Capítulo

2. Conceitos Básicos

2.1 Sistemas de Banco de Dados

Um sistema de banco de dados (SBD) é composto por um programa de software chamado sistema gerenciador de banco de dados (SGBD) e por um conjunto de dados, chamado banco de dados (BD) [EP90]. Um dos objetivos dos BDs é oferecer algum nível de abstração, escondendo os detalhes de armazenamento dos dados dos usuários dos BDs. Para se obter esta abstração, é utilizado o conceito de **modelo de dados**, que é uma coleção de ferramentas conceituais utilizadas para descrever a estrutura do BD. A estrutura de um BD engloba a descrição dos seus dados, os relacionamentos entre eles, a semântica e as restrições que atuam sobre estes dados. A maioria dos modelos de dados também inclui um conjunto de operações utilizadas para especificar consultas e atualizações nos BDs [EN89].

Os SGBDs apresentam gerenciamento automático de erros, capacidade de consultas "ad-hoc", consistência global de dados (ou seja, a ausência de dados inconsistentes), autorização para acesso e eficiente manipulação das informações armazenadas. É através da **linguagem de consulta** oferecida pelo SGBD que os usuários podem submeter consultas ao sistema. A linguagem de consulta oferecida é altamente dependente do modelo de dados utilizado.

Um SBD pode ser tanto centralizado quanto distribuído [SL90]. Um SBD centralizado consiste em um único SGBD centralizado que gerencia o BD localizado em um único sistema de computação. Já um SBD distribuído consiste em um único SGBD distribuído que gerencia múltiplos BDs, garantindo aos usuários transparência de distribuição. Os BDs podem residir em um único sistema de computação ou em vários sistemas de computação, os quais podem diferir em hardware, software ou suporte de comunicação. O SGBD distribuído suporta apenas um único modelo de dados e uma única linguagem de consulta, além de um único mecanismo de gerenciamento de transação e de otimização de consultas, entre outros.

O uso de vários SBDs em uma organização (por exemplo, o caso da UNICAMP) cria a necessidade de compartilhamento de dados. Uma primeira solução para este problema é a conversão de todos os sistemas para um único modelo de dados com um único método de acesso. Porém, esta conversão não é aceitável, porque além de requerer alto custo, há também a necessidade de modelos de dados específicos para representar as diferentes aplicações do mundo real.

Desta forma, a solução encontrada é a **integração** ou a **interoperabilidade** dos diversos SBDs. Há uma diferença sutil entre estes termos. Integração significa combinação de BDs e sistemas como um todo, ao passo que interoperabilidade significa o compartilhamento dos dados entre os BDs, sem uma integração global.

[Hsi92] destaca cinco requisitos necessários ao compartilhamento global de dados:

- **transparência de acesso aos diversos BDs:** os usuários devem acessar os diferentes BDs como se estivessem acessando os SGBDs a que estão acostumados.

- **autonomia local de cada um dos BDs:** os diversos BDs existentes devem compartilhar os seus dados sem comprometer suas restrições de integridade, aplicações específicas ou mecanismos de segurança.

- **característica multimodelo e multilinguagem:** os BDs que participam do compartilhamento de dados global devem possuir diferentes modelos de dados e, conseqüentemente, diferentes linguagens de manipulação de dados.

- **existência de uma arquitetura paralela e de propósito geral para suportar os diversos BDs:** os BDs e os seus respectivos SGBDs devem ser suportados por computadores e dispositivos de armazenamento secundário dedicados, organizados de uma maneira paralela.

- **mecanismos de controle de concorrência eficientes:** os diversos acessos concorrentes aos BDs devem ser controlados eficientemente para que as consultas realizadas pelos usuários sejam executadas corretamente.

Surge, então, o conceito de **sistemas de banco de dados heterogêneos**.

2.2 Sistemas de Banco de Dados Heterogêneos

Um sistema de banco de dados **heterogêneo** (SBDH) é um pacote de software construído sobre vários SGBDs heterogêneos preexistentes, os quais são integrados em um único SBDH de

uma maneira cooperativa. Estes SGBDs são preexistentes no sentido que foram criados independentemente, de uma forma não coordenada e sem considerar que um dia seriam integrados. Eles são heterogêneos no sentido que operam em diferentes ambientes (hardware, protocolos de comunicação, sistemas operacionais) e podem utilizar diferentes SBDs. No ambientes SBDH, cada SGBD preexistente recebe o nome de SGBD componente. Cada SGBD componente pode ser, por sua vez, um SBD centralizado, um SBD distribuído ou um outro SBDH. Além disso, um SGBD componente pode participar de um ou mais SBDHs distintos.

O SBDH cria uma integração lógica dos diversos SGBDs componentes, oferecendo aos usuários acesso uniforme aos dados contidos nos vários BDs, sem migrar os dados e sem requerer que os usuários conheçam tanto a localização quanto as características dos diferentes SGBDs. Além disso, o uso da abordagem SBDH permite que aplicações preexistentes permaneçam operacionais, e que novas aplicações possam acessar os dados localizados nos vários SGBDs componentes. Em outras palavras, um usuário pode acessar os dados dos SGBDs que participam do SBDH tanto indiretamente através da interface do SBDH quanto diretamente através da interface dos diversos componentes.

