Consultas em Sistemas Gerenciadores de Bases de Dados Heterogêneos

Fernanda Lima
e-mail: ferlima@inf.puc-rio.br

Rubens Nascimento Melo
e-mail: rubens@inf.puc-rio.br

PUC-RioInf.MCC 19/99 August, 1999

Abstract

Query Processing is a vital part of any Database Management System (DBMS). The presence of heterogeneity is systems that need to be integrated, insert new query processing issues that need to be addressed. Several research project were conducted in the filed of Heterogeneous Database Management System (HDBMS). More recently, due to the intense use of the Internet, the mediator architecture is being used as alternative, in this environment where the data sources can more unpredictable. This work presents several approaches, as well as their related architectures. Recent research projects using mediators is also described.

Keywords: Query Processing, Heterogeneous DBMS, Mediators

Resumo

O processamento de consultas é parte vital em qualquer Sistema Gerenciador de Bases de Dados (SGBD). A existência de heterogeneidade em sistemas que precisam ser integrados, insere novas questões neste processamento de consultas. Para isto, diversas pesquisas foram desenvolvidas no contexto de Sistemas Gerenciadores de Bases de Dados Heterogêneos (SGBDHs). Atualmente, com a intensa utilização da Internet, a arquitetura de mediadores vem sendo utilizado como alternativa em um ambiente onde as fontes de dados podem ter comportamento menos previsível. Este trabalho apresenta diversas abordagens existenets, assim como suas respectivas arquiteturas. Projetos de pesquisa recentes utilizando mediadores também são comentados.

Palavras-chave: Processamento de Consultas, SGBDs Heterogêneos, Mediadores

1. Introdução

Atualmente existe uma grande quantidade de informação em formato digital disponível em fontes de dados diversificadas, tais como Sistemas de Bancos de Dados (SBDs) Legados, páginas WWW, arquivos convencionais e repositórios de dados. Com a evolução da tecnologia, especialmente com a vasta utilização da Internet, mudanças estão ocorrendo na forma como as pessoas utilizam a informação. O que antes era acessível a apenas uma aplicação, agora deve estar disponível a outras aplicações existentes. Além disto, novas aplicações necessitam de informação armazenada em diferentes fontes de dados.

Existem diversas propostas para atender a estas demandas. Dentre elas pode-se destacar sistemas de bancos de dados federados, sistemas de múltiplos bancos de dados e a arquitetura composta por mediadores e tradutores. Os novos cenários que surgiram com a era da Internet requerem novas tecnologias e pesquisas. Problemas como escalabilidade das fontes de dados e processamento de consultas devem ser investigados. Em um ambiente como a Internet onde a sobrecarga de informação é comum, estes aspectos se tornam cada vez mais relevantes.

As pesquisas em Sistemas Gerenciadores de Bases de Dados Heterogêneos (SGBDHs) no contexto de federação e múltiplos bancos de dados têm como ponto importante a obtenção de um esquema global, onde ocorre a integração dos dados e os conflitos semânticos podem ser resolvidos. Entretanto, esta são tarefas extremamente difíceis e pouco viáveis para os tempos atuais onde o crescimento de fontes de dados é rápido. Muitas pesquisas recentes buscam soluções estes problemas com o uso da arquitetura de mediadores.

O objetivo deste trabalho é apresentar um estudo das diferentes técnicas de processamento de consultas utilizadas no contexto de SGBDHs. O restante deste trabalho está organizado da seguinte forma. Na seção 2 são apresentadas importantes arquiteturas utilizadas em Sistemas Gerenciadores de Bases de Dados Heterogêneos. Na seção 3, as etapas do processamento de consultas são enumeradas de acordo com as arquiteturas. Alguns projetos atuais relevantes são enumerados na seção 4. E finalmente na seção 5 são apresentadas as conclusões deste estudo.

2. Características de Sistemas Gerenciadores de Bases de Dados

Conforme encontrado na literatura [ÖV91, Khos93], existem três dimensões ortogonais que podem ser usadas para classificar bancos de dados: autonomia, distribuição e heterogeneidade. A dimensão da autonomia refere-se à distribuição do controle e indica o grau de independência com que cada SGBD individual trabalha. Distribuição é a dimensão da taxonomia que está ligada aos dados, que podem estar fisicamente distribuídos em vários nós ou armazenados em um único local físico. Já a heterogeneidade pode ocorrer de muitas formas variando da heterogeneidade do *hardware* e diferentes protocolos de comunicação a sistemas gerenciadores de bases dados diversificados. Pode-se destacar os modelos de

dados, linguagens de consulta, interfaces e protocolos de gerenciamento de transações como fatores importantes.

