What is a Data Warehouse?

By Susan L. Miertschin

W. H. Inmon

"A data warehouse is a subject oriented, integrated, time variant, nonvolatile, collection of data in support of management's decision making process."

https://www.business.auc.dk/oekostyr/file/What_is_a_Data_Warehouse.pdf

Ralph Kimball – *The Data Warehouse Toolkit*

What is a Data Warehouse?

"A copy of transaction data specifically structured for query and analysis"

What is a Data Warehouse?

"Data Warehousing is the coordination, architected, and periodic copying of data from various sources, both inside and outside the enterprise, into an environment optimized for analytical and informational processing"

- Alan Simon

Data Warehousing for Dummies

Business Intelligence (BI)

- "…implies thinking abstractly about the organization, reasoning about the business, organizing large quantities of information about the business environment." p. 6 in Giovinazzo textbook
- Purpose of BI is to define and execute a strategy

Strategic Thinking

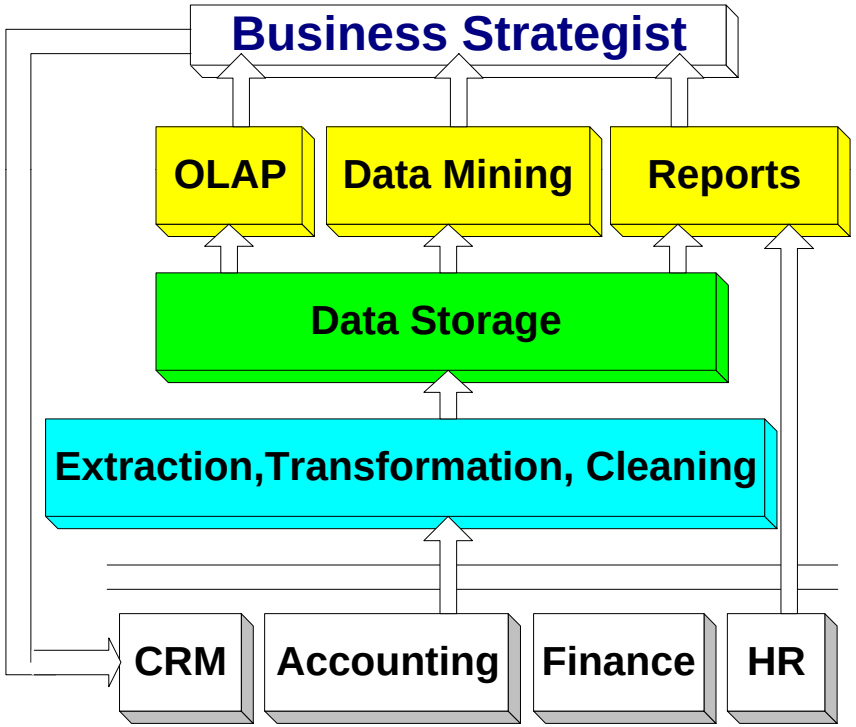
- **Business strategist**
  - Always looking forward to see how the company can meet the objectives reflected in the ***mission statement***
- Successful companies
  - Do more than just react to the day-to-day environment
  - Understand the past
  - Are able to predict and adapt to the future

Business Intelligence Loop

Business Intelligence

Figure 1-1 p. 2 Giovinazzo

- Encompasses entire loop shown
- Data Storage + ETC = Data Warehouse
- Data Warehouse + Tools (yellow) = Decision Support System



