

Sistemas de Banco de Dados Projeto, implementação e gerenciamento

Capítulo 13

Business intelligence e data warehouses

Objetivos

- **Neste capítulo, você aprenderá:**
 - Como a business intelligence fornece um amplo modelo de suporte à tomada de decisões de negócios
 - Como os dados operacionais diferem dos dados de suporte a decisões
 - O que é data warehouse, como preparar os dados para ele e como implementá-los

Objetivos (cont.)

- O que são esquemas estrela e como são construídos
- O que é mineração de dados e que papel executa no suporte a decisões
- Sobre o processamento analítico on-line (OLAP)
- Como extensões de SQL são utilizadas para dar suporte a manipulações de dados do tipo OLAP

Necessidade da Análise de Dados

- A maioria dos gerentes deseja rastrear as transações diárias para avaliar o fluxo dos trabalhos
- Recorrendo ao banco de dados operacional, a gerência pode desenvolver estratégias que atendam às metas organizacionais
- A análise de dados pode fornecer informações sobre estratégias e avaliações táticas de curto prazo

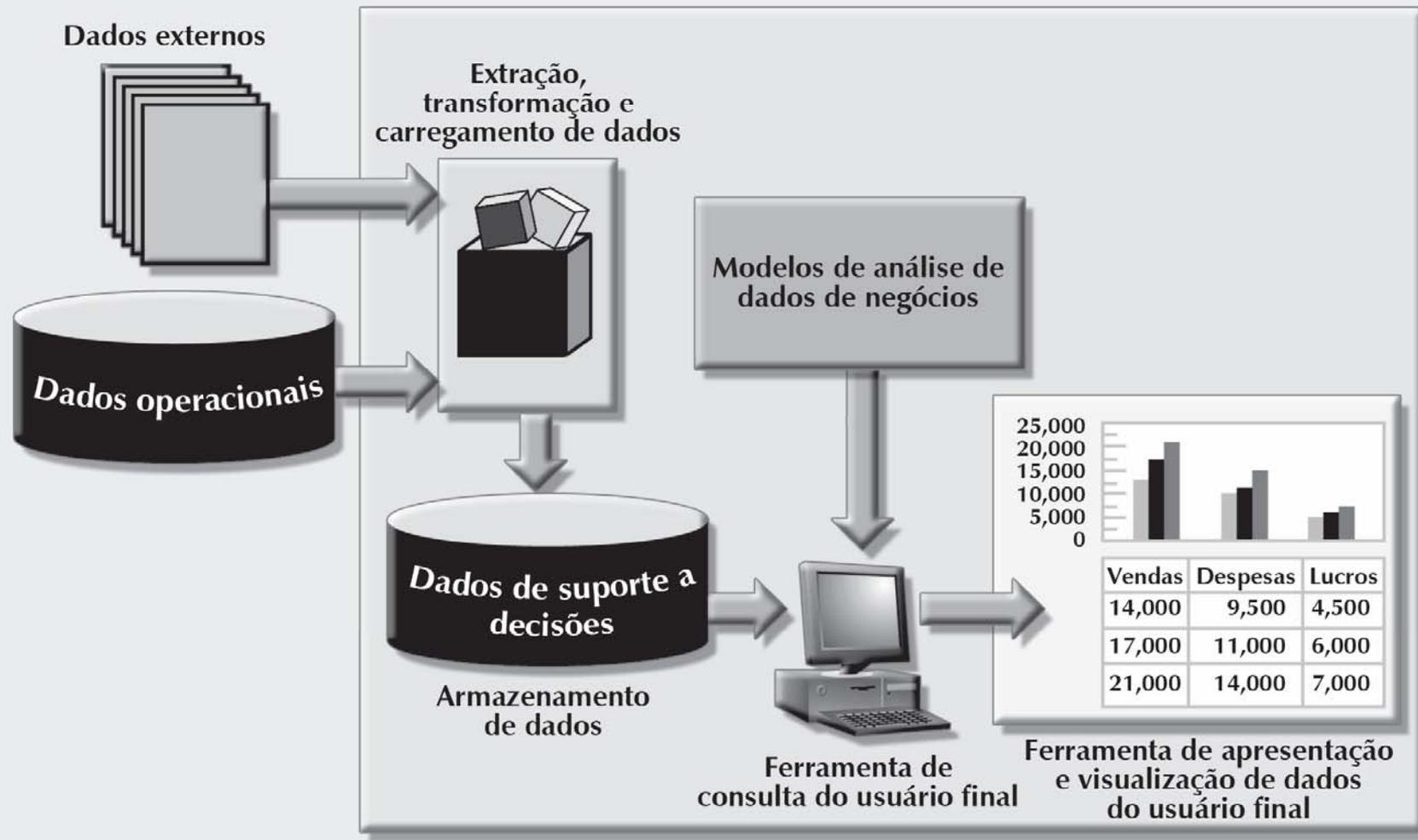
Business Intelligence

- Conjunto amplo, coeso e integrado de ferramentas e processos
 - Captar, coletar, integrar, armazenar e analisar dados
 - Geração e a apresentação de informações que deem suporte à tomada de decisões de negócios
- Modelo que permite à empresa transformar:
 - Dados em informações
 - Informações em conhecimento
 - Conhecimento em sabedoria

