The Journal of Korean
Institute of CALS/EC
Vol 7, No. 3, Dec. 2002

웹 웨어하우징을 위한 신개념의 저장장치 전용네트워크

홍순구*

The SAN for Web Warehousing: An Alternative Data Repository

Soongoo Hong

Abstract

The combination of data warehousing and Internet technology produces a new concept - web warehousing. Due to the availability of web technologies and the need to make prompt decisions with timely information, web warehousing is emerging as a key strategic business weapon. Yet despite the many promising benefits of web warehousing, researchers have also identified several challenges, including scalability and availability. With the rise of the Internet and data centric computing applications, the use of new Storage Area Network (SAN) technology has been spotlighted for the possibility of a new data repository for web warehousing. In this article, the two new concepts of web warehousing and storage area networks are introduced. In particular, a SAN is discussed in detail as an alternative data repository to overcome the current limitations of web warehousing.

Key Word: Web Warehousing, Data Warehouse, Internet, Storage Area Network

---

* This Paper was Supported by the Dong-A University Research Fund in 2001
* 동아대학교 경영정보과학부 교수

1. Introduction

To be competitive in the global market, firms must be able to collect and store detailed customer data. In fact, to cope with severe competition, world-class organizations today implement customer relationship management (CRM) to provide the ready and immediate access to customer data that sales representatives must have to satisfy customer expectations. As a result, data warehousing has become a critical issue since the concept was introduced in 1985. In today's increasingly competitive business environment, the effective management of all data has become "mission critical."

In addition, the Internet and the Web have been the dominant topics of interest in the information systems field during the last decade and are expected to continue to be of primary concern in the future. An ideal medium for data transfer and information exchange, the Internet emerged as a substitute for expensive Local Area Networks (LANs) and Wide Area Networks (WANs). It is now possible to form business networks via the Internet that enable linkages between suppliers, partners, and customers for a variety of activities.

The combination of data warehousing and Internet technology produces a new concept - web warehousing. Almost all researchers view the Web as a natural medium and/or infrastructure for information exchange. Findings of the Meta Group survey indicate that by the year 2001, web-based information systems will be developed for both sharing and distribution of information within the organization and among multiple organizations[6]. Information in the data warehouse can be deployed inexpensively and easily through the Web to decision makers, such as sales representatives and customer service personnel. In addition, customers and others can be allowed access to the data via the Web, which can result in lower costs associated with client support and maintenance. With this popularity, the concept of web warehousing is gaining fame in the industry. Web warehousing can be used not only for information sharing but also for creating "innovative relationships with suppliers, partners, and customers"[3].

Despite its promising benefits, several factors such as scalability and availability must be considered and well managed to ensure the successful implementation of web warehousing. Chen and Frolick [3] stated that "a number of challenges still exist in implementing Web-based data warehousing due both to the immaturity of this

technology and to some management concerns." Some of these challenges are associated with storage management. As a possible solution to storage-related problems, the Storage Area Network (SAN) that interconnects servers and storage at high speed is now available on the market. The purpose of this paper is to provide an overview of web warehousing, including its current challenges. Storage Area Networks are then discussed as a potential new data repository option to that can overcome the current limitations of web warehousing.

2. Web Warehousing

The radical change of globalized business environments and the recent developments in information technology and the Internet have forced all levels of managers to use Decision Support Systems(DSS) for making prompt decisions[16]. As a result, the concept of data warehousing has been introduced into almost every industry. Data warehousing is built for enterprise-wide decision support and thus stores summarized historical data that is consolidated from various operational information systems.

Since the company views data as a critically valuable asset, the implementation of data warehousing has received much attention. In today's environment, data is the underlying resource on which all computing processes are supported. Firms collect increasing volumes of information from customers, suppliers, partners, and the economic environment. The ability to combine these diversified streams of information to make better decisions, gain insight into potentially useful innovations, and streamline operations without being overwhelmed by information overload remains both a major challenge and opportunity.

