JITTA

JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

# BUILDING A DATA WAREHOUSE AND DATA MINING FOR A STRATEGIC ADVANTAGE

**RYAN NEARY, Indiana University**
*Bloomington Indiana, USA.* *Email:* mailto:ryneary@indiana.edu

## ABSTRACT

*Technology is fundamentally changing the way companies do business. Consolidations, globalization, and deregulation have put increased pressure on managers to better understand their businesses and take them to the next level. Given the fast-paced business environment today, decision-making cycles have been shortened and managers need accurate information in a timely manner in order to make quality decisions. A properly designed and populated data warehouse can provide the relevant data necessary to make good decisions. Significant advances in computer hardware and end user software have made it easy to access, analyze, and display information at the desktop. The data companies continue to collect from their current information system provides a great source of information about its customers and processes. Data mining software programs are powerful tools that can be used to interrogate the massive amounts of data contained in the data warehouse in order to uncover relationships. To help business leaders and decision makers manage their companies effectively, companies need to make as much information as possible available and give decision-makers the tools they need to explore it according to Kapstone (1995). By implementing a data warehouse and using data mining tools companies can uncover relationships that can be used to achieve strategic advantages.*

*First, I will explain data warehouses, why they are built, and how to build them. Second, I will cover data mining tools and the benefits companies are experiencing by using them. Finally, I will focus on the strategic advantages of building a data warehouse and extracting valuable data using sophisticated data mining tools.*

## DATA WAREHOUSING

### What is a Data Warehouse?

A data warehouse is a database designed to store relevant information in a central location to support decision-making needs. These warehouses are integrated, provide current and historical data, as well as detailed and summary information. Integration into the company's existing information system network is important for feeding current data, consistent historical data, and providing easy user access. "Integration is complex, painful, and requires much thinking-but it pays off handsomely. John Ladley of the Meta Group argues that the integration takes up to 75% of the development dollars," according to Inmon (1998). A warehouse typically will have historical data stored for five to ten years. By having this much data available it is often easier to spot trends across years and seasons. Summary data is great for management to see the big picture quickly, and then the detailed data can be used to further analyze the issue. With increases in information technology and its use, companies are collecting massive amounts of data from on-line transaction processing systems and other software systems that they use. The goal of the data warehouse is to be an integral part of the organizational information system by efficiently, consistently, and reliably storing large amounts of data in a central location. By providing a single source of corporate truth, data warehouses bring together valuable corporate information for decision makers.

Why Build a Data Warehouse; can't the current operational information system be used?

On-line analytical transaction systems are designed for processing the day-to-day activities of the businesses. These operational systems can be very efficient for their primary purpose of processing transactions, but they typically do not have the historical data nor the capabilities needed to optimally store, summarize, and retrieve data like data warehouses. The operational system is good at telling us what happened and the data warehouse can help show why.

Operational systems are focused on transaction processing with a limited time horizon for storing data. Data warehouses are designed to contain data that is stored specifically for decision making needs and contains a vast amount of historical data. The following table summarizes the major differences between operational systems and data warehouses.

**Table 1. Operational Systems vs. Data Warehouses**

| Operational System | Data Warehouse |
|---|---|
| Time horizon of data 60-90 days | Time horizon of data 5-10 years |
| Limited query capacity | Indexed for efficient query searches |
| Data updated and changed regularly | Data is not changed once loaded,only accessed |
| Designed for transaction processing | Designed for storing data |

Although the operational systems typically are not the best place to access stored historical data, the majority of the information that is contained in the data warehouse comes from the operational environment. It is important to keep in mind that there may be many systems that comprise the "operational system" for an organization. With many different systems being used, compatibility, comparability, and access may become an issue as information becomes fragmented all over the company. Flanagan and Safdie (1998) believe information contained in these disparate systems is often redundant, and since the systems were likely designed without a master Enterprise Data Model in mind, the data is duplicated and usually maintained according to different data needs.

The data warehouse is an effective way to pool this information consistently in a central location for easy access.

Currently these two systems are linked but independent. Many software vendors are working on packages to integrate these into one common system for the enterprise. This will make for an easier flow of consistent data into the warehouse and provide more control over the data. The main point is that this data is collected and stored in an efficient, consistent, and reliable manner for easy access in a data warehouse. Once the data is collected the real value can be derived from the data mining tools that extract valuable information from this warehouse.

## Initiatives

### *Needs Assessment*

"Today's rapidly changing business environment demands increasing amounts of timely information to support decision-making needs," according to Haley and Watson (1997). In order for a data warehouse to be of value to the users, these needs must be determined and evaluated. This is the first step in defining what type of data will be stored, the types of tools that will be used on the data, and what type of access will be needed. Current needs and expected future needs should be considered. The need for better access to accurate and timely information is often a major driver. It is important to keep in mind that although the technology may be exciting, the organization must be able to benefit from its use. The needs of the organization will change over time and the warehouse will need to be updated as part of the maintenance role once it is on-line.

### *Project Champion*

It is critical that a project champion heads up the project. Usually this person is from the IS department who is familiar with the systems as well as company business objectives. The champion is needed to coordinate the needs, determine technology to meet those needs, evaluate what resources are necessary, and determine if the project is in line with overall company objectives. It is important that the technology is taking the company in the direction where it wants to go and that the company is ready to embrace this change. Currently, building information technology capabilities are at the forefront of the rapidly changing business environment.