Na dimensão da distribuição, um SGBD pode ser distribuído ou centralizado, enquanto que na dimensão da heterogeneidade, ele pode ser homogêneo ou heterogêneo. No entanto, quanto se trata da dimensão da autonomia, as nomenclaturas utilizadas: sistemas de bancos de dados federados e sistemas de múltiplos bancos de dados encontram definições diferentes.

Em [SL90], Sheth e Larson definem sistemas de bancos de dados federados como um sub-caso de sistemas de múltiplos bancos de dados. Sendo que os sistemas federados podem ser fracamente ou fortemente acoplados, dependendo do gerenciamento da federação. Em [LMR90], Litwin et al utilizam o termo sistemas de múltiplos bancos de dados para representar o que em [SL90] foi chamado de sistemas de bancos de dados federados com acoplamento fraco. Já em [BRP92], Bright et al descrevem uma taxonomia de seis diferentes níveis com base no tipo de acoplamento. A classificação define: SBDs distribuídos, sistemas de múltiplos bancos de dados com esquema global, sistemas de bancos de dados federados, sistemas de linguagem de múltiplos bancos de dados, sistemas de linguagem de múltiplos bancos de dados homogêneos e, finalmente, sistemas interoperáveis. Em [MY95], ou autores utilizam os termos (federação e multidatabases) como sinônimos, significando os sistemas de bancos de dados acima dos SGBDs existentes responsáveis pela manutenção do esquema global.

Neste trabalho a dimensão da heterogeneidade é a mais relevante. Portanto, o termo Sistemas Gerenciadores de Bases de Dados Heterogêneos (SGBDH) será utilizado para englobar as diversas definições de sistemas federados e de múltiplas bases.

Tradicionalmente, sistemas centralizados utilizam uma arquitetura em três níveis, composta de esquema conceitual, interno e externo. Esta arquitetura, no entanto, não é adequada para os ambientes de SGBDs não centralizados, como o caso dos sistemas federados ou nos sistemas de múltiplos bancos de dados s. A seguir, diferentes propostas de arquitetura serão brevemente descritas.

2.1 Arquitetura genérica conforme [MY95]

Uma solução para permitir a interoperabilidade entre os sistemas heterogêneos é o uso de sistemas de múltiplos bancos de dados [MY95]. Estes sistemas “residem” acima de bancos de dados existentes e apresentam uma ilusão de um único banco de dados para o usuário.

A arquitetura em um sistema de múltiplos bancos de dados utilizada por [MY95] possui um único nó controlando diversos nós locais. Um sistema deste tipo utiliza a arquitetura cliente/servidor, onde múltiplos clientes interagem com um único servidor. O sistema de múltiplos bancos de dados controla todas as informações globais com o esquema global, submetendo e gerenciando que envolvem um ou mais SBDs locais. Para permitir a comunicação entre o sistema de múltiplos bancos de dados e os SBDs locais, *drivers* são

utilizados junto ao sistema local, porém sem necessitar nenhuma alteração nas aplicações ou nos sistemas locais. O acesso as bases locais através dos sistemas locais não é afetado.

O sistema de múltiplos bancos de dados mantém o esquema global único que os usuários utilizam para realizar consultas e atualizações. Cabe ao sistema realizar a manutenção do esquema global, porem os dados referentes aos sistemas de bancos de dados preexistentes são gerenciados localmente por cada sistema. O esquema global é construído através da integração dos esquemas dos bancos de dados locais. Esta atividade de integração requer que os conflitos semânticos sejam homogeneizados.

2.2 Arquitetura em cinco camadas conforme [SL90]

Em [SL90], Sheth e Larson apresentam uma arquitetura de tipos de esquemas utilizada para SGBDs Heterogêneos e Federados. Esta arquitetura consiste em cinco camadas compostas de esquemas diferenciados, estendendo a arquitetura padrão ANSI/SPARC de três níveis, para lidar com autonomia, distribuição e heterogeneidade.

No nível mais baixo encontra-se o esquema local, que é o esquema conceitual de um banco de dados componente. Um esquema local é expresso no modelo de dados nativo do componente. Portanto, pode-se ter esquemas locais diferentes com modelos de dados diversos.

