

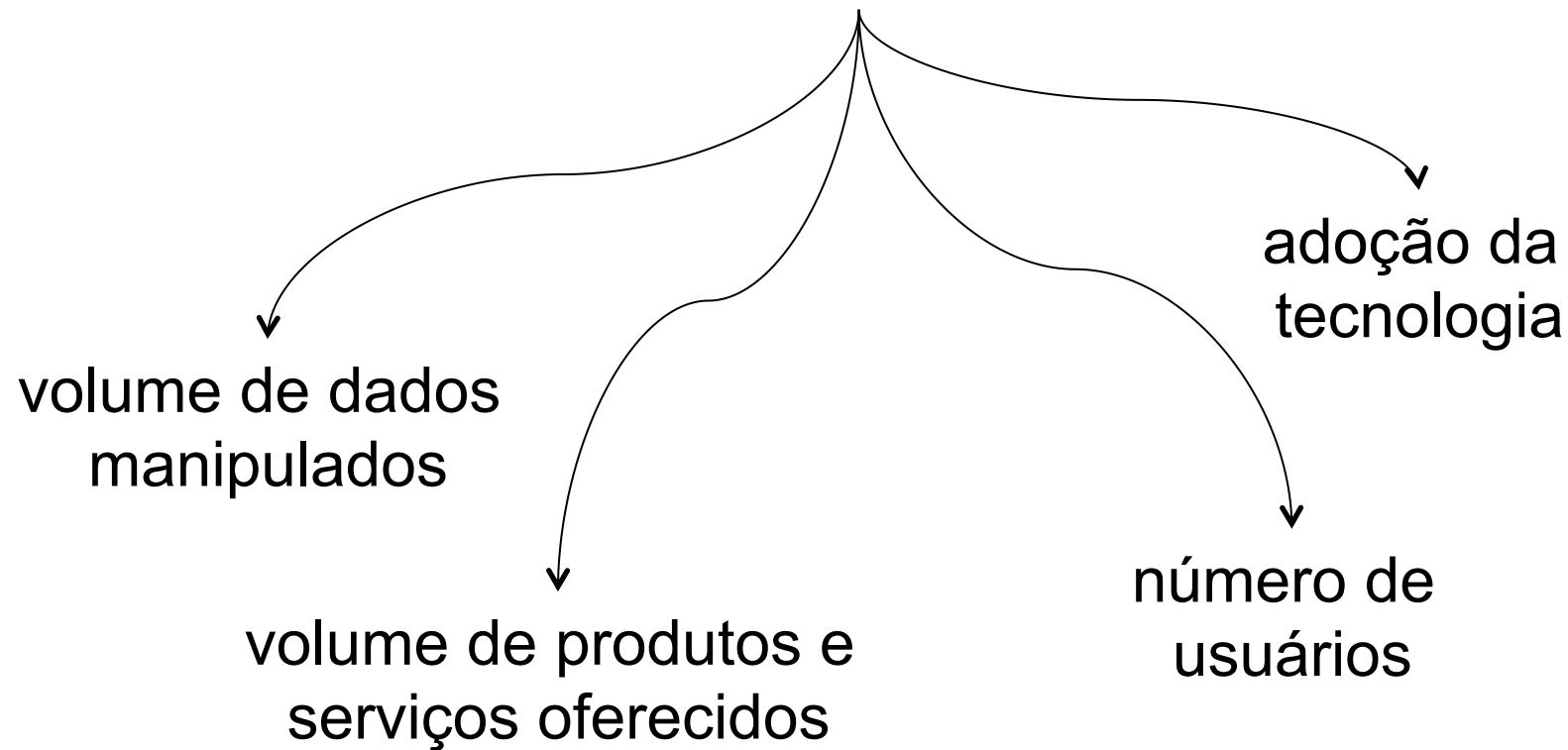
# *Data Warehousing: Conceitos Básicos e Arquitetura*

Profa. Dra. Cristina Dutra de Aguiar Ciferri

Prof. Dr. Ricardo Rodrigues Ciferri

# Visão do Mercado

Crescimento explosivo do uso da tecnologia de data warehousing



# Pensamento Motivacional

A obtenção de **informações estratégicas**, relativas ao contexto de **tomada de decisão**, é de suma importância para o sucesso de uma empresa. Tais informações permitem à empresa um **planejamento rápido** frente às mudanças nas condições do negócio, essencial na atual conjuntura de um mercado globalizado.