Arquitetura de Business Intelligence

- É composta de dados, pessoas, processos, tecnologia e gerenciamento desses componentes
- O foco principal de BI é coletar, integrar e armazenar dados de negócios para fins de criação de informações
- **Indicadores chaves de desempenho (ICD)**
 - Medidas quantificáveis que avaliam a eficiência ou o sucesso da empresa em alcançar suas metas
- É possível integrar várias ferramentas de diferentes fornecedores em um único modelo de BI

FIGURA 13.2 Componentes de business intelligence

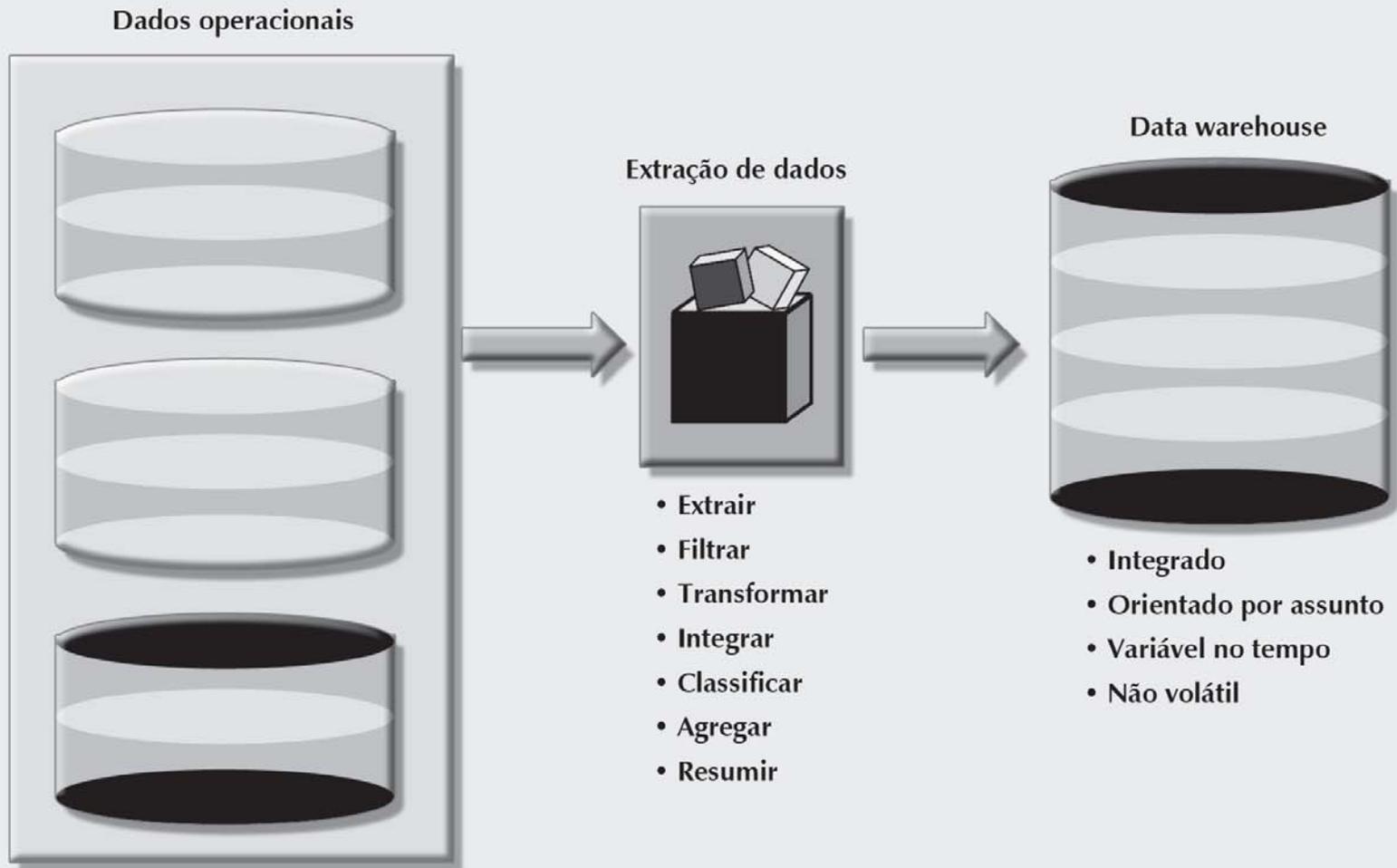


Dados de Suporte a Decisões

- Dados operacionais
 - A maioria é armazenada em bancos relacionais
 - O armazenamento é otimizado para dar suporte a transações que representam as operações diárias
- Os dados de suporte a decisões diferem dos operacionais em três áreas principais:
 - Período de tempo
 - Granularidade (nível de agregação)
 - Dimensionalidade

**FIGURA
13.4**

Criação de um data warehouse



Necessidades de Banco de Dados de Suporte a Decisões

- Um banco de dados de suporte a decisões é um SGBD especializado em fornecer respostas rápidas a consultas complexas
- Quatro necessidades principais:
 - Esquema de bancos de dados
 - Extração e carregamento de dados
 - Interface analítica do usuário final
 - Tamanho do banco de dados

Necessidades de Banco de Dados de Suporte a Decisões (cont.)

