Integração de Banco de Dados

IN1128/IF694 – Bancos de Dados Distribuídos e Móveis

Ana Carolina Salgado – acs@cin.ufpe.br
Bernadette Farias Lóscio – bfl@cin.ufpe.br

Centro de Informática — Universidade Federal de Pernambuco — CIn.ufpe.br

Formas Básicas de BD Distribuído

- **Abordagem *top-down***
  - Adequada para bancos de dados homogêneos
- **Abordagem *bottom-up***
  - Adequada para sistemas de múltiplos bancos de dados

Abordagem *Top-down*

- **Projeto de um novo banco de dados**
  - Busca da redução dos custos e da melhoria do desempenho das aplicações
  - Limites da capacidade de um banco de dados centralizado atingidos
    - Armazenamento
    - Processamento
    - Tempo de acesso

Abordagem *Top-down* (Distribuição de Dados)

- **Particionar o BD em unidades lógicas (fragmentos)**
  - Fragmentação
    - É um processo para o particionamento de BD em banco de dados menores sem replicação
    - O resultado é um conjunto de fragmentos de relações, os quais devem ser alocados aos diferentes bancos de dados locais
  - Replicação
    - Útil para melhorar a disponibilidade de dados
      - Replicação Total x Replicação Parcial

Abordagem *Bottom-up*

- **Processo de integração de bancos de dados já existentes e independentes**
  - Ambientes onde proliferaram os bancos de dados departamentais e individuais
  - Integração gera bancos de dados globais virtuais
  - Preserva os investimentos já feitos em aplicações, possibilitando ainda uma visão integrada dos dados fisicamente dispersos

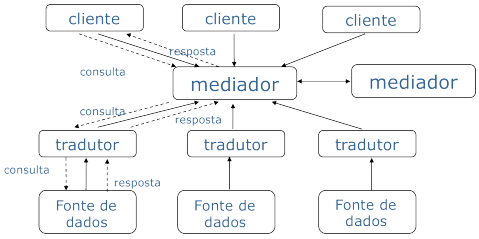
Abordagem *Bottom-up*

- **Razões**
  - Sistemas Legados
    - Organizações possuindo sistemas desenvolvidos há algum tempo (até 1990, aproximadamente)
    - Tecnologias aplicadas bastante diferentes das atuais
  - Migração entre Plataformas
    - Evolução de hardware e software disponibilizando novas facilidades
  - Mudanças Organizacionais
    - Fusão de duas ou mais empresas
      - Visão integrada da nova organização
  - Evolução da Tecnologia de Redes de Comunicação
    - Integração dos ambientes computacionais anteriormente isolados

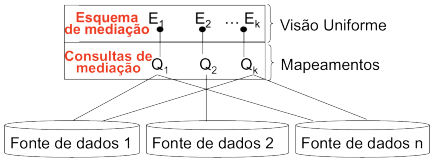
Abordagem *Bottom-up*

- **Integrar os bancos de dados e esquemas locais existentes para um único banco de dados global com seu respectivo esquema global**
- **Principais arquiteturas**
  - Mediadores
  - Data Warehouse

Arquitetura de Mediadores



Arquitetura de Mediadores



Arquitetura de Mediadores

- **O mediador integra dados de diferentes fontes dedados**
- **O tradutor é um software que converte os dados das fontes de dados para o modelo de dados comum e converte consultas de aplicações em consultas específicas da fonte de dados correspondente**
  - Um tradutor provê uma interface única para uma ou mais fontes de dados que tenham uma interface comum de acesso
  - Fontes de dados baseadas em um mesmo modelo de dados e em uma mesma linguagem de acesso são associados a um mesmo tradutor