# Ambiente Operacional *versus* Ambiente Informacional

	<b>Ambiente Operacional</b>	<b>Ambiente Informacional</b>
<b>Principal Característica</b>	voltado ao processamento de transações OLTP	voltado ao processamento de consultas OLAP
<b>Tipos de Operação mais Frequentes</b>	atualização remoção inserção	leitura (consulta)

o termo OLAP (on-line analytical processing) foi introduzido em 1993 por Codd *et al.* para definir a categoria de processamento analítico sobre um banco de dados histórico voltado para os processos de gerência e tomada de decisão

# Ambiente Operacional *versus* Ambiente Informacional

	<b>Ambiente Operacional</b>	<b>Ambiente Informacional</b>
<b>Volume de Transações</b>	relativamente alto	relativamente baixo
<b>Características das Transações</b>	pequenas e simples, acessam poucos registros por vez	longas e complexas, acessam muitos registros por vez e realizam várias varreduras e junções de tabelas

# Ambiente Operacional *versus* Ambiente Informacional

	<b>Ambiente Operacional</b>	<b>Ambiente Informacional</b>
<b>Tipos de Usuários</b>	administradores do sistema, projetistas, usuários de entrada de dados	usuários de SSD por exemplo: executivos, analistas, gerentes, administradores
<b>Número de Usuários Concorrentes</b>	grande (geralmente milhares)	relativamente pequeno (geralmente centenas)
<b>Interações com os Usuários</b>	pré-determinadas estáticas	<i>ad-hoc</i> dinâmicas

# Ambiente Operacional *versus* Ambiente Informacional

	<b>Ambiente Operacional</b>	<b>Ambiente Informacional</b>
<b>Volume de Dados</b>	<i>megabytes a gigabytes</i>	<i>gigabytes a terabytes</i>
<b>Projeto do Banco de Dados</b>	normalizado para suporte às propriedades ACID	multidimensional, refletindo as necessidades de análise dos usuários de SSD
<b>Granularidade dos Dados</b>	detalhado	detalhado e agregado

# Ambiente Operacional *versus* Ambiente Informacional

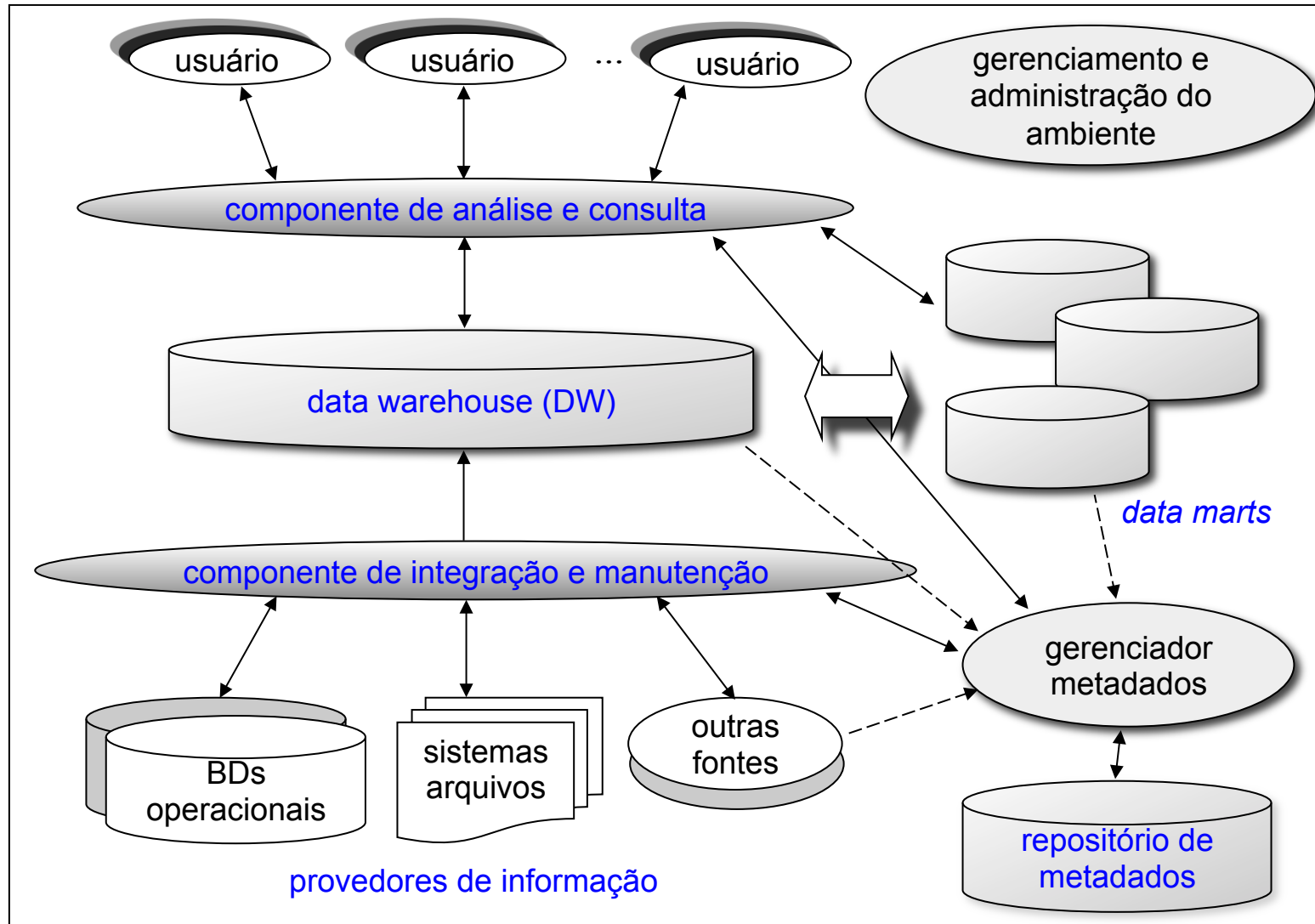
	<b>Ambiente Operacional</b>	<b>Ambiente Informacional</b>
<b>Principal Questão de Desempenho</b>	produtividade da transação	produtividade da consulta
<b>Tempo de Resposta</b>	geralmente poucos segundos	de minutos a horas
<b>Exemplos de aplicações</b>	transações bancárias, empréstimos de livros, contas a pagar	planejamento de <i>marketing</i> , análise financeira



