# Sistemas de Gerenciamento de Banco de Dados

Tradução da
Terceira Edição

Mc Graw Hill

Ramakrishnan • Gehrke

The McGraw-Hill Companies

**Sistemas de Gerenciamento de Banco de Dados**
ISBN 978-85-7726-027-0



Tradução da terceira edição do original em inglês Database Management Systems.
© 2003, 2000, 1998 de The McGraw-Hill Companies, Inc.
ISBN da obra original: 0-07-246563-8

Diretor-Geral: *Adilson Pereira*
Editora: *Gisélia Costa*
Supervisora de Produção: *Guacira Simonelli*
Preparação de Texto: *Lucrécia Freitas e Mônica de Aguiar*
Design da Capa: *Mick Wiggins*
Editoração Eletrônica: *Crontec Ltda.*

**Dados Internacionais de Catalogação na Publicação (CIP)**
**(Câmara Brasileira do Livro, SP, Brasil)**

Ramakrishnan, Raghu
Sistemas de bancos de dados / Raghu Ramakrishnan, Johannes Gehrke ; tradutores Acauan Pereira Fernandes, Celia Taniwake, João Tortello ; revisão técnica Elaine Parros Machado de Sousa. -- 3. ed. -- São Paulo : McGraw-Hill, 2008.

Título original: DataBase management systems
Bibliografia
ISBN 978-85-7726-027-0

1. Banco de dados 2. Banco de dados – Gerência I. Gehrke, Johannes. II. Sousa, Elaine Parros Machado de. III. Título.

07-7008 CDD-005.74

**Índice para catálogo sistemático:**

1. Banco de dados : Gerenciamento : Ciência da computação 005.74


# 25

# DATA WAREHOUSING E APOIO À DECISÃO

- Por que os SGBDs tradicionais são inadequados para o apoio à decisão?
- O que é modelo de dados multidimensional e quais tipos de análise ele facilita?
- Quais recursos do padrão SQL:1999 suportam consultas multidimensionais?
- Como o padrão SQL:1999 suporta análise de seqüências e tendências?
- Como os SGBDs estão sendo otimizados para produzir respostas antecipadas à análise interativa?
- Quais tipos de índice e organizações de arquivo os sistemas OLAP exigem?
- O que é data warehousing e por que ele é importante para o apoio à decisão?
- Por que as visões materializadas se tornaram importantes?
- Como podemos manter visões materializadas eficientemente?
- **Conceitos-chave:** OLAP, modelo multidimensional, dimensões, medidas; roll-up, drill-down, rotação, tabulação cruzada, CUBE; consultas de janela, quadros, ordem; consultas N mais, agregação online; índices de mapa de bits, índices de junção; data warehouse, extração, atualização, eliminação; visões materializadas, manutenção incremental, mantendo visões de data warehouse.

> Nada é mais difícil e, portanto, mais precioso, do que ter o poder de decidir.
>
> —Napoleão Bonaparte

Os sistemas de gerenciamento de banco de dados são amplamente usados pelas organizações para manter dados que documentam operações diárias. Em aplicações que atualizam tais *dados operacionais*, as transações normalmente fazem pequenas alterações (por exemplo, adicionar uma reserva ou depositar um cheque) e um grande número de transações devem ser processadas confiável e eficientemente. Tais aplicações de **processamento de transação online (OLTP** — Online Transaction Processing) fomentaram o crescimento do setor de SGBD nas últimas três décadas e, sem dúvida, continuarão a ser importantes. Tradicionalmente, os SGBDs têm sido extensivamente otimizados para terem bom desempenho em tais aplicações.

Recentemente, entretanto, as organizações têm dado cada vez mais ênfase às aplicações nas quais dados atuais e históricos são amplamente analisados e explorados, identificando tendências úteis e criando resumos dos dados para apoiar a tomada de decisões de alto nível. Tais aplicações são referidas como **apoio à decisão**. Os principais fabricantes de SGBD relacional reconheceram a importância desse segmento de mercado e estão adicionando recursos em seus produtos para suportá-lo. Em particular, a SQL foi estendida com novas construções, e novas técnicas de indexação e otimização de consultas estão sendo adicionadas para suportar consultas complexas.

O uso de visões ganhou popularidade rapidamente, devido à sua utilidade em aplicações envolvendo análise complexa de dados. Embora as consultas em visões possam ser respondidas pela avaliação da definição da visão, quando a consulta é feita, o cálculo prévio dessa definição pode tornar as consultas muito mais rápidas. Levando a motivação das visões previamente calculadas um passo adiante, as organizações podem consolidar informações de vários bancos de dados em um *data warehouse*, copiando tabelas de muitas fontes em um único local ou materializando uma visão definida sobre várias fontes. Data warehousing tornou-se difundido e agora estão disponíveis muitos produtos especializados para criar e gerenciar data warehouse de vários bancos de dados.