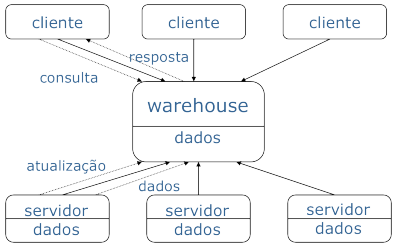
Abordagem Virtual

- **Os dados são mantidos nas fontes e as consultas são decompostas em tempo real e submetidas às diversas fontes**
  - Vantagem: os dados não são replicados e tem-se a garantia de estarem atualizados no momento da consulta
  - Desvantagem: como as fontes de dados são autônomas, são necessários métodos para otimização de consultas para garantir uma performance adequada

Abordagem Virtual

- **O Enfoque Virtual é mais apropriado para as seguintes situações:**
  - o número de fontes é grande
  - os dados são atualizados frequentemente
  - existe pouco controle sobre as fontes de dados

Arquitetura com Data Warehouse



Abordagem Materializada

- **Os dados gerados das diversas fontes são carregados em um repositório (warehouse) e as consultas são aplicadas a estes dados.**
  - Vantagem: performance garantida no momento da consulta
  - Desvantagem: atualização do repositório sempre que houver mudança nos dados

Metodologia *Bottom-up*

- **O esquema global pode ser definido a partir dos requisitos dos usuários ou pode ser obtido a partir da integração dos esquemas locais**
- **Integração de esquemas**
  - É o processo de gerar um esquema integrado (global) a partir de um conjunto de esquemas de entrada (esquemas locais) resolvendo as diversidades estruturais e semânticas existentes entre eles

Integração de Esquemas - Heterogeneidade

- **Uma organização pode ter múltiplos SGBDs e diferentes departamentos**
  - Dentro de uma mesma organização podemos ter diferentes requisitos para um mesmo conjunto de dados
- **Diferentes perspectivas podem levar a diferentes representações (estrutura e restrições) e interpretações de um mesmo dado**
- **Diferentes modelos de dados oferecem primitivas diferentes para a modelagem dos dados**
  - Exemplo: um endereço pode ser modelado como uma entidade em um esquema e como um atributo composto em outro esquema

Integração de Esquemas - Heterogeneidade

- **Problema crítico!**
  - Heterogeneidade terminológica
    - Termos diferentes usados para representar os mesmo conceitos
    - Termos iguais usados para representar conceitos diferente
  - Heterogeneidade estrutural
    - Conceitos similares representados através de construtores divergentes

Integração de Esquemas - Heterogeneidade

- **Heterogeneidade terminológica - Exemplo**
  - Entidades semanticamente relacionadas podem ser representadas com nomes diferentes em bancos de dados distintos

Sinonímia

Esquema	Definição
S1	medico1 (num_conselho, nome, especialidade)
S2	doutor2 (CRM, nome)

Integração de Esquemas - Heterogeneidade

- **Heterogeneidade terminológica - Exemplo**
  - Dois atributos são semanticamente semelhantes e apresentam nomes diferentes (sinônimos) ou quando atributos com mesmo nome podem não ter qualquer relação semântica (homônimos)

Sinonímia

Esquema	Definição
S1	medico1 (num_conselho, nome, especialidade)
S2	doutor2 (CRM, nome)

Homonímia

Esquema	Definição
S1	medico1 (num_conselho, nome, especialidade)
S2	enfermeiro2 (num_conselho, nome)

Integração de Esquemas - Heterogeneidade

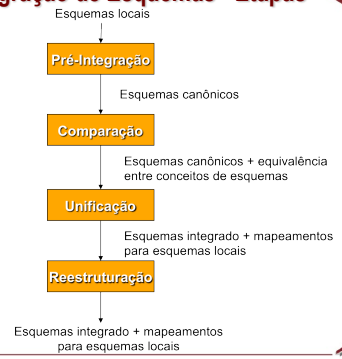
- **Heterogeneidade estrutural - Exemplo**
  - Duas entidades são representadas em dois bancos de dados com níveis diferentes de abstração (generalização)

Esquema	Definição
S1	enfermeiro1 (num_conselho, nome)
S1	medico1 (num_conselho, nome, especialidade)
S2	agente-saude2 (num_conselho, nome, tipo_agente)