As the computing environment changes rapidly, a number of implementation problems in the current data warehouse systems have been reported. Chen and Frolick[3] point out limitations of "traditional data warehousing" as follows; 1) the client/server system that is mostly employed for implementation of data warehousing in organization is expensive, 2) scalability is becoming a big problem due to the rapid growth in number of end users, 3) the desire for system compatibility (i.e., combining heterogeneous systems) increases management costs. According to a 1997 survey by International Data Corp., the low usage rate is the main reason for negative return on investment of data

warehousing projects. In fact, only "10 to 25 percent of knowledge workers" are accessing data warehousing for their tasks[9].

Internet-related technology can cut the gap between the actual usage and deployment of data warehousing. Recognizing the impact of the Internet, many data warehousing software manufacturers began offering web-based data warehousing products in 1996[15]. Moreover, almost all data warehousing vendors currently support web access tools, including network functions and protocols[15]. Due to the availability of web technologies and the need to make prompt decisions with timely information, web warehousing is emerging as a key strategic business weapon.

As shown in Figure 1, the basic architecture and components of web warehousing include a data warehouse, meta-data, query facility, and a Web server. The data warehouse stores the enterprise-wide or departmental data while meta-data permits identification of the contents and location of information in the database. The query facility supports the capability to access data. It transforms user requests into Structured Query Language (SQL) query to the data warehouse, formulates the detailed requests, and returns the outcomes to the end-users. Query, analysis (e.g., OLAP, data mining), and reporting tools as decision support mechanisms are all utilized by end-users in web warehousing. The Web server responds to end-user's requests, but its function is to pass requests on to other application servers such as a HTML server, security server, and/or file server.

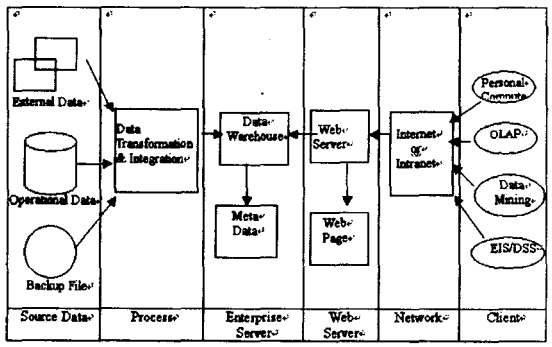

Figure 1. Web Warehousing Architecture

While web warehousing affords many benefits to organizations, various challenges such as scalability and availability have been addressed by many researchers (e.g., [3, 5, 16, 8]) and deserve further attention. These concerns are more fully discussed in the following section.

3. Challenges in Implementing Web Warehousing

The successful players in the new business environment will be those organizations that can keep up with

rapidly changing market requirements and customer demands. The e-driven business environment can be extremely unforgiving as it generates a tremendously high volume of data and places alarming demands on server capacity and network bandwidth. These pressures create challenges of scalability and availability in the implementation of web warehousing.

3.1 Scalability

System scalability is defined as the capability to grow in various dimensions, including the total number of users, the number of simultaneous users, and the transaction volume needed to support desired service levels. As the size of data warehouses grows due to the increasing number of users (transaction volume) and data (database size), with often unplanned and unpredictable surges, scalability become increasingly critical in web warehousing implementation. According to the findings of Survey.com, "the average data warehouse will contain 1.2 terabytes of useful data by 2002, an increase of 290% in just two years"[11]. In addition, Crofts[5] states that "surveys consistently show that the average number of users on a data warehouse doubles in three months, doubles again in six months, doubles again by the first year, and then doubles again in the next six months." Thus, the system should be able to deal with a rapidly increasing number of web warehousing users without substantial amount of delay.

3.2 Availability with High Speed

Organizations must not only handle extremely high growth of rates of data but also manage large volumes of data. As everyone goes online, the number of users accessing the information is increasing dramatically. Those situations invoke the problem of speed when data is delivered. Today's Internet users expect a web site to provide fast web page loads (including those pages which contain detailed graphics) and quick searches. Lynch and Horton[12] noted that for most computing tasks the threshold of frustration is about ten seconds. If the users feel the system is too slow, they will find other means for their decision making and purchases. Thus, the system connected to the organization's networks must provide fast response time.