# Data Warehousing

- Transforma dados operacionais em informação voltada à tomada de decisão estratégica
- Oferece funcionalidades que
  - possibilitam que dados de diferentes provedores de informação sejam extraídos, traduzidos, filtrados, integrados e armazenados no **data warehouse** (DW)
  - permite que usuários de sistemas de suporte à decisão manipulem com flexibilidade e eficiência os dados do DW

# Arquitetura Típica



# Componente: DW

- Coração do ambiente de data warehousing
- Banco de dados
  - voltado para o suporte aos processos de gerência e tomada de decisão
  - tem como principais objetivos prover eficiência e flexibilidade na obtenção de informações estratégicas e manter os dados sobre o negócio com alta qualidade

# Características dos Dados

- Orientados a assunto
  - relativos aos temas de negócio de maior interesse da corporação
  - *exemplos*: clientes, produtos, promoções, contas e vendas
- Integrados
  - dados obtidos dos provedores de informação corrigidos para eliminar possíveis inconsistências

# Características dos Dados

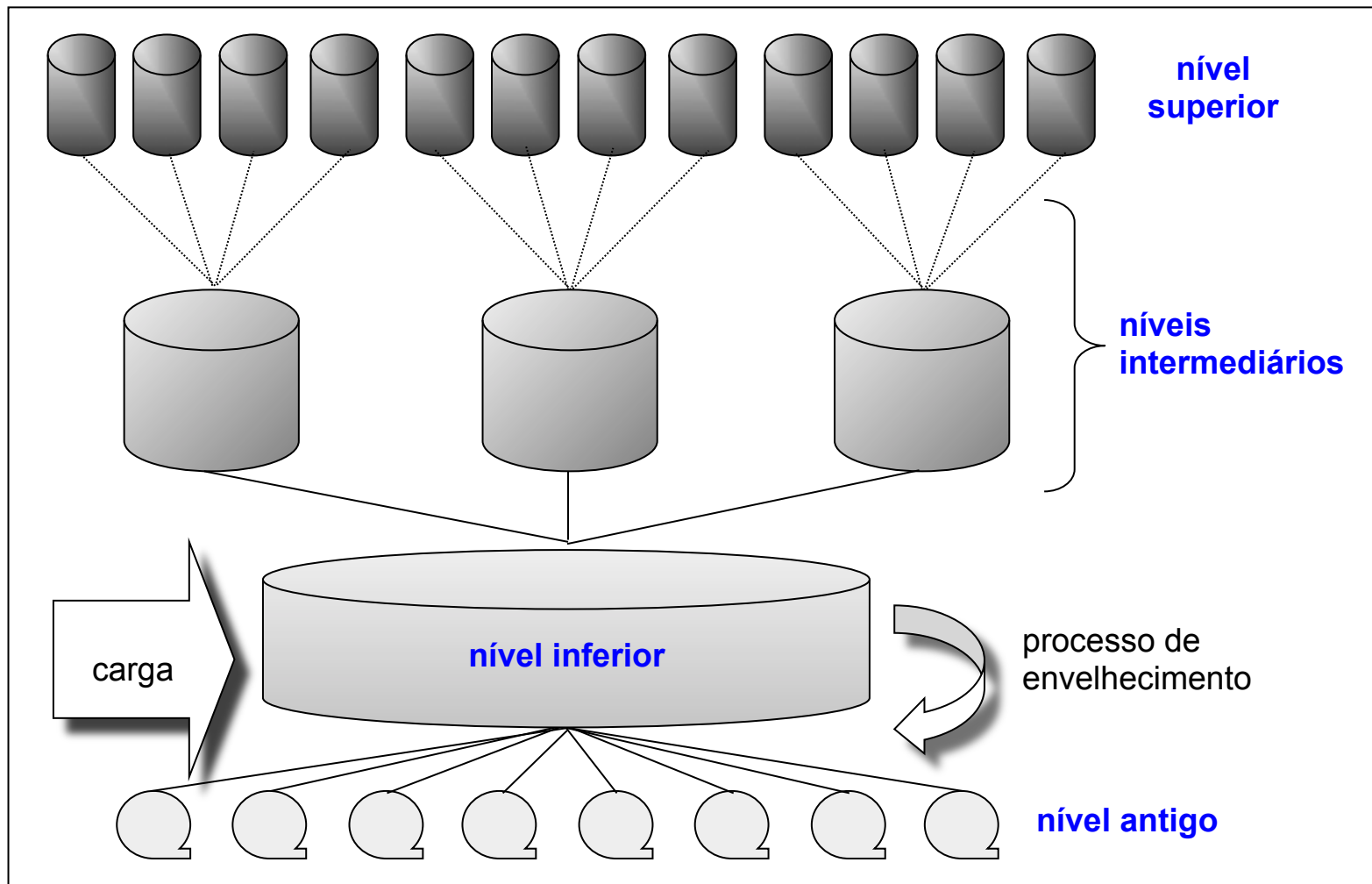
- Não-voláteis
  - o conteúdo do DW permanece estável por longos períodos de tempo
- Históricos
  - relevantes a algum período de tempo
  - *exemplo*: usualmente dados relativos a um grande espectro de tempo (5 a 10 anos) encontram-se disponíveis