Integração de Esquemas

- **Ferramentas de integração de esquemas são semi-automáticas**
- **Sugerem conceitos candidatos a serem integrados**
  - O especialista aceita ou rejeita

Integração de Esquemas - Etapas



Pré-Integração

- **Nesta fase os esquemas locais são traduzidos para o modelo de dados comum (ex:modelo ER)**
- **Análise dos esquemas locais a fim de definir a política de integração, por exemplo, a ordem a ser seguida na integração de esquemas, se a integração é parcial ou total**

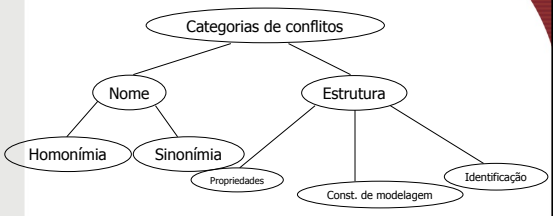
Tradução de Esquemas

- **Um sistema de bancos de dados múltiplos deve oferecer suporte para a tradução entre modelos de dados locais e globais**
- **Quando bancos de dados heterogêneos são integrados os esquemas locais são traduzidos para um modelo de dados comum**

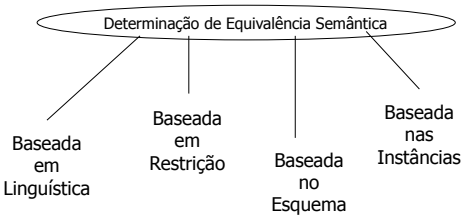
Comparação

- **Identificação de correspondências e conflitos entre esquemas diferentes através da análise de suas características como atributos, relacionamentos e restrições de integridade**
- **O resultado final desta etapa é um conjunto de correspondências entre os conceitos dos esquemas**

Comparação



Comparação



Comparação

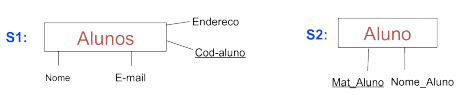
- **Diferentes técnicas podem ser aplicadas na determinação de equivalências semânticas entre esquemas**
- **Baseada em linguística: considera afinidade de nomes ou descrições textuais entre conceitos**
  - Utiliza informações adicionais fornecidas por um dicionário

Comparação

- **Baseada em restrição: tem como base características estruturais e de domínio entre conceitos e suas propriedades**
- **Baseada no esquema: apenas informações sobre os esquemas são consideradas**
- **Baseada nas instâncias: usa informações baseadas no conteúdo dos dados**

Comparação - Identificação das correspondências

- **Dois elementos de dois bancos de dados são equivalentes se eles descrevem o mesmo objeto do mundo real**

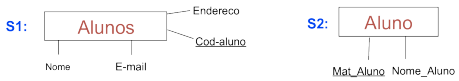


Comparação - Identificação das correspondências

- De tipos
- De atributos
- De caminho

Comparação - Identificação das correspondências

- Entre elementos correspondentes com estruturas semelhantes



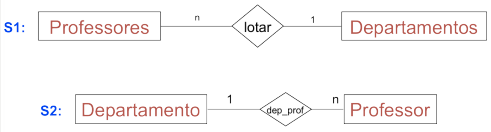
Comparação - Identificação das correspondências

- Correspondência de Entidades

S1.Cursos ≅ S2.Curso

Comparação - Identificação das correspondências

- Correspondência de Relacionamento



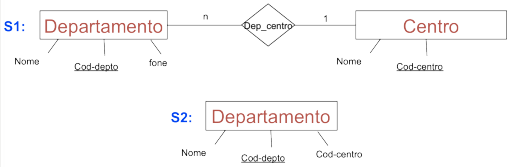
S1.lotar ≅ S2.dep_prof

Comparação - Identificação das correspondências

- Correspondência de Atributos
- S1.Cursos.Cod-curso ≅ S2.Curso.cod_curso
- S1.Cursos.Grade ≅ S2.Curso.desc_curso