These projects often drive change in organizations in several ways. With the increase of electronic commerce and advances in information technology, companies are strategically rethinking the way they conduct business. The company currently may not have an up-to-date operational information system and may want to put in a new system to better capture information, serve customers, and provide increased decision support capabilities. Enterprise Resource Planning software tools such as SAP R/3 and PeopleSoft are currently very popular. Many companies have been taking this opportunity to build information technology capabilities as they make changes to their systems and tackle Y2K issues.

With company-wide changes like these, jobs and roles may change. Often, there is resistance to change in organizations. In addition, the IS department gives up some control over the data once users can access, edit, and utilize data at their desktops. In the past, reports and requests for certain data had to be obtained through the IS department because they had access to the data and knew how to write the queries. Now, with client-server technology and the advances in software that is easier to use, the IS department is quickly losing that power. In most instances the current IS department does not have the staff or the knowledge to tackle such a project on their own, and consultants are needed to help with the process. As the consultants enter and ask questions, people often think they are being downsized out of a job.

The changes that information technology can have on an organization can be enormous. It is important to keep these considerations in mind when evaluating the project. In addition to buy-in, other political issues can come into play within organizations. These changes often push marketing and other management professionals into new roles and often force changes in relationships between vendors and buyers. If the company is not ready to deal with the changes it will be a difficult challenge to make the project successful.

A proposal highlighting the business needs, technology, costs/benefits, and implementation plan must be developed. The ultimate goal is to set up the framework for the project and get it approved by the appropriate parties. The project champion takes the lead role in moving the initiatives forward. The approval process varies by company. However, given the expense and strategic nature of these projects, support from top management is essential. This leadership is important in getting the resources necessary to implement the project properly, especially in a company that may not be as strong in the technology area currently and may be resistant to change. If top management makes the project a strategic priority and helps see the project through, the organization will be better positioned to adapt to the new environment.

The project champion is needed to pull things together and take charge of the project. Changes such as these may get companies to rethink their information systems and the way the company does business. Thus, these projects often have a strategic impact on the business. Time must be taken to think through the impacts of implementing such a change in an organization. Therefore, it is essential to have a project champion and top-level support to see the initiative through completion.

*Constructing a Data Warehouse*

The project champion will play a critical role in setting the standards and getting the data warehouse on-line. The key to success in the data warehouse is to design the system that contains the required data and provides access to the desired data in an efficient manner. By taking the time to set up the initial warehouse properly, organizations will position themselves to exploit the benefits now and in the future. The following factors should be considered:

*Flexibility.* By setting up the data warehouse properly, many headaches can be eliminated in the future. It is important to design flexibility into the warehouse to allow for changes as the needs evolve. Companies may change through mergers, acquisitions, restructurings, or growth. Scalability allows the system to increase capacity as users demand more, as data stores grow, as more users are added to the system, and as more applications are developed against the warehouse, according to Flanagan and Safdie (1997). Therefore, it is important that the warehouse be designed with as much flexibility as possible to allow it to adapt to the changing needs of the business. In determining the flexibility necessary, the scope of the initial warehouse needs to be considered. If flexibility is inherent in the design, more storage capacity and capabilities can be added later as the needs of the organization evolve and as the benefits begin to show.

*Performance.* Flanagan and Safdie (1997) also believe how well the system performs will be the ultimate arbiter in the success or failure of the project. The intent of the warehouse is to help people do their jobs more effectively and efficiently. In the initial stages after the data warehouse is on-line it is critical to allow easy access and fast processing time for simple query searches. This will encourage more use and build confidence in the system. The memory and storage capacity of the system must be sufficient to respond to user requests quickly depending on the nature and complexity of the requests. By building a flexible system, attributes can be altered easily to optimize performance. In addition, indexing and compression tools within the database software can be used to optimize performance.

*Technology.* The type of technologies employed for the hardware architecture and database architecture play important roles in the flexibility and performance of the system. By making informed decisions in the initial planning and analysis stages, common problems can be avoided later. The technology must fulfill the business needs, and meet expected performance standards if the project is going to be effective, utilized, and show a good return over the longer term.

*Hardware.* Two common types of hardware are Massively Parallel Processing (MPP) and Symmetric Multi Processing (SMP) systems. The top end MPP system processing capacity does outstrip the top end capabilities of SMP. Currently, however there is little need for this level of capacity according to Flanagan and Safdie (1997). SMP architectures have been very popular because of their scalability. They can be effective in using as many as 64 processors or as few as four. With the work divided among multiple processors results can be achieved quickly. The ability to scale the processing capacity makes the SMP architecture particularly attractive for most projects. The two architectures I highlighted above are parallel processing systems. These currently are very efficient and are preferred due to the processing capacity. However, non-parallel processing systems can be effective for smaller scale data marts. Keep in mind the compatibility issues with computer hardware, current and future needs, and emerging technologies when deciding which technologies to implement.

