SOLUTIONS JOURNAL

The *IDUG Solutions Journal*
March 1998 - Volume 5, Number 1

Data Warehouse Administration: The Challenges Never Stop

The editors of IDUG Solutions Journal *have invited two DB2 experts to duel peacefully on key issues facing DB2 today. In this regular column, Willie Favero of BMC Software and Craig Mullins of PLATINUM technology will face off. Sometimes they even agree.*

---

by Willie Favero

The issues facing today's data warehouse administrator can be divided into "before" - those involved in creating or implementing a data warehouse, and "after" - those involving maintenance issues that arise in a data warehouse once it is actively used.

Before

How large will your data warehouse be? Can you settle for a series of data marts? For an administrator, both options have good and bad points. A full data warehouse may become completely unmanageable in size. Data marts, on the other hand, may grow independently, creating the same mainframe data issues you are trying to solve by implementing a data warehouse. In either case, the amount of raw and summarized data may be tremendous. The rapid growth of the data may raise capacity planning to an art.

How many data sources will you combine to form a data warehouse? From where will they come? Are they all on the mainframe or scattered across multiple platforms? Are they in databases or flat files? Is, in fact, all the information you need for a successful warehouse even in an automated system? Might some of the information be on paper documents, or is some of it in an end user's head?! Regardless of the source, you will need to combine all pertinent information into the warehouse. Once you have selected those data sources that need combining, how do you clean the data? I hope you won't assume that your data is already clean and that you can simply combine it and load it into a warehouse. Your company will quickly learn that its warehouse is only as good as the information it holds. If, for example, the analysts believe the data is unreliable, your company could end up with a warehouse that no one will use. Remember - the reason for cleaning data is not always because something is wrong with it. If data used by

multiple OLTP systems is being combined, it may contain information, such as indicators and flags, that is simply of no use in a warehouse. The warehouse may even need data that does not currently exist. In some cases, the data warehouse could actually end up being larger than the sum of its original inputs.

Now that you have determined which data to combine to form the warehouse and how to scrub the data, you face a tough decision. Where does it all happen? Do you perform all activities on the mainframe where you have the horsepower, or on the workstation where a wealth of tools to assist you now resides? Either way, there are still questions: which tools to use, how to unload the data for combining and cleaning, and finally, how to reload the data into the warehouse. This could be a huge ongoing task if your warehouse is at all large. The right tools will make a dramatic impact on how you maintain the correct information in the warehouse and how you keep that information fresh. And remember, all these tasks must be predictable and repeatable. Eventually, the warehouse will have to be reloaded, probably a number of times.

Some of the new hardware now available prompts yet another question. Must you move the data at all? At least one hardware vendor allows you to move the data between two different relational databases *in the hardware*, completely eliminating the need to unload and reload any data at all.

After

So you built it, and they really did come. Now, how do you keep them coming back? This is a major "after" issue. All this data consolidation and movement has been accomplished at tremendous corporate expense. It must now be treated like any other corporate investment, any other corporate resource. If an analyst believes the information in the warehouse is accurate, and the data are used to make important corporate decisions, the administrator has a heavy responsibility to ensure that data remains accurate.

First among the issues she faces are the volumes (literally) of data. Moving all this information into a data warehouse is clearly not a one time event. The warehouse's data has meaning and accuracy only to a certain point in time. How do you plan to keep it accurate? Will you use data replication/propagation or perhaps a complete data refresh? Meanwhile, as more users find more ways to use the data, the "success" of the warehouse prompts a never-ending cycle of demands for more data. Now the administrator must determine what to do with the data once it arrives at the warehouse. Should it be archived or simply deleted when it becomes old? How should the data be summarized? Indeed, what level of summary is necessary?

Because the data held in a warehouse is not changed, a warehouse is administered differently from an OLTP system. Here, the administrator won't have to worry about: backups and recoveries, for example, or time-outs and deadlocks. Referential integrity and check constraints may be less of a concern because the data is not being modified.

Instead the administrator can focus on ensuring that the data is clean and sufficiently summarized; his task will be to devise better and faster ways to access the data.

