8-16-1996

The New DSS: Data Warehouses, OLAP, MDD, and KDD

Paul Gray
University of Georgia

The New DSS: Data Warehouses, OLAP, MDD, and KDD

Paul Gray

Claremont Graduate School

Hugh J.Watson

The University of Georgia

Origins

The new DSS is the result of two software solutions needing and finding one another:

Data base firms developed data warehouses and were looking for applications

EIS and DSS software developers and vendors needed to deal with ever increasing data bases.

Three to four years ago, the two groups started interacting with the results described here.

Data Warehouses

Database developers long understood that their software was required for both transactional and analytic processing. However, their principal developments were directed to ever-larger transactional data bases at the expense of informational data bases. This process occurred even though operational and analytic data are separate with different requirements and different user communities. Once these differences were understood, new data bases were created specifically for analysis. These separate databases, given the name data warehouses, have the following characteristics:

subject oriented	data are organized by how users refer to it
integrated	inconsistencies are removed in both nomenclature and conflicting information. That is, the data are 'clean'
non-volatile	Read only data. Data do not change over time.
time series	Data are time series not current status
summarized	Operational data are mapped into decision usable form
larger	Time series implies much more data is retained.
Not normalized	DSS data can be redundant
metadata	Metadata =data about the data.

That is, these data bases hold aggregated data for management decision use separate from the databases used for On Line Transaction Processing (OLTP).

Data warehouses are not cheap. Multimillion dollar costs are common. Their design and implementation is still an are and they require considerable time to create. Being designed for the enterprise so that everyone has a common data set, they are large and increase in size with time. Typical storage sizes run from 50 gigabytes to over a terabyte. Because of the large size, some firms are using parallel computing to speed data retrieval.

A scaled-down version of the data warehouse is the 'data mart.' A data mart is a small warehouse designed for the SBU or department level. It is often a way to gain entry and provide an opportunity to learn.

Data warehouses are a major industry. Estimates vary but it is clear that many more than half of the Fortune 500 have data warehouse projects underway or planned.

On-line Analytical Processing

This term, abbreviated OLAP, was introduced by E.F. Codd, the father of relational databases in 1993 in a major article in Computerworld. Codd came to the conclusion that relational databases for OLTP had reached the maximum of their capabilities in terms of the views of the data they provided the user. The problem stemmed principally from the massive computing required when relational data bases were asked to answer relatively simple SQL queries. He also came to the view that EIS and DSS people had known for a long time; namely, that operational data are not adequate for answering managerial questions. He therefore advocated the use of multi-dimensional data bases (see below). Codd's conversion to the DSS/EIS viewpoint gave legitimacy to the data warehouse based concepts.

The basic idea in OLAP is that managers should be able to manipulate enterprise data models across many dimensions to understand changes that are occurring. He promulgated the following set of 12 rules for OLAP. At this writing, it does not appear that implementations are available for which all rules are implemented. In fact, it may not be possible to implement them simultaneously.

1. Multi-dimensional view	7. Dynamic sparse matrix handling
2. Transparent to the user	8. Multi-user support
3. Accessible	9. Cross-dimensional operations
4. Consistent reporting	10. Intuitive data manipulation
5. Client-server architecture	11. Flexible reporting
6. Generic dimensionality	12. Unlimited dimensions, aggregation

Many EIS vendors claim to be OLAP compliant. Most have part, but not all the rules implemented. Many simply relabeled what they had to be OLAP.

Multidimensional Data Bases (MDD)

Multidimensional databases (MDD) are not new. For about 20 years, the EXPRESS software package has features MDD. By the early 1990's a number of other vendors were aboard. The definition of OLAP gave these vendors a boost. Relational database vendors responded by upgrading their products so that they could also handle multiple dimensions.

MDD stores data as an n-dimensional cube. This arrangement implies very sparse matrices. It lets you deal simultaneously with data views defined by such combinations of quantities as product, region, sales,

actual/budget. More important, MDD add time as a dimension. MDD's advantage over relational databases is that they are optimized for speed and ease of query response.