Storage capacity needs vary depending on the amount of data, how often it is accessed, and the cost/capability trade-off of the media used to store data. For instance, older detail data that is infrequently accessed may be stored on a mass storage media such as tape drives. Current data that is frequently used should be on optical disks or hard drives which offer fairly large capacity with quick access. Writeable optical disks, zip drives, and 3+ gigabyte hard drives have declined in price and can store large amounts of data for quick access. The storage capacity needed varies by organization. The important thing to keep in mind is that you want to have the capacity to store the information needed and also be able to access it quickly. Data warehouses can range from a few gigabytes of information to terabytes of data. Therefore, to meet user needs, price, performance, and capacity should be considered.

In choosing the hardware, companies should determine how the warehouse architecture will integrate with the information system in place. It is important for the end-users to have easy desktop access to the warehouse. The warehouse must also have the capacity and the speed required by the users. If the organization currently has competencies with certain types of hardware such as IBM machines then this knowledge can be applied to the data warehouse if it uses the same hardware, saving valuable learning time. The process can be made easier by matching up as many current competencies as possible.

*Software.* First, an operating system must be selected to utilize the hardware. Again, this decision should focus on integration, performance, and current competencies. Second, database management software must be purchased to develop, use, and maintain the databases. This software determines how data is stored, indexed, and accessed. Data may be stored in a relational database, a multidimensional format, or an indexed combination of the two methods. Traditional methods use relational databases such as Oracle or Sybase. Sybase for instance offers newer techniques such as bit mapped indexes, vertical data storage and compression techniques with their software package. In-house expertise should be used to determine business needs and then what software best meets those needs.

These software packages are used to partition, access and store the data, and manage the data warehouse resources. The software and database design also influence the on line analytical processing capabilities that support decision-making. The software should support decision-making needs by being able to create different views, to "drill down" into the details of the data, and to "roll up" or summarize the data. Data mining software tools to exploit this data will be discussed in the second section of the paper. Therefore, careful consideration should be taken in selecting the most effective software packages.

As we know the technology is changing rapidly. It is often beneficial to align with a "big player" in the industry for the hardware and software needs. The advantage is that they usually set the standards, continue to develop new technologies, and are often able to upgrade your system to the latest and greatest. By determining and communicating your needs to the various industry leaders you can match up the best system to create and integrated data warehouse. The level of in-house expertise and the ability to integrate the new systems will be major factors in the decisions for hardware and software. The success of the project relies on these architectures being in place and meeting the needs of the users.

IBM main frames are the leading platform with nearly one-third of the market, followed by HP, IBM, and DEC in that order. Unix operating systems are the most popular with a 75% of the market followed by NT, MVS and VMS. The most popular warehouse software is Oracle with a 52% market share followed by Sybase, DB2, SQL and Informix. The following table highlights these usage percentages.

**Table 2. Platforms, Operating Systems, and Warehouse Software**

**Journal of Data Warehousing survey 1997**

| | | | | **Warehouse** | |
|---|---|---|---|---|---|
| **Platform** | **%** | **O/S** | **%** | **Software** | **%** |
| IBM-main frame | 31% | Unix | 75% | Oracle | 52% |
| HP | 27% | NT | 10% | Sybase | 16% |
| IBM | 12% | MVS | 6% | DB2 | 13% |
| DEC | 11% | VMS | 5% | SQL Server | 10% |
| Other | 19% | Other | 5% | Informix | 4% |
| | | | | Other | 5% |

### Populating the Data Warehouse

Once the data warehouse is set-up it needs to be populated with valuable data from internal and external sources. The operational transaction processing system in most companies is the main source of data. Often, there are several of these legacy systems used to collect and process information. These fragmented sources of inconsistent data often drive the need for one integrated system to access current and historical data. In addition, outside sources of data can be very relevant and should be considered. Some examples of outside data include rating agency data, industry research data, and media coverage. The data warehouse should be populated with relevant information to meet the needs of the intended users.

The process of populating the data warehouse with historical data can be time consuming, especially if the source data is from several different systems. Several software packages are available which use Cobol, Prism, SQL Server, and Oracle for example to help with this extraction process, but the process is still very time consuming. There are several steps that must be performed to prepare the data before it is loaded into the warehouse.

1. All relevant sources of information should be examined and relevant data should be selected.

2. Data that is going into the warehouse needs to be standardized so that data from all sources are measured in the same manner. For example, the sales figures from division 1 should be the same type of figures as the sales figures from division 2. If division 2 stores sales data net of discounts this must be translated back to a gross sales number to ensure compatibility and comparability. This process is termed "scrubbing" of the data and it is used to get consistent and reliable data into the database. Once the historical data is scrubbed and entered it will not change.

3. Meta data is also formulated. This is data about the data. It keeps track of the fields and how they relate to each other, where the data is stored, the business meaning of the data, and rules about the data. The meta data is very important information and is critical for making use of the data stored in the warehouse.

4. Controls are put in place in the operational systems that feed the warehouse. These controls ensure standardization of current data flowing into the warehouse and require minimal maintenance once set-up.

This process often takes most of the project time and requires cooperation from people all over the company to agree on the standards. It is important for the project champion to play an active role at this stage. In addition, this stage can take a considerable amount of time and effort depending on how integrated and standardized the systems are for the organization. This is a critical step because once the warehouse is set-up users throughout the company will access this centrally located data. The time and effort required at this stage of the process should not be underestimated. The usefulness of the data warehouse depends heavily on getting this stage correct.

