Data Warehousing
Concepts and Strategies

Stefan M. Neikes
Sumit Sircar
Bijoy Bordoloi

Many IT organizations are increasingly adopting data warehousing as a way of improving their relationships with corporate users. Proponents of data warehousing technology claim the technology will contribute immensely to a company’s strategic advantage. According to the Gartner Group, by 1998, the $2 billion data warehouse market quadrupled to an incredible $8 billion.

Companies contemplating the implementation of a data warehouse need to address many issues concerning strategies, type of data warehouse, front-end tools, and even corporate culture. Other issues that also need to be examined include who will maintain the data warehouse and how often and most of all, which corporate users will have access to it.

After defining the concept of data warehousing, this chapter provides an in-depth look at design and construction issues, types of data warehouses and their respective applications, data mining concepts, techniques, and tools, and managerial and organizational impacts of data warehousing.

HISTORY OF DATA WAREHOUSING

The concept of data warehousing is best presented as part of an evolution that begins about 35 years ago. In the early 1960s, the arena of computing was limited by punch cards, files on magnetic tape, slow access times, and an immense amount of overhead. About the mid-1960s, the near explosive growth in the usage of magnetic tapes increased the amount of data redundancy. Suddenly, new problems, ranging from synchronizing data after
updating to handling the complexity of maintaining old programs and
developing new ones, had to be resolved.

The 1970s saw the rise of direct access storage devices and concomitant
technology of database management systems (DBMSs). DBMSs made it
possible to reduce the redundancy of data by storing it in a single place for
all processing. Only a few years later, databases were used in conjunction
with online transaction processing (OLTP). This advancement enabled the
implementation of such applications as automated teller machines and res-
ervations systems used by travel and airline industries to store up-to-date
information. By the early 1980s, the introduction of the PC and fourth-gen-
eration technology let end users innovatively and more effectively utilize
data in the database to guide decision making.

All these advances, however, engendered additional problems, such as
producing consistent reports for corporate data. It was difficult and time-
consuming to accomplish the step from pure data to information that gives
meaning to the organization and a lack of integration across applications.
Poor or non-existent historical data only added to the problems of trans-
forming raw data into intelligent information.

This dilemma led to the realization that organizations need two funda-
mentally different sets of data. On the one hand, there is so-called primitive
data, which is detailed, can be updated, and is used to run the day-to-day
operations of a business. On the other hand, there is summarized or
derived data, which is less frequently updated and is needed by manage-
ment to make higher-level decisions. The origins of the data warehouse, as
a subject-oriented collection of data that supports managerial decision
making are therefore not surprising.

Many companies have finally realized that they cannot ignore the role of
strategic information systems if they are to attain a strategic advantage in
the marketplace. CEOs and CIOs throughout the U.S. and the world are
steadily seeking new ways to increase the benefits that IT provides. Data is
increasingly viewed as an asset with as much importance in many cases as
financial assets. New methods and technologies are being developed to
improve the use of corporate data and provide for faster analyses of busi-
ness information.

Operational systems are not able to meet decision support needs for
several reasons. First, most organizations lack online historical data. Sec-
ond, the data required for analysis often resides on different platforms and
operational systems, which complicates the issue further. Third, the query
performance of many operational systems is extremely poor, which in turn
affects their performance. Fourth, operational database designs are inap-
propriate for decision support.
For these reasons, the concept of data warehousing, which has been around for as long as databases have existed, has suddenly come to the forefront. A data warehouse eliminates the decision support shortfalls of operational systems in a single, consolidated system. Data is thus made readily accessible to the people who need it, especially organizational decision makers, without interrupting online operational workloads.

The key of a data warehouse is that it provides a single, more quickly accessible, and more accurately consolidated image of business reality. It lets organizational decision makers monitor and compare current and past operations, rationally forecast future operations, and devise new business processes. These benefits are driving data warehousing’s popularity and have led some advocates to call the data warehouse the center of IS architecture in the years ahead.