# Características dos Dados

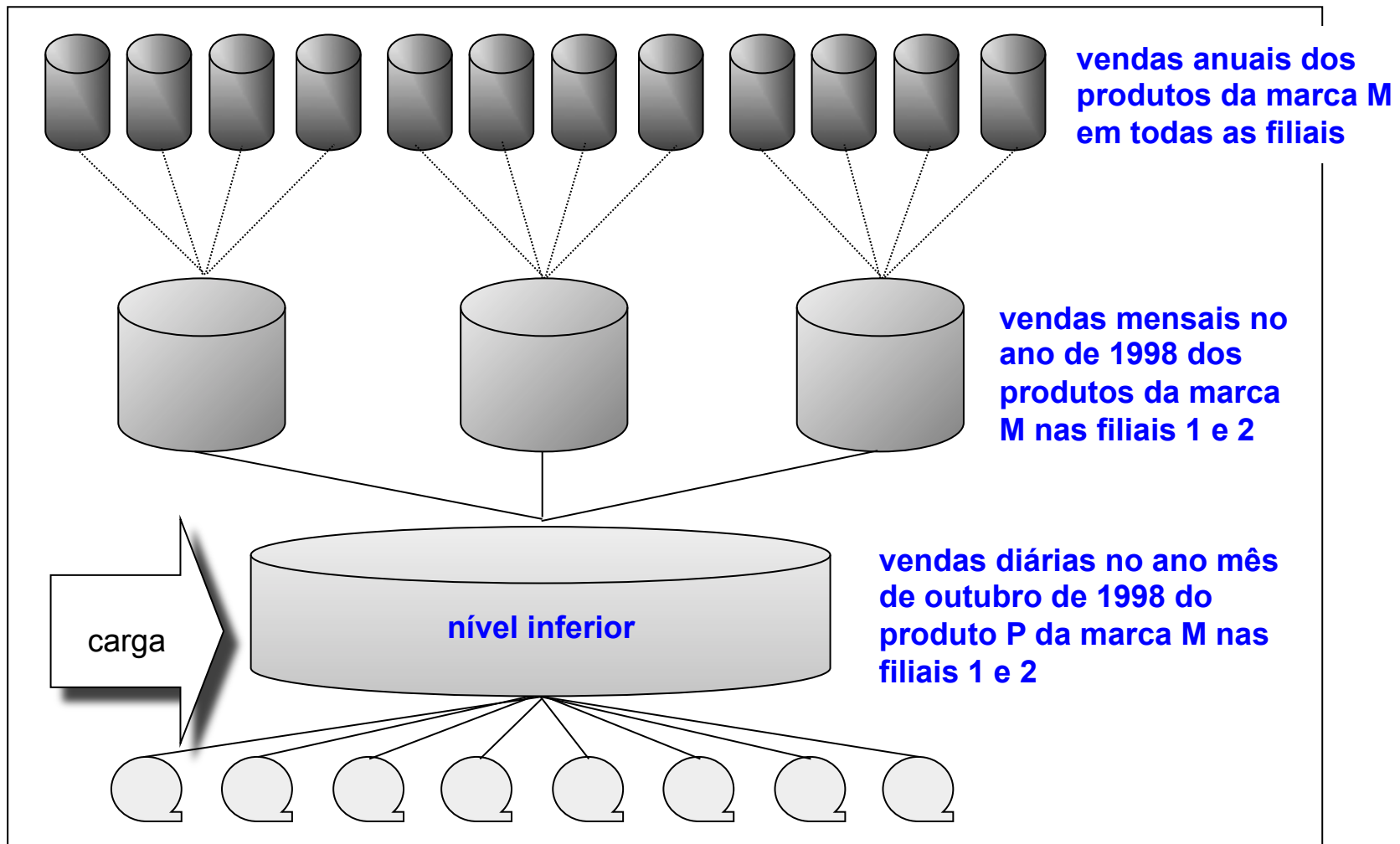
- Organizados em diferentes níveis de agregação
  - nível inferior: dados primitivos coletados do ambiente operacional
  - níveis intermediários: dados com graus de agregação crescente
  - nível superior: dados altamente resumidos (agregados)

devido ao volume de dados armazenados no DW, esses dados podem ser transferidos periodicamente para o nível antigo

# Níveis de Agregação



# Níveis de Agregação



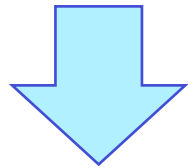


# Granularidade

- Grau de detalhamento em que os dados são armazenados em um nível
- Questão de projeto muito importante
  - impacta no volume de dados armazenado
  - afeta as consultas que podem ser respondidas

# Granularidade

grão muito pequeno

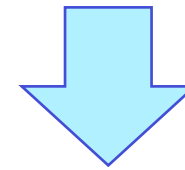


tamanho do data  
warehouse é muito  
grande

+

praticamente qualquer  
consulta pode ser  
respondida

grão muito grande



tamanho do data  
warehouse é menor

+

número de consultas que  
podem ser respondidas  
é menor

# Componente: Provedores de Informação

- Fontes de dados
  - autônomas
  - heterogêneas
  - distribuídas
- Contêm **dados operacionais**
- Exemplos
  - SGBD relacionais, objeto-relacionais, ...
  - documentos HTML, SGML, ...

# Componente de Integração e Manutenção

- Carregamento dos dados
  - atividade mais complexa, cara e demorada
  - essencial ao bom funcionamento do ambiente de data warehousing
  - processos
    - extração
    - tradução
    - limpeza
    - integração
    - armazenamento
    - recuperação de falhas

fluxo de informação: provedores de informação → DW

# Carregamento dos Dados

- Extração
  - **quais** dados são extraídos de quais provedores
  - **como** esses dados são extraídos
- Tradução
  - **conversão** dos dados do formato nativo dos provedores de informação para o formato utilizado pelo ambiente de data warehousing
  - manutenção da **temporalidade** dos dados

# Carregamento dos Dados

- Limpeza
  - garante a **corretude** e a **qualidade** dos dados, de forma que esses dados atendam às restrições de integridade impostas pelas regras de negócio
- Integração
  - geração de um **dado único** a partir de várias cópias do mesmo dado extraídas de diferentes provedores