Acima do esquema local vem o esquema componente, derivado da tradução do esquema local em um modelo de dados chamado canônico (ou comum) da federação. Existem pelo menos duas razões para definir este nível de esquema. Nos esquemas componentes são descritos os esquemas locais possivelmente divergentes e também as semânticas que não existem no esquema local. Ou seja, eles facilitam a negociação e integração entre tarefas executadas como integração ou especificação de visões, conforme o caso.

O esquema de exportação representa o subconjunto do esquema componente que está disponível para a federação. O objetivo de definir esta camada de esquemas é facilitar o controle da autonomia.

O esquema federado é a integração de diversos esquemas de exportação. O esquema federado também inclui informação da distribuição dos dados gerada quando esquemas de exportação são integrados. Conceitos similares ao esquema federado são esquema de importação, esquema global, esquema conceitual global, esquema unificado e esquema corporativo.

O esquema de nível mais alto é o esquema externo, que define o esquema para um usuário ou aplicação. O uso deste tipo de esquema permite a criação de um subconjunto da federação que seja relevante a um usuário, podendo ser alterado com mais facilidade. O modelo de dados deste esquema pode ser diferente do modelo da federação e novas restrições de integridade podem ser definidas. Além disto, os esquemas de exportação podem fornecer controle de acesso aos dados dos sistemas componentes.

2.3 Arquitetura com mediadores conforme [Wied92, Wied95]

Diversas pesquisas foram realizadas com o intuito de permitir a geração de um esquema global único para acessar SGBDs heterogêneos. No entanto, novos fatores surgiram nos tempos atuais. Com o uso da Internet, o escopo dos SGBDs heterogêneos cresceu imensamente. Agora, os SGBDs heterogêneos não pertencem apenas a empresas residindo localmente, existem SGBDs remotos e é necessário realizar consultas em sistemas legados, páginas HTML residindo em servidores remotos, arquivos convencionais. Ou seja, os dados podem ser estruturados, semi-estruturados ou até mesmo não estrurados.

Neste contexto, a proposta apresentada por Wiederhold em [Wied92, Wied95] apresenta uma possível solução para estas dificuldades. A idéia é criar um camada intermediária entre as aplicações e os bancos de dados. Esta camada composta por mediadores tem como objetivo simplificar, abstrair, reduzir, reunir dados e torná-los compreensíveis. Os mediadores são módulos de *software* que ocupam uma camada explícita e ativa em uma arquitetura de compartilhamento e permitem que as aplicações dos usuários sejam independentes dos recursos de dados. Estes módulos exploram o conhecimento codificado sobre subconjuntos de dados criando informação para uma níveis mais altos de aplicação

Na arquitetura de sistemas com mediadores, usuários finais interagem com aplicações escritas por programadores. Aplicações acessam uma representação uniforme das fontes de dados através de linguagens de consulta declarativa. Os mediadores encapsulam a representação das múltiplas fontes de dados para esta linguagem de consulta, fornecendo acesso uniforme. Cabe a estes mediadores resolver conflitos envolvendo representação de conhecimento diferentes como modelos de dados e esquemas, e conflitos devido a diferenças no poder de processamento de cada fontes de dados.

3. Processamento de Consultas

Existem diferentes abordagens para lidar com o problema de processamento de consultas envolvendo SGBDs heterogêneos. Uma solução possível é o uso de sistemas de múltiplos bancos de dados, que suportam um modelo de dados comum e uma linguagem de consulta global acima dos diferentes tipos de sistemas de bancos de dados existentes. Estes sistemas de múltiplos bancos de dados utilizam um esquema global que é o resultado da integração dos esquemas exportados dos bancos de dados locais. Cada esquema de exportação, por sua vez, pode ser obtido através de uma transformação do esquema local.

A linguagem de consulta global deve ser utilizada para realizar consultas globais acessando o esquema global. Segundo [MY95], uma consulta global pode ser executada em três passos, a saber. Primeiro, ela é decomposta em sub-consultas de modo que os dados necessários para cada sub-consulta estejam disponíveis em uma base local. Vale ressaltar que, após a decomposição, as sub-consultas ainda estão na linguagem de consulta global. Em seguida, cada sub-consulta é traduzida para uma ou mais consultas e enviada para a bases de

dados correspondente. E por fim, no terceiro passo, os resultados provenientes das consultas são combinados na resposta.