- Esquema de banco de dados
 - Representação de dados complexos
 - Dados agregados e resumidos
 - Consultas extraídas em períodos de tempo multidimensionais
- Extração e filtragem de dados
 - Deve dar suporte a diferentes fontes de dados:
 - Banco de dados de arquivo
 - Hierárquicos, em rede e relacional
 - De vários fornecedores
 - Deve verificar dados inconsistentes

Necessidades de Banco de Dados de Suporte a Decisões (cont.)

- Interface analítica do usuário final
 - É um dos componentes mais importantes do SGBD
 - Permite que o usuário navegue pelos dados, simplificando e acelerando o processo de tomada de decisões
- Tamanho do banco de dados
 - Em 2005, o Walmart, a maior empresa do mundo, possuía 260 terabytes de dados em seus data warehouses
 - O SGBD deve ser capaz de dar suporte a **bancos de dados muito grandes (VLDBs – *very large databases*)**

Data Warehouse

- Integrado, orientado por assunto, variável no tempo e não volátil
 - Fornece suporte à tomada de decisões
- Normalmente é um banco de dados apenas de leitura, otimizado para processamento de análises e consultas
- Sua criação exige tempo, dinheiro e considerável esforço gerencial

Data Warehouse (cont.)

- Pequeno subconjunto de um data warehouse, sobre um único assunto
- Pode ser criado a partir de dados extraídos de um data warehouse maior com a finalidade específica de dar suporte a um acesso mais rápido por determinado grupo ou função
- Tem custo mais baixo e leva menos tempo para ser implementado

Doze Regras que Definem um Data Warehouse

- Os ambientes operacional e de data warehouse são separados
- Os dados em data warehouse são integrados.
- O data warehouse contém dados históricos por um longo tempo
- Os dados em data warehouse constituem um retrato instantâneo tirado em determinado ponto do tempo
- Os dados em data warehouse são orientados por assunto

Doze Regras que Definem um Data Warehouse (cont.)

- Os dados em data warehouse são essencialmente apenas para leitura
 - Atualizações periódicas em batch dos dados operacionais
 - Não são permitidas atualizações on-line
- O ciclo de vida do desenvolvimento do data warehouse difere do desenvolvimento dos sistemas tradicionais
- O data warehouse contém dados com vários níveis de detalhe:
 - Dados atuais em detalhes, dados antigos em detalhes, dados levemente resumidos e dados altamente resumidos

Doze Regras que Definem um Data Warehouse (cont.)

- Transações de apenas leitura para conjuntos de dados muito grandes
- O ambiente de data warehouse possui um sistema que rastreia fontes, transformações e armazenamento
- Os metadados de data warehouse são um componente fundamental desse ambiente
- O data warehouse contém um mecanismo de retorno da utilização de recursos que leva à aplicação ideal dos dados pelos usuários finais

Estilos de Arquiteturas de Suporte a Decisões

- Fornecem recursos avançados
- Algumas são capazes de dar acesso à análise de dados multidimensionais
- Sua arquitetura completa inclui:
 - Capacidade de armazenamento de dados de suporte a decisões
 - Filtro de extração e integração de dados
 - Interface de apresentação especializada

Processamento Analítico On-line

- Ambiente avançado de análise de dados que dá suporte:
 - Tomada de decisões
 - Modelagem comercial
 - Pesquisa operacional
- Quatro características principais:
 - Utilizam técnicas de análise de dados multidimensionais
 - Proporcionam suporte avançado a bancos de dados
 - Fornecem interface fácil de utilizar para os usuários finais
 - Dão suporte a arquitetura cliente/servidor

Técnicas de Análise de Dados Multidimensionais

- Os dados são processados e visualizados como parte de uma estrutura multidimensional
- As técnicas de análise de dados multidimensionais são ampliadas pelas seguintes funções:
 - Funções avançadas de apresentação de dados
 - Funções avançadas de agregação, consolidação e classificação de dados
 - Funções computacionais avançadas
 - Funções avançadas de modelagem de dados

FIGURA 13.5

Visualização operacional versus visualização multidimensional de vendas

Nome da tabela: DW_INVOICE

INV_NUM	INV_DATE	CUS_NAME	INV_TOTAL
2034	15-May-08	Dartnik	1400.00
2035	15-May-08	Summer Lake	1200.00
2036	16-May-08	Dartnik	1350.00
2037	16-May-08	Summer lake	3100.00
2038	16-May-08	Trydon	400.00

