (Custos)	Soma (Vendas - Custos)
1995	Jan	R\$ 79.913,71	R\$ 27.213,20	R\$ 52.700,50
	Feb	R\$ 67.437,00	R\$ 23.136,92	R\$ 44.300,08
	Mar	R\$ 76.582,96	R\$ 25.918,40	R\$ 50.664,56
	Apr	R\$ 63.099,22	R\$ 21.592,00	R\$ 41.507,22
	May	R\$ 65.549,38	R\$ 22.517,38	R\$ 43.032,00
	Jun	R\$ 62.315,26	R\$ 21.489,03	R\$ 40.826,23

Cálculo de porcentagens



Comparação de Vendas			
Ano	1995	1996	% Vendas
Jan	R\$ 79.913,71	R\$ 85.214,57	6,63%
Feb	R\$ 67.437,00	R\$ 69.892,45	3,64%
Mar	R\$ 76.582,96	R\$ 74.111,34	-3,23%

REGIÃO	CIDADE	QTDE VENDIDA
Central	Chicago	6655
	Cincinnati	30001
	Dallas	6702
	Louisville	27260
	Minneapolis	6736
	Nashville	6747
	St. Louis	15109



REGIÃO	CIDADE	QTDE VENDIDA
Central	Chicago	6655
	Dallas	6702
	Minneapolis	6736
	Nashville	6747
	St. Louis	15109
	Louisville	27260
	Cincinnati	30001

Entender

- **Slice and Dice**
 - **Consultas**
 - **Visualizações**
- **Mineração de Dados (Data Mining)**

Características:

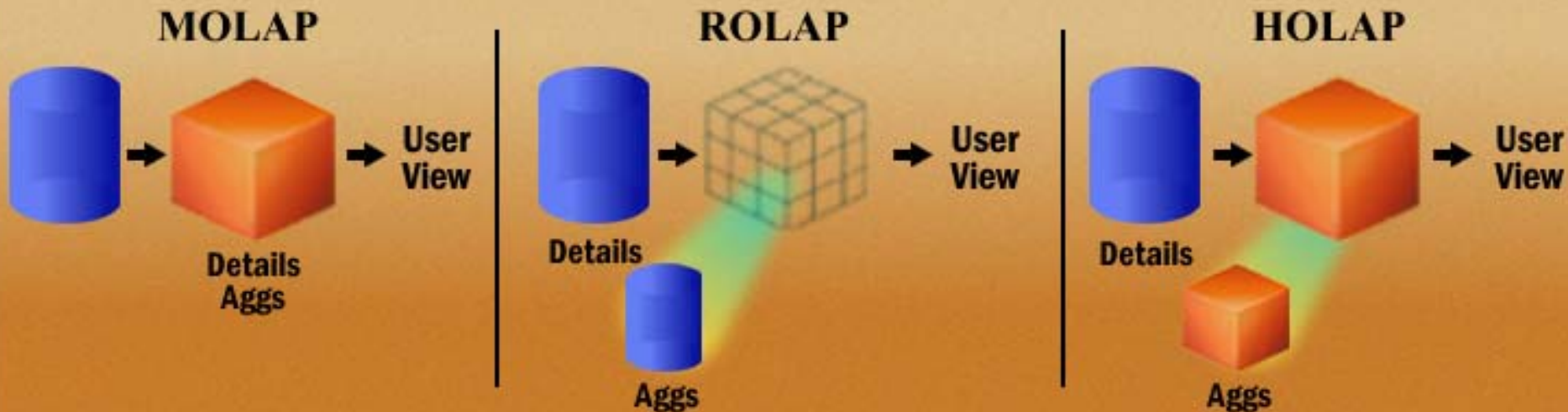
- Buscar padrões novos, úteis e compreensíveis em grandes volumes de dados;
- Padrão = estrutura de relacionamento entre atributos e seus valores;
- Dados detalhados;
- Auxiliar os decisores a ampliar seu espaço de investigação de hipóteses;
- Técnicas de mineração (classes de problemas);
- Tipo de análise mais complexa (analista de dados).

