Information Integration

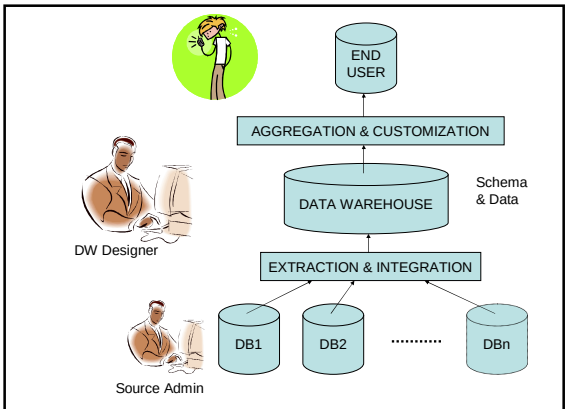
Modes of Information Integration

- Applications involved more than one database source
- Three different modes
  - Federated Databases
  - Data Warehouses
  - Mediation

Three Different Modes

- Federated Databases
  - Similar to multiple databases system – DBs are connected to each other but no global schema – Everything depends on application programmer
- Data Warehouses
  - Global schema for different sources – application programmers can develop/optimize the execution of queries issued to the data warehouse
- Mediator
  - A software program that provides users with a virtual global database – does not store data like data warehouse but interaction between users and data sources is similar to the data warehouse

Data Warehouse



Overview of Data Warehouse

- Components
- Designing issues
- Loading data from sources into DW
- Extracting data from DW
- Physical structure of DW
- Metadata management
- DW project management

Components

- Local data sources (schema & data)
- Global database (global schema & data)
- Users' requests

Designing Issues

- Top-down (design a global schema, then integrate data) vs bottom-up (develop smaller, specialized data marts)
- Changes in local schema need to be reflected in the global schema
- Integration issues
  - Syntactic difference
    - Data type
    - Value
  - Semantic difference
  - Missing value

Getting Data into DW

- Requirement: access to different information source
- Need: wrappers, loaders, mediators (programs that load data of the information sources into DW)
  - Wrappers & loaders: loading, transforming, cleaning, and updating the data from the source to the data warehouse
  - Mediators: integrate the data into the warehouse by resolving inconsistencies and conflict between different information souces

Getting Data into DW

- Wrappers, loaders, mediators: called Extract-Transform-Load Tools (ETL Tools) try to automate/support the following tasks:
  - Extraction
  - Cleaning
  - Transformation
  - Loading
  - Replication
  - Analyzing
  - High-speed data transfer
  - Checking for data quality
  - Analyzing metadata

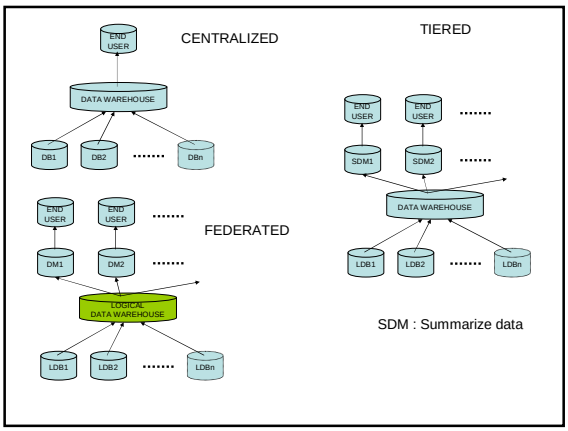
Extracting data from DW

- Basic OLAP operations
  - Roll up
  - Drill down
  - Slice and dice
  - Pivot
- Others
  - Report and query tools
  - GIS
  - Data mining
  - Decision support s
  - Executive information system
  - Statistics

Require SUPER relational DBMS

Physical structure of DW

- Centralized (single database, one location)
- Federated (single *virtual* database, several locations)
- Tiered (combined both centralized & federated)



Metadata management

- Provide the information necessary for accessing the data warehouse efficiently
- May include
  - Data dictionary (definitions of the dbs being maintained and the relationship between data elements)
  - Data flow (direction and frequency of the data feed)
  - Data transformation (transformations required when data is moved)
  - Version control (changes to the metadata are stored)
  - Data usage statistics (profile of data in the warehouse)
  - Alias information (alias names for a field)
  - Security (who is allowed to access the data)

DW project management

- Important issues
  - Design (no standard methodology as of now)
  - Technical (Hardware, Software)
  - Procedural (deployment: training end users!)

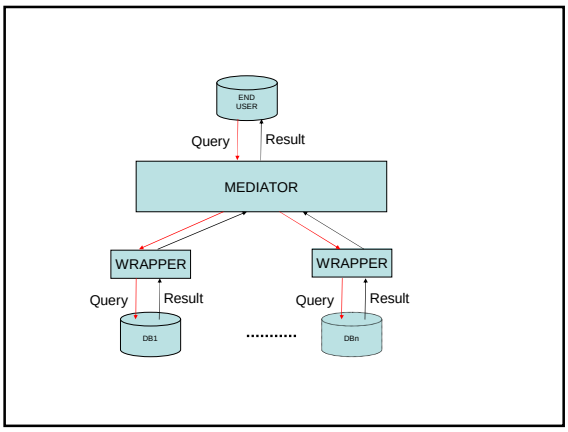
DW Research: Issues

Issues

- Data extraction and reconciliation (difficult to automate)
- Data aggregation and customization: requires a richer language for representation of hierarchies
- Query optimization
- Update propagation
- Modeling and measuring data warehouse quality (what can be used to measure the quality of a data warehouse?)
  - Accessibility (more accessible to users)
  - Interpretability (help users understand the data they get)
  - Usefulness (help users in their work)
  - Believability (trust of users in the data, makes data from less reliable source becomes more trustworthy)
  - Validation (how to ensure that all of the above quality issues have actually been adequately addressed)

Mediators

- A software program that provides users with a virtual global database – does not store data like data warehouse but interaction between users and data sources is similar to the data warehouse


END USER
Query
Result
MEDIATOR
WRAPPER
WRAPPER
Query
Result
Query
Result
DB1
DBn