THE BASICS OF DATA WAREHOUSING TECHNOLOGY

According to Bill Inmon. Author of *Building the Data Warehouse* (NY: John Wiley. 1993), a data warehouse has four distinguishing characteristics:

1. Subject-orientation.
2. Integration.
3. Time-variance.

As depicted in Exhibit 1, the subject-oriented database characteristic of the data warehouse organizes data according to subject, unlike the application-based database. The alignment around subject areas affects the design and implementation of the data found in the data warehouse. For this reason, the major subject areas influence the most important part of the key structure. Data warehouse data entries also differ from application-oriented data in the relationships. Although operational data has relationships among tables based on the business rules that are in effect, the data warehouse encompasses a spectrum of time.

A data warehouse is also integrated in that data is moved there from many different applications (Exhibit 2). This integration is noticeable in several ways, such as the implementation of consistent naming conventions, consistent measurement of variables, consistent encoding structures, and consistent physical attributes of data. In comparison, operational data is often inconsistent across applications. The preprocessing of information aids in reducing access time at the point of inquiry.

Exhibit 3 shows the time-variant feature of the data warehouse. The data stored is about five to ten years old and used for making consistent comparisons, viewing trends, and providing a forecasting tool. Operational environment data reflects only accurate values as of the moment of access.
The data in such a system may change at a later point in time through updates or inserts. On the contrary, data in the data warehouse is accurate as of some moment in time and will produce the same results every time for the same query.

The time-variant feature of the data warehouse is observed in different ways. In addition to the lengthier time horizon as compared to the operational environment, time-variance is also apparent in the key structure of a data warehouse. Every key structure contains — implicitly or explicitly — an element of time, such as day, week, or month. Time-variance is also evidenced by the fact that data warehouse is never updated. Operational data is updated as the need arises.


Exhibit 1. The data warehouse is subject-oriented.
Exhibit 2. Integration of data in the data warehouse.

Exhibit 3. The data warehouse is time-variant.
The nonvolatility of the warehouse means that there is no inserting, deleting, replacing, or changing of data on a record-by-record basis, as is the case in the operational environment (Exhibit 4). This difference has tremendous consequences. At the design level, for example, there is no need to be cautious about update anomaly. It follows that normalization of the physical database design loses its importance, because the design focuses on optimized access of data. Other issues that simplify data warehouse design involve the nonpresence of transaction and data integrity as well as detection and remedy of deadlock, which is found in every operational database environment.

Effective and efficient use of the data warehouse necessitates that the data warehouse run on a separate platform. If it does not, it will slow down the operations database and reduce response time by a large factor.

**DESIGN AND CONSTRUCTION OF A DATA WAREHOUSE**

**Preliminary Considerations**

Like any other undertaking, a data warehouse project should demonstrate success early and often to upper management. This ensures high visibility and justification of the immense resources commitment and costs associated with the project. Before undertaking the design of the data warehouse, however, it is wise to remember that a data warehouse project is not as easy as copying data from one database to another and

---

**Exhibit 4. The data warehouse is nonvolatile.**

Data is regularly updated on a record-by-record basis. Data is loaded into the warehouse and is accessed there, but once the snapshot of data is made, the data in the warehouse does not change.

Developers should not underestimate the many complex issues involved in data warehousing. These include architectural considerations, security, data integrity, and network issues. According to one estimate, about 80% of the time that is spent constructing a data warehouse is devoted to extracting, cleaning, and loading data. In addition, problems that may have been undetected for years can surface during the design phase. The discovery of data that has never been captured as well as data that has been altered and stored are examples of these types of problems. A solid understanding of the business and all the processes that have to be modeled are also extremely important.

Another major consideration important to up-front planning is the difference between the data warehouse and most other client/server applications. First, there is the issue of batch orientation for much of the processing. The complexity of processes (which may be executed on multiple platforms), data volumes, and resulting data synchronization issues must be correctly analyzed and resolved.

Next, the data volume in a data warehouse, which can be in the terabyte range, has to be considered. New purchases of large amounts of disk storage space and magnetic tape for backup should be expected.