The Data Warehouse

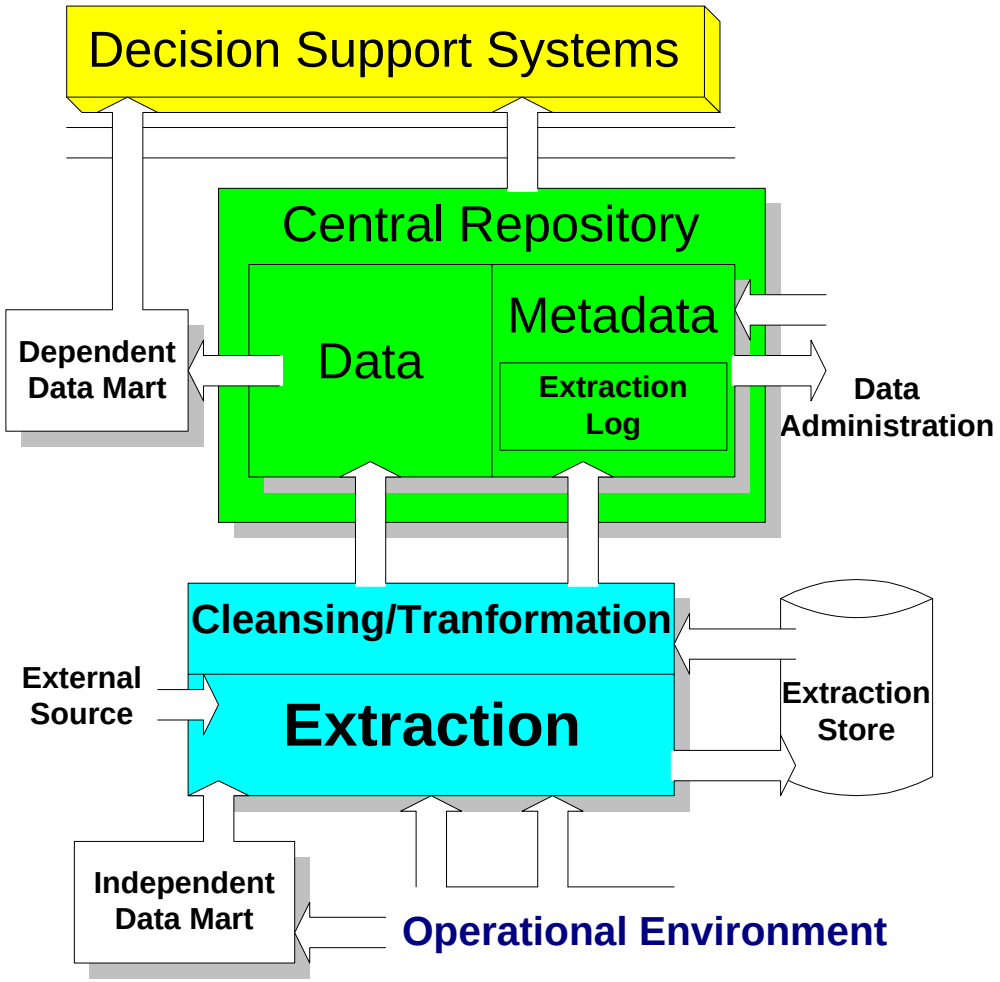
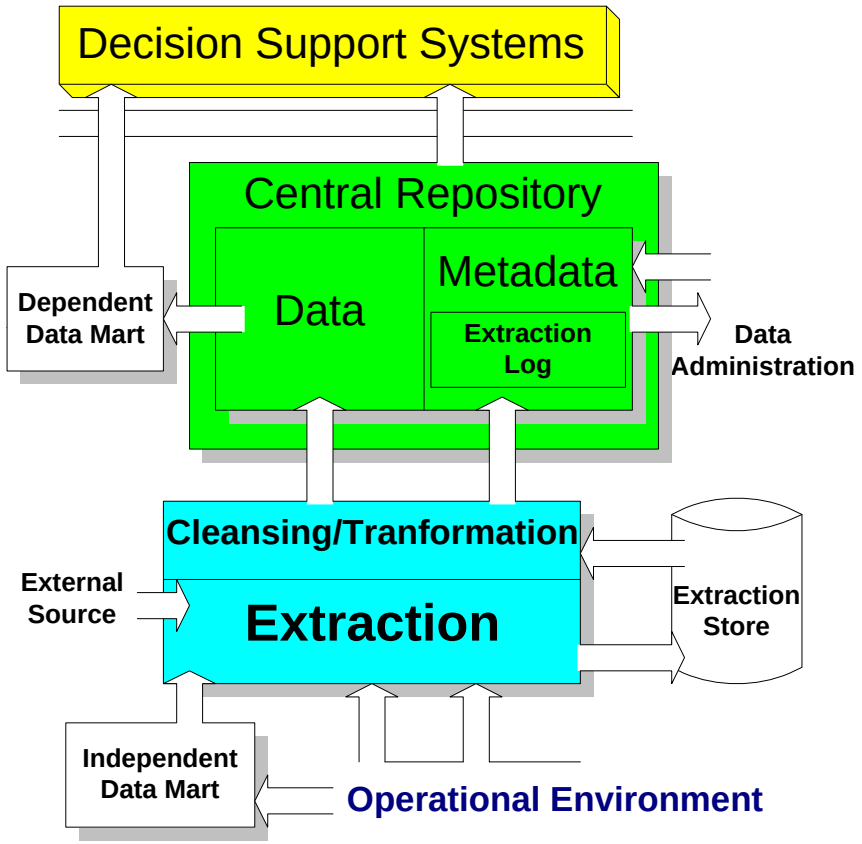


