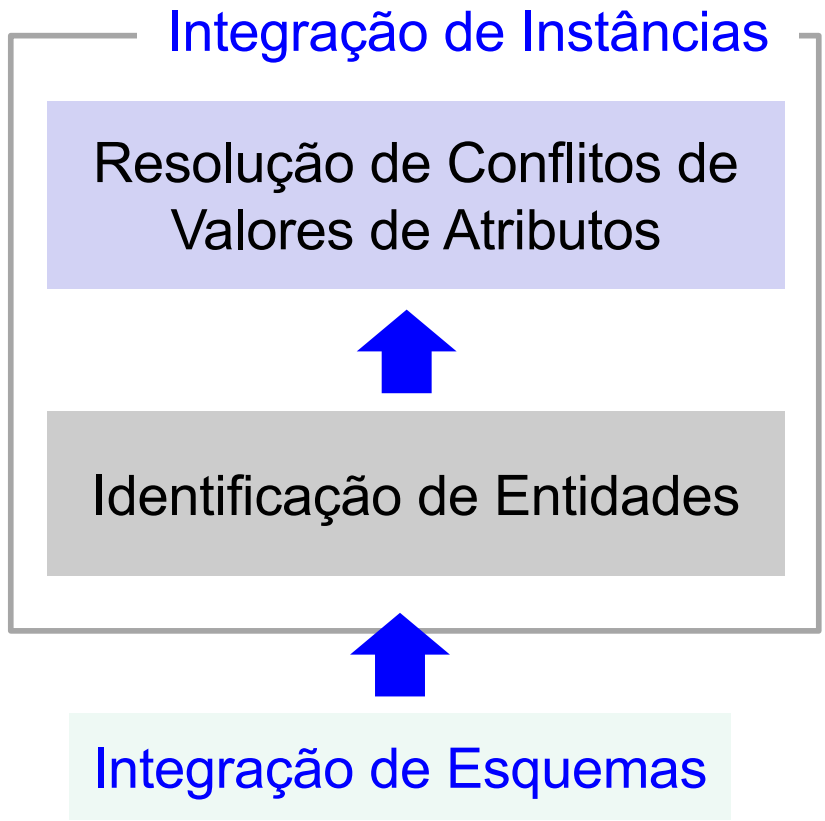
Processo ETL

Processamento Analítico de Dados

Profa. Dra. Cristina Dutra de Aguiar

Etapas

- Extração
- Transformação
  - Tradução
  - Limpeza
  - Integração
- Carregamento



Extração

- Extração de dados de interesse a partir das fontes de dados
- Tarefas
  - quais dados são extraídos de quais fontes
  - como esses dados são extraídos
  - com qual frequência esses dados devem ser periodicamente extraídos
  - qual técnica empregar para identificar dados das fontes que foram alterados

Tradução

- Conversão dos dados do formato nativo das fontes de dados para o formato do *data warehousing*
- Engloba aspectos de integração
  - transformação de esquema
  - transformação de instâncias

Limpeza

- Visa a produção de dados corretos e de qualidade
- Exemplos
  - comprimentos de campos inválidos
  - dados incompletos ou em branco,
  - violação de restrições de integridade
  - associações de valores inconsistentes
  - abreviações de valores não padronizadas
  - ...

Integração de Dados

- Problema: dados armazenados nos provedores
  - são heterogêneos
    - seguem diferentes modelos de dados
    - são representados por conceitos diferentes
    - possuem diferentes formatos
    - etc
  - são redundantes, inconsistentes e até mesmo complementares
- Dois níveis: esquema e instância

Integração de Esquemas

- Definição
  - especificação de mapeamentos que descrevem os relacionamento semânticos entre os esquemas dos provedores heterogêneos
- Relativismo semântico
  - conflitos entre duas ou mais representações é relacionado ao fato de que diferentes usuários modelam o mesmo pedaço do mundo real de diferentes formas, de acordo com as suas percepções