It is also vital to plan and provide for the transport of large amounts of data over the network. The ability of data warehousing to support a wide range of queries, from simple ones that return only limited amounts of information to complex ones that might access several million rows, can cause complications. It is also necessary to incorporate the availability of corporate metadata into this thought process. The designers of the data warehouse have to remember that metadata is likely to be replicated at multiple sites. This point to the need for synchronization across the different platforms to avoid inconsistencies.

Finally, security must be considered. In terms of location and security, data warehouse and non-data warehouse applications must appear seamless. Users should not need different IDs to sign on to the different systems, but the application should be smart enough to provide users the correct access with only one password.

**Designing the Warehouse**

After having addressed all the preliminary issues, the design task begins. There are two approaches to designing a data warehouse: the top-down approach and the bottom-up approach. In the top-down approach, all of an organization’s business processes are analyzed to build an enterprisewide
data warehouse in one step. This approach requires an immense commitment of planning, resources, and time and results in new information structure from which the entire organization benefits.

The bottom-up approach, on the other hand, breaks the task down and delivers only a small subset of the data warehouse. New pieces are then phased in until the entire organization is modeled. The bottom-up approach lets data warehouse technology be quickly delivered to a part of the organization. This approach is recommended because its time demands are not as rigorous. It also allows development team members to learn as they implement the system, identify bottlenecks and shortfalls, and find out how to avoid them as additional parts of the data warehouse are delivered.

Because a data warehouse is subject-oriented, the first design step involves choosing a business subject area to be modeled and eliciting information about the following:

- The business process that needs to be modeled.
- The fact that need to be extracted from the operational database.
- The level of detail required.
- Characteristics of the facts (e.g., dimension, attribute, and cardinality).

After each of these areas has been thoroughly investigated and more information about facts, dimension, attributes, and sparsity has been gathered still another decision must be made. The question now becomes which schema to use for the design of the data warehouse database. There are two major options: the classic star schema and the snowflake schema.

The Star Schema. In the star design schema, a separate table is used for each dimension, and a single large table is used for the facts (Exhibit 5). The fact table's indexed key comprises the keys of the different dimensions tables. With this schema, the problem of sparsity, or the creation of empty rows, is avoided by not creating records where combinations are invalid. Users are able to follow paths for detailed drilldowns and summary rollups. Because the dimension tables are also relatively small, precalculated aggregation can be imbedded within the fact table, providing extremely fast response times. It is also possible to apply multiple hierarchies against the same fact table, which leads to the development of a flexible and useful set of data.

The Snowflake Schema. The snowflake schema is depicted in Exhibit 6 is best used when there are large dimensions such as time. The dimension tables are split at the attribute level to provide a greater variety of combinations. The breakup of the time dimension into a quarter entity and a
month entity provides more detailed aggregation and also more exact information.

Decision Support Systems and Data Warehousing

Because many vendors offer decision support systems (DSS) products and information on how to implement them abounds, insight into the different technologies available is helpful. Three concepts should be evaluated in terms of their usability for decision support and relationship to the so-called real data warehouse. They are virtual data warehouses, multidimensional online analytical processing (OLAP), and relational OLAP.

The Virtual Data Warehouse

The virtual data warehouse promises to deliver the same benefits as a real data warehouse but without the associated amount of work and difficulty. The virtual data warehouse concept can be subdivided into the surround data warehouse and the OLAP/data mart warehouse. In a surround data warehouse, legacy systems are merely surrounded with methods to access data without a fundamental change of the operational data. The surround concept


Exhibit 5. The star design schema.
thus negates a key feature of the real data warehouse, which integrates operational data in a way that allows users to make sense of it.

In addition, the data structure of a virtual data warehouse does not lend itself to DSS processing. Legacy operational systems were built to ease updating, writing, and deleting and not with simple data extraction in mind. Another deficiency with this technology is the minimal amount of historical data that is kept, usually only 60 to 90 days worth of information. A real data warehouse, on the other hand, with its two-to-five years worth of information, provides a far superior means of analyzing trends.