Another significant challenge relates to availability. E-business has forced companies to be in a 24*7*365 high-availability world. In other words, the on-line site must always be available whenever users attempt to access to the

data and service with zero downtime. Information access to applications at any time is critical and has increased the importance of ensuring data availability with high speed.

In sum, due to the immaturity of both web warehousing and network technology, a number of challenges have been identified. As one of the possible solutions to overcome these challenges, the concept of SAN technology is introduced and discussed in detail in the following section.

4. Storage Area Network (SAN)

Facing high storage growth needs, organizations are seeking an effective and efficient way to handle data management issues. Due to their inherent scalability and uncompromising availability, SANs are becoming increasingly attractive to companies managing a large volume of data. The Storage Networking Industry Association (SNIA) defined a SAN as "a network whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements"[13]. In other words, the SAN is a high-speed network or system that allows different kinds of storage devices, including tape libraries and disk arrays, to be shared by all users through network servers.

The SAN represents a major shift from the traditional server to disk data. In the vast majority of enterprise networks, file servers are at the center of data access. Driven by end-user requests, servers are perpetually reading and writing data files on a disk and wrapping the data in the appropriate network protocol for delivery to the user. Built on the Small Computer Systems Interfaces (SCSI) parallel bus architecture, this legacy model links to storage arrays with fixed, dedicated connections[4]. For example, in the past, the buying decisions on storage hardware were often completely tied to the servers (e.g., if you buy a Sun server, you then buy Sun storage). Accordingly, the amount of data available to that server is limited by the number of disks supported by the bus and the number of buses supported by the server. If the server or any of its SCSI connections to disks fail, access to data is lost. For mission-critical networks, this potential for catastrophic loss is unacceptable[4].

The fact that this server-centric model does not lend itself to the high availability and high-volume requirements of enterprise networks is the primary impetus for changing the server/storage relationship. SANs are in the forefront of this change. SANs

introduce the flexibility of networking to servers by eliminating the dedicated connection between server and disk and enable a data-centric universe to expand[2]. The speed, capacity, and network flexibility of SANs are based primarily on the fibre channel architecture. A fibre channel provides scalable bandwidth, redundant data paths, and long distance (refer Table 1).

Table 1. SAN and SCSI

	SAN (Fibre Channel)	SCSI Interface
Transfer Speed	100 MB/sec	80 MB/sec
Number of Devices	126	15
Distance Covered	6 miles	25 meters
Fault Tolerance of Wiring	More	Less

*source: Dot Hill Systems (2000)

The Figure 2 shows the comparison of a traditional model (server-centric) and a SAN. The SAN may be physically located in a data center, alongside an enterprise's mainframe, but is better distributed across a metropolitan area so that it can survive physical damage to one site.

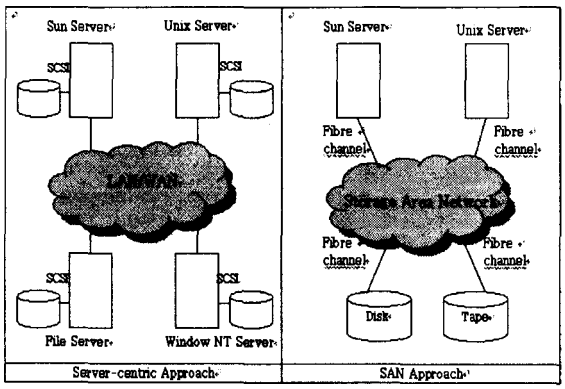

Figure 2. Server-centric Model vs. SAN

5. San for Web Warehousing: A Scalable Storage Alternative

The pressures of globalization require more data to be incorporated into complicated decision-making, and result in more organizational users becoming involved in the use of web warehousing. Accordingly, the system should be available to end users with a reasonable speed throughout the network to satisfy the internal and external customers who hold greater expectations than ever before. If a company is unable to support this demand due to system failures and network bottlenecks it will likely lose customers, productivity, and profits.