Além da **heterogeneidade** e **distribuição**, uma terceira característica inerente aos SBDHs é a **autonomia** dos seus componentes ([SL90, Man+92]). Existem quatro tipos diferentes de autonomia: de projeto, de comunicação, de execução e de associação.

De acordo com a autonomia de projeto ou decisão, os SGBDs são independentes na escolha dos seus modelos de dados e linguagem de consulta, na forma de gerenciamento dos dados e na definição das restrições que atuam sobre este gerenciamento, na interpretação semântica dos seus modelos, na determinação das funcionalidades oferecidas ao sistema global e na implementação de suas estruturas e algoritmos. Este tipo de autonomia é um dos principais causadores da heterogeneidade entre SGBDs componentes.

Já a autonomia de comunicação permite que os SGBDs sejam independentes para determinar se irão ou não se comunicar com os outros SGBDs componentes, além de decidir quando e como isso será realizado.

A autonomia de execução, por sua vez, determina que os SGBDs são independentes para executar consultas locais e externas, definindo sua ordem de execução. Uma vez que os SGBDs componentes têm total controle sobre a execução das suas transações, eles podem abortá-las a qualquer momento durante as suas execuções, desde que estas não satisfaçam as restrições locais. Adicionalmente, as operações locais não são afetadas logicamente pela participação do SGBD componente no ambiente SBDH.

Finalmente, de acordo com a autonomia de associação, os SGBDs são independentes para determinar qual informação compartilharão com o sistema global, quais consultas globais irão atender, quando irão iniciar a sua participação com o sistema global e quando irão finalizá-la.

A preservação da autonomia é um ponto chave no ambiente SBDH, uma vez que a integração dos SGBDs em um SBDH não deve infringir o direito destes gerenciarem os seus recursos sem a interferência do SBDH.

Um SBDF (SBD federado) é sinônimo de um SBDH [EP90]. Um SBDF pode ser classificado como fracamente ou fortemente acoplado ([SL90, Cle+93, Oli93]), dependendo de quem gerencia a federação e de como os componentes são integrados.

A abordagem SBDF fracamente acoplado parte do princípio que os diversos SBDs devem ser interoperáveis. Desta forma, neste ambiente, é responsabilidade dos usuários criar e manter a federação, uma vez que o SBDF não exerce nenhum controle sobre esta. Para tanto, os usuários devem conhecer tanto a localização quanto os problemas de heterogeneidade dos SGBDs componentes. As vantagens apresentadas por esta abordagem são que nenhum esforço é exigido ao administrador do BD (ABD) para resolver a heterogeneidade semântica existente entre os SGBDs componentes e que, além disso, não é necessário que o ABD antecipe as necessidades dos usuários da federação, uma vez que estes conseguem acessar os dados desejados quando necessário. Por outro lado, esta abordagem possui algumas desvantagens, tais como a ausência da transparência de localização dos dados dos diversos BDs e a dificuldade da realização das operações de atualização. Outros termos utilizados para SBDF fracamente acoplados são SBDMs (SBD Múltiplos) e sistemas interoperáveis.

Já a abordagem SBDF fortemente acoplado parte do princípio que os esquemas que representam os vários SGBDs componentes devem ser integrados. Desta forma, nesta abordagem é responsabilidade da federação e dos seus administradores criar, controlar e manter a federação. A vantagem apresentada advém do fato de que o SBDF fortemente acoplado esconde dos usuários da federação a localização e a heterogeneidade semântica dos BDs locais, através de esquemas globais. Por outro lado, a integração de esquemas é uma tarefa difícil de ser realizada.

Em suma, todos os termos acima citados descrevem um sistema distribuído que inclui um componente global para acessar as informações globalmente compartilhadas, e múltiplos componentes locais que gerenciam as suas informações. As diferenças estão na estrutura do componente global e como ele interage com os componentes locais [BHP92].

2.2.1 Problemas Apresentados pelo Ambiente SBDH

A integração de diversos SGBDs em um ambiente SBDH origina vários problemas, devido às características de heterogeneidade, autonomia e distribuição destes componentes.

Os principais problemas do ambiente SBDH são: o **gerenciamento de transações** ([BK91, Geo91, GRS91, BGS92, BST92, JS92, Raz92, Cle+93, Dre+93, HHS93, ST93, Woe+93]), a **autonomia local**, o **gerenciamento de consultas** ([HB91, DKS92, Cle+93, HK93, LS93a,

RCD93]), a **transparência de localização** e a **manutenção do esquema** ([BHP92]), o **gerenciamento de autorização**, a **integração de esquemas** ([BLN86, Tho+90, BHP92, Oli93]) e a **diferença na representação dos dados** ([AP93, Cle+93, Dre+93, Lim+93, LS93b, LSS94]).