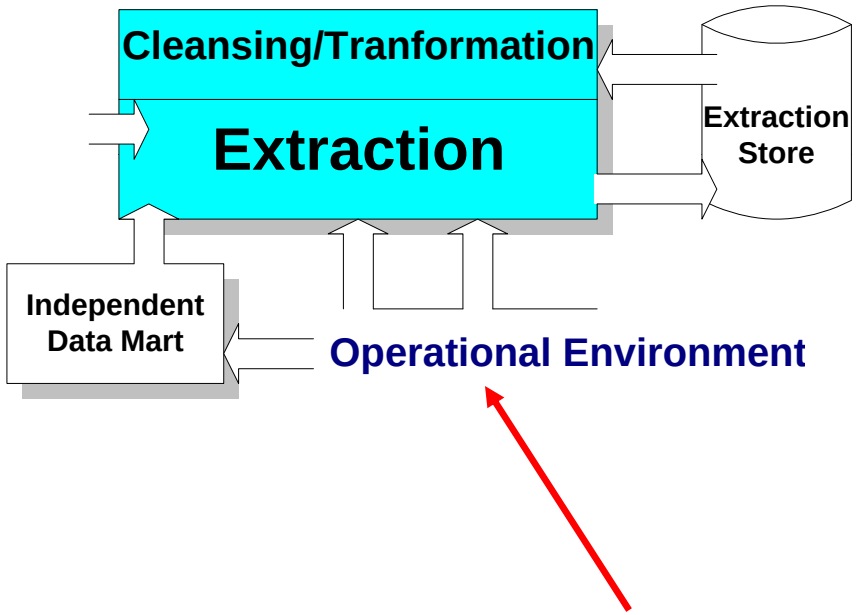
Figure 1-2 p. 9 Giovinazzo

Data Path

- The path to get data from the operational environment to the business strategist is complex
- There is much more to a data warehouse than the Central Repository