Nome da tabela: DW_LINE

INV_NUM	LINE_NUM	PROD_DESCRIPTION	LINE_PRICE	LINE_QUANTITY	LINE_AMOUNT
2034	1	Optical Mouse	45.00	20	900.00
2034	2	vWireless RF remote and laser pointer	50.00	10	500.00
2035	1	Everlast Hard Drive, 60 GB	200.00	6	1200.00
2036	1	Optical Mouse	45.00	30	1350.00
2037	1	Optical Mouse	45.00	10	450.00
2037	2	Roadster 56KB Ext. Modem	120.00	5	600.00
2037	3	Everlast Hard Drive, 60 GB	205.00	10	2050.00
2038	1	NoTech Speaker Set	50.00	8	400.00

Visualização Multidimensional de Vendas

Dimensão de Cliente	Dimensão Temporal		Total
	15-Maio-08	16-Maio-08	
Dartnik	US\$ 1.400,00	US\$ 1.350,00	US\$ 2.750,00
Summer Lake	US\$ 1.800,00	US\$ 3.100,00	US\$ 4.900,00
Trydon		US\$ 400,00	US\$ 400,00
Total	US\$ 3.200,00	US\$ 4.850,00	US\$ 8.050,00

As vendas localizam-se na intersecção de uma linha cliente com uma coluna tempo.

As agregações são fornecidas por ambas as dimensões.

Suporte Avançado de Banco de Dados

- Recursos avançados de acesso a dados:
 - Acesso a vários tipos de SGBDs, arquivos fora do banco de dados e fontes de dados internas e externas
 - Acesso a dados agregados de data warehouse
 - Recursos avançados de navegação de dados
 - Tempo rápido e consistente de resposta a consultas

Suporte Avançado de Banco de Dados

- Recursos avançados de acesso a dados (cont.):
 - Capacidade de mapear solicitações de usuários finais, expressas em termos de negócios ou de modelo, para a fonte adequada de dados e, em seguida, para a linguagem adequada de acesso aos dados
 - Suporte a bancos de dados muito grandes

Interface Fácil de Utilizar para os Usuários Finais

- Os recursos avançados OLAP são mais úteis quando o acesso a eles permanece simples
- Muitos recursos de interface foram “emprestados” de gerações anteriores de ferramentas de análise de dados
 - Já familiares aos usuários finais
 - Torna a OLAP facilmente aceita e prontamente utilizada

Arquitetura Cliente/servidor

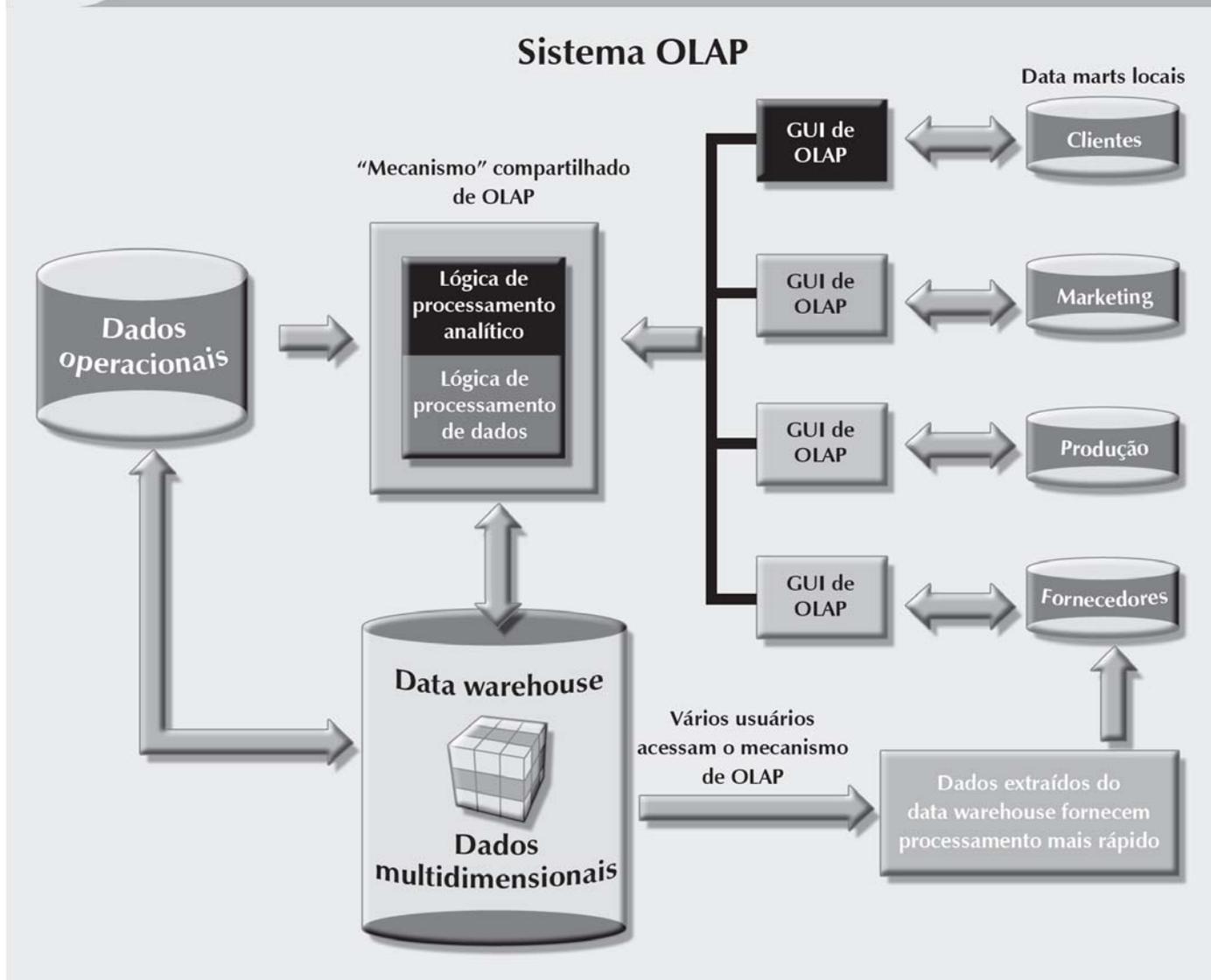
- Fornece um modelo em que novos sistemas podem ser projetados, desenvolvidos e implementados
 - Possibilita que um sistema OLAP seja dividido em vários componentes que definem sua arquitetura
 - OLAP é projetado para atender a exigências de facilidade de utilização, ao mesmo tempo em que mantém a flexibilidade do sistema

