

Data integration

Why database integration?

- Combine database when two companies merge
- enhance information using data from different sources
- access data in legacy databases

Related Issues

- Data linkage
  - Identifying records referring to the same real-world entity
  - Computing data similarity
  - Applicability of different similarities
- Info fusion
  - Extract info from different data source
- Data cleaning
  - Remove noise in original data
  - Remove noise in integrated data

inconsistency and redundancy
- Data quality
  - Data augmentaion
  - Data constraints
  - Data provenance
- Data privacy
  - Share data with the assurance that private info can not be derived

Difference between Data integration and Info fusion

- Common goal: Integrate and organize data from multiple sources in order to present a unified view of data to derive actionable insights
- Data integration: focuses on combining data to create a bigger and consistent data set
- Info fusion: focuses on deriving insight from real-time streaming data with semantic context from other big data source

Global Info systems

- Federated database(FDB)
  - Semi-autonomous database systems
  - global view is provided
- Multi-databases(MDB)
  - Autonomous database systems, no or limited global view is provided
- Interoperable info systems
  - Definition
    - Only API provided to communicate with different database
    - No global virtual view
  - Interoperability
    - Ability for an application to access multiple distinct systems
    - Not necessarily for database only
  - Interoperable systems
    - Exchange messages and requests
    - Receive services and operate as a unit towards a common goal
  - Different types of interoperability
    - Syntactic
      - languages and data formats
    - Semantic
      - meaning
    - System
      - machines, networks, database models
    - Structural
      - data structures and data models

Challenges in DB integration

- Schema heterogeneity
- Type heterogeneity
- Value heterogeneity
- Semantic heterogeneity

Process of DB Integration

- Schema mapping
  - mapping of structures
- Data mapping
  - matching based on content
- Data fusion
  - reconciliation of mismatching content