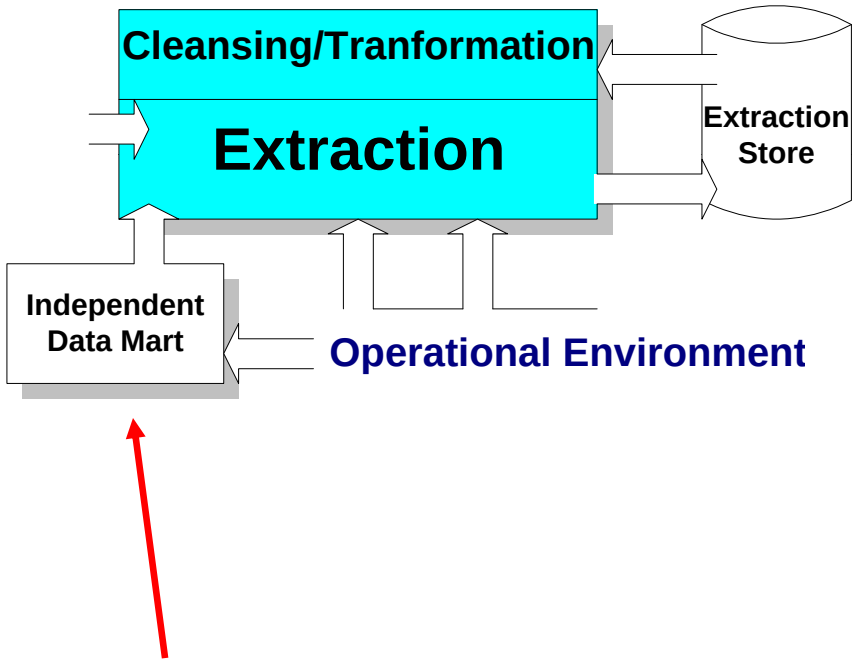


Operational Environment



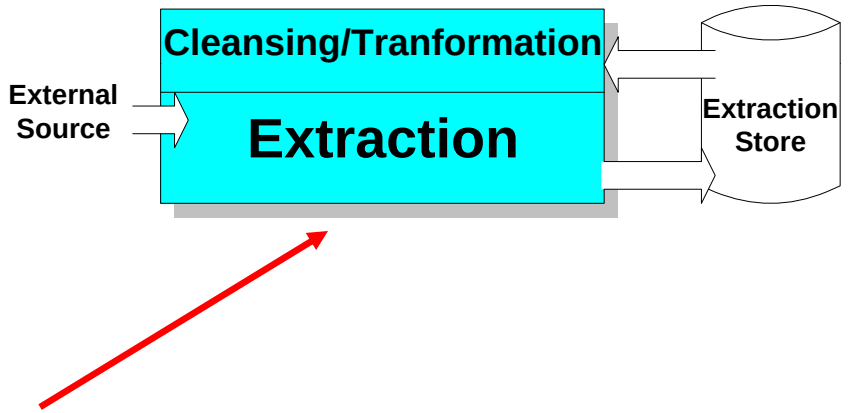
- Operational environment runs day to day activities of the organization
- Systems contain "raw" data – transactional data
- Data describes the current state of the organization

Independent Data Mart



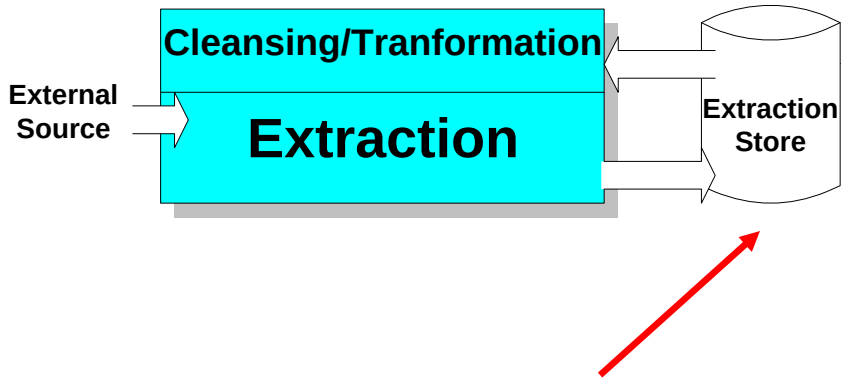
- Data Mart focuses on one subject area within the organization
- Data warehouse focuses on the entire organization