The important thing to consider here is that the users must believe in the integrity of the data they are relying on to make decisions or the system will not fulfill its needs. Therefore, attention must be focused on the needs and uses of the data to make sure they consistently and efficiently fulfill those needs.

Because this process is time consuming, companies may outsource the bulk of this work. The existing staff of the IT department of most companies will not have the necessary skills and people to tackle this project. Therefore, additional people will have to be hired or consultants will be needed to partner on the project. Outsourcing can be effective when there is a company team working with the consultants to ensure that the objectives of the project are being met.

### Updates

The warehouse needs to be updated periodically to meet the changing needs of the users. The time frame varies by company, industry, and preferences. Also, the frequency of use, changes in transactional data, and staff availability are factors. Batch processing, daily updates, and scheduled updates are common patterns. For a manufacturing company that warehouses production information daily, this information could be updated once the line is shut down for the day. This assumes that they need that information in the warehouse quickly for a desired use and that they have the IS staff available to manage the process. For some companies it is a 24-hour turnaround; in others it may be weekly or monthly depending on the nature of the business and the use of the warehouse. This process is automated but needs to be monitored to continue to make the data relevant for the users. The update schedule is part of maintaining the warehouse, which is discussed below.

### Access

All interested users should be allowed to access information contained in the data warehouse. Use should be encouraged and not hindered by access problems. The internet and intranets have become popular access points for users. These can allow access from multiple locations and are effective to give the appropriate users the appropriate information at the appropriate time. This can allow decentralizing of some tasks. Passwords and encryption technology can be used for higher security. There is further discussion about access and training in the data mining section because the data mining tools are what will be used by the end user to access this vast amount of data. Access and availability help is often maintained by the IS help desk staff.

### Maintaining Data Warehouses

Designing the warehouse and standardizing the feeds of information into the warehouse takes the most effort. However, there are several maintenance roles that must continue once the warehouse is built. First, the frequency of data access and types of data being access must be monitored. It is important to keep the current and frequently used data quickly available. Less used and older data can be moved to more archived storage areas, keeping the frequently used data in the quickest access positions. Second, the current data flow into the warehouse must be monitored. This includes making sure the proper data is being loaded on time and is still relevant. Third, the levels of summary data must be updated to reflect new data. Fourth, access must be available for the intended users. This includes making sure the systems are up and running as well as users having the access rights to reach the data. Fifth, the warehouse must be backed up for any unanticipated events. Sixth, security of the system must be monitored and adapted as necessary. Finally, with the growth of warehouses into the hundreds of gigabytes to terabytes in size, the systems must be fine-tuned for efficiency.

Over time, data warehouses tend to grow at an amazing rate due to the high flow of data into the warehouse. This growth can become an obstacle to success because the performance slows down. Therefore, an active role must be taken in the maintenance function. This involves taking into account user needs, summarizing data, archiving dormant data, and purging unnecessary data.

Some data may be kept forever such as legal data that may be needed to defend the company in a lawsuit. This proactive approach will continue to streamline the warehouse and make it most useful. Reindexing and repartitioning the data on a regular basis can optimize performance.

One of the keys to the acceptance of the data warehouse is that the information is available and can be utilized efficiently. It is important that the warehouse be tailored to the users needs. Therefore, maintaining the warehouse once it comes on-line is important and should not be overlooked. Often there will be dedicated IS staff to this function, depending on the magnitude of the warehouse and the stage of acceptance throughout the organization. Because the company has made a considerable investment in the warehouse it must be maintained and continually improved to provide rapid access to this valuable data.

Figure 1, below, summarizes the structure of a data warehouse and the types of data that could be found in a typical data warehouse. The foundation is made up of older, less frequently accessed, detail data stored in a cost/use effective manner. The next level is current detail data that is frequently used. These two levels of data stores contain the data that is accessed through the software tools. The software tools then parse out the data for different levels of detail for frequently accessed information. Examples include lightly summarized data for regional sales data by week and highly summarized data for national sales by month.

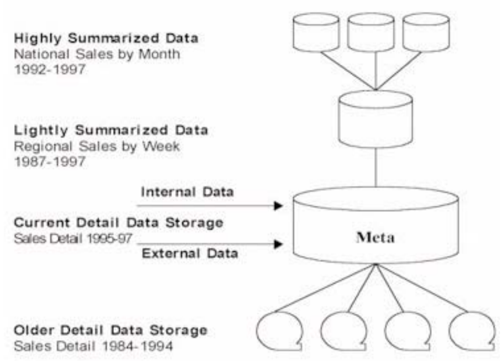


**Figure 1. The structure of a data warehouse**

Examples of Internal Data are inputs from the operational or transactional processing system, Excel data analysis, and Word documents. External Data can be gathered from sources such as Dow-Jones News Retrieval Service for current company news, analyst reports, and stock prices.

This system is then connected to the end user computers on their desktops through the current network IS backbone which could include a network, intranet, or internet connection. This enables access to the warehouse by the people who can take advantage of this massive amount of data.