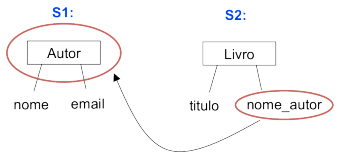
Comparação - Identificação das correspondências

- Correspondência de Caminhos
- S1.Departamento-Dep_centro-Centro-Cod-centro ≅ S2.Departamento-Cod-centro



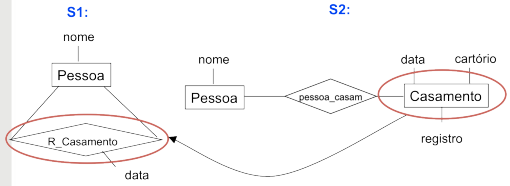
Comparação - Identificação das correspondências

- **Entre elementos correspondentes com estruturas diferentes**
  - Entidade e Atributo
  - **S1.Autor ≃ S2.Livro.nome_autor**



Comparação - Identificação das correspondências

- **Entidade e Relacionamento**
- **S1.RCasamento ≃ S2.Casamento**



Comparação - Identificação das correspondências

Passo 1: Identificação de correspondências entre elementos com estruturas correspondentes

1.1 Identificação das entidades correspondentes
- 1.1.1 Identificação dos atributos correspondentes
- 1.1.2 Especificação dos identificadores
- 1.1.3 Identificação da correspondência entre as extensões

Comparação - Identificação das correspondências

(cont.)

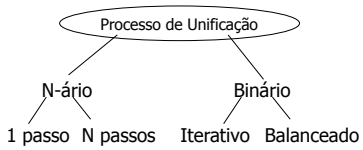
1.2 Identificação dos relacionamentos correspondentes
- 1.2.1 Identificação dos atributos correspondentes
- 1.2.2 Identificação das entidades participantes correspondentes
- 1.2.3 Identificação da correspondência entre as extensões

Passo 2: Identificação de correspondências entre elementos com estruturas diferentes
- Ex: Entidade/Relacionamento e Entidade/Atributo

Unificação

- **Resolução de conflitos entre os conceitos dos esquemas locais e produção do esquema integrado ou global**
- **Informações de mapeamento entre os conceitos dos esquemas locais e os conceitos do esquema global são identificadas**
  - Estes mapeamentos são importantes para a transformação de consultas

Unificação



Unificação

- **Binária: a unificação ocorre entre pares de esquemas locais**
  - Redução da complexidade de comparação e unificação em cada passo
  - Iterativa: unifica-se seqüencialmente pares de esquemas
  - Balanceada: unifica-se em paralelo pares de esquema

Unificação

- **N-ária**
  - A unificação ocorre entre mais de dois esquemas, podendo ser realizada em um ou vários passos
  - Redução do número de passos de unificação

Unificação

- **Os seguintes critérios devem ser atendidos:**
  - Completude e corretude: o esquema global deve conter todos os conceitos dos esquemas locais de forma semanticamente correta
  - Minimalidade: um conceito representado em mais de um esquema local deve ser representado uma única vez no esquema global
  - Compreensão: o esquema global deve ser fácil de ser compreendido (a representação mais clara deve ser escolhida)

Reestruturação

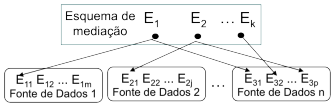
- **Nesta etapa o esquema global é analisado e reestruturado para remover redundâncias**
- **Ao final do processo, são gerados:**
  - Esquema global
  - Mapeamentos entre o esquema global e os esquemas locais

Abordagens para definição dos mapeamentos entre o esquema global e os esquemas locais

- **Visão Global (Global as View - GAV)**
- **Visão Local (Local as View - LAV)**
- **Abordagens que combinam características de GAV e LAV**
  - GLAV
  - BAV (Both as view)

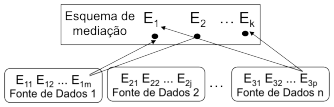
Abordagens para definição dos mapeamentos

