Leveraging the Information Asset

Janet Perna
Director, Database Technology IBM Toronto Laboratory
jperna@vnet.ibm.com

Abstract Data is a corporate asset, and being able to derive more information from data can provide database users with a competitive advantage. For example, catching on to trends quickly can reduce unwanted store inventory and lower capital outlay for the same profit. If you have store sales data by product analyzed on a daily basis, that can make a 2-3% difference in margin -- and in a business where margins might be 4%, this is a significant competitive edge. This paper will cover what technology is needed by customers to leverage their information assets. Real-time access to production point of sale information, database mining for analysis to detect trends immediately, high performance, and multi-vendor database connectivity, cooperation among heterogeneous clients and servers are among the customer needs we are seeing in the marketplace.

1. Introduction

At the midpoint of the decade of the 1990's, the environment for information technology use by corporations includes an increasingly global marketplace, an increasingly diverse population base in languages and cultures, and a shortage of skilled labor. Coupling these global aspects with a mobile work force and corporations undergoing significant organizational restructuring results in an atmosphere of diversity and change, emphasizing teamwork, empowerment that rewards adaptability and flexibility.

The environment for information technology reflects this atmosphere. We have today the most diverse array of options for computer information systems ever, from mobile hand-held personal digital assistants, laptops, desktop PCs and workstations, and larger LAN-based and WAN-based servers. Customers and vendors alike are faced with the reality of architecturally heterogeneous hardware and software environments, and at any given time a mixture of corporate cultures for making business decisions.

In this paper, we focus on what this means for database technology, and provide examples of how one specific vendor, IBM™, is providing customer solutions in this environment. More specifically, we will address only one aspect of information technology, leveraging corporate data to derive more information in order to provide database users with a competitive advantage.

2. Information Management Framework

In today's global, highly-competitive economy, turning an organization's data into an information asset is key to gaining competitive advantage. To create and manage this information asset, customers need to store, retrieve, access, analyze, and distribute their data. The core componentry IBM offers customers to do this includes database engines together with middleware for distributed databases, access enabling tools for data replication and decision support, tools for systems management and administration, and tools for application development. These data management offerings integrate with complementary components for transaction management, networking, and workgroup solutions. To summarize, what we database vendors need to do is deliver the right information to the right place from any source at the right time.

2.1. Database servers

Databases form the heart of many customers' businesses. For example, fifteen million users perform more than 7.7 billion transactions a day using IBM's databases, and DB2™ is in use by every corporation in the Fortune 100. The data stored in these databases is essential to running the customer's business.

Customers are looking for an optimal mix of heterogeneous platforms, operating systems, and technologies as they try to reduce costs and improve efficiencies of business processes. For many organizations, the computing environment contains a variety of platforms including PC and workstation database servers and high-end and mid-range server systems evolving to high capacity, highly available WAN- and LAN-connected superservers and repositories of enterprise data. Parallel hardware and software including symmetrical multi-processors (SMP) and massively parallel systems (MPP) allow customers to exploit their data in ways that were not possible or affordable before. For example, retail customers can use data mining to derive information about buying patterns in order to manage store inventories more profitably.

Open database vendors, including IBM, provide one or more products across a variety of these platforms. IBM provides the DB2 family of products on both IBM and non-IBM hardware. To address the growing diversity of data and applications, this core relational database technology is being expanded to deal with the explosion of information types such as text, sound, image, and video. We are incorporating new technologies into our data management solutions to improve price/performance and capacity, to enhance ease-of-use and productivity, and to enable new application types. For example, IBM's Ultimedia Manager provides the capability of finding images based on their content. In addition to storing and manipulating image files users can sort through thousands of pictures, analyze them, and get feedback about their

SIGMOD ' 95,San Jose , CA USA

™ DataJoiner, DB2, IBM, IMS/ESA, VSAM, DataHub, DataPropagator, DataGuide, and Ultimedia are trademarks of the International Business Machines Corp. ORACLE and Oracle7 are registered trademarks of Oracle Corp. Sybase is a registered trademark of Sybase, Inc.

content. Images may be classified and queried by color, shape, and texture features.