Iniciamos este capítulo com um panorama do apoio à decisão na Seção 25.1. Apresentamos o modelo de dados multidimensional na Seção 25.2, e consideramos os problemas do projeto de banco de dados na Seção 25.2.1. Na Seção 25.3, discutimos a rica classe de consultas que ele suporta naturalmente. Na Seção 25.3.1, discutimos como as novas construções do padrão SQL:1999 nos permitem expressar consultas multidimensionais. Na Seção 25.4, discutimos as extensões do padrão SQL:1999 que suportam consultas sobre relações como coleções ordenadas. Consideramos a otimização para a geração rápida de respostas iniciais na Seção 25.5. As muitas extensões de linguagem de consulta exigidas no ambiente OLAP inspiraram o desenvolvimento de novas técnicas de implementação; discutimos essas técnicas na Seção 25.6. Na Seção 25.7, examinamos os problemas envolvidos na criação e na manutenção de um data warehouse. Do ponto de vista técnico, um problema importante é como manter as informações do data warehouse (tabelas ou visões replicadas), quando as informações de origem subjacentes mudam. Após abordarmos a importante função desempenhada pelas visões no ambiente OLAP e em data warehousing na Seção 25.8, consideramos a manutenção de visões materializadas, nas Seções 25.9 e 25.10.

## 25.1 INTRODUÇÃO AO APOIO À DECISÃO

A tomada de decisão organizacional exige uma visão abrangente de todos os aspectos de uma empresa, de modo que muitas organizações criaram **data warehouses** consolidados, que contêm dados extraídos de vários bancos de dados mantidos por diferentes unidades empresariais, com informações históricas e de resumo.

A tendência para o uso de data warehouses é complementada por uma maior ênfase em ferramentas de análise poderosas. Muitas características das consultas de apoio à decisão tornam os sistemas SQL tradicionais inadequados:

- A cláusula `WHERE` freqüentemente contém muitas condições `AND` e `OR`. Conforme vimos na Seção 14.2.3, as condições `OR` em particular são tratadas deficientemente em muitos SGBDs relacionais.
- As aplicações exigem uso extensivo de funções estatísticas, como desvio padrão, que não são suportadas na SQL-92. Portanto, as consultas em SQL freqüentemente precisam ser incorporadas em um programa de linguagem hospedeira.

- Muitas consultas envolvem condições ao longo do tempo ou exigem agregação ao longo do tempo. O padrão SQL-92 fornece suporte deficiente para tal análise de seqüência de tempo.
- Os usuários freqüentemente precisam fazer várias consultas relacionadas. Como não existe nenhuma maneira conveniente de expressar essas famílias de consultas que ocorrem comumente, os usuários precisam escrevê-las como uma coleção de consultas independentes, o que pode ser maçante. Além disso, o SGBD não tem nenhum modo de reconhecer e explorar as oportunidades de otimização que surgem na execução em conjunto de muitas consultas relacionadas.

Estão disponíveis três classes amplas de ferramentas de análise. Primeiro, alguns sistemas suportam uma classe de consultas estilizadas que normalmente envolvem operadores de agrupamento e agregação, e fornecem excelente suporte para condições booleanas complexas, funções estatísticas e recursos para análise de seqüência de tempo. As aplicações dominadas por tais consultas são chamadas processamento analítico online **(OLAP — Online Analytic Processing)**. Esses sistemas suportam um estilo de consulta no qual os dados são melhor considerados como um array multidimensional e são afetados por ferramentas de usuário final, como planilhas eletrônicas, além de linguagens de consulta de banco de dados.

Segundo, alguns SGBDs suportam consultas estilo SQL tradicionais, mas são projetados para também suportar eficientemente consultas OLAP. Tais sistemas podem ser considerados SGBDs relacionais otimizados para aplicações de apoio à decisão. Muitos fabricantes de SGBDs relacionais estão atualmente aprimorando seus produtos nessa direção e, com o passar do tempo, a distinção entre sistemas OLAP especializados e SGBDs relacionais aprimorados para suportar consultas OLAP provavelmente vai diminuir.