Uma outra abordagem para a consulta aos dados dos sistemas de múltiplos bancos de dados é a criação de visões temporárias referentes às consultas de usuários [LMR90], ao invés de construir esquemas globais. Nesta abordagem, cabe à linguagem que acessa as bases fornecer possibilidade de definição das visões temporárias. Tanto a criação quanto a manutenção destas visões são responsabilidade do usuário.

A seguir serão descritas as diferentes etapas do processamento de consultas conforme apresentado em Meng e Yu [MY95], Sheth e Larson [SL90] e em diversos projetos recentes utilizando mediadores.

3.1 Processamento de Consultas em SGBDs Heterogêneos conforme [MY95]

O sistema de múltiplos bancos de dados traduz consultas ou atualizações globais de forma apropriada para cada sistema local, realiza o envio para os SGBDs locais que realizam o processamento, agrupa os resultados e gera o resultado final para o usuário. Além disto, cabe ao sistema coordenar a confirmação ou o cancelamento da transação global que pode ser uma consulta ou atualização.

As etapas do processamento de consultas em sistemas múltiplos bancos de dados envolvem decomposição de consultas, tradução de consultas e otimização global de consultas [MY95].

3.1.1 Decomposição

Quando uma consulta global é submetida, ela é decomposta em dois tipos de consulta por um decompositor de consultas. O primeiro tipo é chamado sub-consultas do esquema de exportação ou somente sub-consultaconsultas, enquanto que o segundo é chamado de consultas de pós-processamento.

A decomposição de consultas normalmente é realizada em duas fases: a modificação e em seguida a decomposição propriamente dita. A consulta global utiliza nomes globais que são modificados para nomes que somente pertençam aos esquemas de exportação. Este nomes podem referenciar mais de uma base de dados, tornando necessária a segunda fase: a decomposição da consulta. A transmissão de dados pode auxiliar a decomposição, pois dados podem ser enviados para nós específicos dependendo da consulta a ser realizada. Estas questões estão interligadas a otimização global de consultas.

Quanto à fase de modificação, alguns aspectos podem aumentar a complexidade a ser tratada, dentre eles: a linguagem de consulta global e o modelo de dados global, o método utilizado para integrar esquemas de exportação (generalização ou *outerjoin*), sobreposição entre diferentes SBDs, inconsistência de dados, diferenças semânticas ou outras incompatibilidades.

Em [MY95] o modelo de dados orientado a objetos é utilizado para apresentar exemplos. Quando não há sobreposição entre os SBDs nem inconsistência, a modificação ocorre de forma relativamente simples. A idéia principal é representar cada nome global como uma classe ou atributo do esquema global em nomes dos esquemas de exportação e, durante a modificação da consulta, substituir estas representações. Os autores apresentam discussões a respeito de modificações na presença de sobreposição e inconsistência, descrevendo um algoritmo de tra6es passos.

3.1.2 Tradução

Após a decomposição, cada sub-consulta necessita de dados pertencentes a apenas uma base. Entretanto estas sub-consultas estão expressas na linguagem de consulta global, que pode não ser a mesma de algum sistema local. Desta forma, é necessário utilizar um tradutor de consultas que será responsável por transformar a sub-consulta do esquema de exportação em uma sub-consulta local a um sistema específico.

O processamento e a dificuldade da tradução depende da sintaxe e da expressividade das linguagens de origem e destino. Se a linguagem de origem tem maior poder de expressividade, então algumas consultas não poderão ser traduzidas ou serão traduzidas com o auxílio de alguma linguagem de alto nível. Como exemplo, pode-se citar uma consulta recursiva a um SGBD orientado a objetos não pode ser traduzida em uma consulta relacional utilizando apenas SQL.

Para alcançar melhor desempenho através da otimização de consultas, diferentes técnicas podem ser utilizadas. Se, ao realizar a tradução, apenas uma consulta destino é gerada e o SGBD destino possui otimizador, então este será responsável pela otimização.

Entretanto, otimizadores atuais não conseguem realizar otimização em um conjunto de consultas considerando o custo de processamento total. É proposta uma otimização em duas fases: a seleção do conjunto mínimo de consultas destino corretas e a otimização individual.

Diversos estudos de tradução de consultas relacionais para consultas hierárquicas, em rede e orientadas e objeto são mencionadas. Além das publicações comentadas em [MY95], pode-se citar pesquisas atuais como [QR95, Qian96, Y+95].