In the case of direct OLAP/data marts, legacy data is transferred directly to the OLAP/data mart environment. Although this approach recognizes
the need to remove data from the operational environment, it too falls short of being a real data warehouse. If only a few, small applications were feeding a data mart, the approach would be acceptable. The reality is, however, that there are many applications and thus many OLAP/data mart environments, each requiring a customized interface, especially as the number of OLAP/data marts increases.

Because the different OLAP/data marts are not effectively integrated, different users arrive at different conclusions when analyzing the data. It is thus possible for the marketing department to report the business is doing fine and another department to report just the opposite. This drawback does not exist with the real data warehouse, where all data is integrated. Users who examine the data at a certain point in time would all reach the same conclusions.

**Multidimensional OLAP**

Multidimensional database technology is a definite step up from the virtual data warehouse. It is designed for executives and analysts who want to look at data from different perspectives and have the ability to examine summarized and detailed data. When implemented together with a data warehouse, multidimensional database technology provides more efficient and faster access to corporate data. Proprietary multidimensional databases facilitate the organization of data hierarchically in multiple dimensions, allowing users to make advanced analyses of small portions of data from the data warehouse. The technology is understandably embraced by many in the industry because of its increased usability and superior analytical functionality.

As a standalone technology, multidimensional OLAP is inferior to a real data warehouse for a variety of reasons. The main drawback is that the technology is not able to handle more than 20 to 30 gigabytes of data, which is unacceptable for most of the larger corporations, whose need range in the 100 gigabyte to several terabyte range. Furthermore, multidimensional databases does not have the flexibility and measurability required of today’s decision support systems because they do not support the necessary ad hoc creation of multidimensional views of products and customers. Multidimensional databases should be considered for use in smaller organizations or on a department level only.

**Relational OLAP**

Relational OLAP is also used with many decision support systems and provides sophisticated analytical capability in conjunction with a data warehouse. Unlike multidimensional database technology, relational OLAP lets end users define complex multidimensional views and analyze them.
These advantages are only possible if certain functionalities are incorporated into relational OLAP.

Users must be removed from the process of generating their own structured query language (SQL). Multiple SQL statements should be generated by the system for every analysis request to the data warehouse, in this way, a set of business measurements (e.g., comparison and ranking measurements) is established, which is essential to the appropriate use of the technology.

The shortcoming of relational OLAP technology works well in conjunction with a data warehouse, by itself, the technology is somewhat limited.

Examination of the three preceding decision support technologies leads to the only correct deduction—that data warehouse is still the most suitable technology for larger firms. The benefit of having integrated, cleansed data from legacy systems together with historical information about the business makes a properly implemented data warehouse the primary choice for decision support.

**Benefits of Warehousing for Data Mining**

The technology of data mining is closely related to that of data warehousing. It involves the process of extracting large amounts of previously unknown data and then using the data to make important business decisions. The key phrase here is unknown information buried in the huge mounds of operational data that, if analyzed, provides relevant information to organizational decision makers.

Significant data is sometimes undetected because most data is captured and maintained by a particular department. What may seem irrelevant or uninteresting at the department level may yield insights and indicates patterns important at the organizational level. These patterns include market trends, such as customer buying patterns. They aid in such areas as determining the effectiveness of sales promotions, detecting fraud, evaluating risk and assessing quality, or analyzing insurance claims. The possibilities are limitless and yield a variety of benefits ultimately leading to improved customer service and business performance.

Data that is needed but often located on several different systems, in different formats and structures, and somewhat redundant provides no real value to business users. This is where the data warehouse comes into play. As a source of consolidated and cleansed data facilitating analysis than do regular flat files or operational databases.

Three steps are thus needed to identify and use hidden information:
1. The captured data must be incorporated into view of the entire organization instead of only one department.
2. The data must be analyzed or mined for valuable information.
3. The information must be specially organized to simplify decision making

**DATA MINING TASKS**

In data mining, data warehouses, query generators, and data interpretation systems are combined with discovery-driven systems to provide the ability to automatically reveal important yet hidden data. The following tasks need to be completed to make full use of data mining.