On-Line Analytical Processing (OLAP)

- designação genérica para as atividades de acesso e apresentação de dados provenientes de um DW
- baseado em representação multidimensional dos dados
- Tecnologias:
 - **MOLAP**
 - **ROLAP**
 - **HOLAP: MOLAP + ROLAP**
 - **DOLAP: Desktop OLAP**

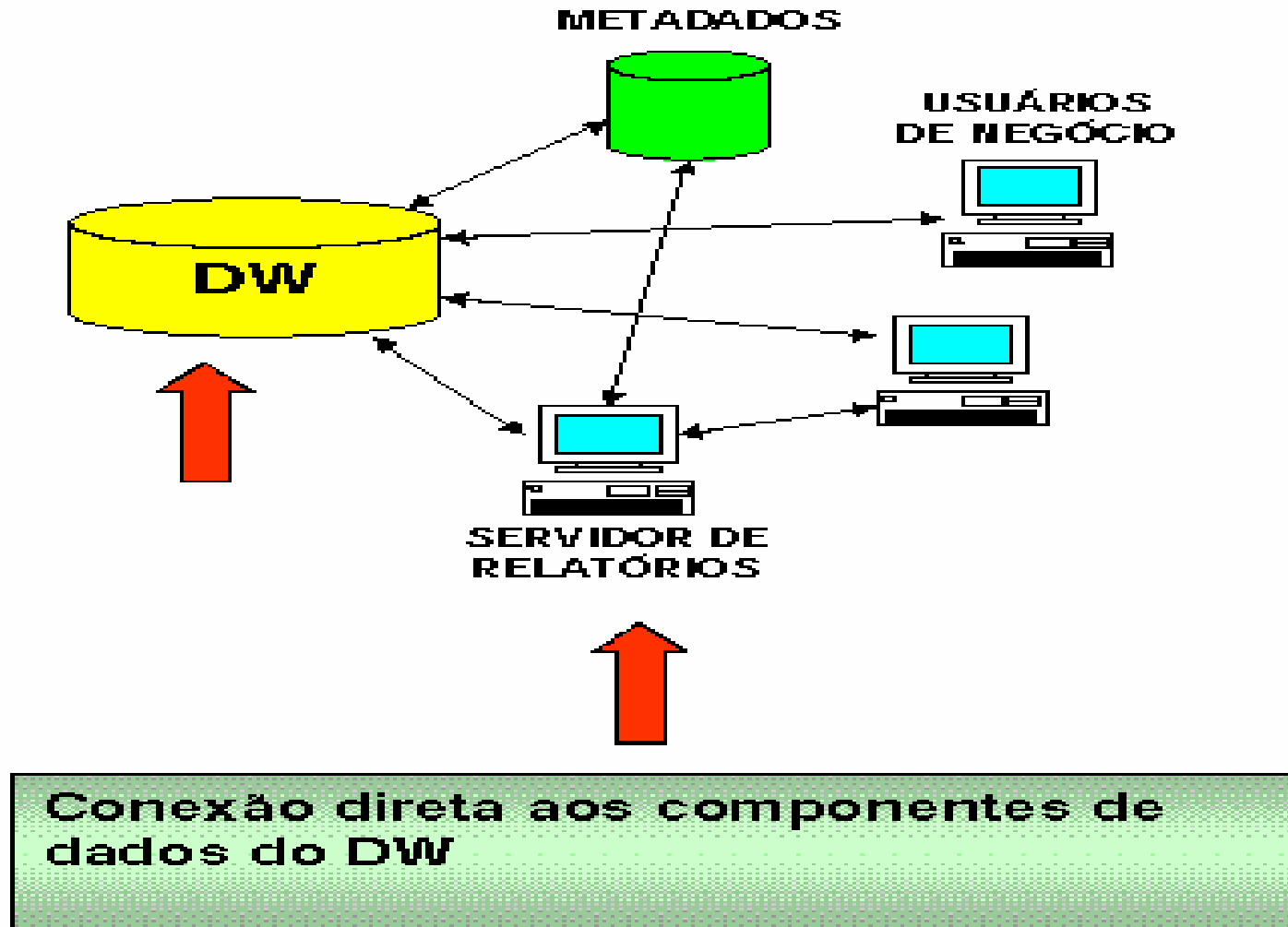
OLAP Services

- Hierarquias Múltiplas e não equilibradas
- Particionamento de dados
- Junção virtual de cubos
- Monitoração de utilização
- Membros calculados
- Múltiplas estratégias de armazenamento
MOLAP, ROLAP, HOLAP, DOLAP