- **Visão Global**
  - Cada elemento no esquema de mediação é definido como uma visão sobre os esquemas locais
    - A reformulação de consultas torna-se mais simples
    - Não é adequada para sistemas em evolução



Abordagens para definição dos mapeamentos

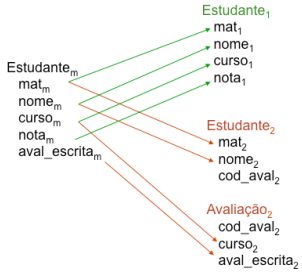
- **Visão Local**
  - Cada elemento em uma fonte local é definido como uma visão sobre o esquema de mediação
    - Torna mais fácil a manutenção das fontes de dados
    - O processo de decomposição de consultas é mais complexo



Abordagens para definição dos mapeamentos – Exemplo

- **Fonte de dados1:**
  - $Estudante_1(mat_1, nome_1, curso_1, nota_1)$
- **Fonte de dados2:**
  - $Estudante_2(mat_2, nome_2, cod\_aval_2)$
  - $Avaliação_2(cod\_aval_2, curso_2, aval\_escrita_2)$
- **Esquema de mediação:**
  - $Estudante_m(mat_m, nome_m, curso_m, nota_m, aval\_escrita_m)$

Exemplo – Visão Global



Exemplo – Visão Global

- **Semântica dos mapeamentos - Fonte 1**

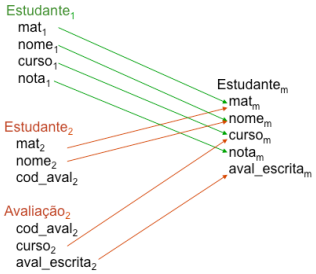
Para cada **e** em $\mathbf{Estudante_m}$
Existe $\mathbf{e_1}$ em $\mathbf{Estudante_1}$
Com $\mathbf{e_1.mat_1 = e.mat_m}$ e
$\mathbf{e_1.nome_1 = e.nome_m}$ e
$\mathbf{e_1.curso_1 = e.curso_m}$ e
$\mathbf{e_1.\ nota_1 = e.nota_m}$

Exemplo – Visão Global

- **Semântica dos mapeamentos - Fonte 2**

Para cada **e** em $\mathbf{Estudante_m}$
Existe $\mathbf{e_2}$ em $\mathbf{Estudante_2}$ e $\mathbf{a_2}$ em $\mathbf{Avaliação_2}$
Onde $\mathbf{e_2.cod\_aval = a_2.cod\_aval}$
Com $\mathbf{e_2.mat_2 = e.mat_m}$ e
$\mathbf{e_2.nome_2 = e.nome_m}$ e
$\mathbf{a_2.curso_2 = e.curso_m}$ e
$\mathbf{a_2.aval\_escrita_2 = e.aval\_escrita_m}$

Exemplo – Visão Local



Exemplo – Visão Local

- **Semântica dos mapeamentos – Fonte 1**

Para cada e_1 em **Estudante**$_1$
Existe **e** em **Estudante**$_m$
Com **e.mat**$_m$ = **e**$_1$**.mat**$_1$ e
e.nome$_m$ = **e**$_1$**.nome**$_1$ e
e.curso$_m$ = **e**$_1$**.curso**$_1$ e
e.nota$_m$ = **e**$_1$**. nota**$_1$

Exemplo – Visão Local

- **Semântica dos mapeamentos - Fonte 2**

Para cada e_2 em **Estudante**$_2$ e a_2 em **Avaliação**$_2$
Onde **e**$_2$**.cod_aval** = **a**$_2$**.cod_aval**
Existe **e** em **Estudante**$_m$
Com **e.mat**$_m$ = **e**$_2$**.mat**$_2$ e
e.nome$_m$ = **e**$_2$**.nome**$_2$ e
e.curso$_m$ = **a**$_2$**.curso**$_2$ e
e.aval_escrita$_m$ = **a**$_2$**.aval_escrita**$_2$