A terceira classe de ferramentas de análise é motivada pelo desejo de encontrar tendências e padrões interessantes ou inesperados em grandes conjuntos de dados, em vez das características de consulta complexa que acabamos de listar. Na **análise de dados exploratória**, embora um analista possa reconhecer um 'padrão interessante' ao ver tal padrão, é muito difícil formular uma consulta que capture a essência de um padrão interessante. Por exemplo, talvez um analista que esteja examinando históricos de utilização de cartão de crédito queira detectar atividade incomum, indicando o mau uso de um cartão perdido ou roubado. Talvez um atacadista queira ver os registros de um cliente para identificar prováveis compradores para uma nova promoção; essa identificação dependeria do nível da renda, dos padrões de compra, das áreas de interesse demonstrado etc. Em muitas aplicações, o volume de dados é grande demais para permitir uma análise manual ou mesmo uma análise estatística tradicional, e o objetivo da **mineração de dados** é suportar a análise exploratória sobre conjuntos de dados muito grandes. Discutiremos melhor a mineração de dados no Capítulo 26.

Claramente, é provável que a avaliação de consultas OLAP ou de mineração de dados em dados distribuídos globalmente seja excruciantemente lenta. Além disso, para tal análise complexa, freqüentemente de natureza estatística, não é fundamental que seja usada a versão mais atual dos dados. A solução natural é criar um repositório centralizado de todos os dados; isto é, um data warehouse. Assim, a disponibilidade de um data warehouse facilita a aplicação de ferramentas OLAP e de mineração de dados e, inversamente, o desejo de aplicar tais ferramentas de análise é uma forte motivação para a construção de um data warehouse.

### 25.7.1 Criando e Mantendo um Data Warehouse

Muitos desafios devem ser superados na criação e manutenação de um data warehouse grande. Um bom esquema de banco de dados deve ser projetado para conter uma coleção integrada de dados copiados de diversas fontes. Por exemplo, o data warehouse de uma empresa poderia incluir os bancos de dados de inventário e do departamento pessoal, junto com bancos de dados de vendas mantidos por escritórios em diferen-

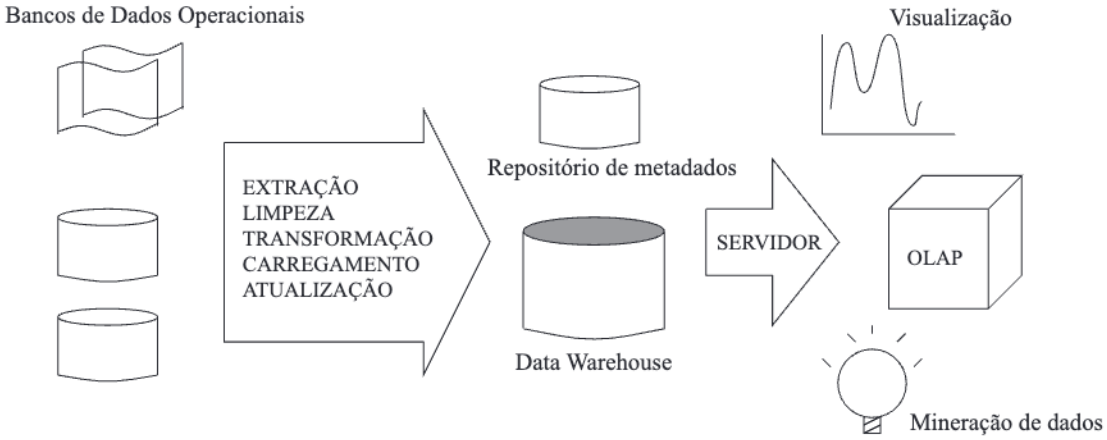


**Figura 25.10** Uma arquitetura de data warehousing típica.

tes países. Como os bancos de dados de origem são freqüentemente criados e mantidos por grupos diferentes, existem várias discordâncias semânticas entre esses bancos de dados, como diferentes unidades de moeda corrente, nomes diferentes para o mesmo atributo e diferenças em como as tabelas são normalizadas e estruturadas; essas diferenças precisam ser harmonizadas quando os dados são trazidos para o data warehouse. Depois de seu esquema ser projetado, o data warehouse precisa ser preenchido e, com o passar do tempo, ele deve se manter consistente com os bancos de dados de origem.

Os dados são **extraídos** dos bancos de dados operacionais e de fontes externas, **limpos** para minimizar os erros e preencher as informações ausentes quando possível, e **transformados** para harmonizar discordâncias semânticas. A transformação dos dados normalmente é feita pela definição de uma visão relacional sobre as tabelas nas origens de dados (os bancos de dados operacionais e outras fontes externas). O **carregamento** dos dados consiste em materializar tais visões e armazená-las no data warehouse. Ao contrário de uma visão padrão em um SGBD relacional, portanto, a visão é armazenada em um banco de dados (o data warehouse) que é diferente do(s) banco(s) de dados que contém(êm) as tabelas sobre as quais ela é definida.