Extraction



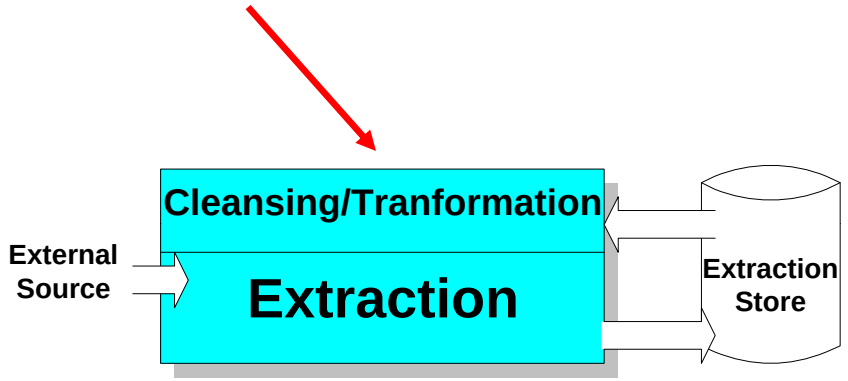
- Extraction engine retrieves/receives data from the operational environment
- Data from other external sources may also be collected during extraction

Extraction Store



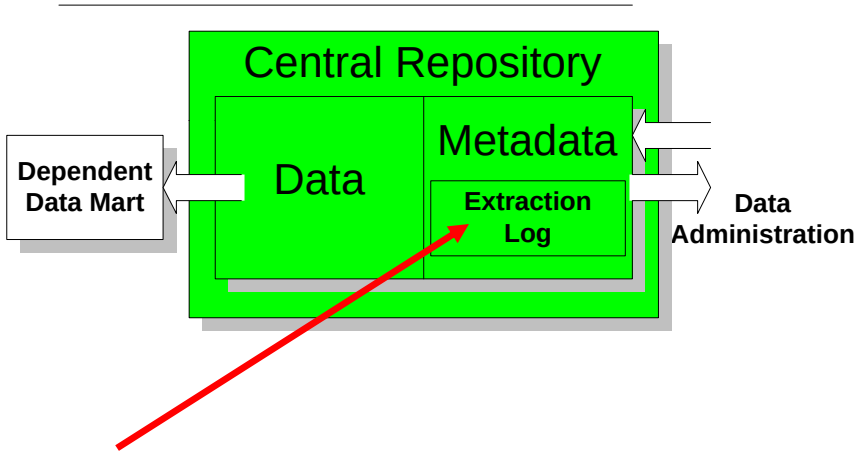
- Holding area for the collected data until it can be cleaned and transformed into the correct format

Transformation/Cleansing



- Scrubbing = data transformation + cleansing
- Transformation = converting data to a common format
- Cleansing = removing errors from data

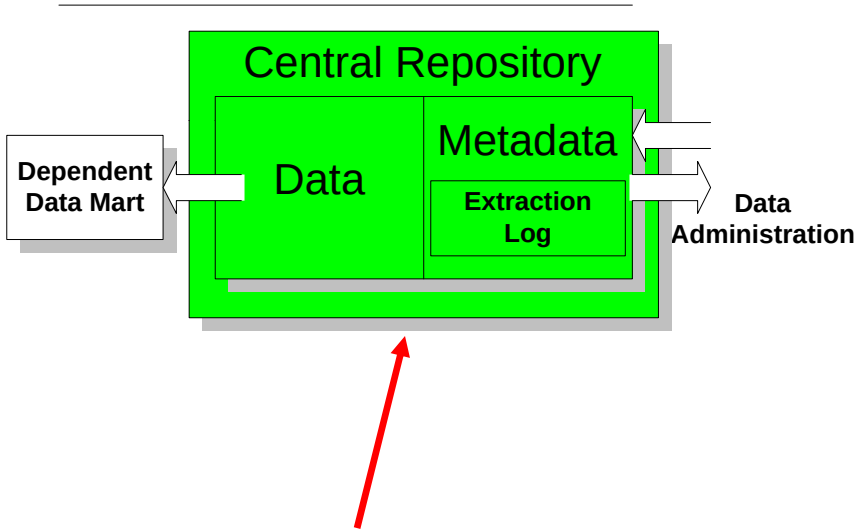
The Extraction Log



- The extraction log records success/failure of extraction process steps (+ more)
- The log is part of the Metadata
- Used to verify quality of data placed in the warehouse

