Data warehousing

Mihai Enescu
Signal Processing and Computer Technology Laboratory
Helsinki Univ. of Technology

November 29, 1999

1 Introduction

Data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. A data warehouse is a relational database management system (RDMS) designed specifically to meet the needs of transaction processing systems. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats.

2 Characteristics of a data warehouse

According to Bill Inmon, author of "Building the Data Warehouse", who is considered to be the originator of the data warehousing concept, there are generally four characteristics that describe a data warehouse:

- *subject oriented:* data are organized according to subject instead of application e.g. an insurance company using a data warehouse would organize their data by customer, premium, and claim, instead of by different products (auto, life, etc.). The data organized by subject contain only the information necessary for decision support processing.
- *integrated:* when the data resides in many separate applications in the operational environment, encoding of data is often inconsistent. For instance, in one application, gender might be coded as "m" and "f" in another by 0 and 1. When data are moved from the operational environment into the data warehouse, they assume a consistent coding conversion e.g. gender data is transformed to "m" and "f".
- *time-variant:* the data warehouse contains a place for storing data that are five to 10 years old, or older, to be used for comparisons, trends, and forecasting. These data are not updated.
- *non-volatile:* data are not uploaded or changed in any way once they enter the data warehouse, but are only loaded and accessed.

3 Processes in data warehousing

The data warehouse retrieves data from a variety of heterogeneous operational databases. The data is then transformed and delivered to the data warehouse based on a selected model (or mapping definition). The data transformation and movement process are executed whenever an update to the warehouse data is required so there should some form of automation to manage and execute these functions. The information that describes the model and definition of the source data elements is called *metadata*. The metadata is the means by which the end user finds and understands the data in the warehouse and is an important part of the warehouse. The metadata should at the very least contain:

- the structure of the data
- the algorithm used for summarization
- the mapping from the operational environment to the data warehouse

Data cleansing is an important aspect of creating an efficient data warehouse in that it is the removal of certain aspects of operational data, such as low-level transaction information, which slow down the query times. The cleansing stage has to be as dynamic as possible to accommodate all types of queries even those which may require low-level information.

Once the data has been cleaned it is then transferred to the data warehouse which typically is a large database on a high performance box either Symmetric Multi-Processing (SMP) or Massively Parallel Processing (MPP). A data warehouse can be used in different ways for example it can be used as a central store against which the queries are run or it can be used to like a data mart. Data marts which are small warehouses can be established to provide subsets of the main store and summarized information depending on the requirements of a specific group/department. The central store approach generally uses very simple data structures with very little assumptions about the relationship between data whereas marts often use multidimensional databases which can speed up query processing as they can have data structures which are reflect the most likely questions.

Another approach to data warehousing is Parsaye's Sandwich Paradigm. This paradigm or philosophy encourages acceptance of the probability that the first iteration of a data warehousing effort will require considerable revision. The Sandwich Paradigm advocates the following approach:

- pre-mine the data to determine what formats and data are needed to support a data-mining application
- build a prototype mini-data warehouse i.e. the meat of the sandwich, with most of the features envisaged for the end product
- built the final warehouse

4 Data warehousing and OLTP systems

A database which is built for on line transaction processing (OLTP), is generally regarded as unsuitable for data warehousing as they have been designed with a different set of needs in mind i.e. maximizing transaction capacity and typically having hundreds of tables in order not to lock out users etc. Data warehouses are interested in query processing as opposed to transaction processing.

OLTP systems cannot be repositories of facts and historical data for business analysis. They cannot quickly answer ad hoc queries and rapid retrieval is almost impossible. The data is inconsistent and changing, duplicate entries exist, entries can be missing and there is an absence of historical data which is necessary to analyze trends. Basically OLTP offers large amounts of raw data which is not easily understood. The data warehouse offers the potential to retrieve and analyze information quickly and easily. Data warehouse do have similarities with OLTP as shown in Table 1.

