Managing The Data Warehouse
An Irreverent View

Alan Paller
Director of Education and Research
The Data Warehousing Institute

June, 1997

Executive Overview
For decades, decision support systems were not managed – they just happened – and if they were unavailable, people waited. Today, on the other hand, the decision support functions offered by data warehouses are mission critical. High-performance, twenty-four hour by seven-day-per-week operations are demanded by an increasing number of companies. Data warehousing managers are starting to lose sleep over questions such as "Are all the updates being made before the system goes live?" "Is the performance going to be a problem for the key users?" "How will we rebuild the database when we have a crash?" "Can users find the data they need as the warehouse grows?" and on and on.

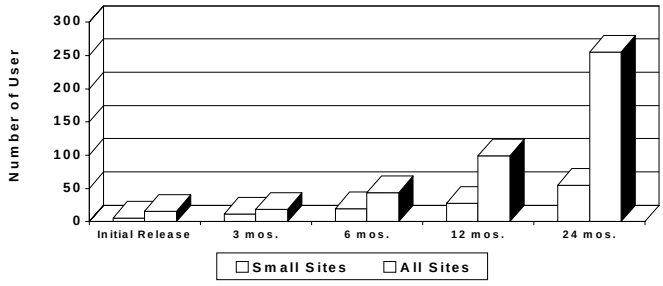
Management tool vendors over-promise and under-produce at a rate that would embarrass a huckster. Smart companies ignore all the unfinished tools; they set up their own management systems. This paper describes some of the techniques that leading user organizations have found or created to solve the most difficult problems involved in managing data warehouses

Section I: Performance

According to The Data Warehousing Institute's "Ten Mistakes To Avoid," slow performance is the primary challenge that causes pain for warehousing managers who have had operational systems for at least 90 days. Three common methods used to speed performance in a data warehouse (though common, these are not the ultimate solution; we'll cover that one in the presentation at HP World in Chicago) are:

1. Increase the size of the hardware – memory, CPU, channels.
2. Rewrite the queries to use a more efficient path and rearrange the data on the disk drives to allow more rapid retrieval.
3. Move to parallel hardware, 64-bit hardware, or bit-mapped indexing of the data.

Hardware upgrades are inevitable if you have a successful data warehouse. The number of users grows approximately 100 percent every six months and the average query rises at the same time the number of queries is rising. In combination, these two forces push the demand for processing



far above any initial use and create massive demand for new capacity.

Section II. Metadata Management

When companies spend money on metadata projects, they are generally attempting to solve two data warehousing management problems:

1. Helping the data warehousing team track the origins and transformations for each piece of information stored in the data warehouse and
2. Helping users of the warehouse identify the information they want and gain confidence.

Metadata software tools rarely solve either of these problems completely, and in their incompleteness, they cause so much additional effort for the data warehousing team, that their use may have a negative impact on productivity and value to the user.

The common solution to these challenges is the do-it-yourself data directory that includes both the source/transformation data and the user directory. One of the most innovative models for the user directory is the L.L. Bean catalogue. Using that model, a company builds a catalogue of data that is easy to navigate, easy to understand (written in plain English), active in giving advice about which is best under various circumstances, and able to deliver the information without forcing the user to switch to another tool. Many companies use intranets and even extranets to deliver this type of user information directory.

Section III. Usage Monitoring

Some data is unimportant, but you'll never know which data that is without monitoring tools. You'll want to monitor which data are used, how often, and through what queries. Your goals are (1) to reduce the amount of data you keep by excising information that is never used and (2) to identify the queries that can be reused and then either cache the results or optimize the queries themselves or both. Hewlett Packard's Intelligent Warehouse tool is quite effective for these purposes, but you'lw ant to make sure it runs on all the hardware you plan to use for data warehousing before considering it.

Section IV. Managing End-User Tools

There are two widely used approaches to the management of end-user tools: ignorance and control. In the ignorance model (which is probably more intelligent than the alternative) the IT department allows user organizations to acquire whatever end-user data access tools they consider valuable. The IT department's job is limited to supplying the data. The alternative approach sets the IT department up as czar of end user tools. Every update must be installed on every computer (these people still haven't heard of easy to manage "thin-client tools), and the semantic layer (the tables that translate English requests into table names) must be constantly maintained.

There's no "right way" to manage end user computing, but one change coming during late 199 and early 1998 will ease the burden. Much as the "office" products brought together multiple end-user tools, so will families of information access tools bring together this category. Every forward-looking data warehousing tools vendor is working on a strategy to provide clients with a complete set of reporting, query, OLAP, ROLAP and even mining – all from one vendor and all with a common user interface. As these comprehensive offerings mature, companie will begin to standardize on one or two families and thus ease the management burden.

Section V. Serving Warehouse Data On The Web

The killer application for data warehousing is the extranet where clients of the company use the web to access the warehouse to get answers to questions that would have taken hours or days or even longer to answer. But opening up the warehouse to outside users creates two problems – security and performance contention. We'll discuss security in the next session. Here we talk about performance contention.

Hundreds or thousands of outside users can swamp a data warehousing server just at the time that inside users get their work done. To meet this challenge data warehousing managers are designing an extra tier into their warehouses to serve as a resource allocation controller. They set the new machine up between the users and the application server and monitor loads on both the application server and database servers. Then as loads start to rise toward capacity, the resource allocator limits access from outside or otherwise lower priority) users.

Section VI. Security For Web Delivery

When your boss asks you whether the corporate information will be secure over your extranet, the only rational answer is "No! But delivering the data offers so much value that we'll just have to live with the threats." To minimize the threats, you'll want to use physical security devices such as tokens or fingerprint machines, encryption, and tight policies. However, no security system can completely stop the threats such as denial of service or social engineering.

Smart warehousing/web managers have teams of security professionals working constantly to stay a few steps ahead of the hackers. The bible for most such "watchers" is the SANS Network Security Digest which provides the consensus view from all twelve top security gurus about the most important new threats and any solutions that have been found.

Section VI. Data Warehousing Roles and Responsibilities

It might seem as if the data warehouse is the easiest of fields to find qualified people. That's not so. Both back end and front end data warehousing professionals must have such a broad knowledge and skill set that few people ever completely fill the position. Back end managers, for example, must know every legacy data source, all transformations, scheduling and quality control procedures, database management software selection, dbms tuning and data partitioning, hardware acquisition, management and maintenance, data cleaning, and dozens of other similarly complex tasks.

Similarly the front end manager must know C++, Visual Basic, and/or Powerbuilder, as well as the popular data warehousing query tools. It isn't sufficient to provide your users with query or OLAP tools exclusively. You'll find you need to develop specific applications to meet two goals: (1) limiting complexity for users who don't want to learn query tools just to get simple answers and (2) keeping users from creating queries that time out after hours of fruitless machinations.

The Bottom Line

Practical data warehousing managers have stopped going to expositions to listen to vendor claims . Instead, they are focusing their efforts on users in similar situations who have solved the tough problems. The Data Warehousing Institute brings such users together four times each year (the next one is in Washington in November) to share the lessons they have learned and to discuss the trends shaping the future of this fast-moving area. For information on these meetings email info@dw-institute.com or call 301-947-3731.