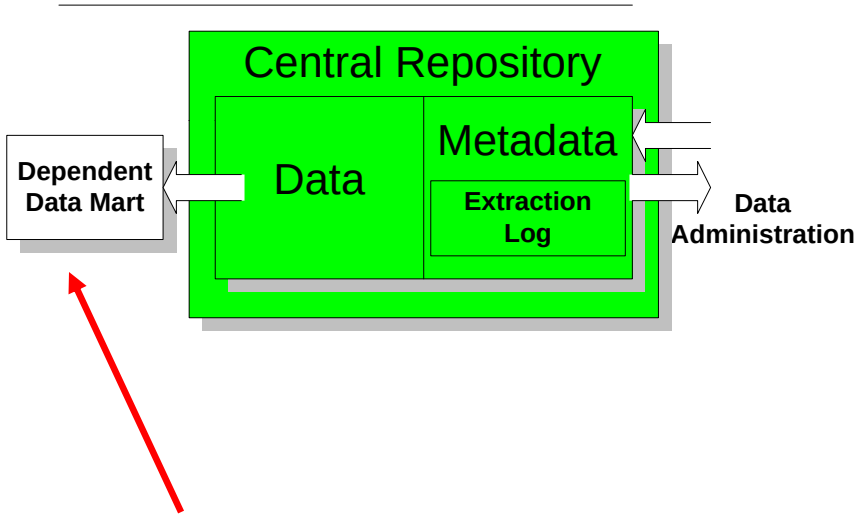
Central Repository

- Cornerstone of the data warehouse architecture
- Stores all the data and metadata for the data warehouse