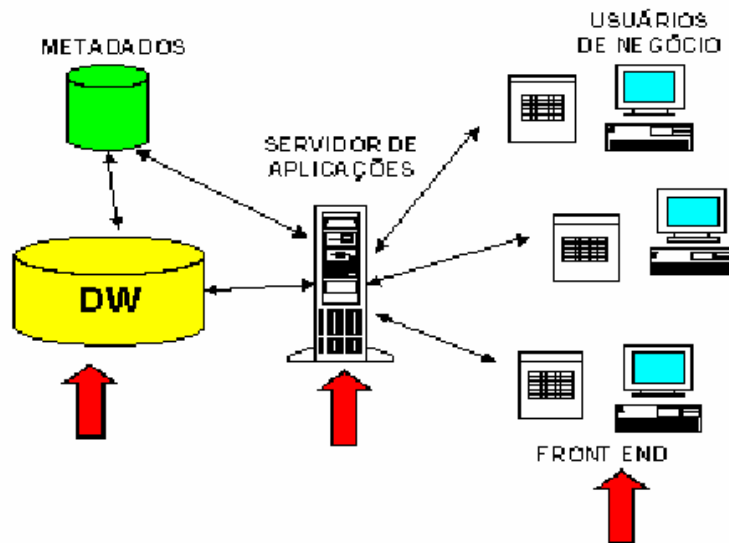
Recuperação e Exploração de Dados

Duas Camadas



Recuperação e Exploração de Dados

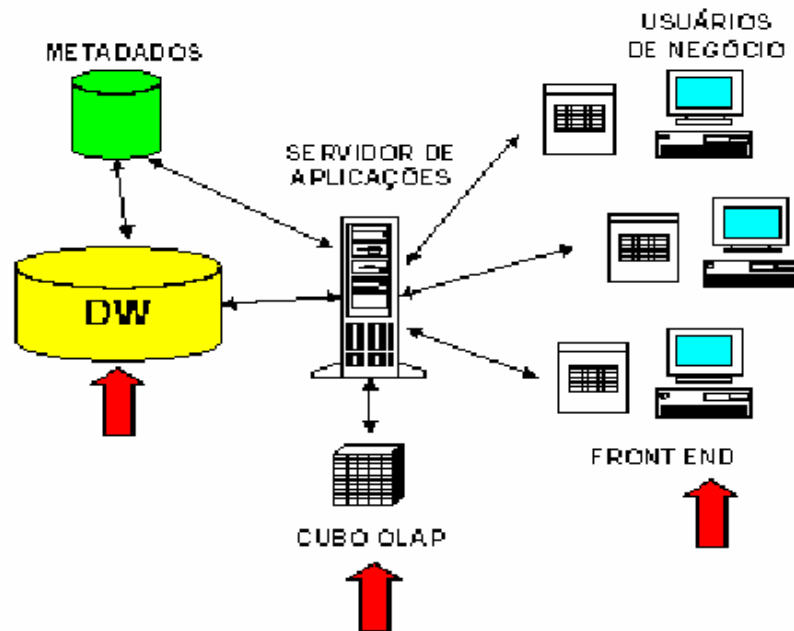
Três Camadas - ROLAP



- Separação das funções de gerenciamento de consultas das ferramentas de front-end e centraliza-as em um servidor de aplicações
- Apresenta o BD analítico para o cliente como um ambiente multidimensional
- Tornando-se comum no ambiente da WEB
- Uso intensivo de metadados, facilitando o gerenciamento de modificações em componentes e estruturas (tabelas fato, dimensões, hierarquias, agregados, etc.)

Recuperação e Exploração de Dados

Três Camadas - MOLAP



- Similar a arquitetura ROLAP
- A camada intermediária inclui um banco de dados de cubo multidimensional
- O cubo OLAP constitui-se de pré-agregados do BD do DW
- Consultas dos usuários são gerenciadas pelo Servidor de Aplicações que as envia inicialmente ao cubo OLAP e, caso não possa atendê-las, são destinadas ao BD do DW

Aplicações para o Usuário Final

Uma coleção de ferramentas que consulta, analisa e apresenta informações desejáveis para apoiar uma necessidade de negócio. São ferramentas para acesso aos dados, planilhas, pacotes gráficos e uma interface amigável.