- Creating prediction and classification models.
- Analyzing links.
- Segmenting databases.
- Detecting deviations.

**Creating Models.** The first task makes use of the data warehouse’s contents to automatically generate a model that predicts desired behavior. In comparison to traditional models that use statistical techniques and linear and logical regression, discovery-driven models generate accurate models that are also more comprehensible, because of their sets of if-then rules. The performance of a particular stock, for example, can be predicted to assess its suitability for an investment portfolio.

**Analyzing Links.** The goal of the link analysis is to establish relevant connections between database records. An example here is the analysis of items that are usually purchased together, like a washer and dryer. Such analysis can lead to a more effective pricing and selling strategy.

**Segmenting Databases.** When segmenting databases, collections of records with common characteristics or behaviors are identified. An example is the analysis of sales for a certain time period, such as President’s Day or Thanksgiving weekend, to detect pattern in customer purchase behavior. For the reasons discussed earlier, this is an ideal task for data warehouse.

**Detecting Deviations.** The fourth and final task involves detection of deviation, which is the opposite of data segmentation. Here, the goal is to identify records that vary from the norm, or lie outside of any particular cluster with similar characteristics. This discovery from the cluster is then explained as normal or as a hint of a previously unknown behavior or attribute.
DATA MINING TECHNIQUES

At this point, it is important to present several techniques that aid mining efforts. These techniques include the creation of predictive models, and the performing of supervised induction, association, and sequence discovery.

Creating Predictive Models. The creation of a so-called predictive model is facilitated through numerous statistical techniques and various forms of visualization that ease the user’s recognition of patterns.

Supervised Induction. With supervised induction, classification models are created from a set of records, which is referred to as the training set. This method makes it possible to infer from a set of descriptors of the training set to the general. In this way, a rule might be produced that states that a customer who is male, lives in a certain zip code area, earns $25,000 and $30,00 is between 40 and 45 years of age, and listens more to the radio than watches TV might be a possible buyer for a new camcorder. The advantage of this technique is that the patterns are based on local phenomena, whereas statistical measures check for conditions that are valid for an entire population.

Association Discovery. Association discovery allows for the prediction of the occurrence of some items in a set of records if other items are also present. For example, it is possible to identify the relationship among different medical procedures by analyzing claim forms submitted to an insurance company. With this information the prediction could be made, within a certain margin of error, that for treatment usually the same five medicine are required.

Sequence Discovery. Sequence discovery aids the data miner by providing information on a customer’s behavior over time. If a certain person buys a VCR this week, he or she usually buys videotapes on the next purchasing occasion. The detection of such a pattern is especially important to catalog companies, because it helps them better target their potential customer base with specialized advertising catalogs.

Tools

The main tools used in data mining are neural networks, decision trees, rule induction, and data visualization.

Neural Networks. A neural network consists of three interconnected layers: an input and output layer with a hidden layer in between (Exhibit 7). The hidden processing layer is like the brain of the neural network because it stores or learns rules about input patterns and then produces a known set of outputs. Because the process of neural networks is not transparent,
it leaves the user without a clear interpretation of the resulting model, which, nevertheless, is applied.

**Decision Trees.** Decision trees divide data into groups based on the values that the different variables take on (see Exhibit 8). The result is often a complex hierarchy of classifying data, which enables the user to deduct possible future behavior. For instance, it might be deducted that for a person who only uses a credit card occasionally, there is a 20% probability that an offer for another credit card would be accepted. Although decision trees are faster than neural networks in many cases, they have drawbacks. One of theses is the handling of data ranges as in age groups, which can inadvertently hide patterns.

**Rule Induction.** The method of rule induction is applied by creating nonhierarchical sets of possibly overlapping conditions. This is accomplished by first generating partial decision trees. Statistical techniques are then used to determine which decision trees to apply to the input data. This method is specially useful in cases where there are long and complex condition lists.

**Data Visualization.** Data visualization is not really a data mining tool. However, because it provides a picture for the user with as many as four graphically represented variables, it is a power tool for providing concise 

![Exhibit 7. Neural network.](image-url)
information. The graphics products available make the detection of patterns much easier than is the case when more numbers are analyzed.