Two ways of organizing the data base are being pushed by vendors, MOLAP and ROLAP. The former includes vendors such as Arbor Essbase, and Pilot Executive Software who offer MDD products for OLAP. However, these vendors must compete in a world in which the predominant legacy databases are relational. The relational data base vendors are protecting their client base by creating relational multidimensional data bases to compete. They have invented a new design schema, called the star schema, which creates two types of tables: (1) fact tables that contain information being queried about the business (and which may have millions of rows) and (2) smaller dimension tables which hold descriptive data about the dimensions of the business.

At present, the final form of multidimensionality, that is, whether MOLAP or ROLAP will dominate is unclear. ORACLE, with its huge relational customer base has

hedged its bets by purchasing IRI's MDD capabilities.

Knowledge Data Discovery

Knowledge data discovery is usually abbreviated as KDD but is also known as 'data mining'. The mining terminology refers to finding answers about a business from the data warehouse that the executive or analyst had not thought to ask. One definition of EIS is that it allows managers to obtain managerial information from the legacy systems they have long been paying for.

KDD applies techniques mostly from artificial

intelligence to discover new information. That is, it is

designed to find information that queries and reports

don't reveal effectively.

KDD seeks to find patterns in data and to infer rules. Techniques include:

- statistical analysis of data
- neural networks, expert systems, intelligent agents
- multidimensional analysis; data visualization
- decision trees

The software associated with these approaches is called 'siftware'

Data mining, which is still in its early stages deals with five kinds of data:

Associations	things done together (buy groceries)
Sequences	events over time (house, refrigerator)
Classifications	pattern recognition (rules)
Clusters	define new groups
Forecasting	predictions from time series

A Survey of Data Warehousing

To understand more about what is happening in data warehousing, an industry survey of current practices was undertaken at the January 1996 conference of the Data Warehousing Institute (DWI). The Institute, an organization whose goal is to foster a greater understanding of the various issues associated with data warehousing and to share this information with the data warehousing community, holds conferences on data warehousing throughout the year.

The survey findings provide insights about:

- the factors that motivate the development of a data warehouse.
- who championed the project
- whether a formal proposal was prepared
- how the benefits and the costs were assessed, both prior to and after implementation
- the number of person-years that were expended on the project
- whether outside consultants were used.
- how long it took to build the data warehouse
- the critical success factors
- the biggest obstacles to a successful project
- the hardware, software, and operating systems used
- the size of the user base
- who the users are
- the applications that use the data warehouse
- the size of the warehouse
- the costs of the project
- the organizational units that funded the project
- measures of project success

Research and Publishing Opportunities for AIS Members

The state of the data warehousing and related fields is almost entirely the result of work in industry. The survey described above is one of the initial research efforts. Almost no other research has been done. As a result, major opportunities exist for AIS members to apply their research methodologies to the data warehousing problem.

Fortunately, the data warehousing field has progressed to the point where it merits its own journal. Edited by Hugh Watson, the refereed Journal of Data Warehousing began publication in May 1996 and covers operational, conceptual, and research questions in the field. Thus, there is a vehicle for publication for AIS members

An example of the research and publication possibilities is the extension of the survey results reported above as well as additional survey research. Many of the respondents indicated an interest and willingness to participate in further data warehouse studies. The database of names is available from co-author of this paper, Hugh Watson, under specific conditions.

(1. Prepare A proposal for the study which appears promising. Include proposed cover letters, survey instruments, telephone interview scripts, etc.

2. Once the proposal is approved, the database will be made available for use only in the proposed study.

3. Researchers agree to submit a manuscript describing their findings to the Journal of Data Warehousing.)

Conclusions

A related set of developments (data warehouses, OLAP, MDD, KDD) are leading to new ways of performing decision support and creating executive information systems for data rich environments. They have created a new category of data base and a major set of applications. Yet, these developments have received almost no attention from academics either in research or in teaching. They are a fertile fields for both. We hope that this paper will stimulate our colleagues to join with us in studying this emerging field.