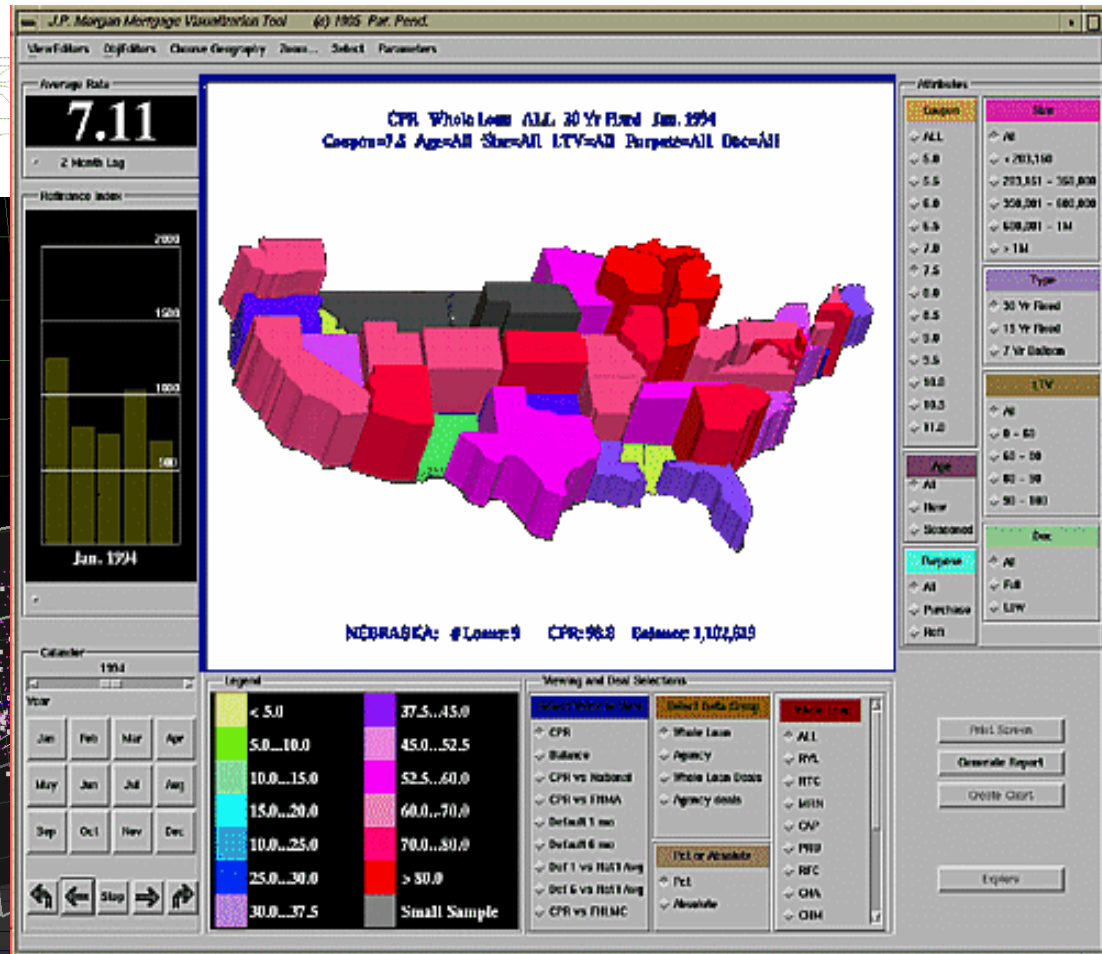
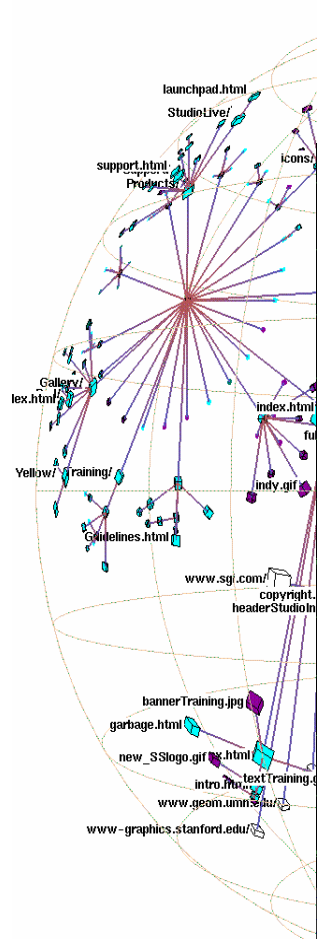