3.1.3 Otimização Global de Consultas

A otimização de consultas globais em SGBDs heterogêneos está diretamente ligada a otimização de consultas globais em SGBDs distribuídos homogêneos. Entretanto, a aplicação direta dos algoritmos para sistemas distribuídos homogêneos só é possível com o cumprimento de alguns pre-requisitos, que nem sempre ocorrem. Pode-se citar alguns deles: ausência de inconsistência de dados, possibilidade de transmitir dados entre diferentes bases locais, e disponibilidade de informações do sistema local para o otimizador global, como estatísticas de cardinalidades e seletividade.

Para resolver problemas de falta de informação sobre entidades locais como cardinalidades, pode-se enviar consultas de amostragem para as bases locais com o objetivo de coletar e atualizar estatísticas sobre as bases locais, o que acrescentará custo a otimização global.

3.2 Processamento de Consultas em SGBDs Heterogêneos conforme [SL90]

Em [SL90], os autores descrevem a operação de processamento de consultas em três etapas chamadas: formulação de consultas, transformação de comandos e processamento e otimização de consultas.

3.2.1 Formulação de Consultas

Na primeira etapa, uma única linguagem pode ser utilizada para formular consultas em SGBDs heterogêneos. Como os autores trabalham com a arquitetura de sistemas federados com subdivisões de acoplamento forte e fraco, é importante comentar que, no caso de acoplamento forte, o uso de uma única linguagem é possível devido as transparências de localização, distribuição e replicação oferecidas. Já no caso de acoplamento fraco, o sistema federado oferece uma linguagem de acesso com funcionalidade adicionais. Nesta linguagem é possível definir esquemas federados como visões sobre os esquemas de exportação e formular consultas a estas visões. Além disto, a linguagem lida com problemas de integração de esquemas como conflitos de nomes e estruturas de dados.

3.2.2 Transformação de Comandos

A segunda etapa corresponde às atividades de um processador de transformação de comandos, que traduz os comandos de uma linguagem fonte em uma linguagem destino. Na época da publicação do artigo, os maiores interesses eram conversões de linguagens procedimentais em linguagens não procedimentais e vice-versa.

3.2.3 Processamento da Consulta e Otimização

O processamento de consultas envolve a conversão da consulta realizada a um esquema federado em uma consulta a esquemas de exportação. Em [SL90], os autores afirmam que a otimização de consultas pode ser muito explorada em sistemas fortemente acoplados, enquanto que em sistemas fracamente acoplados não há suporte adequado. Na próxima seção será possível verificar que atualmente existem propostas para resolver este problema.

O processamento de consultas em SGBDs heterogêneos é semelhante em SGBDs Heterogêneos, mas existem complexidades adicionais a serem tratadas. O custo de uma operação pode ser diferentes nos diversos bancos de dados componentes. Devido a autonomia de cada nó, o custo de executar uma operação em um determinado SGBD pode ser desconhecido ou apenas parcialmente conhecido. Além disto, este custo pode variar ao

longo do tempo. Outro fator dificultador é a possível diferença de capacidade de otimização de consultas de cada SGBD componente. E também há a diferença nas operações suportadas por cada SGBD componente, pois uma mesma operação pode ter implementações diferenciadas em cada um dos nós.

3.3 Processamento de Consultas em SGBDs Heterogêneos com uso de mediadores

Nas pesquisas mencionadas nos itens anteriores, o escopo principal era o processamento de consultas em SGBDs heterogêneos que faziam parte de um ambiente controlado, como por exemplo, uma empresa. Isto significa que a dinamicidade do ambiente era pequena, fazendo com que o número de SGBDs componentes aumentasse pouco.

Por outro lado, com o surgimento da Internet, uma quantidade de informação digital foi se tornando disponível um uma velocidade muito grande. Estas informações estão armazenadas em diferentes fontes de dados que podem ser tanto SGBDs de modelos diferentes, como páginas HTML ou até mesmo arquivos convencionais.

Quando os SBDs Distribuídos Heterogêneos aumentam sua quantidade de fonte de dados, surgem muitas questões importantes. O uso do sistema fica mais difícil tanto para usuários finais como para os programadores. A resposta a uma consulta requer que todas as fontes de dados envolvidas estejam disponíveis. A manutenção do sistema fica mais complexa para o administrador de banco de dados, pois para adicionar novas fontes de dados é necessário alterar esquemas, atualizar catálogos e adicionar definições. Para os implementadores de banco de dados, o sistema se torna mais complicado para programar e ajustar, uma vez que para cada nova fonte é preciso gravar novas informações de custo.