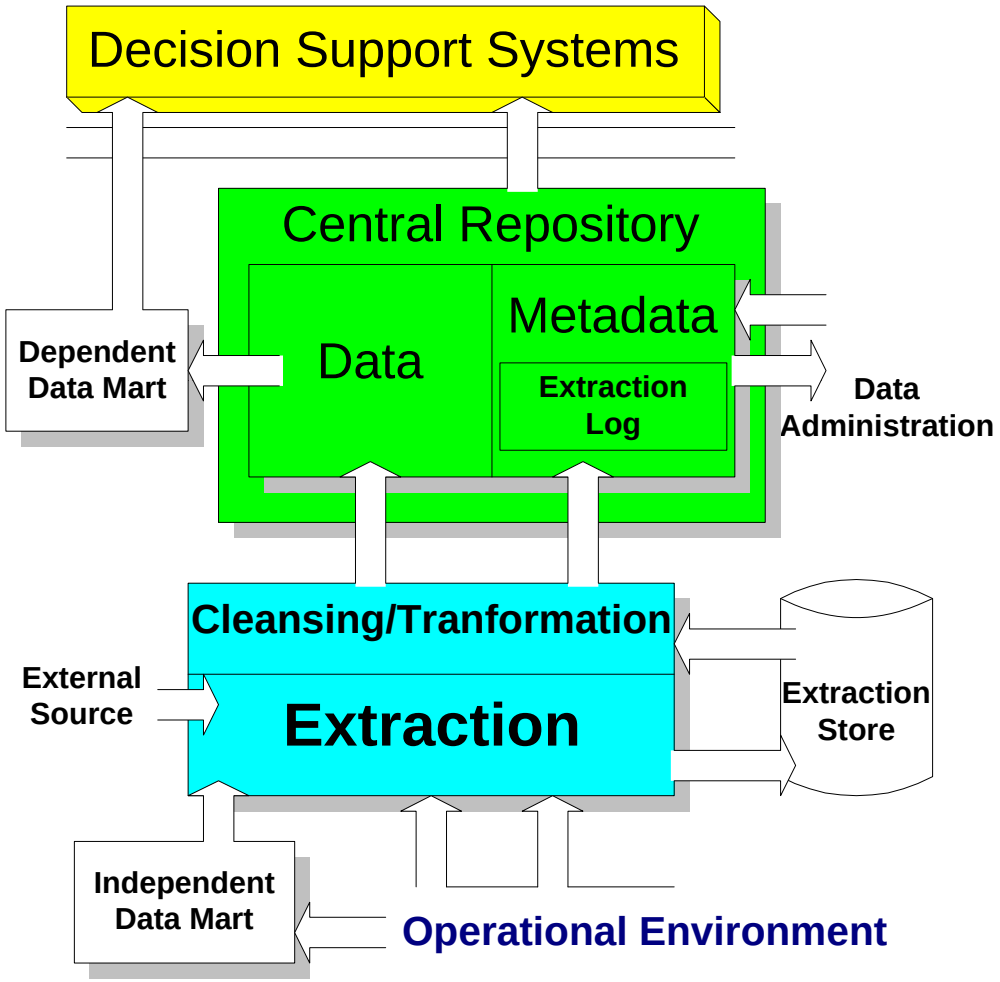


Dependent Data Mart

- Different from an independent data mart
- Dependent data mart relies on the data warehouse as the source of its data



Business Intelligence Infrastructure



Subject Orientation of DW

Data Warehouse

- Subject oriented
- Focuses on the what things drive operational transactions

Operational Database

- Focuses on day-to-day transactions
- Normalized and optimized for this purpose

Data Warehouse vs. Operational Db

Data Warehouse

- Gathers distributed data together into one place
- Facilitates analysis processes

Operational Database

- Distributed across multiple tables within an application
- Distributed across multiple applications

Integrating Transactional Data into DW

- Most time-consuming and problematic process
- Two steps
  - Data transformation
  - Data Cleansing

Data Cleansing

- Remove errors from data extracted from the operation environment
- Critical
- What should be done with data that contains errors?
  - Send the data back to be fixed at the operational level and resubmitted
  - Fix the data and inform the operational system of the errors

Data Transformation

- Operational environment consists of numerous applications and databases
- Data definitions will not be consistent
- Must be consistent format for DW
- Four issues to address
  - Description
  - Encoding
  - Units of Measure
  - Format

Description

- Same things may be described differently across systems
- Map each different description into a single description
- Example: customer, client, user

Encoding

- Nominal scale: number or letter assigned as a label, a category name
  - ordering is arbitrary
- Example:
  - R = Red Red = Red 36 = Red
  - B = Blue Blue = Blue 45 = Blue

Units of Measure

- Measurement system must be common
- Precision can cause problems

- Example:
  - Values in metric vs. English units
- Example:
  - 1/3 = .333 = .333333333333

Format

- Different operational systems store data in different formats
- Example:
  - SS#: 999999999
    - Char(9)
    - Int

Is Data a “Political” Issue?

- Can be
- Operational level designers work hard to make the data match their needs
- Arguments can arise over whether customer name should be 30 characters or 32 characters long

DW Contains a Snapshot

- Once data is placed in the central repository – it becomes read-only
- Time becomes an important dimension for the *data*
- DW: A series of organizational snapshots over time

Decision Support Systems (DSS)

- Data placed in the data warehouse must be easy to access for business strategists
  - Timely
  - Support their mission
- Three common DSS tools
  - Reports
  - OLAP
  - Data Mining

Reports

- One of the most basic DSS tools
- Present summary information
- Reporting tool should support
  - Rapid development
  - Easy maintenance
  - Easy distribution
  - Internet enabled

On-Line Analytical Processing (OLAP)

- OLAP environment allows business strategist to interact directly with data
- OLAP tool should support
  - Multiple dimensional presentation of data
  - Rotation {Data Cube}
  - Drill-down / Roll-up
  - “What if” analysis

Data Mining

- “ Data Mining is defined as a process of identifying hidden patterns and relationships within data.” - Robert Groth
- Business strategists use OLAP to help them find answers to their questions
- Data mining supplies answers without knowing the questions (sometimes)

In Summary …

- DW is the heart of BI
- DW is more than just an archive of operational data
- Data must be formed according to business needs for strategic information
- Time becomes a dimension of the data
- DSS tools used to analyze the data

What is a Data Warehouse?

By Susan L. Miertschin