Arquitetura/Usuário

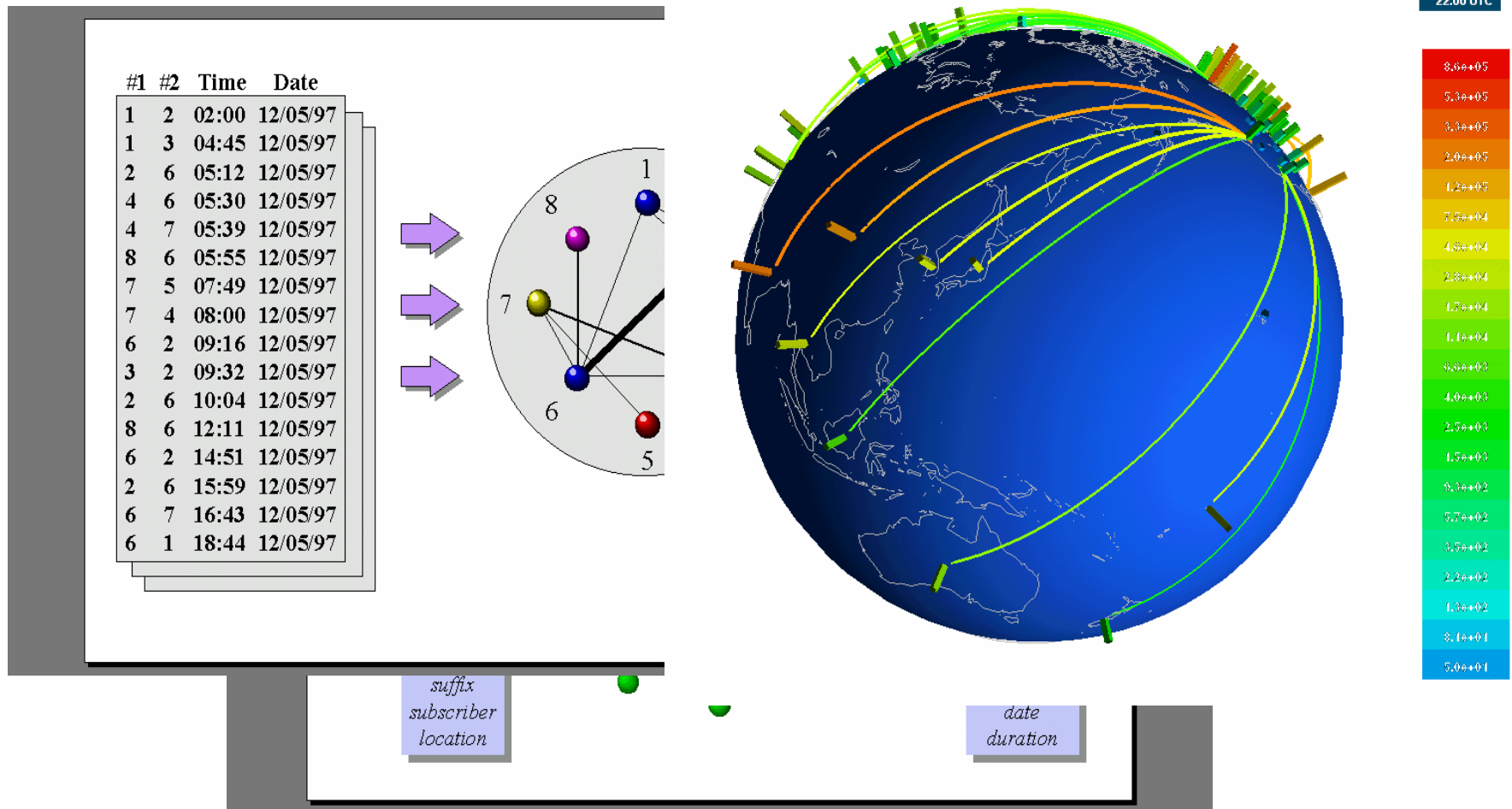


Arquitetura/Usuário