O problema de integração de esquemas está relacionado à maneira segundo a qual o usuário pode ver logicamente os dados localizados nos múltiplos BDs. Cada BD tem um esquema, o qual descreve a estrutura dos seus dados, e cada usuário tem uma visão, a qual descreve que porção do dado é de seu interesse. A visão particular do usuário não pode ser afetada pela participação do SGBD no ambiente SBDH.

Resumidamente, este problema pode ser definido como o processo de desenvolvimento de um esquema conceitual global que engloba uma coleção de esquemas locais. Existem duas abordagens básicas para este problema: abordagem de esquema global e abordagem de esquema federado.

Na abordagem de esquema global, os diversos esquemas locais que participam do sistema são integrados em um único esquema global. Em outras palavras, os esquemas locais desenvolvidos independentemente são integrados, suas diferenças semânticas são resolvidas e um esquema global final é oferecido ao usuário. Este esquema global é, portanto, independente da heterogeneidade dos SGBDs e das representações de dados. As visões que os diferentes grupos de usuários e suas respectivas aplicações podem possuir podem ser definidas sobre o esquema global. Quando esta abordagem é utilizada, o problema de integração de esquemas também é conhecido como problema de integração de visões.

Já na abordagem de esquema federado, não há a criação de um esquema global. A tarefa de integrar os dados localizados nos múltiplos BDs é deixada a cargo dos usuários. Para tanto, estes devem conhecer tanto as diferenças de representação dos dados quanto as suas localizações. O sistema global apenas oferece diversos esquemas locais, os quais descrevem os dados dos diversos SGBDs componentes que podem ser compartilhados. Geralmente, a linguagem oferecida por estes tipos de esquemas possui uma grande variedade de funções e é poderosa o suficiente para oferecer maior controle sobre a informação sendo manipulada.

Uma lista abrangente de projetos que utilizam as abordagens acima citadas pode ser encontrada em [BHP92].

2.2.2 Identificação de Conflitos no Processo de Integração de Esquemas

A identificação de conflitos é de essencial importância para o projeto de BDs heterogêneos. Segundo [BLN86], um conflito entre duas representações R_1 e R_2 do mesmo conceito, pertencentes a esquemas distintos, surge quando estas representações não são idênticas. Duas representações são

idênticas quando são exatamente as mesmas, ou seja, os mesmos construtores foram utilizados e as mesmas restrições foram aplicadas.

A existência de conflitos entre duas ou mais representações está relacionada ao fato de que diferentes usuários modelam o mesmo pedaço do mundo real de maneiras distintas, de acordo com as suas percepções. A esta multiplicidade de representações de um determinado problema do mundo real dá-se o nome de relativismo semântico [SP94].

Desta forma, o processo de integração depende da identificação de similaridades e diferenças existentes entre os elementos dos diferentes esquemas, além da identificação de conjuntos de elementos distintos que são relacionados entre si por alguma propriedade semântica. De acordo com [SP94], os conflitos podem ser divididos em três grupos: conflitos de nome, conflitos semânticos e conflitos estruturais.

O primeiro tipo de conflito, **conflito de nome**, refere-se aos nomes utilizados para representar os diferentes elementos existentes nos esquemas a serem integrados. Diferentes nomes podem ser aplicados ao mesmo elemento (problema dos sinônimos) ou o mesmo nome pode ser aplicado a diferentes elementos (problema dos homônimos). Conflitos de nome não estão relacionados apenas a elementos do mesmo tipo, ou seja, podem ocorrer, por exemplo, entre um objeto e um atributo. Um exemplo de sinônimo ocorre quando o nome *cliente* é utilizado para representar, em um esquema, todos os clientes atendidos por uma loja, enquanto que o nome *comprador* é utilizado em outro esquema para representar a mesma situação.

Após a determinação dos conflitos de nome, devem ser identificados os **conflitos semânticos**. Este tipo de conflito surge quando o mesmo elemento é modelado nos diferentes esquemas, porém representando conjuntos que se sobrepõem. Em outras palavras, o conjunto de instâncias do elemento de um esquema é mais abrangente do que o conjunto de instâncias do elemento do outro esquema. Por exemplo, em uma visão, uma classe *estudante* representa todos os alunos de uma universidade, ou seja, alunos de graduação, pós-graduação e doutorado. Já em uma segunda visão, uma classe *estudante_grad* representa apenas os alunos de graduação da universidade.

Finalmente, **conflitos estruturais** surgem sempre que diferentes construtores estruturais são utilizados para modelar o mesmo conceito representado em diferentes visões. Como exemplo, o mesmo conjunto de objetos do mundo real pode ser representado como um tipo-entidade em uma visão e como um atributo de um tipo-entidade em outra visão. A diversidade de conflitos estruturais que podem aparecer em um problema de integração depende da semântica do modelo de dados utilizado.

Em geral, as discrepâncias existentes entre os esquemas apresentam mais do que um tipo de conflito.