As discussed earlier, the main problem that web warehousing designers confront is how to handle the rapid growth in number of end users and storage capacity (i.e., scalability and availability). Chen and Frolick[3] suggest

that designers "increase the server capability, ensure server reliability, and improve network bandwidth" to meet the scalability challenge of web warehousing. However, given a rapidly growing customer base, this solution which is mainly focused on improvement of server and network capability may not be enough to meet the never-ending increasing demand. The introduction of a SAN mitigates the limitations of traditional server-based solutions. The SAN's nature of "interoperability of heterogeneous servers and storage," makes it simple and easy to add more servers or storages whenever they are necessary. Instead of putting storages directly to the server, a new storage is installed to the networks connecting heterogeneous storages and servers. Some additional advantages that SANs provide to address scalability problems include:

1. Storage Pool: A storage pool can be accessed via a SAN, which reduces the total extra storage needed for projected growth; and
2. Long-distance Coverage: SAN technology breaks the physical distance limitations by enabling storage units to be located miles, rather than feet, away from each other.

The challenge of availability in implementing web warehousing can also be the use of SANs. This issue is also highly related to speed (one of the metrics measuring the quality of web warehousing). As the delivery time increases, the system and networks will be overloaded and slowed down, resulting in low availability. The speed of data delivery can be improved by increasing network performance (e.g., increasing bandwidth, installing a cache memory); however, network technology is limited due to the requirement of operating 24 hours a day if something fails on the network device. A SAN can be utilized to improve the availability and performance. First, a SAN can be used to bypass traditional network bottlenecks. IBM pointed out that a SAN supports direct, high-speed data transfers between servers and storage devices in the following three ways[14]:

1. Server to Storage: it is a traditional approach to transfer data between server and storage;
2. Server to Server: it is a communication among servers at high speed with a large volume of data; and
3. Storage to Storage: it is a new approach to communicate among storages. SANs allow applications that move data to perform better, for example, by having the data

sent directly from source to target devices without any server intervention.

The storage to storage communication, in which data is backed and restored while the LAN and servers continue their operations, supports the following additional benefits of SANs:

1. LAN Free Backup: The advent of the Internet and 24*7 e-business creates more data than ever before, and less time for backup, which requires that businesses devise new and innovative ways to backup "mission-critical" data. SANs simplify the data backup and recovery process and result in a faster, more scalable, and more reliable backup and recovery solution with an effective utilization of storage, server, and LAN resources; and
2. Server Free Backup: With server-free backup, data is transferred directly between storage devices without using host servers. This unique backup approach is enabled by "a third party copy," which is implemented in SAN appliances (e.g., bridges), host systems, or storage devices themselves. As a result, operating efficiency and automation are improved[1].

Also, SANs currently allow data transfer rates of up to 100 MB, and initiatives are underway to further increase this throughput. Finally, the SAN provides redundant paths to the servers and storages. Storages in the storage area network can be switched and shared among the servers if necessary. With use of a SAN, companies are better equipped to guard against equipment failures, keep critical systems online, and meet increased data access expectations.

6. Conclusions and Implications

The marriage of the world-wide connectivity of the Internet with data warehousing to enable business analysis has provided both opportunities and challenges. While web warehousing can be a cost effective way to deliver information to employees and customers, it invokes the need for managing high volumes of data. Furthermore, enterprise-wide databases have been developed to support the many new management paradigms, such as CRM (Customer Relationship Management), ERP (Enterprise Resource Planning), SCM (Supply Chain Management), KM (Knowledge Management), and even e-

business applications. These management tools involve the use of huge corporate database systems in which data are stored and accessed by numerous users for customer satisfaction and decision making via the Internet. This increasing reliance on the access to enterprise data is challenging the limitations of a traditional server-centric approach. Consequently, the SAN has emerged as a viable new data repository option for today's web economy, as shown in Figure 3.

