ETL and Meta Data

• ETL
• Meta data
• External data
ETL\textsubscript{1}

- Data Extraction
  - Extract data from data source
    - Affecting current operational database?
  - Extraction methods
    - Incremental: for large and continuous data
    - Complete: for frequently changing data
  - Extraction time
    - How long?
  - Extraction period
    - How often?
ETL\textsubscript{2}

- Data Transformation and cleaning
  - From different sources, different format
    - Year/month/day
    - Month/day/year
    - Day/month/year
  - When?
    - During extraction
    - Loaded in files, then process by files
    - Loaded into DW, then process
• Data Loading
  – Loading from data sources to data warehouse
• Two techniques
  – By loading tool provided by DW vendors
  – By API programs
• Load strategies
  – Load update directly
  – Complete overwrite
  – Check difference then update
ETL\textsubscript{4}

• Tools and vendors
  – IBM: Warehouse manager
  – Oracle: Oracle Warehouse Builder
  – Microsoft: DTS
  – Informix: Ardent Data Stage
  – CA: Inforbump
Meta Data

• Meta data: data describing data
• Two types of meta data in DW
  – Management meta data
  – User meta data
• Management meta data
  – Used by DW developer and maintenance
Management Meta Data

• Description of DW structure:
  – Definition of warehouse model, visualization diagrams, dimensions, hierarchy structure, export data, data mart location and content
• Service system, DW and DM structure and model
• Aggregation algorithms, including measure and dimension definition algorithms; data granularity, subject definition, query and report definition
Management Meta Data

• Describing the mapping between source data and DW data
  – Source of data and their contents
  – Data partition, extraction, cleaning, transformation rules
  – Data update rules
  – User authorization and access control rules
User Meta Data

• Data used by users of DW
  – Data describing how to connect to DW
  – Data describing accessible data in DW
  – Data describing data coming from which source
Standardization of Meta Data

- Two main standards
  - OIM model from Meta Data Coalition (MDC)
  - CWM model from OMG organization
- OIM (Open Information Model)
  - Use UML (Universal Modeling Language) to describe the meta data for many subject areas such as Analysis and Design, Object and Component, Database and Warehousing, Business Engineering, Knowledge Management
  - Microsoft Repository; CA Repository
Standardization of Meta Data

- **CWM (Common Warehouse Meta model)**
  - Developed by OMG organization
  - Provide meta data exchange between different platforms and data warehousing tools
  - Includes 4 standards
    - CWM Meta model: for data warehouse model
    - CWM XML: for XML
    - CWM DTD: for DW/BI meta data exchange
    - CWM IDL: for DW/BI API
External Data

• Meta data
  – File ID
  – Date entered
  – File description
  – File source
  – File classification
  – Index
  – Cleaning date
  – File length
  – Reference

Unstructured data (image, voice)

External data (from other company)

Meta data

DW