Arquitetura OLAP

- As características operacionais de OLAP podem ser divididas em três módulos principais:
 - Interface gráfica de usuário (GUI)
 - Lógica de processamento analítico
 - Lógica de processamento de dados
- Projetados para utilizar tanto dados operacionais como de data warehouse
- Na maioria das implementações, o data warehouse e o OLAP constituem ambientes complementares inter-relacionados
- Alguns sistemas OLAP fundem as abordagens de data warehouse e data mart

FIGURA 13.10

Servidor de OLAP com pequenos data marts locais



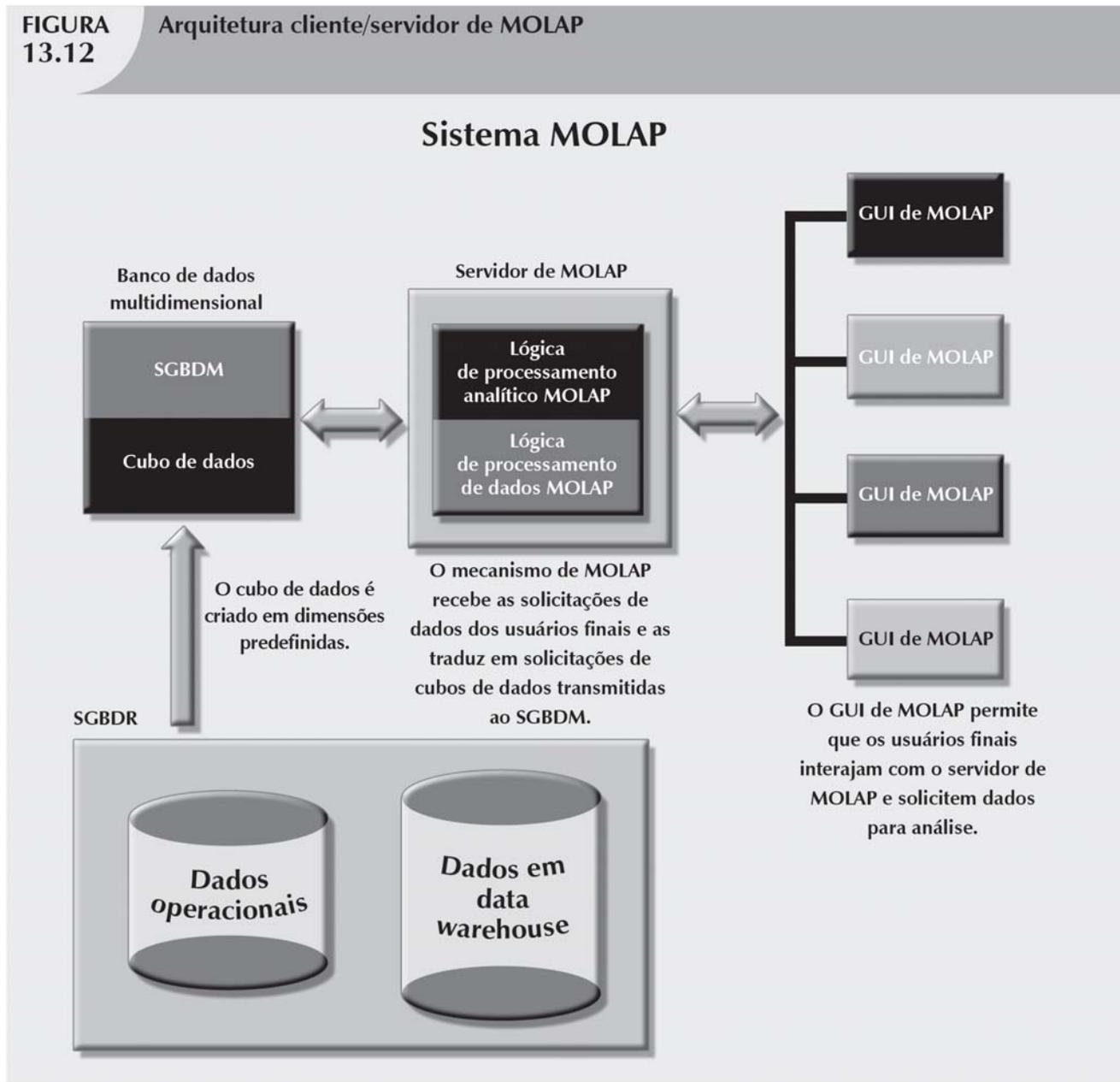
OLAP Relacional

- Utiliza bancos de dados relacionais e ferramentas familiares de consulta relacional para armazenar e analisar dados multidimensionais
- Adiciona as seguintes extensões à tecnologia de SGBDR tradicional:
 - Suporte a esquemas de dados multidimensionais no SGBDR
 - Linguagem de acesso a dados e desempenho de consulta otimizados para dados multidimensionais
 - Suporte a bancos de dados muito grandes (VLDBs)

OLAP Multidimensional

- Amplia os recursos de OLAP para **sistemas de gerenciamento de banco de dados multidimensionais (SGBDMs)**
 - Usuários finais de SGBDM visualizam os dados armazenados como um **cubo de dados**
 - Os cubos podem crescer até um número n de dimensões, tornando-se, assim, hipercubos
 - Para acelerar o acesso aos dados, os cubos normalmente são mantidos na memória do chamado cache de cubos

FIGURA 13.12



OLAP Relacional *versus* Multidimensional

- A seleção de um ou outro costuma depender do ponto favorável ao avaliador
- Uma avaliação adequada do OLAP deve incluir preço, plataformas de hardware suportadas, compatibilidade com SGBD existente, etc.
- Os fornecedores de ROLAP e MOLAP têm trabalhado em direção à integração de suas respectivas soluções em um modelo unificado de suporte a decisões
- Os bancos de dados relacionais utilizam com sucesso o projeto de esquema estrela para tratar de dados multidimensionais

Esquema Estrela

- Técnica de modelagem de dados
 - Mapeia dados multidimensionais de suporte a decisões em um banco de dados relacional
- Cria um equivalente próximo do esquema de banco de dados multidimensional a partir do banco relacional existente
- Produzem um modelo de fácil implementação para a análise de dados multidimensionais
 - Preservam as estruturas relacionais em que o banco operacional foi criado