2.2. Data Access Middleware

Data access middleware shields applications and users from the underlying complexity of distributed environments through consistent and transparent application programming interfaces. Providing transparent access to heterogeneous data servers enables customers to leverage all of their information assets. For example, IBM's DataJoiner™ product allows a user to join data from disparate databases through a single SQL statement, hiding the various database interfaces from the user and the user's application. It supports distributed queries like multi-site joins transparently across all IBM relational databases, key competitive relational databases (such as ORACLE™ and Sybase™), IMS/DB™, and VSAM™ file systems through standardized interfaces. DataJoiner incorporates advanced global optimization techniques for highly efficient distributed query processing to provide outstanding SQL performance for both simple and complex SQL queries. Furthermore, datasource-specific requests can be passed directly to the proper backend server, and if there are functions (e.g. text search) or datatypes (e.g. money) supported by DataJoiner that are not supported by all the data sources, then DataJoiner will "compensate" and provide that function for those servers which do not have it.

2.3. Decision Support Technology

One key customer issue is finding out where the relevant data for solving a customer's business problem is located. IBM has the DataGuide™ product that provides client/server catalogs on a LAN for sharing and browsing and allows a user to launch applications to view information quickly. DataGuide provides extractors to gather information descriptions from many sources, including DBMSs and desktop tools, to populate DataGuide catalogs.

A large number of organizations have created ultra large data bases of business data, such as sales records, purchase history, etc. Such data forms a potential goldmine of valuable business information. In Quest, an IBM research project on data mining, we are inventing new technology to discover patterns of useful information embedded in large databases. Given a database of sales transactions (each transaction giving the items bought by a customer in a visit), Quest can find what sells together. That is, it discovers ALL associations such that the presence of one set of items implies another item (e.g. 90% of transactions that purchase bread and butter also purchase milk). Knowledge of such associations can be used for retailing decisions and help in understanding customer buying behavior.

2.4. Data Replication

One of the essential elements of leveraging customers' information assets is to be able to analyze data from a variety of sources without affecting the performance of operational systems running the company's business. If gateways and data access middleware are inappropriate, for example because they place too much load on operational systems, then the alternative is to replicate the data. Replication may not just consist of straightforward copying but also might require point in time capture (e.g. end of day summary), aggregation, enhancement, and other transformations.

IBM's DataPropagator™ Relational provides these functions between relational databases. Copies are automatically maintained at user-specified intervals or can be initiated through application events (for example, end-of-business day reconciliation). DataPropagator Relational provides powerful data enhancement capabilities for tailoring copies to specific applications or end-user requirements. Enhancements are specified through standard SQL and include joins, subsetting, summarization, new data calculations, and other transformations as part of the replication process. DataPropagator Relational supports continued operation in the event of network or system failures and recovers automatically, identifying where it left off and how best to continue processing. IBM's DataPropagator NonRelational provides replication between IMS/DB and DB2 with synchronous two-way propagation allowing two master copies, while it also supports point-in-time copies or complete histories.

2.5. System Management Tools

Customers can't productively implement applications without tools to manage complex, heterogeneous environments cost effectively. In order to get maximum value from their data, customers want the cost reduction and productivity benefits of easy-to-use and integrated tools for systems administration, and systems management while choosing from different tool providers. Tool and application builders play an increasing role as organizations of all sizes increase purchases of applications and tools rather than building and maintaining them internally. Suites of integrated tools and applications are gaining popularity at all levels - desktops through enterprise systems.

For example, IBM's DataHub™ provides powerful functions DBAs can use to manage a variety of databases, all from a single control point that can run on a variety of IBM and non-IBM platforms. From the control point, DataHub can manage members of the DB2 family running on IBM and non-IBM platforms, ORACLE 6, Oracle7™, Sybase, and other vendor's databases. Examples of DataHub services include display, create, drop, and alter cataloged database objects and relationships between those objects; copy objects, such as tables and their associated authorizations between databases; invoke and schedule database utilities and commands, such as BACKUP and RECOVER, for remote or local databases; monitor database management tasks for certain thresholds and events; and interoperate with network management.

3. Summary

Keys to success for information technology vendors include supporting diverse technologies, bridging disparate systems, providing language support according to user's preferences, offering tools and technology to manage the information overload that corporate decision-makers face and above all allowing flexibility in the use of information systems. To summarize, what we database vendors need to do is deliver the right information to the right place from any source at the right time.