Craig S. Mullins

Return to Home Page

March 1998

From IDUG Solutions Journal...

Data Warehouse Administration

The Challenges Never Stop

By Craig S. Mullins

Data warehousing is indeed becoming common place in large organizations. According to a Forrester Research survey of executives at large firms 62% percent have data in, on average, three data warehouses or data marts. The same survey indicates that the pace of data warehousing will increase before it slows down; with the average growth showing the number of data warehouses and marts to double to nearly six by 1999 and increase in size from approximately 130 GB to approximately 260 GB.

Dealing With Aggregates

Additionally, most data warehouses require denormalized data in the form of aggregate tables. Aggregate tables contain redundant data that is summarized from other data in the warehouse. The purpose of the aggregate tables is to optimize performance and increase data availability — both

noble goals. However, these tables add to the size of the data warehouse and the complexity of the environment that must be managed.

Data warehouse administrators (DWAs) need to be able to control the creation and management of aggregate tables. This will eventually take the form of intelligent agents that manage aggregate tables. These agents will recommend when to create and remove aggregate tables, estimate their space usage, automatically create the tables, and asynchronously load data and propagate updates. This will be required until the RDBMS vendors provide optimization technology that can handle dynamic data summarization and aggregation from normalized structures. Until these intelligent agents arrive, DWAs will need to tweak the tools used by DBAs to do the task. For example, a SQL performance monitor can be used to determine which SQL queries using a GROUP BY are run most often. These are good candidates for summarization into aggregate tables. Or perhaps the aggregate table already exists. In the absence of an aggregate-aware query generator that routes queries to aggregate tables if they exist, the DWA can review usage using the monitor and suggest alternate query formulations for frequently run queries.

Consistent Data Acquisition

As the data in operational systems changes, so must the data warehouse. Over time, fields will be eliminated, meanings will change, international growth

occurs, sizes change, and more. The business reacts and adapts to respond to industry trends. You must plan to keep track of physical data changes, as well as changes to the semantics of the data. Regardless of the type of change you will need utilities and tools as well as processes to allow you to keep on top of these issues and respond appropriately.

Backup and Recovery

Backup and recovery needs special consideration within the context of the data warehouse. The data warehouse should have a backup and recovery strategy that will enable the organization to recover essential data in an emergency. Depending on the size of the data warehouse, you may choose not to do a backup, because you can refresh the data more efficiently. Review the cost/benefit of each warehouse and mart, keeping in mind how often the data is updated or refreshed and how long recovery will take to implement. Additionally, disaster recovery requirements must not be overlooked. Organizations are becoming more dependent on the information that a data warehouse provides thereby raising the importance of the warehouse application. This means the warehouse must be treated like any other critical system in terms of disaster recovery planning.

Financial Chargeback

In most organizations, data warehouse projects are managed by multiple departments, each of which has

its own financial goals. Data warehouse managers should ensure that they can charge back appropriate costs to business units and users so that they can meet financial reporting requirements. An integrated solution is required that monitors IT costs by providing critical chargeback services that track information resources used organization-wide.

Scalability

As a data warehouse becomes accepted in an organization, demand for its services grows. The need for new reports and aggregate tables increases, and the data warehouse can explode to several times its original size. Industry surveys indicate that 60 to 70 percent of data warehouses are filled with duplicate or redundant data such as summary tables and indexes. This can more than double the size of the disk subsystem required to store the data. The more users on the system, the more simultaneous queries, and the more potential there is to frustrate users with delays in response time. It is important therefore to architect the system so that it will be able to scale linearly with demand. Parallel processors, parallel databases, bit mapped indexes, data compression, and other techniques can been applied to these issues.

Performance

System performance is closely linked to scalability and can be viewed from three perspectives:

1. extract performance - how smoothly data is updated and refined
2. data management - quality, maintenance, and query performance
3. server performance - hardware performance and maintenance

The server on which data warehouses reside requires peak performance around the clock. However, performance may have a different definition for analytical warehouse access than for OLTP. A query may realistically execute for hours in the warehouse environment, but not so for OLTP. Organizations should seek an agent-based performance monitor that collects, analyzes, and stores thousands of performance measures, is configurable for multiple environments, and offers both a real-time and a historical perspective on viewing all critical metrics. In this manner organizations can implement an integrated database performance solution which is capable of monitoring and managing the performance of: relational databases in Windows NT, UNIX, and MVS environments, servers in distributed environments, the entire enterprise network, and distributed client/server transactions. Additionally, it is imperative to optimize the speed by which the data warehouse is loaded, unloaded, reorganized, and accessed. High speed database utilities can be used to optimize the flow of data throughout the lifecycle of the data warehouse.

Synopsis

Each of the areas covered impacts the structure and operation of the data warehouse, which means the warehouse must be able to react to these changes in order to maintain its value to the organization and to leverage the substantial financial and resource investment. To ensure stability in the face of change DWAs need to use enterprise technology that manages the applications, systems platforms, and data to keep pace with demanding business requirements. Operating a data warehouse smoothly is as challenging as running any sophisticated OLTP system in a day to day operational environment. In the end it is people, tools, and methods and their interaction that will get it built and make it last.

From *IDUG Solutions Journal*, March 1998.

Home. Phone: 281-494-6153 Fax: 281-491-0637