- O esquema estrela básico possui quatro componentes: fatos, dimensões, atributos e hierarquias de atributos

Fatos

- Medidas numéricas (valores) que representam um aspecto ou atividade específica dos negócios
 - Costumam ser armazenados em uma tabela de fatos que constitui o centro do esquema estrela
- A **tabela de fatos** contém fatos vinculados por meio de suas dimensões
- **Métricas são** fatos computados e derivados

Dimensões

- Características de qualificação que fornecem perspectivas adicionais a um determinado fato
- Os dados de suporte a decisões são quase sempre vistos relacionados a outros dados
- As dimensões são as lentes de amplificação por meio das quais são estudados os fatos
- As dimensões normalmente são armazenadas em **tabelas de dimensões**

Atributos

- Costumam ser utilizados para buscar, filtrar e classificar fatos
- Dimensões fornecem características descritivas sobre os fatos por meio de seus atributos
- Não há limite matemático para o número de dimensões utilizadas
- **Detalhamento:** a capacidade de focar em “fatias” do cubo para executar uma análise mais detalhada

Hierarquias de Atributos

- Fornece uma organização vertical utilizada para duas finalidades principais: agregação e análise
- Determinam como os dados do data warehouse são extraídos e apresentados
- É armazenada no dicionário de dados do SGBD
- É utilizada pela ferramenta OLAP para acessar o data warehouse adequadamente

Representação em Esquema Estrela

- Os fatos e dimensões normalmente são representados por tabelas físicas no banco de dados do data warehouse
- Várias linhas de fatos se relacionam a cada linha de dimensão
 - A chave primária da tabela de fatos é composta
 - A chave primária da tabela de fatos é sempre formada combinando-se as chaves estrangeiras que apontam para as tabelas de dimensões à qual está relacionada
- As tabelas de dimensões são sempre menores do que as tabelas de fatos
- Cada registro de dimensão é relacionado a milhares de registros de fatos

Técnicas de Aprimoramento do Desempenho do Esquema Estrela

- Quatro técnicas para otimizar o projeto de data warehouse:
 - Normalização de tabelas dimensionais
 - Manutenção de várias tabelas de fatos para representar diferentes níveis de agregação
 - Desnormalização de tabelas de fatos
 - Particionamento e replicação de tabelas

Técnicas de Aprimoramento do Desempenho do Esquema Estrela (cont.)