Table 1: Similarities and Differences in OLAP and data warehousing

	OLTP	Data Warehouse
Purpose	Run day-to-day operations	Information retrieval and analysis
Structure	RDBMS	RDBMS
Data Model	Normalised	Multi-dimensional
Access	SQl	SQL plus data analysis extensions
Type of Data	Data that runs the business	Data that analyzes the business
Condition of Data	Changing, incomplete	Historical, descriptive

5 The Data Warehouse model

Data warehousing is the process of extracting and transforming operational data into informational data and loading it into a central data store or warehouse. Once the data is loaded it is accessible via desktop query and analysis tools by the decision makers. The data warehouse model is illustrated in Figure 1.

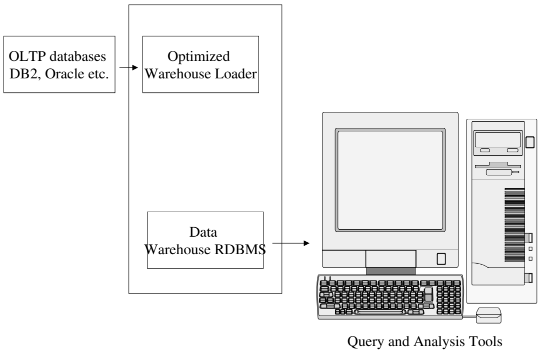

Figure 1: A data warehouse model

The data within the actual warehouse itself has a distinct structure with the emphasis on different levels of summarization as shown in Figure 2.

The current detail data is central in importance as it:

- reflects the most recent happenings, which are usually the most interesting
- it is voluminous as it is stored at the lowest level of granularity
- it is always (almost) stored on disk storage which is fast to access but expensive and complex to manage.

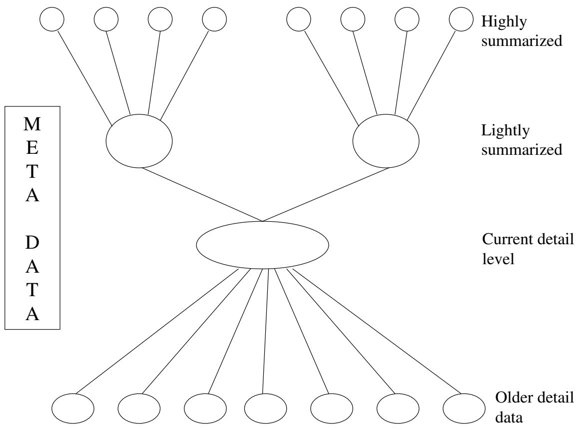

Figure 2: The structure of data inside the data warehouse

Older detail data is stored on some form of mass storage, it is infrequently accessed and stored at a level detail consistent with current detailed data.

Lightly summarized data is data distilled from the low level of detail found at the current detailed level and generally is stored on disk storage. When building the data warehouse have to consider what unit of time is summarization done over and also the contents or what attributes the summarized data will contain.

Highly summarized data is compact and easily accessible and can even be found outside the warehouse.

Metadata is the final component of the data warehouse and is really of a different dimension in that it is not the same as data drawn from the operational environment but it is used as:

- a directory to help DSS analyst locate the contents of the data warehouse,
- a guide to the mapping of the data as the data is transformed from the operational environment to the data warehouse environment,
- a guide to the algorithms used for summarization between the current detailed data and the lightly summarized data and the lightly summarized, etc.

The basic structure has been described but Bill Inmon fills in the details to make the example come alive as shown in Figure 3.

The diagram presented in Figure 3 assumes the year is 1999 hence the current detail data is 1998-99. Generally sales data doesn't reach the current level of detail for 24 hours as it waits until it is no longer available to the operational system i.e. it takes 24 hours for it to get to the data warehouse. Sales details are summarized weekly by sub product and region to produce the lightly summarized detail. Weekly sales are then summarized again to produce the highly summarized data.

6 Criteria for a Data Warehouse

In order to avoid problems with Data Warehousing some criteria has been proposed. The requirements for data warehouse RDBMSs begin with the loading and preparation of data for query and analysis. If a product fails to meet the criteria at this stage, the rest of the system