# Integração dos Dados

- Problema: dados armazenados nos provedores
  - são heterogêneos
    - seguem diferentes modelos de dados
    - são representados por conceitos diferentes
    - possuem diferentes formatos
    - etc
  - são redundantes, inconsistentes e até mesmo complementares
- Dois níveis: **esquema** e **instância**

# Carregamento dos Dados

- Armazenamento
  - realização de **processamentos adicionais**, como verificação de restrições de integridade, geração de agregações, construção de índices, etc
- Recuperação de Falhas
  - **evita** que tanto leituras desnecessárias aos dados dos provedores de informação quanto computações cujos resultados já foram armazenados no DW **sejam realizadas novamente**



# Componente de Integração e Manutenção

- Atualização dos dados
  - periodicidade
    - necessidades dos usuários de SSD
    - nível de consistência desejado
  - manutenção dos dados
    - **recomputação**: conteúdo do DW é descartado e os dados são carregados novamente a partir dos provedores de informação operacionais
    - **atualização incremental**: apenas as alterações nos dados dos provedores são refletidas no DW

# Componente de Integração e Manutenção

- Expiração dos dados
  - remoção de dados do DW visando diminuir o volume de dados armazenado
  - pode ocorrer quando
    - dados atingem o limite de tempo no qual tornam-se inválidos
    - dados não são mais relevantes ou necessários ao ambiente de data warehousing
    - espaço de armazenamento é insuficiente

# Componente de Análise e Consulta

- Permite a interação do usuário com o ambiente de data warehousing por meio de **ferramentas** dedicadas à análise e consulta dos dados
- Ferramentas
  - oferecem facilidades de navegação e de visualização
  - possuem diferentes classificações, com base nas funcionalidades oferecidas

# Exemplos de Ferramentas

- Geradores de relatório
  - tipos mais simples de ferramentas
  - permitem a produção de relatórios periódicos
- OLAP
  - oferecem capacidades analíticas sofisticadas, permitindo que os dados sejam analisados usando visões multidimensionais complexas e elaboradas
  - oferecem navegação facilitada nessas visões

# Exemplos de Ferramentas

- De mineração de dados
  - permitem que informações, padrões e tendências de negócio “escondidas” nos dados sejam descobertas

**IMPORTANTE:** Independentemente da ferramenta utilizada, um fator primordial refere-se à **visualização dos resultados obtidos**. Técnicas de visualização dos dados devem determinar a melhor forma de se exibir relacionamentos e padrões complexos em um monitor bidimensional, de modo que o problema inteiro e/ou a solução sejam claramente visíveis usuários de SSD

# Componente: Data Mart

- DW que possui escopo limitado
- Armazena dados que compartilham as mesmas características dos dados do DW
- Enfoques
  - subconjunto dos dados do DW
  - política no projeto de construção de um DW corporativo

# Componente: Repositório de Metadados

- Dados de nível mais alto que descrevem dados de nível mais baixo
- Características
  - permite que os usuários de SSD conheçam a **estrutura** e o **significado** dos dados
  - representa o principal recurso para a administração dos dados no ambiente de data warehousing

# Exemplos de Metadados

<p><b>Metadados Administrativos</b></p>	<p>contêm informações relacionadas à construção e à utilização do data warehousing, tais como os esquemas dos provedores de informação e do DW, além dos mapeamentos existentes entre os diversos esquemas; regras de extração, de tradução, de limpeza e de atualização dos dados, em adição às regras de mapeamento utilizadas para a solução de problemas de heterogeneidade existentes entre os dados dos diversos provedores de informação que participam do ambiente; especificações sobre grupos de usuários e privilégios a eles associados, incluindo políticas de controle de acesso, autorização e perfis; ferramentas de integração e manutenção, e regras associadas aos processos envolvidos; ferramentas de análise e consulta; consultas, agregações e relatórios pré-definidos</p>
---	---



# Exemplos de Metadados

<b>Metadados Específicos da Aplicação</b>	incluem um conjunto de terminologias específicas ao domínio da aplicação, além de restrições da aplicação e outras políticas
<b>Metadados de Auditoria</b>	mantêm informações relacionadas à linhagem dos dados, à geração de relatórios de erros, às ferramentas de auditoria empregadas e às estatísticas de utilização do ambiente de data warehousing, incluindo dados sobre a frequência das consultas, os custos para se processar uma determinada consulta, o tipo de acesso aos dados e o desempenho do sistema

classificação baseada em Wu, M.-C., Buchmann, A.P. Research Issues in Data Warehousing. In *Proceedings of The German Database Conference*, pages 61-82, Ulm, Germany, March 1997.