Surge, então, a necessidade de investigar as dificuldades e soluções para o acesso uniforme e otimizado a diversas fontes de dados utilizando uma linguagem de consulta (preferencialmente declarativa e única). É neste contexto que se inserem as pesquisas em processamento de consultas com o uso de mediadores a ser descrito a seguir.

Para que múltiplas fontes de dados possam ser acessadas de maneira uniforme, o mediador aceita uma consulta, transforma-as em sub-consultas que são distribuídas pelas fontes de dados. Quando as sub-respostas retornam, o mediador as combina gerando resposta final para a aplicação. Esta arquitetura permite que os mediadores sejam desenvolvidos de forma independente e possam ser combinados, fornecendo um mecanismo para lidar com a complexidade introduzida pelo crescente número de fontes de dados.

Para lidar com a heterogeneidade das fontes de dados, *wrappers* (tradutores) fornecem uma visão estruturada da fonte de dados e transforma sub-consultas do mediador na linguagem particular da fonte de dados. O tradutor possui as funcionalidades de transformar consultas para uma fonte de dados em particular e reformatar as respostas apropriadas para cada mediador. Eles contém informações sobre a estrutura do objeto desejado para integração e seu mapeamento para a fonte de informação. É possível ter vários tradutores para um fonte de dados, se múltiplos objetivos devem ser atendidos. É necessário

que o implementador de bancos de dados escreva tradutores para cada tipo de fonte de dados.

4. Processamento de Consultas em alguns Projetos de Pesquisa

Diversos projetos de pesquisa atuais utilizam a arquitetura básica de mediadores apresentada na seção anterior. Estes projetos abordam o processamento de consulta de formas diferentes, destacando tópicos específicos. Alguns projetos dão maior ênfase a problemas de escalabilidade, enquanto que outros abordam aspectos como tratamento de dados não estruturados. A seguir são apresentados tópicos de pesquisa em projetos recentes, dentro do contexto de processamento de consultas.

Os sistemas SIMS e IRO-DB suportam capacidade de mediador a partir de esquema global unificado que integra cada base de dados remota e resolve conflitos entre estas bases através deste esquema unificado. Estes projetos oferecem contribuições importantes na resolução de conflitos entre diferentes esquemas e modelos de dados. Entretanto, escalabilidade não foi tratada explicitamente e traz problemas, uma vez que o esquema integrado deve ser modificado substancialmente a medida que novas fontes de dados são integradas. Também não são considerados servidores com capacidade de consulta limitada.

No sistema SIMS, o compartilhamento de informações proveniente de diversos esquemas relacionais é facilitado através do uso do esquema de representação de conhecimento para construir o esquema global para cada domínio de aplicação. Aqui, a linguagem de consulta é LOOM. O projeto SIMS desenvolve um servidor de conhecimento que realiza as funções de mediador entre fontes de informação (bases de dados e de conhecimento) e aplicações. Ele fornece acesso de forma independente da organização da informação, da linguagem de consulta utilizada e da localização das fontes de dados. Na fase inicial do projeto, SIMS lidou com casos onde o conhecimento era completo, pois o modelo da informação encaixava com o dado disponível. Entretanto, redes não confiáveis podem causar a indisponibilidade de determinadas fontes. Além disto, fontes com dados não estruturados ou semi-estruturados existem e são difíceis de modelar. Em SIMS II, a abordagem do projeto foi estendida para lidar com incompletude tanto de modelo, quanto de acesso [AKH96, AKS96]. Foi construído um mediador flexível que recebe consultas a nível de domínio de aplicação e seleciona dinamicamente as fontes de dados, com base em seus conteúdos e disponibilidade. O plano de execução da consulta é gerado especificando as operações e a otimização semântica é executada para minimizar o tempo final de execução.

Os principais focos do projeto TSIMMIS são: integração de fontes de dados estruturados a não estruturados, técnicas para prototipação rápida de tradutores [P+95]e técnicas para implementação de mediadores [PGU96]. O modelo de dados comum é o modelo dinâmico auto-descritivo chamado Object Exchange Model (OEM) [PGW95] utilizado para troca de informações baseado em objetos com um especificação simples. Uma linguagem de consulta correspondente, LOREL, é proposta. TSIMMIS possui componentes que extraem propriedades de objetos não estruturados, transformam informação em um

modelo comum, combinam informação de diversas fontes [PGA96], permitem navegação de informação e gerência de restrições através de nós heterogêneos. O aspecto de diferença na capacidade de processamento de consulta das diferentes fontes de dados é tratado e são propostas técnicas para reformulação de consultas visando resolver este problema. Para a prototipação rápida de tradutores, são descritas técnicas de transformação de consultas [P+95].