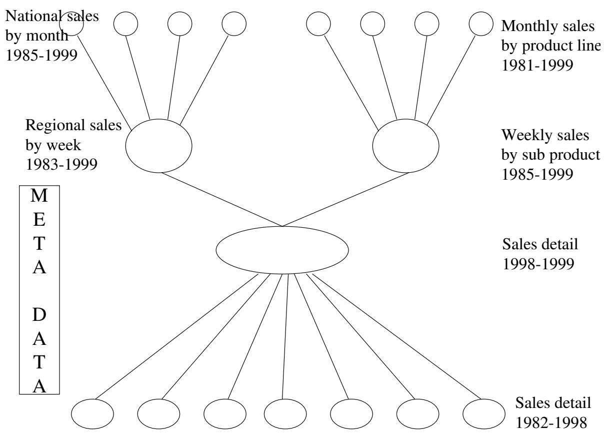

Figure 3: An example of levels of summarization of data inside the data warehouse

will be inaccurate, unreliable and unavailable. The criteria for data warehouse RDBMS are as follows:

- *Load Performance* - Data warehouse require incremental loading of new data on a periodic basis within narrow time windows; performance of the load process should be measured in hundreds of millions of rows and gigabytes per hour and must not artificially contain the volume of data required by the business.

- *Load Processing* - Many steps must be taken to load new or updated data into the data warehouse including data conversion, filtering, reformatting, integrity checks, physical s-torage, indexing and metadata update. These steps must be executed as a single, seamless unit of work.

- *Data Quality Management* - The shift to fact-based management demands the highest data quality. The warehouse must ensure local consistency, global consistency, and referential integrity despite "dirty" sources and massive database size. While loading and preparation are necessary steps, they are not sufficient. Query throughput is the measure of success for a data warehouse application.

- *Query Performance* - Fact-based management and ad hoc analysis must not be slowed or inhibited by the performance of the data warehouse RDBMS; large, complex queries for key business operations must complete in seconds not days.

- *Terabyte Scalability* - Data warehouse sizes are growing at astonishing rates. Today these range from a few to hundreds of gigabytes, and terabyte-sized data warehouses are a near-term reality. The RDBMS must not have any architectural limitations. It must support modular and parallel management. It must support continued availability in the event of a point failure, and must provide a fundamental different mechanism for recovery. It must support near-line mass storage devices such as optical disk and Hierarchical Storage Management devices. Lastly, query performance must not be dependent on the size of the data base, but rather on the complexity of the query.

- *Mass User Scalability* - Access to warehouse data must no longer be limited to the elite few. The RDBMS server must support hundreds, even thousands, of concurrent users while maintaining acceptable query performance.

- *Networked Data Warehouse* - Data warehouses rarely exist in isolation. Multiple data warehouse systems cooperate in a larger network of data warehouses. The server must include tools that coordinate the movement of subsets of data between warehouses. Users must be able to look at and work with multiple warehouses from a single client workstation. Warehouse managers have to manage and administer a network of warehouses from a single physical location.

- *Warehouse Administration* - The very large scale and time-cyclic nature of the data warehouse demands administrative ease and flexibility. The RDBMS must provide controls for implementing resource limits, chargeback accounting to allocate costs back to users, and query prioritization to address the needs of different user classes and activities. The RDBMS must also provide for workload tracking and tuning so system resources may be optimized for maximum performance and throughput.

- *Integrated Dimensional Analysis* - The power of multidimensional views is widely accepted, and dimensional support must be inherent in the warehouse RDBMS to provide the highest performance for relational OLAP tools. The RDBMS must support fast, easy creation of precomputed summaries common in large data warehouses. It also should provide the maintenance tools to automate the creation of these precomputed aggregates. Dynamic calculation of aggregates should be consistent with the interactive performance needs.

- *Advanced Query Functionality* - End users require advanced analytic calculations, sequential and comparative analysis, and consistent access to detailed and summarized data. Using SQL in a client/server point-and-click tool environment may sometimes be impractical or even impossible. The RDBMS must provide a complete set of analytic operations including core sequential and statistical operations.

References

[1] J. Srivastava, P-Y. Chen, “Warehouse Creation - A Potential Roadblock to Data Warehousing”, IEEE Transactions on Knowledge and Engineering, Vol 11, No. 1, 1999, pp. 118-126.

[2] V.R. Gupta, “An introduction to Data Warehouse”,
http://system-services.com/dwintro.htm

[3] “The Data Warehousing Information Center”, http://pwp.starnetinc.com/larryg/

[4] Information Discovery Inc., “Data Mines for Data Warehouses”,
http://www.datamining.com/dm4dw