- As tabelas dimensionais são normalizadas para:
 - Se obter simplicidade semântica
 - Facilitar a navegação do usuário final pelas dimensões
- A desnormalização de tabelas de fatos aprimora o desempenho de acesso a dados e poupa espaço de armazenamento
- O **particionamento** separa a tabela em subconjuntos de linhas ou colunas
- A **replicação** faz uma cópia da tabela e a coloca em uma localização diferente

Implementação de um Data Warehouse

- Várias restrições, como:
 - Fundos disponíveis
 - Visão da gerência sobre o papel executado por um departamento de SI
 - Extensão e profundidade das necessidades de informações
 - Cultura corporativa
- Nenhuma fórmula simples é capaz de descrever o desenvolvimento perfeito de um data warehouse

Data Warehouse como Modelo Ativo de Suporte a Decisões

- Data warehouse:
 - Não é um banco de dados estático
 - É um modelo dinâmico de suporte a decisões que é, quase por definição, sempre um trabalho em andamento
- O data warehouse é um componente fundamental do ambiente moderno de BI, mas certamente não é o único
- Seu projeto e implementação devem ser examinados à luz da infraestrutura toda

Um Esforço de toda a Empresa que Exige o Envolvimento dos Usuários

- Os dados do data warehouse atravessam os limites departamentais e as fronteiras geográficas
- Para construir um data warehouse, o projetista deve:
 - Envolver os usuários finais no processo
 - Garantir o engajamento dos usuários finais desde o início
 - Solicitar *feedback* contínuo dos usuários finais
 - Gerenciar as expectativas dos usuários finais
 - Estabelecer procedimentos para a resolução de conflitos

Satisfação da Trilogia: dados, análise e usuários

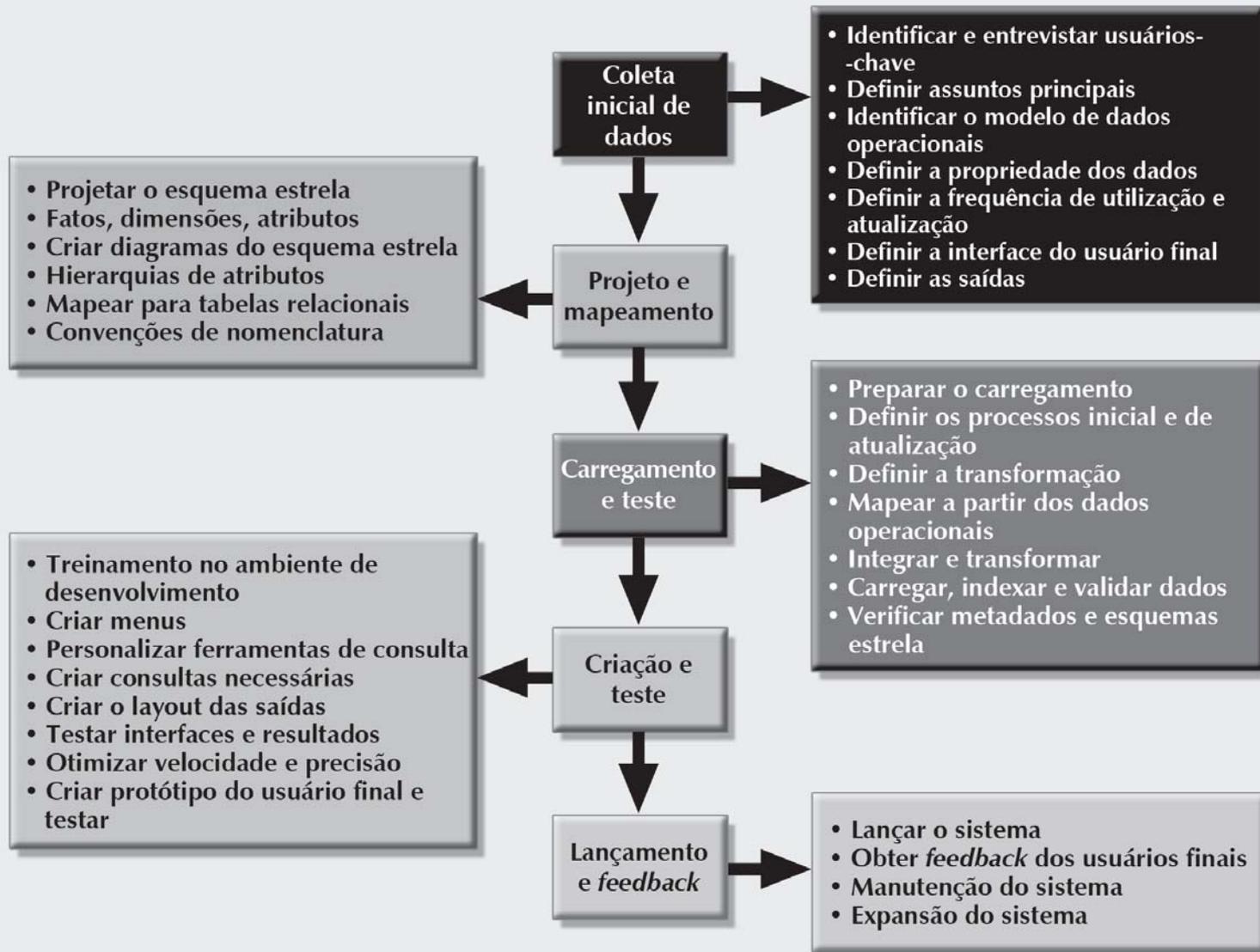
- O projetista do data warehouse deve satisfazer:
 - Critérios de integração e carregamento de dados
 - Recursos de análise de dados com desempenho aceitável de consulta
 - Necessidades de análise de dados do usuário final