Because of the pros and cons of the varied data mining tool, software vendors today incorporate all or some of them in their data mining packages. Each tool is essentially a matter of looking at data with different means and from different angles.

One of the potential problems in data mining is performance-related. To speed up processing, it might be necessary to subset the data either by the number of rows access or by the number of variables that examined. This can lead to slightly different conclusions about data set, consequently, in most cases it is better to wait for the correct answer using a large sample.

**Managerial and Organizational Impacts of Data Warehousing**

Although organizational managers eagerly await the completion of a data warehouse, many issues must be dealt with before the fruits of this new technology are harvested. This is especially true in today's fast changing enterprise with its quick reaction times.

The subject of economic benefit also deserves mentioning when dealing with data warehousing because some projects have already acquired the reputation of providing little or no payback on the huge investments involved. Data warehouses are sometimes accused of being pits into which data disappears never to be seen again.

---

*Exhibit 8. Decision tree.*

---

Managers have to understand at the outset that the quality of the data is of extreme importance in a data warehousing project. The sometimes difficult challenge for management is to make data entering the data warehouse consistent. In some organizations, data is stored in flat, VSAM, IMS, IDMS, or SA files and a variety of relational databases. In addition, different systems that were designed for different functions contain the same terms, used different meanings.

If care is not taken to clean up this terminology during data warehouse construction, misleading management information results. The logical consequence of this requirement is that management has to agree on the data definition for elements in the warehouse. This is yet another challenging task. People who use the data in the short term and the long term must have input into the process and know what the data means.

The manager in charge of loading the data warehouse has four ways to handle erroneous data. If the data is inaccurate, it must be completely rejected and corrected in the source system. Data may also be accepted as is, if it is within a certain tolerance level and if it is marked as such. An example here is a nine-digit code from the operational database.

A third option for handling erroneous data is the capture and correction of the data before it enters the warehouse. Capture and correction are handled programmatically in the process of transforming data from one system to the data warehouse. An example might be a field that was in lowercase and needs to be stored in uppercase. A final means of handling errors is to replace erroneous data with a default value. If, for example, the data Feb. 29 of a nonleap year is defaulted to Feb. 28, there is no loss in data integrity.

Another way that data warehousing affects management and organizations in general concerns today’s business motto of working smarter, not harder. Today’s data warehouse users can become more productive, because they will have the tools to analyze the huge amounts of data that they store instead of just collecting it.

Organizations are also affected by the invalid notion the implementing data warehousing technology simply consists of integrating all pertinent existing company data into one place. Managers need to be aware that data warehousing implies changes in the job duties of many people. For example, in an organization implementing a data warehouse, data analysis and modeling become much more prevalent than just requirements analysis. The database administrator position does not merely involve the critical aspects of efficiently storing data but takes on the central role in the development of the application. Furthermore, because of it data model-oriented methodology, data warehouse design requires a development life cycle that reverses traditional development approaches, the development of a
data warehouse virtually begins with a data model, from which the warehouse is built.

In summary, it must be noted that data warehouses are high-maintenance systems that require their own staffs. In this way, future changes are implemented in a timely manner by experienced personnel. It is also important to remember that a technically advanced and fast warehouse that adds little value will probably be abandoned by users from the start, reiterating the immense importance of clean data.

One of the most important issues that is often disregard during the construction and implementation of a data warehouse is data quality. This is not surprising because in many companies the concern for data quality in regard to legacy and transaction systems is not a priority. Accordingly, when it comes to ensuring the quality of data being moved into the warehouse many companies continue with their old practices. This can turn out to be a costly mistake and has already lead to many failures of corporate warehousing projects. As more and more companies are making use of these strategic database systems, data quality must become the number one prerogative of all parties involved with data warehousing effort.