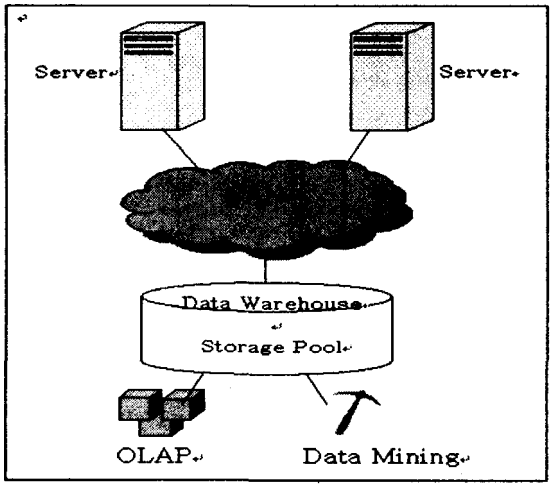

Figure 3. Data Warehousing with a SAN

In this article, the two new concepts of web warehousing and storage area networks were introduced and discussed as a way to help both managers and researchers better understand the current challenges in data management. This article further asserts that SANs can be a viable solution to storage-related problems.

References

[1] Brocade Communication Systems, Inc., *SAN Solutions: Scalable Storage for the Web*, 1999.

[2] Burniece, T. and Hart, L., "Network Storage: The Challenge, *Computer-World*," October 2000, pp5-8.

[3] Chen, L. and Frolick, M., "Web-Based Data Warehousing: Fundamentals, Challenges, and Solutions," *Information Systems Management*, Spring 2000, pp80-86.

[4] Clark, T., *Designing Storage Area Networks: A Practical Reference for Implementing Fibre Channel SANs*, Addison-Wesley, 1999.

[5] Crofts, S., "How Data Warehousing Turns Information into Competitive Advantage," *Fortune*, August 18, 1997, ppS1 - S9.

[6] DCI, "A Data Warehousing Forecast," February 5, 1998, http://www.dci.com/news/1998/feb/dwtrends.htm

[7] Dot Hill Systems Corp, *Storage Area Networks: The Superior Storage Solution*, October 2000.

[8] Eckerson, W., "Criteria for Evaluating Business Intelligence Tools," *Journal of Data Warehousing*, Vol. 4, No. 1,

Spring 1999, pp27-36.

[9] Eberlin, R., "Data Warehousing Breaking Out on Web," *National Underwriter*, December 1997, pp3-4.

[10] Gray, P., "What's New in Data Warehousing 1999," *Journal of Data Warehousing*, Vol. 4, No. 2, Summer 1999, pp12-14.

[11] Hoss, D., "Data Warehousing Trend Report: Get Ready for "Big Data," dataWarehouse.com, September 1, 2000. http://datawarehouse.com/iknowledge/articles/article.cfm?ContentID=352

[12] Lynch, P.J. and Horton, S., *Web Style Guide: Basic Design Principles for Creating Web Sites*, Yale University Press, 1999.

[13] Mcintyre, S., "Demystifying SANs and NAS," *Enterprise Systems Journal*, July, 2000, pp10~17.

[14] Nystrom, K., *Introduction to Storage Area Network, SAN*, IBM International Technical Support Organization, September, 1999.

[15] Row, H., "Just Browsing, Thanks," *CIO*, October 1, 1996, Vol. 10, No. 1, pp98-106.

[16] Watson, R., "A Design for an Infrastructure to Support Organizational Decision-Making," *Proceedings of the Twenty-third Annual Hawaii International Conference on System Sciences*, Los Alamition, CA: IEEE Computer Society Press, January 1990, pp 111-119.

[17] Watson. H. and Gray, P., "New Developments in Data Warehousing-1998," *Journal of Data Warehousing*, Vol. 3, No. 2, Summer, 1998, pp8-11.

저자소개

홍순구(e-mail : shong@daunet.donga.ac.kr)

홍순구 교수는 현재 동아대학교 경영정보학과의 전임강사로 재직중임. 미국 네브라스카 주립대학교에서 석,박사를 마친 후 Texas A&M International University에서 조교수로 2년간 근무. 유학 전에는 한국은행 대구지점 및 전산정보부에 재직. 연구 관심 분야로는 Data Warehousing, Knowledge Management, e-commerce, IS Evaluation, Healthcare Information Systems 등임.