Aplicação de Procedimentos de Projeto de Banco de Dados

- O desenvolvimento do data warehouse é um esforço de toda a empresa, exigindo muitos recursos
 - A quantidade enorme e confusa de dados de suporte a decisões provavelmente exigirá o software e hardware mais recente
 - São necessários procedimentos muito detalhados para coordenar o fluxo de dados dos bancos operacionais para o data warehouse
 - Para implementar e dar suporte à arquitetura de data warehouse, é necessário pessoal com habilidades avançadas em projeto de banco de dados, integração de software e gerenciamento

FIGURA
13.22

Mapa do percurso de projeto e implementação de data warehouses



Mineração de Dados (Data Mining)

- **A mineração de dados** refere-se às seguintes atividades:
 - Análise os dados
 - Descoberta dos problemas e oportunidades ocultos em seus relacionamentos
 - Formação de modelos computacionais com base nessas descobertas
 - Utilização desses modelos para prever o comportamento do negócio
- Exige a mínima intervenção do usuário final

Extensões de SQL para OLAP

- A proliferação de ferramentas OLAP promoveu o desenvolvimento de extensões de SQL
- Muitas inovações encontraram seu espaço em SQL-padrão
- Todos os comandos de SQL funcionarão no data warehouse conforme esperado
- A maioria das consultas executadas em data warehouse incluem muitos agrupamentos e agregações de dados em várias colunas

Extensão ROLLUP

- É utilizada com a cláusula GROUP BY para gerar agregados por diferentes dimensões
- A cláusula GROUP BY gerará apenas um agregado para cada nova combinação de valores de atributos
- A extensão ROLLUP permite obter um subtotal para cada coluna listada, exceto para a última, que obtém um total
- A ordem da lista de colunas é muito importante

Extensão CUBE

- A extensão CUBE também é utilizada com a cláusula GROUP BY para gerar agregados por colunas listadas, inclusive a última
- Permite a obtenção de um subtotal para cada coluna listada na expressão, além de um total final para a última coluna da lista
- É particularmente útil quando se deseja computar todos os subtotais possíveis de agrupamentos
- As tabulações cruzadas são candidatas especialmente adequadas à aplicação da extensão CUBE

Visualizações Materializadas

- A **visualização materializada** é uma tabela dinâmica que contém não apenas o comando de consulta de SQL para gerar as linhas
 - Também armazena as próprias linhas
- É criada quando a consulta é executada pela primeira vez, e as linhas resumidas são armazenadas na tabela
- São atualizadas quando da atualização das tabelas de base

Resumo

- **Business intelligence (BI)** gera e apresenta informações que dão suporte à tomada de decisões
- O BI cobre uma faixa de tecnologias e aplicações para o gerenciamento de todo o ciclo de vida dos dados
- O sistema de suporte a decisões foi o precursor original da geração atual de sistemas de BI
- Os dados operacionais não são adequados ao suporte a decisões

Resumo (cont.)

- As necessidades um SGBD de suporte a decisões estão divididas em quatro categorias principais:
 - Esquema de bancos de dados
 - Extração e carregamento de dados
 - Interface analítica do usuário final
 - Necessidades de tamanho do banco de dados
- O data warehouse é um conjunto de dados que fornecem suporte à tomada de decisões
 - Geralmente é um banco de dados apenas de leitura
 - Otimizado para processamento de análises e consultas

Resumo (cont.)

- Os sistemas OLAP apresentam quatro características principais:
 - Utilização de técnicas de análise de dados multidimensionais
 - Suporte avançado a banco de dados
 - Interfaces de usuário final fáceis de utilizar
 - Arquitetura cliente/servidor
- O ROLAP fornece recursos do OLAP utilizando bancos de dados relacionais
- O MOLAP fornece recursos do OLAP utilizando os SGBDMs

Resumo (cont.)

- O esquema estrela é uma técnica de modelagem
 - Mapeia dados multidimensionais de suporte a decisões em um banco relacional
- O esquema estrela possui quatro componentes:
 - Fatos
 - Dimensões
 - Atributos
 - Hierarquias de atributos

Resumo (cont.)

- Quatro técnicas para otimizar o projeto de data warehouse:
 - Normalização de tabelas dimensionais
 - Manutenção de várias tabelas de fatos que representem níveis de agregação diferentes,
 - Desnormalização de tabelas de fatos
 - Particionamento e replicação de tabelas
- A mineração de dados automatiza a análise de dados operacionais
- A SQL foi aprimorada com extensões que dão suporte a processamento e geração de dados do tipo OLAP