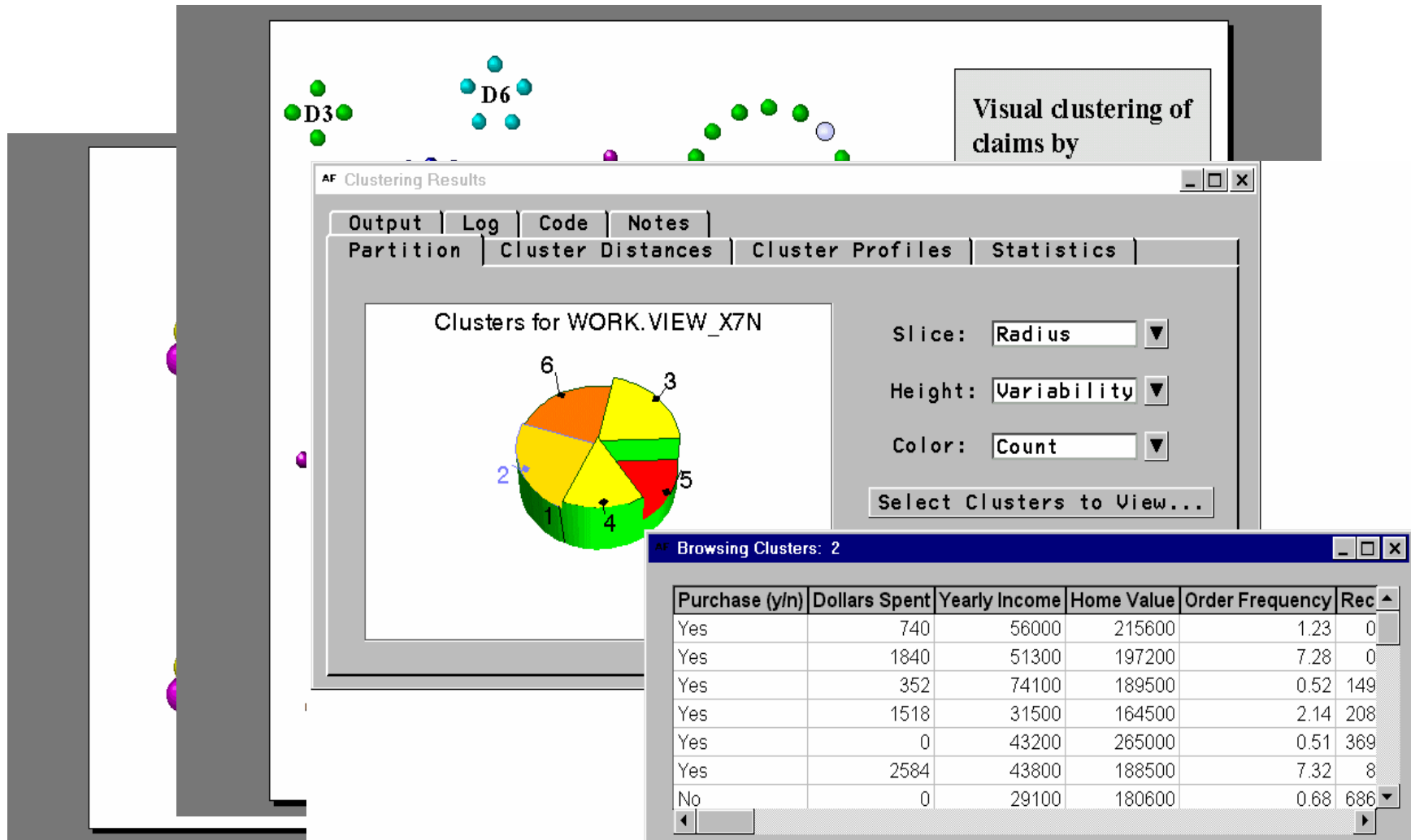
Ferramentas para interpretar um mundo complexo