*Lessons Learned*

- Top Management and Project Champions are important success factors
- Responsiveness to use and users needs is critical in developing and maintaining
- Building flexibility in the warehouse architecture is important for changing needs
- Take the necessary time and resources to populate and set-up the data warehouse to ensure consistency, reliability, and usability of the data
- Need Integration to operating systems for current data additions and company-wide accessibility
- Training to use systems and allowing access to decision-makers to encourage use

## Data Marts

I have mentioned data marts several times throughout the data warehouse section. Data marts are smaller versions of data warehouses. These are often used as data warehouses within a single department. These can be very effective for limited computing needs. However, they become data islands and miss out on the real usefulness of data warehouses by not being integrated with all the company data or keeping consistency in one database. They can be cost effective, are often easier to get approved in budgets, and are a step in the right direction. However, they miss out on some of the major advantages that are available through a corporate-wide integrated data warehouse.

These can be effective for small computing needs or these data marts can be used as part of a star network that has a central data warehouse which feeds and updates the data marts. This maintains consistency and access to the data and can often speed up the data mining process. However, additional time is needed to maintain this more complex system. With the needs assessment these issues can be addressed in building a system to meet the needs of the users and help the company achieve its goals.

### Summary of Benefits and Costs of Building and Maintaining Data Warehouses

The major benefit of the data warehouse is that it provides a consistent, reliable, and efficient source for data from all areas of the organization. This is costly and takes a considerable effort to initially set up the structure properly. However, if the data warehouse is utilized the benefits outweigh the costs. In a recent study by Fisher (1997), the average 3-year ROI for a data warehouse project was 401%, with the high being 16,000% and the low of -1,887%. These projects vary in length from 6 months to 3+ years, depending on the current level of technology and integration, in-house expertise, and management buy-in as a strategic priority. Flanagan and Safdie (1998) highlight a survey of 62 companies that revealed the average time to implement a data warehouse project was 3 years and the cost averaged several million dollars. These projects have high up-front costs and uncertain returns at the beginning. In addition, the returns are often the result of strategic changes and hard to quantify to the specific project. However, some marketing returns by directly targeting customers can be determined fairly easily. In summary, traditional methods such as NPV and ROI, which are often used to evaluate projects, may not be as useful for these projects. With the use of sophisticated data mining tools, companies have been able to extract valuable strategic insight from their vast amounts of data to significantly improve their business. The next section will provide an overview of what data mining can do and the final section will discuss strategic uses.

## DATA MINING

### What is Data Mining?

Once the data warehouse is built and the architecture is in place a vast amount of data is readily available. Data mining helps management and decision-makers transform the raw data from the warehouse into meaningful information. Data mining tools can be as general as running simple queries on a data set to using sophisticated tools that automatically search for relationships that we would ordinarily not consider from the data warehouse. These tools vary in complexity and depend on the level of user knowledge of the data and the ability to write queries. Although a data warehouse is not necessary for using these tools, data warehouses are an optimal source of raw data for the entire organization.

These tools should be part of the decision support system for companies. The goal of the DSS is to provide analysts, knowledge workers, and decision-makers with an understandable representation of the data. The hardware and software tools have evolved to allow this analysis to be done at the desktop level instead of submitting a request to the information technology department to have them run a query. The advanced tools contain graphical user interfaces and are user friendly. This allows the users to discover important business patterns, examine relationships between obscure and otherwise unnoticed variables, and measure long-term trends quickly and easily. In the sections that follow I will discuss some of the tools that are used.

### Query Tools

There are many standard software packages that can be used to query data from databases. These are usually associated with data mining tools. Many of these are very powerful tools that can be used to extract data. Gillman (1998) argues however, that they are limited in their ability to discover trends and complex patterns in a database because the user must "think up" a hypothesis and then test it. Also, relevant and important hypothesis may not be obvious or come from patterns obscured within the data. These queries are often limited to a few variables and improper interpretation of the relationships can easily occur. These tools start with the user developing and running a query on the data. An example would be to search a database for all customers over the age of 50 and list their names and addresses. In this example, a group of customers could be targeted for certain mailings. This could be accomplished using the Select and Where commands in SQL to obtain the required data. The ability to program in SQL can be a limiting factor when using this analysis tool. SQR is based on SQL but offers more programming flexibility for selecting the data and preparing reports.

Queries can be used for some of the basic needs of the decision-makers for quick and simple outputs. In addition, Crystal Reports is a popular software tool to create and format reports based on output from SQL. These tools also can be very effective to familiarize the users with the information contained in the data warehouse. When the user can determine the variables needed and the number of different variables and variations is small, these methods are particularly useful. As companies grow and more information becomes available through information technology these methods become inefficient. Because of the massive amounts of data that companies are storing, more sophisticated data mining tools have been developed that are more efficient and effective at handling increasingly complex issues and large amounts of data.

### Statistical Tools

Another type of analysis that is often used is statistical analysis. Regression analysis and other statistical processes can be run on sets of data. This can be done very easily by querying a set of data and using any of the statistical software packages on the market such as SAS or by just using Excel spreadsheet features and tools. Increases in desktop computing power and software capabilities have increased the use of these tools. These analysis techniques often reveal correlations and relationships among variables. Although these can be effective, they typically assume linear relationships, normal distributions, and attributes that have continuous or ordinal values. These assumptions can limit the real-world use of the findings.