Conflito

- Conflito entre duas representações do mesmo conceito pertencentes a esquemas distintos
  - surge quando essas representações não são idênticas
- Representações idênticas
  - usam os mesmos construtores
  - aplicam as mesmas restrições de integridade

Conflito

- Tipos de conflito
  - de nome
  - semântico
  - estrutural

discrepâncias existentes entre os esquemas apresentam mais do que um tipo de conflito

Conflito de Nome

- Definição
  - nomes que representam os diferentes elementos nos esquemas a serem integrados
- Sinônimos
  - diferentes nomes são aplicados ao mesmo elemento
  - ex.: cliente representa, em um esquema, os clientes atendidos por uma loja, enquanto que comprador é usado em outro esquema para representar o mesmo caso

Conflito de Nome

- Homônimos
  - mesmo nome é aplicado a diferentes elementos
  - ex.: nome representa, em um esquema, o nome de um aluno de uma universidade, enquanto que nome representa, em outro esquema, o nome de um produto vendido em uma loja

Conflito Semântico

- Definição
  - surge quando o mesmo elemento é modelado em diferentes esquemas, porém representando conjuntos que se sobrepõem
- Exemplo
  - produto representa, em um esquema, todos os produtos de um supermercado, enquanto que produto é usado em outro esquema para representar apenas os produtos da seção de cosméticos

Conflito Estrutural

- Definição
  - surge sempre que diferentes construtores estruturais são utilizados para modelar o mesmo conceito representado em diferentes aplicações
- Exemplo:
  - o mesmo conjunto de objetos do mundo real pode ser representado como um tipo-entidade em um esquema e como um atributo de um tipo-entidade em outro esquema

Integração de Instâncias

- Tipos
  - ambiguidade na identificação de entidades
    - também conhecido como resolução de entidades, reconciliação de referências e deduplicação de dados
  - resolução de conflitos de valores de atributos
    - também conhecido como fusão de dados

Ambiguidade na Identificação de Entidades

- Objetivos
  - identificar quais entidades dos provedores heterogêneos referem-se à mesma entidade do mundo real
  - agrupar essas entidades em agrupamentos de entidades similares

Resolução de Conflitos de Valores

- Objetivo
  - resolver inconsistências nos valores dos dados das entidades que referem-se à mesma entidade do mundo real, mas que diferem nos valores dos seus atributos

Estratégias

- Objetivo
  - resolver (ou não) os conflitos de valores
- Tipos de estratégia
  - ignorar o conflito
  - evitar o conflito
  - resolver o conflito

BLEIHOLDER, J.; NAUMANN, F. Conflict Handling Strategies in an Integrated Information System. In Proceedings of the International Workshop on Information Integration on the Web, 2006

Ignorar o Conflito

- Objetivo
  - não decidir nada
- Característica
  - não se tem conhecimento de que o conflito exista
- Exemplo
  - PASS IT ON: mostra todos os valores conflitantes ao usuário ou a uma aplicação e deixa o usuário ou a aplicação decidir como resolver os conflitos

Evitar o Conflito

- Objetivo
  - não decidir nada, mas manipular o conflito
- Característica
  - decisão de como manipular o conflito é tomada previamente, sem análise dos dados
- Exemplo
  - TRUST YOUR FRIENDS: permite definir a confiabilidade de cada provedor, e usa as confiabilidades para resolver o conflito

Resolver o Conflito

- Objetivo
  - resolver o conflito
- Característica
  - usa como base os valores de dados
- Exemplos
  - CRY WITH THE WOLVES: escolhe o valor reportado pela maioria dos provedores
  - MEET IN THE MIDLE: escolhe um novo valor, o qual é um valor mediano reportado pelos provedores

Carregamento (Carga)

- Armazenamento dos dados pré-processados no DW
- Inclui processamentos adicionais
  - geração de agregações (ou visões materializadas)
  - construção de índices
  - preenchimento de campos faltantes
  - geração de chaves primárias artificiais
  - ...