The next challenge: how do business rules apply to a data warehouse? Data is being combined in many cases from multiple sources. What if the operational OLTP system has used a set of rules that have no bearing on the information one is attempting to maintain in the data warehouse? What if the OLTP systems use different rules for similar data because the data will eventually be used in different ways? Many facts can be kept about a piece of data because different organizations have different ideas of what the "right" answer is. When you combine this "different" data, how do you decide which is correct? You may not be able to rely on the business rules because there may be more than one rule for the same piece of data!

Then, of course, there is schema management. Here the administrator may need to manage and coordinate schema changes across heterogeneous relational database systems, possibly on different platforms. The tools the administrator uses will have a significant impact on the success of this task. Assuring that changes made to the originating data sources are reflected in the data warehouse is as important as the data itself. For this maintenance task, a tool is needed that makes this task easy and efficient. However, the tool should also be usable by someone who is not as skilled in the RDBMS requiring the schema changes. This allows schema maintenance tasks to be spread across administrators with skills in different relational databases.

And finally there is meta data. Was it there when you began the data warehouse design? Or will the meta data be created as a result of your data warehouse's implementation? Whatever its origin, meta data can play an important role in the future success of your data warehouse. It should describe the relationship between the data in the warehouse and the operational data, contain any formulas or rules used to summarize the data in the warehouse, and include a description of how the data is structured. Once created, it must be kept up to date just like the schema information.

Conclusion

A data warehouse is a completely different beast from the operational OLTP. Its problems and the tools needed to solve them are different. But administrators also need to be concerned with warehouse availability and performance during access. Data size, placement, indexing, and all those performance issues you have been fighting with DB2 will still be there to challenge you. In addition, new functionality is available in the database world that may also enhance how a data warehouse performs. Such features as parallel processing in the RDBMS and parallel processing in hardware are only the beginning. New analytical techniques made available through OLAP, both relational and multi-dimensional, are now on the scene. Then there's object technology - it can improve the operational usefulness of a data warehouse. The articles in this special data warehouse issue of the *IDUG Solutions Journal* and the North American IDUG Conference in May will offer you new insights and helpful answers.

About the Author

Willie Faero has been a database professional for more than 20 years, the last 14 years primarily with DB2. He has been a software consultant for BMC Software, Inc. and a senior DB2 instructor for IBM. Favero is the author of numerous articles, a contributor to several IBM Redbooks, and a regular speaker at regional, national, and international conferences. He can be reached via email at willie_favero@compuserve.com.

---

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 survey suggests that the pace of data warehousing will increase before it slows down; the number of data warehouses and marts is projected to double to nearly six (per organization) by 1999; warehouses themselves are projected to increase in size from approximately 130 GB to approximately 260 GB.

Dealing with aggregates

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 aggregate tables is to optimize performance and to 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. Such agents will be required until RDBMS vendors provide optimization technology that can handle dynamic data summarization and aggregation from normalized structures. Until 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 warehouse's size, you may choose not to do a backup, because you can refresh the data more efficiently. Review the cost and 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, don't overlook disaster recovery requirements. 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 - one that monitors IT costs by providing critical chargeback services to track information resources that are, after all, used throughout the organization.

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 greater the number of simultaneous queries; the result is the potential for frustrating 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 be applied to these issues.

Performance

System performance is closely linked to scalability; it 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 be defined differently for analytical warehouse access than it is 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 way, 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 discussed here impacts the structure and operation of the data warehouse; 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 the warehouse built and make it last.

About the Author

Craig S. Mullins is vice president of marketing and operations for the database tools division of PLATINUM technology, inc. *He is also the author of the popular book,* DB2

Developer's Guide*, now in its third edition; the book includes tips, techniques, and guidelines for DB2 through Version 5. He can be reached via email at cmullinw@platinum.com.*

About IDUG | Conferences | DB2 Resources |
Regional User Groups | Solutions Journal | Vendor Directory

Home | Contact Us

International DB2 Users Group
401 N. Michigan Ave.
Chicago, IL 60611
(312) 644-6610