

---

## Data Warehouse Management

---

Dan Vesset  
September 2003

### INTRODUCTION

In the previous sections of this series we introduced the closed-loop model — and its relationship to business analytics — and the individual software components that support the model. These software components act together as building blocks of a complete business analytics system, providing organizations with a unified platform while maintaining flexibility in addressing unique analytic needs of various end-user groups. The core of the platform is based on the data warehouse management software, which includes data stores specifically designed for analytic processing. Data warehousing process is concerned with storage and preparation of data for use by business intelligence and advanced analytics tools as well as prepackaged analytic applications.

Business analytics projects have traditionally involved departmental solutions. These standalone systems have proliferated in many of today's decentralized organizations. Because operational decision-making processes are unique to specific departments or business units, the software to support such processes has resulted in disconnected data silos. Ongoing mergers and acquisitions further aggravate the problem of disparate data stores. The decentralization of data stores makes it difficult to provide users at all levels of the organization with a consistent view of enterprisewide information. Regardless of the sophistication of any one analytic application, in this case the return on investment of such disparate systems will remain substandard from an enterprise perspective.

A solid data warehousing foundation can overcome the challenge of data silos by providing consistent data and information to users ranging from executives and analysts to partners and suppliers. At the same time, a well-planned and executed data warehousing strategy should not force a change in organizational structure. If a company is managed as a decentralized organization, the unified data warehousing architecture should make its benefits available to end users transparently. For example, managers and analysts in geographically disbursed business units can benefit from consistency of the data while maintaining their independence within the corporate organizational structure.

---

### TRANSACTIONAL VERSUS ANALYTIC DATA STORES

Data warehouses differ from transactional database management systems in being purposely built for analytic processing. The two serve different purposes and require different features and functionality. While transaction processing systems are developed to handle a large number of concurrent users, they access few records at a time. Data warehouses are developed to handle a much smaller number of complex and processing-intensive queries that deal with many records at a time. The overhead costs to manage a transaction processing system should not be part of a data warehouse, nor should analytic processing interfere with daily operational systems.

The market for data warehouse management tools grew to \$3.6 billion in 2002. Much of the data warehouse management tools software is based on relational database management systems (RDBMS). A major trend in this market segment continues to be the embedding of analytic functions into a single relational database. To date, this trend has focused on multidimensional analysis and data mining. However, the addition of reporting and spatial information management functions is also gaining momentum.

A unified data warehouse should support multiple analytic and information delivery needs, from basic reporting to OLAP and advanced analytics such as data mining. To this end, data warehouse solutions should handle relational, multidimensional, and other data storage and management techniques.

The two primary storage schemas, relational and multidimensional, are complementary and should be used either in combination or individually depending on the end users' analytic needs. They are defined as follows:

- ☒ Multidimensional schema is best suited for well-defined analytic needs. In this format, large volumes of data are aggregated and precalculated, providing end users with fast access to summary data. Multidimensional data stores serve OLAP end-user business intelligence tools.
- ☒ Relational schema is utilized when ad hoc query requirements exist. In this format, detailed data is kept in the relational DBMS and summarizations and calculations are performed at the time of query execution.

Increasingly, database vendors are providing data warehouse management solutions that combine the two formats, allowing end users to seamlessly drill down from summaries to the underlying detail. Organizations should choose storage options to match user requirements. Depending on whether these include reporting, OLAP, or data mining and statistical analysis, the data file structures will differ. Besides multidimensional and relational databases, data could be held in flat files, object-relational, hierarchical, or XML databases.

When a data warehouse is being implemented, the most important features to consider are scalability, performance, availability, and security, which are defined as follows:

- ☒ **Scalability.** This refers to the ability to manage both a growing number of users and increasing data volumes. IDC research indicates that user populations are growing both within organizations and outside with suppliers, partners, and customers gaining access to information held in data warehouses. At the same time, data volumes are also growing. A research study conducted in 2002 by DM Review and IDC showed that 65% of all organizations expect their data warehouses to grow by at least 100% over a three-year period. Scalability of the business analytics architecture is one of the cornerstones of long-term success of supporting all the different end-user groups. For example, there are many more information consumers than business analysts or model developers (look for a discussion of the various users groups in the upcoming section on business intelligence). In supporting the former group, user scalability is an important consideration, while for the latter group, data scalability is more significant.

- ☒ **Performance.** This refers to the processing speed of the data warehouse. Organizations should look for data warehouse management solutions incorporating parallel processing techniques to speed data processing requests. The latest software tools also incorporate techniques to decrease I/O (input/output) tasks, which are often a bottleneck in database performance.
- ☒ **Security.** This refers to the management of user access rights. A data warehouse serves as a foundation for decision-making processes for different internal and external users. Each group of users, ranging from executives to information consumers, has different requirements and access rights to the data. In addition to internal users, an increasing number of external stakeholders such as suppliers and partners are provided with access to data warehouses. The addition of external user groups places further importance on security management in business analytics environments.
- ☒ **Availability.** This refers to the uptime of the data warehouse. In the past, the availability of data warehouses was less crucial. When analysis of data was the responsibility of the few, the impact of a delay in bringing a data warehouse back online following a load or a crash was minimal. Today, when the data warehouse serves an increasingly strategic role as the foundation for decision making for a multitude of users, such delays are becoming unacceptable. This is particularly important for global business when users are accessing a central data warehouse across multiple time zones at any hour of the day or night. Reporting functionality that is provided to clients needs to adhere to the same high standards of availability as in transactional systems.

Organizations should carefully evaluate data warehouse management solutions based on all four of these features and plan for ever-increasing demand on the system from growing data volumes and expanding user populations.

---

## METADATA MANAGEMENT

In evaluating best-in-class data warehouse management solutions, organizations should also consider the metadata management facilities of the software. As discussed in the previous section, metadata is an enabler of integration between the data warehouse and other software components of the entire business analytics solution. The data warehouse is the control point of the data integration process for decision support, requiring the ability to integrate the underlying metadata. Solutions based on standard metadata management techniques help to minimize software integration and ongoing data warehouse maintenance costs.

---

## CONCLUSION

It is often difficult for IT to consolidate on a single end-user tool or application given the decentralized nature of many organizations. However, the underlying data warehousing platform can provide a level of uniformity that facilitates enterprisewide consistency in decision making.

Data warehouses should be tightly integrated with ETL and data quality tools. This integration ensures efficiency in data warehouse generation and management thereby decreasing the total cost of ownership of the system. The integration with ETL, business intelligence, and advanced analytics tools should be accomplished through integrated metadata, which opens the specialized features of the warehouse to all the relevant applications built on it. Without this unified metadata view, integration and management costs will curtail the return on investment from business analytics solutions.

Finally, data warehouse management solutions should provide relational, OLAP, and parallel data management options able to support applications ranging from simple reporting and OLAP to advanced forecasting and data mining. These options should be provided transparently to applications through a unified metadata management environment. Metadata not only links individual software components provided by one software vendor, but it also has the potential to open a data warehousing platform from one provider to third-party analytic tools and applications. Emerging industry standards, such as the Common Warehouse Metamodel (CWM) will facilitate this interoperability. The result will be decreased software integration and ongoing maintenance costs for business analytics solutions.

---

#### COPYRIGHT NOTICE

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2003 IDC. Reproduction without written permission is completely forbidden.