The Data Warehouse

1

Stores, Warehouses and Marts

- A data warehouse is a collection of integrated databases designed to support a DSS.
- An operational data store (ODS) stores data for a specific application. It feeds the data warehouse a stream of desired raw data.
- A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users (rather than the entire firm).
- The metadata is information that is kept about the warehouse.

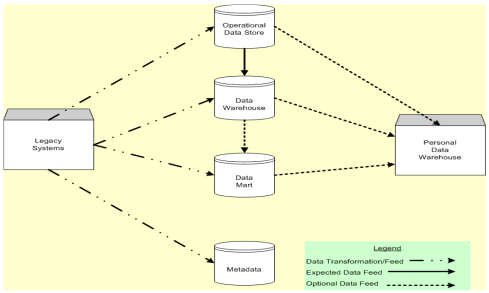
2

The Data Warehouse Environment

- The organization's legacy systems and data stores provide data to the data warehouse or mart.
- During the transfer of data from the various sources, cleansing or transformation may occur, so the data in the DW is more uniform.
- Simultaneously, metadata is recorded.
- Finally, the DW or mart may be used to create one or more "personal" warehouses.

3

Organizational Data Flow and Data Storage Components



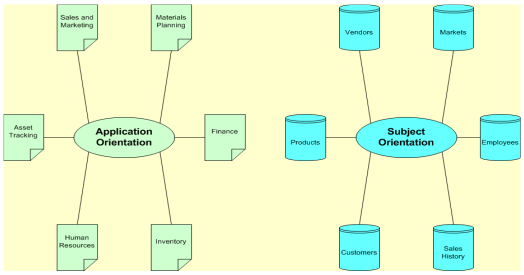
4

Characteristics of a Data Warehouse

- *Subject oriented* – organized based on use
- *Integrated* – inconsistencies removed
- *Nonvolatile* – stored in read-only format
- *Time variant* – data are normally time series
- *Summarized* – in decision-usable format
- *Large volume* – data sets are quite large
- *Non normalized* – often redundant
- *Metadata* – data about data are stored
- *Data sources* – comes from nonintegrated sources

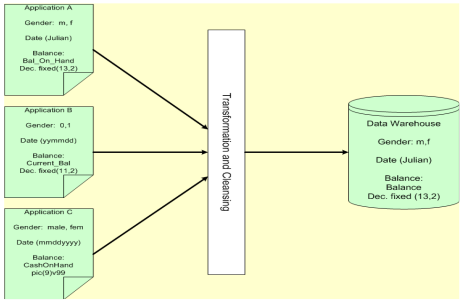
5

A Data Warehouse is Subject Oriented



6

Data in a Data Warehouse are Integrated



7

The Data Warehouse Architecture

The architecture consists of various interconnected elements:

- *Operational and external database layer* – the source data for the DW
- *Information access layer* – the tools the end user access to extract and analyze the data
- *Data access layer* – the interface between the operational and information access layers
- *Metadata layer* – the data directory or repository of metadata information

8

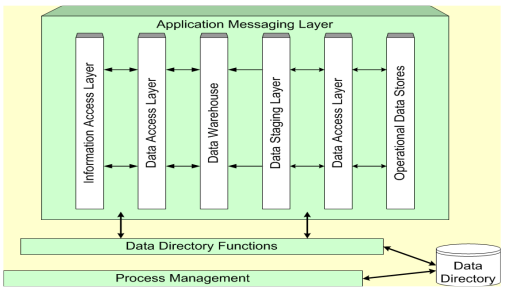
The Data Warehouse Architecture (cont.)

Additional layers are:

- *Process management layer* – the scheduler or job controller
- *Application messaging layer* – the "middleware" that transports information around the firm
- *Physical data warehouse layer* – where the actual data used in the DSS are located
- *Data staging layer* – all of the processes necessary to select, edit, summarize and load warehouse data from the operational and external data bases

9

Components of the Data Warehouse Architecture



10

Data Warehousing Typology

- *The virtual data warehouse* – the end users have direct access to the data stores, using tools enabled at the data access layer
- *The central data warehouse* – a single physical database contains all of the data for a specific functional area
- *The distributed data warehouse* – the components are distributed across several physical databases

11

Data Have Data -- The Metadata

- The name suggests some high-level technological concept, but it really is fairly simple. Metadata is "data about data".
- With the emergence of the data warehouse as a decision support structure, the metadata are considered as much a resource as the business data they describe.
- Metadata are abstractions -- they are high level data that provide concise descriptions of lower-level data.

12

The Metadata in Action

The metadata are essential ingredients in the transformation of raw data into knowledge. They are the "keys" that allow us to handle the raw data.

For example, a line in a sales database may contain: 1023 K596 111.21

This is mostly meaningless until we consult the metadata (in the data directory) that tells us it was store number 1023, product K596 and sales of $111.21.

13

The Need for Consistency in the Metadata

- The data warehouse is set up for the benefit of business analysts and executives across all functional areas.
- In their individual databases, the different areas may define and store data according to their own version of the "truth".
- When data are retrieved from these different areas and placed in the warehouse, the transformation and cleansing process ensures that there is a single, integrated "truth" at the organizational level.

14

Interviewing the Data—Metadata Extraction

Regardless of the nature of a query, certain aspects of the metadata are important to all decision-makers. Some of these are:

- What tables, attributes and keys does the DW contain?
- Where did each set of data come from?
- What transformations were applied with cleansing?

15

Interviewing the Data—Metadata Extraction (cont.)

- How have the metadata changed over time?
- How often do the data get reloaded?
- Are there so many data elements that you need to be careful what you ask for?

16

Components of the Metadata

- *Transformation maps* – records that show what transformations were applied
- *Extraction history* – records that show what data was analyzed
- *Algorithms for summarization* – methods available for aggregating and summarizing
- *Data ownership* – records that show origin
- *Access patterns* – records that show what data are accessed and how often

17

Typical Mapping Metadata

Transformation mapping records include:

- Identification of original source
- Attribute conversions
- Physical characteristic conversions
- Encoding/reference table conversions
- Naming changes
- Key changes
- Values of default attributes
- Logic to choose from multiple sources
- Algorithmic changes

18

Implementing the Data Warehouse

Kozar assembled a list of "seven deadly sins" of data warehouse implementation:

1. *"If you build it, they will come"* – the DW needs to be designed to meet people's needs
2. *Omission of an architectural framework* – you need to consider the number of users, volume of data, update cycle, etc.
3. *Underestimating the importance of documenting assumptions* – the assumptions and potential conflicts must be included in the framework

19

"Seven Deadly Sins", continued

4. *Failure to use the right tool* – a DW project needs different tools than those used to develop an application
5. *Life cycle abuse* – in a DW, the life cycle really never ends
6. *Ignorance about data conflicts* – resolving these takes a lot more effort than most people realize
7. *Failure to learn from mistakes* – since one DW project tends to beget another, learning from the early mistakes will yield higher quality later

20

Data Warehouse Technologies

- No one currently offers an end-to-end DW solution. Organizations buy bits and pieces from a number of vendors and hopefully make them work together.
- SAS, IBM, Software AG, Information Builders and Platinum offer solutions that are at least fairly comprehensive.

21

The Future of Data Warehousing

As the DW becomes a standard part of an organization, there will be efforts to find new ways to use the data. This will likely bring with it several new challenges:

- Regulatory constraints may limit the ability to combine sources of disparate data.
- These disparate sources are likely to contain unstructured data, which is hard to store.
- The Internet makes it possible to access data from virtually "anywhere". Of course, this just increases the disparity.

22