New Developments in Data Warehousing

Paul Gray
Claremont Graduate School

Hugh J. Watson
University of Georgia

Abstract

The authors attended the Second Annual Implementation Conference of TDWI in San Diego, California from January 20-25, 1997. They describe the things they did and did not see that caught their attention. Their observations provide insight on recent developments in data warehousing.

Introduction

The Second Annual Implementation Conference of TDWI was held January 20-25, 1997 in San Diego, California. As we attended sessions, visited the vendors’ exhibits, and talked with people, we frequently found ourselves saying, “Things are changing quite quickly.” We thought that it would be interesting and useful to describe what seem to be the most important and/or interesting developments. We also include comments on things that we did not see or hear, which surprised us by their absence. We are not predicting the future, Alan Paller does this nicely in the July 1996 issue of the Journal, but rather, are commenting on some of the changes that have already taken place.

Things That Are New in 1997

1. Internet/Intranet/Extranet capabilities for data warehouses are arriving at a much faster clip than had been imagined, even six months ago.
   • The browser is becoming a primary vehicle for end user access.
   • The thin client approach is being seen more and more as the Web comes to the fore.
   • The Web is being used for distributing and installing new software.
   • Companies are allowing external parties to access data warehouses, such as the EPA allowing the public to access environmental data through a browser.
   • There is integrated use of browsers, Web servers, and application servers to post reports to intranets and to support end-user queries.
   • Old and new applications alike are moving to the Web; it is increasingly becoming the favored application environment.

2. Data mining is starting to be imbedded in data warehouse and DSS/EIS products.
   • Data mining is still in its infancy. Despite the high-profile examples reported in the press, only a handful of companies are actually doing data mining.
   • Data mining is largely a product being developed by small vendors (companies of 100 people or less). The notable exception is IBM, who has invested heavily in this area.
   • Some vendors of DSS/EIS products (e.g., Pilot) are licensing the rights to integrate data mining capabilities developed by other vendors (e.g., Thinking Machines).
   • The data mining products that are currently available include only a subset of the available data mining techniques.
   • The challenge for vendors is to develop an interface that allows non-technical people to use and understand the “rocket science” algorithms that are used in data mining.

3. Suites of products are coming.
   • It is clear that people think that users are not specialized. A user may be working in executive information systems mode seeking a point-and-click answer at one time, directing a complex query at another, and doing OLAP-type analysis at another. In other words, every user needs multiple tool capabilities.
   • As complete suite products emerge, it will be interesting to see how many companies standardize on such a product or continue to go with multiple, best of breed products.
4. Metadata continues to be a big, unresolved issue.
   - The efforts of the Metadata Council, which appeared promising in 1996, seem not to have converged to concrete results.
   - The view of metadata is expanding. Rather than just being information about the contents of the data, it now includes information about the management of the data such as when the data was entered, when a job using particular data is scheduled, etc.

5. Managing the warehouse is an issue that has come to the fore.
   - Organizations that have data warehouses are becoming aware that their warehouse is a major resource for an organization and, like all resources, must be managed if its full value is to be obtained.

6. There are two opposite views of the role of the information systems department with respect to data warehousing.
   - Many speakers and vendors talked about data warehouses reducing the role of IS in organizations. That is, the products and concepts were touted as being suitable for direct implementation by end users and not requiring much support from IS.
   - On the other hand, there seems to be a trend toward decentralization of data, with IS having the responsibility for it. For example, the managed query environments require considerable IS support in order to allow users to make queries without knowing about data tables and SQL, and protecting against "the query from hell."

7. Data warehousing vendors are now international.
   - For many years, software developed in foreign countries came to the U.S. and failed to establish any significant market share. SAP was the first real exception. Now a larger number of DSS/EIS/managed query environment software products are from abroad, and some are doing quite well in the marketplace. The exhibitors included Business Objects, a French company, Andyne (Canada), Holos (UK), Cognos, (Canada), Planning Sciences (UK), Speedware (Canada), and Lingo (Canada).

8. Parallel computing, once thought exotic, is now thought of as standard, thanks to quad Pentium machines.

9. Security was notable by its absence.
   - Although speakers mentioned it occasionally ("of course, security must be considered"), there were no sessions devoted to it and none of the vendors we saw featured security as a key element of their offerings.

10. Databases are getting bigger. This finding has direct implications on the argument between relational OLAP (ROLAP) and multidimensional OLAP (MOLAP).
    - It was generally agreed that MOLAP is preferable if the database is small enough (a few gigabytes) and response time is critical, but only ROLAP could be used for very large databases.
    - One vendor claimed that a breakthrough had occurred, and that in early 1997, 16-terabyte MOLAP would become commercial.

11. Precalculating is faster than ROLAP or 64-bit parallelism.
    - The conventional wisdom on obtaining increased speed of response for large data warehouses is changing. People now believe that the key to faster response is understanding the users' information needs well enough to make precalculations and store them in the warehouse for later retrieval.

12. A new phrase, "We deliver the data mart; you do your own analysis" refers to a data mart that comes in a box. This reduces sales costs and avoids the vendor's issue about what you will do with the data mart once it is installed.

13. Little or no discussion of how to determine data requirements.
    - The vendors and presenters said little about how organizations should go about deciding what data to place in their warehouses. It was assumed that data requirements are easily obtained. This seems to be a serious omission.
    - By contrast, the importance of having an enterprise data model was mentioned repeatedly.
New Developments in Data Warehousing, continued

BIographies

Paul Gray is a Professor and Chair of the Program in Information Science at the Claremont Graduate School in Claremont, California. He worked for many years in decision support and related fields and is currently involved in data warehousing. His experience includes eighteen years in industry.

Paul Gray  
Programs in Information Science  
Claremont Graduate School  
Claremont, CA 91711-6190  
PHONE: 909-621-8209  
Email: grayp@cgs.edu

Hugh J. Watson is a Professor of MIS and holds a C. Herman and Mary Virginia Terry Chair of Business Administration at The University of Georgia. He is the author of 21 books and over 100 journal articles. His research and consulting focus on executive information systems, decision support, and data warehousing. He is the senior editor of the Journal of Data Warehousing.

Hugh J. Watson  
Terry College of Business  
The University of Georgia  
Athens, GA 30602-6256  
PHONE: 706-542-3744  
Email: hwatson@uga.cc.uga.edu