An example would be to run a regression analysis on available data to come up with a formula for predicting Sales based on historical data. These can be effective if the characteristics of variables considered are well understood. However, this is often technical, not easy to explain, and contains many assumptions about the data. For example, the regression output analysis consists of R2 and correlation statistics to predict the reliability of the model. These measures are often difficult to interpret intuitively making it challenging to convince top management and executives of the statistical credibility of the analysis. In addition, the logic used to come up with the formula is not easy to explain. Like traditional query tools, findings uncovered by statistical means can be effective. However, true data mining tools can provide much more insight into the data. Artificial Intelligence tools and genetic algorithm tools are also great ways to interrogate the data using a top-down approach. These are similar to regression analysis but use different logic to look for relationships. These tools continue to be developed for widespread use.

### Real Data Mining

True data mining tools go beyond the user-driven queries and systematically formulate and test alternative hypotheses on the data that users would not ordinarily consider. This is done using sophisticated neural networks and decision tree methods. Data mining tools are clearly superior for data warehouses because they do not require the users to query the information to identify relationships that exist among the data. Their strength, Gillman (1998) asserts, lies in their ability to automatically identify key relationships in a database and to discover rather than confirm trends or patterns in data and to present solutions in usable business formats. These utilize a "bottom-up" approach where traditional statistical and query techniques typically use a "top-down" user-driven approach.

By coupling the massive amounts of data, sophisticated software, and shear computing power, these systems can uncover hidden "treasures" in the data. Companies can use this new insight to fine-tune the way they do business. The findings could trigger an altering of a manufacturing process, changing advertising methods or audiences, or even changing the types of businesses in which the company operates. These tools are fast becoming a business necessity. Two popular data mining methods used to uncover hidden treasures are neural nets and decision tree models.

### Neural Nets

Neural networks form mathematical models of processes studied. These processes are very complex and are not well understood by outsiders. The processes can be similar to statistical modeling techniques. The methods vary by researcher and are particularly applicable under current situations but don't adapt well when assumptions change. Neural nets work best with numerical data and not as well with nominal data. These methods are partially an art and partially a science that, like statistical methods, are hard to describe. These methods can be very effective. However, with the current technology they are not at the level of widespread use. These methods may be standardized and included in future software packages.

### Decisions Trees

These methods segregate data into a set of rules. Then, a decision tree is set up to cover all possible variations and rules for attaining the outcomes. This process can be performed quickly and the rules are stated in English. This provides a significant advantage over other methods that process in a "black box." The drawback is that some decision rules may not be considered because of an implication earlier in the decision tree.

*How Does Data Mining Work?*

- A dependent variable or outcome field from the database is specified. Example: Computer Sales
- Gathering data from various sources such as the data warehouse, on-line transaction systems, and data marts.
- Data "scrubbing" to ensure consistency in the data.
- Rules development stage. This is the processing stage where the software interrogates the data to determine rules that relate to the dependent variable, Computer Sales in our example.
- Output analysis and review. This could explain variables searched, records evaluated, and time. The output report is often listed by rules starting with the highest confidence factor.
- Rule Example:
  IF AGE = 20 through 45
  AND EDUCATION > 12
  AND USAGE = 4
  THEN Potential Customer = Yes with a confidence factor of 78%
  Variables searched = 200
  Records evaluated = 5,000
  Time 0:28:01

By having a data warehouse set up, steps 2 and 3 can be avoided. The data warehouse will already have up-to-date and historical information that has been "scrubbed" for consistency and accuracy. Because everyone is using the same quality data it will ensure consistency of findings and recommendations across all areas of the organization. In addition, this also eliminates the duplication of efforts and saves considerable time by not having each individual research on their own or not use enough data to make a good decision. Thus, the data warehouse provides a tremendous advantage by providing the users the data that they need when they need it. This can play into a strategic advantage for the company.

The advantages of this type of output include:

- Easy to understand input variables
- Computer generated testing
- No need to develop queries
- Easy to understand confidence % and variables output
- Easy to use
- Relationships considered that may not ordinarily be considered

### Software

As the benefits of data mining continue to be seen, more and more software companies are developing software tools to extract valuable information from these growing warehouses. Companies may also develop their own proprietary software to aggregate, analyze, and report the information.

**Table 3. Platforms, Operating Systems and End-User Software**

**Journal of Data Warehousing survey 1997**

| Platform | % | O/S | % | Warehouse Software | % |
|---|---|---|---|---|---|
| PC | 65% | Windows | 54% | Access | 18% |
| Sun | 11% | Unix | 23% | Excel | 17% |
| Mac | 9% | DOS | 18% | Business Objects | 8% |
| IBM | 7% | NT | 15% | Power Play | 8% |
| HP | 4% | OS/2 | 9% | Cognos | 7% |
| Other | 13% | Mac | 8% | DSS Agent | 7% |
| | | Other | 8% | Oracle | 7% |
| | | | | SQL | 7% |
| | | | | Brio | 6% |
| | | | | GQL | 6% |
| | | | | SAS | 6% |

The clear trend is towards the client-server platform with a Windows or NT operating system using a variety of tools that include Access, Excel and Oracle. The use of sophisticated data mining tools to extract valuable information is continuing to grow and become accepted.

## Getting the Most out of your Data Warehouse

The format of the data warehouse is such that decision-makers can use top-level summary data to locate areas of concern and bring into light current issues facing the organization. Then, the detailed data can be analyzed using one of the data mining tools mentioned above. By locating variances and spotting trends, the decision-makers will be better informed to do their jobs.