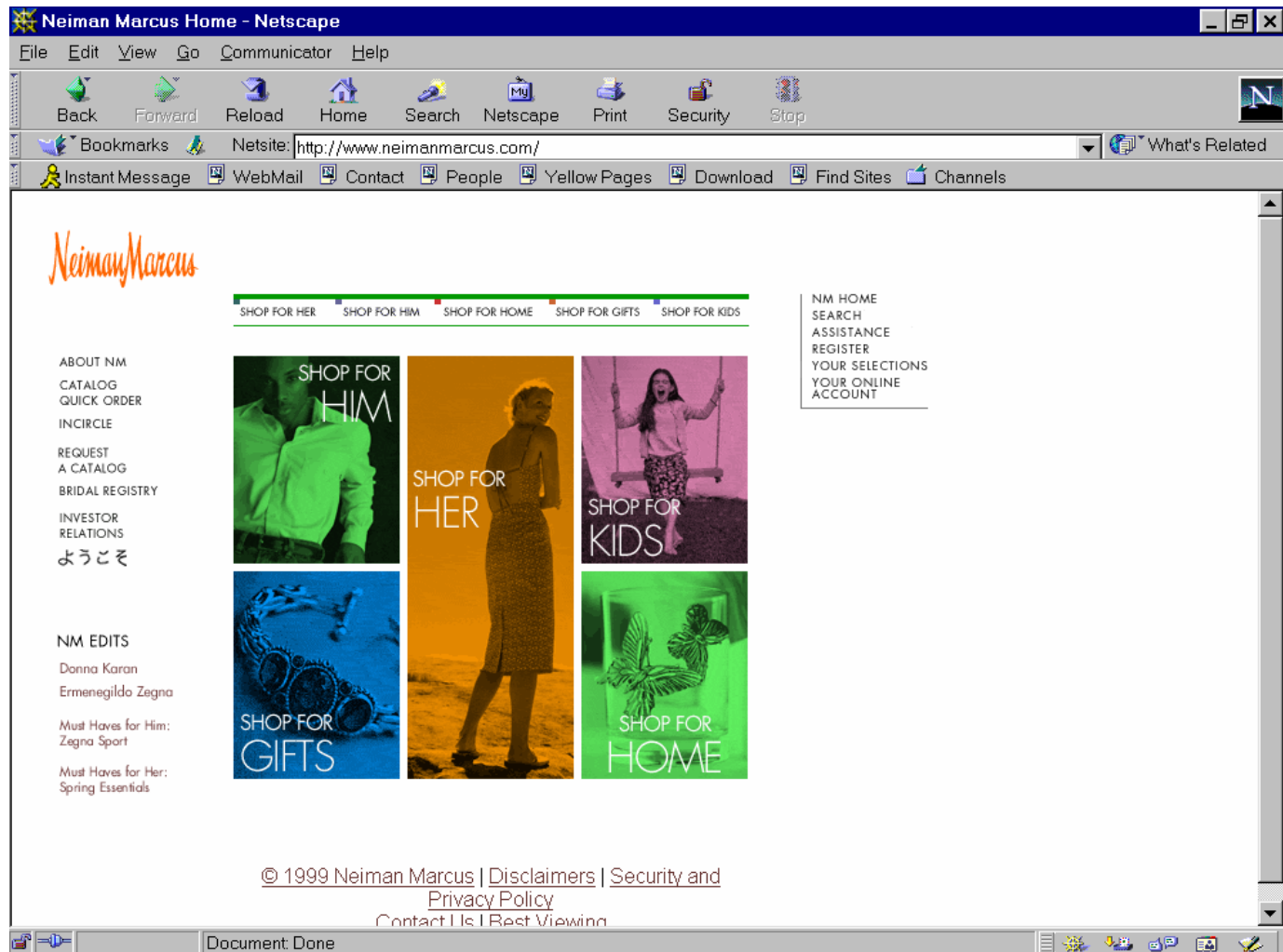
Ferramentas para interpretar um mundo complexo



