Integração de Dados

Prof. Álisson R. Arantes

PUC Minas Virtual

Introdução

- Fontes de dados úteis e disponíveis nos mas variados formatos
- Necessidade da integração de dados
- Processo de integração: combinar dados de diversas fontes de dados (independente de formato e local) para armazenamento em um repositório de destino

Introdução

- Objetivo: tornar os dados disponíveis para os usuários, tipicamente para o processo de tomada de decisão.
- Antes: disponíveis apenas fontes de dados internas
- Hoje: tecnologia para utilização de qualquer conjunto de dados, desde que seja possível acessá-lo

Introdução

- Tipos de fontes de dados:
  - arquivos de texto
  - planilhas eletrônicas
  - bancos de dados transacionais
  - bancos de dados NoSQL
  - arquivos em formato CSV ou JSON
  - dados abertos disponíveis na nuvem
  - dados de redes sociais
  - dados de sensores IoT
  - ...

Introdução

- Levantamento dos requisitos de integração:
  - identificação das fontes de dados disponíveis (internas ou externas)
  - identificação dos dados que serão integrados
  - definição da periodicidade da extração dos dados

Introdução

- Dados gerados e manipulados pelas aplicações da organização se complementam quando integrados com outras fontes de dados

ETL (Extract, Transform, Load)

- Abordagem mais natural para o processo de integração
- Dados extraídos das fontes originais para carga no destino em um ambiente integrado
- Entre a extração e a carga: transformação

ETL (Extract, Transform, Load)

- Extrair, Transformar, Carregar
- Responsável pela conversão dos dados do ambiente operacional para o suporte à decisão
- Lê dados de uma ou mais fontes, transforma o dado de forma que fique compatível com o destino e faz a carga de dados

Tarefas da Transformação

- Remoção de conteúdo desnecessário:
  - remoção de tabelas, colunas ou linhas das fontes de origem que não serão utilizadas no destino

- Renomeação de atributos:
  - mesmo atributo armazenado em fontes diferentes com nomes diferentes (Matricula, Cod, Id, ...)

Tarefas da Transformação

- Adequação do tipo de dados, conforme os valores armazenados e as operações a serem realizadas sobre eles:
  - dados numéricos ou datas armazenados como cadeias de caracteres

Tarefas da Transformação

- Formatação adequada dos valores:
  - datas armazenadas em formatos diferentes (dd/mm/aaaa, aaaa/mm/dd, dd/mm/aa, ...)

- Padronização de valores:
  - o mesmo valor armazenado de formas diferentes em fontes diferentes (BH, BHTE ou Belo Horizonte)

Tarefas da Transformação

- Criação de novos atributos como colunas com valores calculados por alguma sumarização (soma, contagem, média, ...):
  - criação de um atributo ValorTotal a partir do produto entre Quantidade e ValorUnitário
  - criação de um atributo Intervalo a partir da diferença entre duas datas

Tarefas da Transformação

- Unificação de esquema:
  - o mesmo atributo codificado de diferentes formas em fontes diferentes (sexo codificado como ‘H’ e ‘M’ para homem e mulher em uma fonte ou como ‘M’ e ‘F’ para masculino e feminino em outra fonte)
- ...

Outras Abordagens

- ELT (Extract, Load, Transform):
  - carga dos dados no destino sem a etapa de transformação
  - transformação feita nos dados já carregados no destino
  - indicado para carga de enormes volumes de dados (Big Data)

Outras Abordagens

- ELTL (Extract, Load, Transform, Load):
  - carga dos dados em algum meio de armazenamento escalável e de baixo custo
  - transformação e nova carga dos dados para apresentação
  - variedade de fontes de dados utilizadas com objetivos diversos (data lake)

PUC Minas
Virtual