Training will be important once the system is on-line. By making the resources available and then showing users how they can utilize the power from their desktop, the project will begin to show benefits. Training will be essential in getting employees to initially start to use the system and build confidence in the integrity of the system. By giving the people who need the information the tools and the training to use them, end users will benefit and IT professionals will be able to devote time to other areas. According to industry estimates, 70% of IT budgets are dedicated to creating reports for users and querying data for end users. With pressure on IT budgets, improved technology, and the need for accurate information quickly, end users will be able to get the information they need when they need it by taking advantage of these powerful tools.

Continued support of the system and needs monitoring will be important for keeping the proper information in the warehouse as the needs of the organization change. If the company has an IS support forum they should set up a special forum for this function. In addition, the IS staff should be trained and encouraged to help users get past the initial learning curve. New and easier tools continue to be developed. These should be evaluated as part of the maintenance process and added as needed.

## Examples of Companies Using Data Mining Techniques

- AT&T uses these techniques to uncover calling card fraud from their massive amounts of data contained in customer databases.
- Consumer package goods companies and retail organizations such as Procter and Gamble and Wal-Mart were early adopters of the technology. They have used this to better target customers and build cost efficiencies to compete in these narrow profit margin industries. In addition, they are used to forecast demand, relationship market with customers and suppliers, and evaluate effectiveness of targeted marketing campaigns.
- Longs drugstore chain has used these tools to fine-tune product placement, purchasing, and promotions to maximize sales and profits.

### Summary of Benefits for Data Mining from a Data Warehouse

The software tools continue to adapt to the changing needs of the users and are becoming more sophisticated than the simple query models that started out. Companies are collecting massive amounts of data due to increases in information technology. This information contains valuable insight into how the business is operating and can provide managers and decision-makers with the information they need. As I argued above, a data warehouse is the best way to provide optimal access to the data. With all this information it pays to invest in sophisticated data mining tools to get the most out of your data. With the growing use of client server technology and the capabilities of PC's, meaningful information can be extracted quickly from the data. It is much more efficient to have the users performing the analysis and making meaningful interpretations than the IS department providing the reports. Data Mining tools will continue to improve in easily finding relationships among data variables.

### Strategic Uses for Data Warehouses and Data Mining

*Business*

- Direct Marketing-market research, product success prediction
- Insurance-risk analysis, claim predictions, credit and collection models
- Banking-mortgage approval, loan underwriting tasks, fraud analysis and detection
- Finance-analysis and forecasting of business performance, investment analysis
- Market Research-media effectiveness, product segmentation, media selection
- Telecommunication-better determine which customers to market additional products
- Relationship marketing with customers and suppliers

*Manufacturing*

- Quality control, defect analysis, maintenance scheduling, automation analysis

*Medicine and Science*

- Diagnosis, drug interactions, risk factor analysis, toxicology, research

The benefits continue to accrue as companies find more uses for the massive amounts of data they collect on a daily basis. These tools make the data available and also help to make sense of it.

### Competitive Advantages and Competitive Necessities

Today companies, both nationally and globally, are constantly seeking core competitive advantages. Information technology can be a very important tool to gain a competitive advantage in the short- run. Over the long-run as these technologies are copied by competitors, these tools become more of a competitive necessity. In any case these tools will help your business stay in the game.

- Do you know who your most profitable customers are?
- Do you know the attributes of the majority of your customers?
- Do you want insight into the defects caused by a specific manufacturing line?
- Does your marketing campaign target your average customer?
- Would you benefit if you could use Relationship Marketing?
- Which customer segment would this product appeal to?

By knowing more about your customers and internal processes you can better understand your business. As information technology, mergers, acquisitions, and deregulation continue to change the way companies do business, it is critical to have more information on your customers and internal processes. For example: By knowing the characteristics of your primary customers, companies can target those specific customers with advertising and promotions. This can help strengthen the niche in which the company operates. By targeting only that niche, mass advertising campaigns can be avoided. This can help by lowering cost and increasing sales, thus providing a significant advantage to the company.

*Why Not Take Advantage of Data Already Collecting in the Current System?*

The operational or transactional systems that companies are currently using collect valuable information about the customers they serve and the internal processes. By setting up a data warehouse to collect and store all this data, companies can possess a wealth of information from which to data mine for rules about their customers and processes. Since companies are accumulating more and more data a data warehouse is a great way to summarize and make valuable that information. This massive amount of data can be the key to the future success of the company by providing strategic insight into the current situation and new opportunities.

### The MCI Example

MCI's Small Business Sales Unit created a 48 million row data warehouse called SOLD for its Small-Business Online Lead Database. This system is providing substantially increased flexibility and performance than its mainframe predecessor. With the uses of SOLD, MCI can better target this lucrative segment.

According to Flanagan and Safdie (1998), "The legacy system that, until recently, housed the data had grown so ponderous that it threatened the very characteristic that has made MCI what it is today—able to respond quickly to market changes," says Chase Hacker, MCI's manager of Technical Architecture Team (SBS). He goes on to say, "To get anything done, we had to break up the data by region, which meant doing everything seven times." They needed the ability to respond in days not years. Therefore, they moved to client/server technology which offered advantages in responsiveness, expandability, flexibility, scalability, and cost-effectiveness.