No projeto DISCO, o componente de busca de informações distribuídas (*Distributed Information Search COmponent*) é um Sistema de Banco de Dados Distribuído Heterogêneo que acessa diferentes fontes de dados [TRV96, T+97]. O projeto utiliza a arquitetura de mediadores e tradutores para pesquisar problemas relativos a diferenças de funcionalidades de cada fonte de dados [KTV97], variações no custo das operações nas fontes de dados heterogêneas [NGT97] e falhas em consultas devido a fontes de dados não disponíveis [BT97]. Para a questão das diferentes funcionalidades foi definida uma interface que permite definir a capacidade de cada fonte de dados. Quanto aos custos variados, o uso do paradigma de orientação a objetos permite que informações atualizadas sobreponham informações obsoletas para permitir otimização da consultas. E a respeito das fontes de dados indisponíveis, uma solução proposta é permitir que as consultas retornem "respostas parciais", compostas de parte da resposta final produzida pelas fontes de dados disponíveis e uma nova consulta representando as partes realizadas e não realizadas da consulta original. Desta forma, a resposta a consulta pode ser outra consulta.

Pesquisas em reformulação de consultas através de conhecimento semântico podem ser encontradas em [FRV96]. Contrastando com o esquema global unificado que resolve todos os conflitos entre entidades do esquema local, estas pesquisas assumem um ambiente de mediador baseado no modelo de dados comum. Em [FRV96], o modelo de dados comum é o modelo de objetos padrão ODMG, que estende o modelo de dados orientado a objetos OMG. Conhecimento semântico expressa o mapeamento entre a descrição da interface do mediador e as descrições locais correspondentes a cada base de dados local. O conhecimento semântico é expresso em equivalências onde cada consulta é expressa usando a linguagem OQL. Conhecimento semântico inclui: mapeamento do conhecimento na forma de consultas que são visões sobre a união das interfaces do mediador e das fontes de dados; equivalências expressando restrições de integridade nas interfaces locais e do mediador; e equivalências expressando replicação de dados nas interfaces locais. Todas estas equivalências são utilizadas para reformulação de consultas. Também é possível rescrever uma consulta usando visões que são materializadas na interface do mediador.

5. Conclusões

Este estudo abordou as diferentes etapas do processamento de consultas em Sistemas Gerenciadores de Bases de Dados Heterogêneos. Diversas abordagens foram apresentadas assim como suas respectivas arquiteturas. Alguns projetos de pesquisa recentes utilizando mediadores foram comentados.

Atualmente as pesquisas no contexto de processamento de consultas em SGBDs heterogêneos enfrentam novos desafios relacionados a ambientes dinâmicos como a Internet. Não basta mais buscar a integração de SGBDs componentes através de um único esquema global. Como o crescimento do volume de informação é muito rápido, torna-se inviável buscar a tradução e integração dos diversos esquemas componentes em um único esquema global.

O uso de mediadores permite a modularização ao invés da centralização. Esta modularização é naturalmente suportada em ambientes distribuídos e presente nos ambientes computacionais atuais.

É necessário, portanto, buscar novas soluções para problemas tais como: escalabilidade, dados em formatos semi-estruturados e não estruturados, fontes de dados não disponíveis, diferenças nas funcionalidades de cada componente, dentre outros.

Referências Bibliográficas

[AKS96] Arenas, Y. Knoblock, C. Shen, W. "Query Reformulation for Dynamic Information Integration", *Jounal of Intelligent Information Systems*, vol. 6, pp. 99-130, 1996.

[AKH96] Arens, Y. Knoblock, C. Hsu C. "Query Processing in th SIMS Information Mediator", *Advanced Planning Technology, editor, Austin Tate,* AAAI Press, Menlo Park, California, 1996.

[BRP92] Bright, M. Huron, A. Pakzad, S. "A Taxonomy and Current Issues in Multidatabase Systems", *IEEE Computer*, v. 25, n. 2, pp. 50-60, março 1992.