The numerous problems with unreliable and inaccurate data in the data warehouse. First and foremost, the confidence of the users in this technology is shattered and contributes to the already existing rift between business and IT. Furthermore, if the data is used for strategic decision making, unreliable data cannot only hurt the IT department but the entire company. One example are banks that had erroneous risk exposure data on a Texas based businesses. When the oil market slumped in the early 80s, major losses were encountered by those banks that had many Texas accounts. In other cases, manufacturing firms scaled down its operations and took actions to ride itself of excess inventory. Because of inaccurate data it had overestimated the inventory and sold off critical business equipment. Such examples demonstrate the need and the importance of data quality.

Poor quality of data appears to be the norm rather than the exception and points out that many technology managers have largely ignored the issue of quality. This is caused in part by the failure to recognize the need to manage data as a corporate asset. One cannot simply allow just anything to be moved into a data warehouse or it will become useless and might be likened to a “data garbage dump”. In order to avoid data inaccuracies and their potential for harboring disasters, a general data quality awareness has to be made. There are critical success factors that each company needs to identify before moving forward with the issue of data quality.

First and foremost, senior management must make a commitment to the maintenance of the quality of corporate data. This can be achieved by instituting a data administration department that oversees the management of
the corporate data resource. Furthermore, this department will establish data management standards, policies, procedures, and guidelines pertaining to data and data quality.

Second, data quality has to be defined. In order for data to be useful it has to be complete, timely, accurate, valid, and consistent. It does not simply consist of “scrubbing” or auditing to measure its usefulness. The definition of data quality also entails to define the degree of quality that is required for each element being loaded into data warehouse. If, for example, customer addresses are stored it might be acceptable that the four digit extension to the zip code is missing. However, the street address, city, and state are of much higher importance. Again, this must be identified by each individual company and for each item that is used in the data warehouse.

A third factor that needs to be considered is the quality assurance of data. Since data is moved from transactional/legacy systems to the data warehouse, the accuracy of this data needs to be verified and corrected if necessary. This might be the largest task since it involves cleansing of existing data. Since no company is able to rectify all of its unclean data, procedures have to be put in place to ensure data quality at the source. Such a task can only be achieved by modifying business processes and designing data quality into the system. In identifying every data item and its usefulness to the ultimate users of this data, data quality requirements can be established. One might argue that this is too costly, but it has to kept in mind that increasing the quality of data as an after-the-fact task is five to ten times more expensive than capturing it correctly at the source.

If companies want to use data warehouse as a competitive advantage and reap its benefits, the issue of data quality has become one of the most important to ones. Only when data quality is recognized as a corporate asset and treated as such by every member of the organization will the promise benefits of a data warehouse initiative be realized.

**CONCLUSION**

The value of warehousing to an organization is multidimensional. An enterprisewide data warehouse serves as a central repository for all data names used in an organization and therefore simplifies business relationships among departments by using one standard. Users of the data warehouse get consistent results when querying this database and understand the data in the same way without ambiguity. By its nature, the data warehouse also allows quicker access to summarized data about products, customers, and other business items of interest. In addition, the historical aspect of such a database (i.e., information kept for two to five years) allows users to detect and analyze patterns in the business items.
Organizations beginning to build a data warehouse should not undertake the task lightly. It does not simply involve the moving of data from the operational database to the data warehouse but rather the cleaning of data for its future usefulness. It is also important to distinguish the different types of warehouse technologies (i.e., relational OLTP, multidimensional OLTP, and virtual data warehouse) and understand their fundamental differences.

Other issues that need to be addressed and resolved range from creating a team dedicated to the design, implementation, and maintenance of a data warehouse to the need for top-level support from the outset and management education on the concept and benefit of corporate sharing of data.

A further benefit of data warehousing results from the ability to mine the data using a variety of tools. Data mining aids corporate analysts in detecting customer behavior patterns, finding fraud within the organization, developing marketing strategies, and detecting inefficiencies in the internal business processes.

Because the subject of data warehousing is immensely complex, outside assistance is often beneficial. It provides organizational members with training in the technology and exposure, both theoretical and hands-on, that enables them to continue with later phases of the project.

The data warehouse is without doubt the most exciting technologies of our time. Organizations that make use of it increase their chances of improving customer service and developing more effective marketing strategies.