O uso de mapeamentos no processo de reescrita de consultas

- **Os mapeamentos especificam como obter os elementos do esquema de mediação**
- **A partir dos mapeamentos podem ser geradas expressões de consulta que facilitam o processo de reescrita de consultas**

O uso de mapeamentos no processo de reescrita de consultas

- **Passos do processo de execução de uma consulta**
  - Identificar quais relações estão sendo consultadas
  - Descobrir as definições das relações (visões)
  - Reformular e submeter a consulta para as fontes locais
  - Integrar os resultados

Reescrita de Consultas - Exemplo

Consulta 1 - Esquema de mediação
For **e** in **Estudante**$_m$
Return (**e.mat**$_m$, **e.nome**$_m$, **e.curso**$_m$, **e.nota**$_m$, **e.aval_escrita**$_m$)

Consulta 1 - Fonte de Dados1
For e_1 in **Estudante**$_1$
Return (**e**$_1$**.mat**$_1$, **e**$_1$**.nome**$_1$, **e.curso**$_1$, **e**$_1$**.nota**$_1$)

Estudante$_m$.mat$_m$ ≡ Estudante$_1$.mat$_1$
Estudante$_m$.nome$_m$ ≡ Estudante$_1$.nome$_1$
Estudante$_m$.curso$_m$ ≡ Estudante$_1$.curso$_1$
Estudante$_m$.nota$_m$ ≡ Estudante$_1$.nota$_1$

Reescrita de Consultas - Exemplo

Consulta 1 - Esquema de mediação
For **e** in **Estudante**$_m$
Return (**e.mat**$_m$, **e.nome**$_m$, **e.curso**$_m$, **e.nota**$_m$, **e.aval_escrita**$_m$)

Consulta 1 - Fonte de Dados2
For e_2 in **Estudante**$_2$, a_2 in **Avaliacao**$_2$
Where **e**$_2$**.cod_aval** = **a**$_2$**.cod_aval**
Return (**e**$_2$**.mat**$_2$, **e**$_2$**.nome**$_2$, **e**$_2$**.curso**$_2$, **e**$_2$**.cod_aval**$_2$)

Estudante$_m$.mat$_m$ ≡ Estudante$_2$.mat$_2$
Estudante$_m$.nome$_m$ ≡ Estudante$_2$.nome$_2$
Estudante$_m$.curso$_m$ ≡ Estudante$_2$.cod_aval$_2$.Avaliação$_2$.curso$_2$
Estudante$_m$.aval_escrita$_m$ ≡ Estudante$_2$.cod_aval$_2$.Avaliação$_2$.aval_escrita$_2$

Integração de dados

- **Os resultados das consultas locais devem ser integrados para produzir a visão integrada solicitada pelo usuário**
- **Conflitos de dados podem surgir durante o processo de integração e devem ser tratados corretamente**

Integração de dados

- **Tipos de conflito:**
  - Valores de dados
    - Dois atributos semanticamente equivalentes com valores distintos
  - Tipos de dados
    - O atributo CPF é representado como string em uma fonte e como inteiro em um outra fonte
  - Escala
    - O atributo área é definido em m^2 em uma base de dados e em cm^2 em uma outra base

Integração de dados

- **Tipos de conflito:**
  - Precisão dos dados
    - Os atributos nota e conceito (um conceito pode corresponder a um conjunto de notas)
  - Valor default
    - O atributo maioridade pode ter como valor default 18 ou 21 anos
  - Restrições
    - Em uma fonte de dados o intervalo de valores aceitáveis para a freqüência cardíaca em um ser humano é 60-100 bpm (batimentos por minuto), em outra fonte o intervalo é 50 a 110bpm

Integração de dados

- **Funções de transformação podem ser usadas para resolver os conflitos de dados**
- **Exemplos:**
  - converteDataParaFormatoPadrao(string formato, string data)
  - converteParaMetro(string unidade, int valor)
  - converteNotaParaConceito(double nota, string conceito)