Synthesis with Data Warehouse Applications and Utilities

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

Today's fast moving business world faces continuous challenges and abrupt changes in real-life situations at the context of data and information management. In the current trend of information explosion, businesses recognize the value of the information they can gather from various sources. The information that drives business decisions can have many forms, including archived data, transactional data, e-mail, Web input, surveyed data, data repositories, and data marts. The organization's business strategy should be to deliver high-quality information to the right people at the right time.

Business analysis desires that some data must be absolutely current. Other data may be comprised of historical or summary information and are less time sensitive. To overcome the data loss, improve efficiency, make real-time update, and maintain a well-marked path to other data, a high-speed connectivity is always needed. It also needs to protect the information the systems gather, while ensuring that it is readily available, consistent, accurate, and reliable. It also must consider how the software environment has been designed and what impact that design has on the performance, availability, and maintainability of the system. Among all these parameters, defining the basic layout of a storage environment is critical for creating an effective storage system.

With data residing on numerous platforms and servers in a multitude of formats, gaining efficient and complete access to all relevant organizational data is essential. While designing the data warehouse, the network topology, data consistency, data modeling, reporting tool, storage, and enactment of data need to be clearly understood. In recent days, data warehouse database(s) grew at such a pace that the traditional concept of database management needed to be revisited, redesigned, and refocused with increased demand, availability, and frequent updates by putting pressure on data warehouse methodologies and application tools. Innovative applications and techniques have evolved to handle data warehousing more efficiently and to provide easier data access.

BACKGROUND

Information is all about integration and interaction of data sets. Inaccuracies in a single data column may affect the results and directly affect the cost of doing business and the quality of business decisions. Usually, preventive measures are more economical and less tormenting to ensure data quality. It has been found that delaying the inevitable data cleansing dramatically increases the cost of doing so, as well as increases the time delay for the cleansing process.

Data warehousing was formally defined as a separate environment to support analytical processing that is subject-oriented, time-variant, and integrated. A data warehouse that provides accurate, consistent, and standardized data enables organizations to achieve better revenue generation and, at the same time, attain cost optimization. An effective data quality utility and methodology should address its quality at application and data entry levels, during application integration stages, and during the quality analysis level.

Earlier data warehouses used to be mere replacements of MIS systems with limited service facilities. Due to simpler operating environments, they did not justify allocation of significant resources. With incremental demand, especially from the business community, progress of data warehousing concepts triggered tremendous development with sophisticated requirements, increase in database sizes, and complexity in the data warehouse environment. Nowadays, companies are spending thousands of dollars, and a significantly large portion of it goes to the information technology budget in the form of firmware to build sophisticated databases and data warehouses. In the quest for successful business intelligence, various applications and systems have been deployed, and manifold information retrieval processes have been developed.

Traditional database system architectures face a rapidly evolving operating environment, where millions of users store and access terabytes of data (Harizopoulos & Ailamaki, 2003). Database applications that use multi-terabyte datasets are becoming increasingly important for scientific fields such as astronomy and biology (Papadomanolakis & Ailamaki, 2004).

Data warehousing ultimately reconciles the design conflicts best by integrating operational applications and analytical applications into a coherent information architecture (SAS, 2001). In this aspect, fully integrated approaches are need to be realized to improve the data quality and its processes (ETL, extraction, transformation, and loading), including data warehousing techniques with the goal of transforming raw data into valuable strategic assets.

MAIN THRUST

Technological advancements use techniques like data pattern analysis, clustering, algorithms, and other sophisticated capabilities to ensure that data gathered throughout the organization is accurate, usable, and consistent. By intelligently identifying, standardizing, correcting, matching, and consolidating data, specially designed software can offer better solutions to the data quality (SAS, 2002).

The task of ensuring optimal query execution in database management systems is, indeed, daunting (Schindler et al., 2003). To meet the challenges of managing data scalability and handling large volumes of data, the strategic solution should provide a powerful foundation for building a robust and resilient data campus and should integrate the popular access characteristics of the modern day's information economy.

The solutions should be able to run under a wide variety of hardware environments, enabling the choosing of the computing resources by matching the particular needs of the enterprise. On the other hand, the computing environment creates a base for making better business decisions by hosting powerful analysis tools and organizing the information.

Establishment and operation of an efficient data site is a critical component of successful solution implementation in order to deal with the ever-increasing volumes of data associated with customer relationship management (CRM), supplier relationship management (SRM), enterprise performance management (EPM), and hazard analysis. Similarly, inconsistencies in data semantics are among the most difficult tasks in establishing large data warehouses.

A few of the data warehouse applications and utilities are synthesized in the following section:

Direct Data Storage

Direct data storage is an acceptable method of modeling relay time-current characteristics for devices with fixed characteristics. Like the name implies, the direct data storage approach consists of storing data points over a given curve into computer memory. The relay then monitors the line current and compares that value to the current values stored in memory. Forelle (2004) reports that the sales of hard-disk data-storage systems rose steadily in the fourth quarter of 2003, reversing previous slides and giving some hope for a recovery in big-ticket technology spending.

Data Mining

Data mining is the search for patterns and structure in large data sets, and the discovery of information may not be present explicitly in the data. However, one of the most difficult problems in data mining is to concretely define the classes of patterns that may be of interest. Riedel, et al. (2000, p. 3) stated that "the major obstacles to starting a data mining project within an organization is the high initial cost of purchasing the necessary hardware".

Data mining also is employed widely in sales and marketing operations to calculate the profitability of customers or to find out which customers are most likely to leave for the competition. Forrester Research (1999) reported in a study of Fortune 1000 companies that the usage of data mining will grow rapidly. The report also suggested that marketing, customer service, and sales may remain as the major business application areas for data mining.

EMC

EMC (www.emc.com) has a reputation for building a highly resilient information environment and protects valuable information by providing flexibility, as the business requirements change. EMC's Symmetrix information storage systems can be integrated with other computer systems to manage, protect, and share the IT infrastructure. EMC Sysmmetrix storage systems implement a broad range of storage protection and acceleration techniques, including disk mirroring, RAID storage protection and redundancy, data caching, and hot spares and replacements of individual components.

RAID

RAID techniques in hardware or software (or both) can be implemented. Most data sites with high data volumes choose to implement RAID storage options in hardware, using disk arrays. Thus, disk arrays offer additional performance and availability options beyond basic RAID techniques. Mirroring and Parity RAID techniques balance general performance and availability of data for all task-critical and business-critical applications by maintaining a duplicate copy of volumes on two disk devices.