[BRU97] Buneman, P. Raschid, L. Ullman, J. "Mediator Languages - a Proposal for a Standard", *SIGMOD Record*, v. 26, n. 1, pp. 39-44, março 1997.

[BT97] Bonnet, P. Tomasic, A. "Partial Answers for Unavailable Data Sources", *Technical Report RR-3127 INRIA*, Grenoble, França, março 1997.

[FRV96] Florescu, D. Raschid, L. Valduriez, P. "Answering Queries using oql view expressions", *Workshop on Materialized View: Techniques and Applications, with ACM SIGMOD International Conference on Management of Data,* Montreal, Canada, junho 1996.

[Khos93] Koshafian, S. *Object-Oriented Databases*, John Wiley&Sons, Inc., 1993.

[Kim95] Kim, W. *Modern Database Systems - The Object Model, Interoperability, and Beyond,* ACM Press, 1995.

[KTV97] Kapitskaia, O. Tomasic, A. Valduriez , P. "Dealing with Discrepancies in *Wrapper* Functionality", *Technical Report RR-3138 INRIA*, Rocquencourt, França, março 1997.

[LMR90] Litwin, W. Mark, L. Roussopoulos, N. "Interoperability of Multiple Autonomous Databases", *ACM Computing Surveys*, v.22, n.3, pp.267-293, setembro 1990.

[MY95] Meng, W. Yu, C. "Query Processing in Multidatabase Systems", cap. 27 em Kim, W. *Modern Database Systems - The Object Model, Interoperability, and Beyond,* ACM Press, 1995.

[NGT97] Naacke, H. Gardarin, G. Tomasic, A. "Leveraging Mediator Cost Models with Heterogeneous Data Sources", *Technical Report RR-3143 INRIA*, Rocquencourt, França, março 1997.

[SL90] Sheth, a. Larson, J. "Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases", *ACM Computing Surveys*, v.22, n.3, pp.183-234, setembro 1990.

[ÖV91] Özsu, M. Valduriez, P. "Principles of Distributed Database Systems", Prentice Hall, California, 1991.

[PGA96] Papakonstantinou, Y. Garcia-Molina, H. Abiteboul, S. "Object Fusion in Mediator Systems", *Proceedings of the 20th International Conference on Very Large Databases*, Bombay, India, setembro 1996.

[PGW95] Papakonstantinou, Y. Garcia-Molina, H. Widom, J. "Object Exchange across Heterogeneous Information Sources", *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, pp.251-260, março 1995.

[PGU96] Papakonstantinou, Y. Garcia-Molina, H. Ullman, J. "MedMaker: A Mediation System based in Declarative Specification", *Proceedings of the 12th International Conference on Data Engineering, ,* pp. 132-141, 1996.

[P+95] Papakonstantinou, Y. Gupta, A. Garcia-Molina, H. Ullman, J. "A Query Translation Scheme for Rapid Implementation of Wrappers", *Proceedings of the 4th International Conference on Deductive and Object-Oriented Databases*, Singapore, pp. 97-107, agosto 1995.

[QR95] Qian, X. Raschid, L. "Query Interoperation Among Object-Oriented and Relational Databases", *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, pp. 271-278, março 1995.

[Qian96] Qian, X. "Query Folding", *Proceedings of the 12th International Conference on Data Engineering*, pp. 48-55, 1996.

[TRV96] Tomasic, A. Raschid, L. Valduriez, P. "Scaling Heterogeneous Databases and the Design of Disco", *Proceedings of the 16th International Conference on Distributed Computer Systems, Nominated for Best Paper Award*, Hong Kong, pp. 449-457, 1996.

[T+97] Tomasic, A. Amouroux, R. Bonnet, P. Kapitskaia, O. Naacke, H. Raschid, L. "The Distributed Information Search Component (Disco) and The World Wide Web", *Prototype Demonstration Description in Proceedings of the ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, pp.546-548, maio 1997.

[Wied92] Wiederhold, G. "Mediators in the Architecture of Future Infomation Systems", *IEEE Computer*, v. 25, n. 2, pp. 38-49, março 1992.

[Wied95] Wiederhold, G. "Mediation in Information Systems", *ACM Computing Surveys*, v. 27, n. 2, pp.265-267, junho 1995.

[Y+95] Yu, C. Meng, W. Kim W. Tracy, G. Pham T. Dao, S "Translation of Object-Oriented Queries to Relational Queries", *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, pp. 90-97, 1995.