Os dados limpos e transformados são finalmente **carregados** no data warehouse. O pré-processamento adicional, como ordenação e geração de informações de resumo, é realizado nesse estágio. Por eficiência, os dados são particionados e índices são construídos. Devido ao grande volume de dados, o carregamento é um processo lento. Carregar um terabyte de dados seqüencialmente pode demorar semanas, e carregar mesmo um gigabyte pode demorar horas. Portanto, o paralelismo é importante para o carregamento de data warehouses.

Após os dados serem carregados em um data warehouse, medidas adicionais precisam ser tomadas para garantir que esses dados sejam **atualizados** periodicamente, para refletir as atualizações feitas nas origens de dados e periodicamente eliminar dados antigos (talvez para uma mídia de arquivamento). Observe a conexão entre o problema de atualizar tabelas do data warehouse e manter réplicas de tabelas de forma assíncrona em um SGBD distribuído. Manter réplicas das relações de origem é uma parte fundamental do data warehousing e esse domínio de aplicação é um fator importante na popularidade da replicação assíncrona (Seção 22.11.2), mesmo que a replicação assíncrona viole o princípio da independência de dados distribuídos. O problema da atualização de tabelas do data warehouse (que são visões materializadas sobre tabelas nos bancos de dados de origem) também tem interesse renovado na manutenção incremental de visões materializadas. (Discutiremos as visões materializadas na Seção 25.8.)

Uma tarefa importante na manutenção de um data warehouse é monitorar os dados correntemente armazenados nele; essa contabilidade é feita pelo armazenamento de informações sobre os dados data warehouse nos catálogos de sistema. Os catálogos de sistema associados a um data warehouse são muito grandes e freqüentemente são armazenados e gerenciados em um banco de dados separado, chamado **repositório de metadados**. O tamanho e a complexidade dos catálogos se dá em parte devido ao tamanho e à complexidade do próprio data warehouse e em parte porque muitas informações administrativas preecisam ser mantidas. Por exemplo, precisamos monitorar a origem de cada tabela do data warehouse e quando ela foi atualizada pela última vez, além de descrever seus campos.

Em última instância, o valor de um data warehouse está na análise que ele permite. Os dados de um data warehouse normalmente são acessados e analisados usando-se uma variedade de ferramentas, incluindo mecanismos de consulta OLAP, algoritmos de mineração de dados, ferramentas de visualização de informações, pacotes estatísticos e geradores de relatório.

## 25.8 VISÕES E APOIO À DECISÃO

As visões são amplamente usadas em aplicações de apoio à decisão. Diferentes grupos de analistas dentro de uma organização normalmente estão preocupados com diferentes aspectos do negócio e é conveniente definir visões que forneçam a cada grupo idéias dos detalhes do negócio que interessem. Uma vez definida uma visão, podemos escrever consultas ou novas definições de visão que a utilizem, como vimos na Seção 3.6; sob esse aspecto, uma visão é exatamente como uma tabela-base. Avaliar consultas feitas em visões é muito importante para aplicações de apoio à decisão. Nesta seção, consideraremos como essas consultas podem ser avaliadas eficientemente após a criação de visões dentro do contexto das aplicações de apoio à decisão.

### 25.8.1 Visões, OLAP e Data Warehousing

As visões são intimamente relacionadas com OLAP e data warehousing.

Normalmente, as consultas OLAP são consultas de agregação. Os analistas querem respostas rápidas para essas consultas sobre conjuntos de dados muito grandes e é natural considerar a computação prévia de visões (consulte as Seções 25.9 e 25.10). Em particular, o operador `CUBE` — discutido na Seção 25.3 — origina várias consultas de agregação intimamente relacionadas. Os relacionamentos existentes entre as muitas consultas de agregação que surgem de uma única operação `CUBE` podem ser explorados para desenvolver estratégias de computação prévia muito eficientes. A idéia é escolher um subconjunto das consultas de agregação para materialização, de tal modo que as consultas `CUBE` típicas possam ser respondidas rapidamente, usando visões materializadas e realizando-se algum cálculo adicional. A escolha de visões para materializar é influenciada pela quantidade de consultas que elas poderiam acelerar e pela quantidade de espaço exigido para armazenar a visão materializada (pois temos de trabalhar com determinada quantidade de espaço de armazenamento).

Um data warehouse é apenas uma coleção de tabelas replicadas de forma assíncrona e visões sincronizadas periodicamente. Um data warehouse é caracterizado pelo seu tamanho, pelo número de tabelas envolvidas e pelo fato de a maioria das tabelas subjacentes ser composta de bancos de dados externos, mantidos de forma independente. Contudo, o problema fundamental na manutenção do data warehouse é a manutenção assíncrona de tabelas replicadas e visões materializadas (consulte a Seção 25.10).

Encerra aqui o trecho do livro disponibilizado para esta Unidade de Aprendizagem. Na Biblioteca Virtual da Instituição, você encontra a obra na íntegra.