Ferramentas para interpretar um mundo complexo



Mostrando produtos e serviços de maneira visual, interativa e com conteúdo rico



Conclusões

- Data Warehouse é uma base de dados voltada a apoio à decisão
- o processo de alimentação do DW é complexo
- ferramentas de acesso devem levar em conta tipo de usuário e funcionalidades desejadas
- produtos comerciais
 - reaproveitam muitas funcionalidades originalmente projetadas para apoio a criação e gestão de sistemas operacionais
 - inclusão de novas funcionalidades para processamento OLAP
 - mineração é na prática pouco usada em contextos de data warehouse

Algumas Tendências

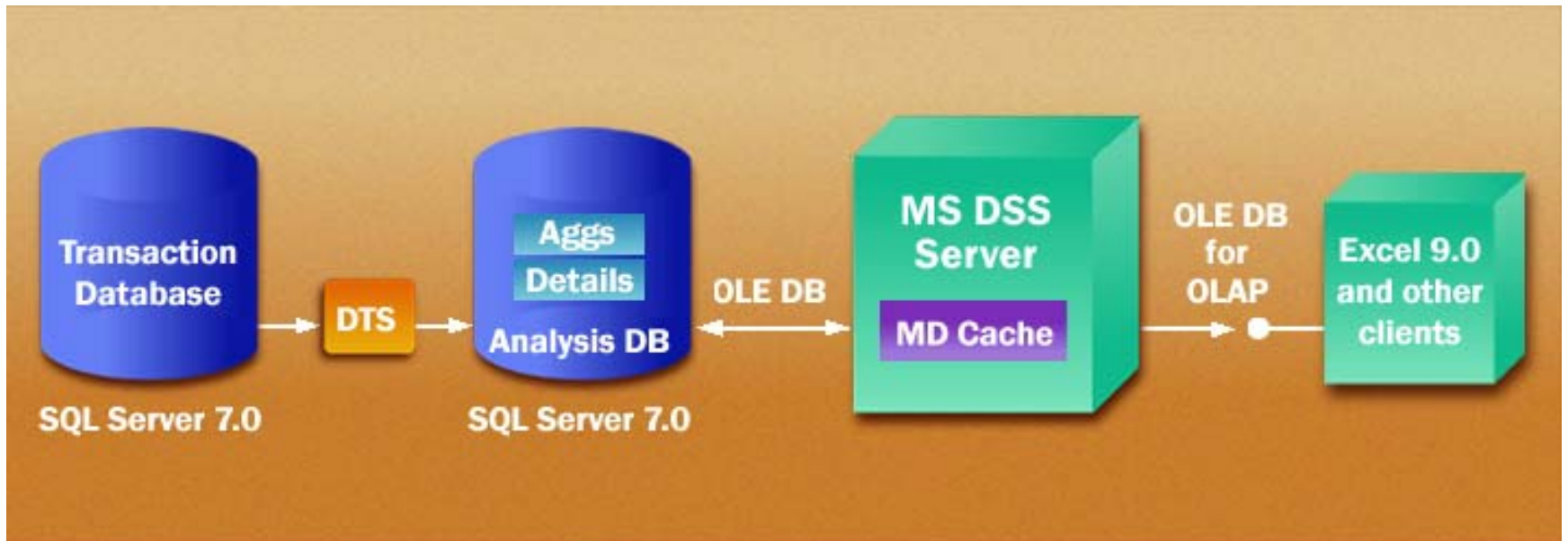
- metodologias de desenvolvimento
- apoio à manutenção
- materialização de versões
- metadados
- sistematização do processo de alimentação do DW e maior integração com os sistemas fonte
- mais recursos para usuário final, considerando seu perfil
- tecnologias para otimização de desempenho e armazenamento
- distribuição
- uso da Web

Investimento Softwares

- Ferramenta ETL
- Ferramenta BD
- Ferramenta OLAP
- Ferramentas Usuário Final

OLAP Services

- Servidor OLAP distribuído com o SQL Server



DATA WAREHOUSE

Professor MSc Ly Freitas Filho

Site: www.lyfreitas.com

E-mail: ly@lyfreitas.com