This is handled by a 60-GB data warehouse with Sybase IQ VII software, which is updated daily with millions of records from dozens of sources. A Sun SPARCcenter 2000 server and a Sun SPARCserver 1000 server are used as the system servers. Windows 95 is used as the PC operating system with PowerBuilder as the primary software tool.

"The best measure of the quality of our calling lists, and of the service we provide when we deliver them, is salesman productivity—sales per individual," according to Hacker. "Since we've modernized our system, that figure has grown 200 to 300 percent. What's more, this has all happened when our sales force grew from 300 to 2,000 strong. Just keeping up with growth would have been impossible before this system, let alone improving the quality of our deliverable", according to Flanagan and Safdie (1998). MCI has been able to reap substantial benefits from implementing such a data warehouse system.

MCI buys calling lists from several companies but they only pay for customers not already in the SOLD database. SOLD not only enables them to store this massive amount of data but also keeps them from paying for it more than once. With the data mining tools MCI can pinpoint good prospective customers and remove from the list customers with poor payment records. This system continues to play an important role as the SBS unit moves into emerging markets. "We predict that five years from now, half of our revenues will come from businesses we're not even in today," according to Hacker from the Flanagan and Safdie (1998) article. SOLD also allows the potential to change the company's sales campaign procedure, so the new products and services will be based on available calling data instead of the other way around.

"With day and night improvement in analysis speed we now have," explains Hacker, "we can try a bunch of things until we identify a really interesting segment, then go back to the marketing geniuses, and say, 'Hey, here's a super list we've got in our pocket. Now, go design a service that appeals to exactly these people.?That's great for the whole organization and frankly, it's great for us too," according to Flanagan and Safdie (1998).

These technologies are enabling MCI to create new opportunities in this intensely competitive market.

### Conclusion

Remember that data warehouses and data mining are not solutions to your problems. They are powerful tools to provide accurate, reliable, and timely information to decision-makers. Given the turbulent global business environment, companies are constantly seeking ways to stay ahead of the competition. Advances in information technology have increased the usability of the massive amounts of information that companies are collecting. In organizations today, the trend is toward employee empowerment. This is accomplished by letting the people closest to the issue solve the problem. By providing access to data warehouses and sophisticated data mining tools, decision-makers will be able to perform data analysis more efficiently and effectively.

Data uncovered from this wealth of information can help companies improve current operations and position themselves for success in the future. Small changes in strategy, provided by the data mining's discovery process, can translate into a difference of millions of dollars to the bottom line, according to Gillman (1998). Advances in hardware and software architectures have significantly contributed to the rapid growth and exploitation of the data warehouse. Data mining tools will continue to make it easier to derive value from this data. If the company takes the proper measures to set up an effective data warehouse and makes the tools available to those who need them, the strategic advantages are limitless.

## REFERENCES

Chilton, David (1998, January 5). Best from Data. Information Week, 103, 3-9.

Data Mining: Discovering Treasures In Databases. (1995). http://www.cwi.nl/~~marcel/applic.html.

Fisher, Lawrence (1997). Along the Infobahn: *Data Warehouses*.

Flanagan, Thomas and Safdie, Elias (1997). A Practical Guide to Getting Started with Data Warehousing. *http://www.techguide.com*.

Flanagan, Thomas and Safdie, Elias (1998). Data Warehouse Technical Guide. *http://www.sybase.com/products/dataware/techguide.html*.

Gilman, Michael (1998, January). Nuggets and Data Mining. *http://www.data-mine.com/white.htm*.

Haley, Barbara and Watson, Hugh (1997). Data Warehousing: A Framework and Survey of Current Practices. *Journal of Data Warehousing*, Volume 2 Number 1, 11-16.

Inmon, Bill (1995). Tech Topic, What is a Data Warehouse? *http://www.cait.wustl.edu/papers/prism/vol1_no1/*.

Inmon, Bill (January 1998). Wherefore Warehouse. *Byte Magazine*, 1-10.

Kapstone (1995, November 15). Does Your Company Have the Insight it Needs to Survive? *http://www.kapstone.com/kapwhite.htm*.

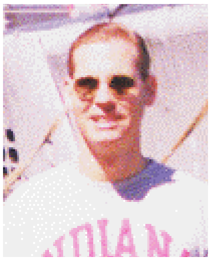
Kelley, Thomas J. (1998). Dimensional Data Modeling. *http://www.sybase.com/services/dwpractice/dimensional.html*.

Sybase (1996). Data Marts and Beyond: The Pragmatic Approach to Enterprise Decision Support. *http://www.sybase.com/products/dataware/dm_beyond.html*.

Sybase (1997). Sybase's Data Warehouse. *http://www.sybase.com/products/dataware/*.

Sybase (1996). Taming the Terabyte. *http://www.sybase.com/products/dataware/terabyte.html*.

## AUTHORS



**Ryan Neary**, MBA (Indiana, 1999) with concentrations in Information Technology and Finance. He will begin working for a leading technology consulting firm in mid-summer of 1999. Ryan attended IU from 1990 to 1994 as an Evans Scholar where he earned a B.S. in Business with an accounting and finance double major. Prior to graduate work, Ryan was a Senior Accountant at Unitrin, Inc. in Chicago with CPA and FLMI certifications.