ISSUE 3, 2009 / VOLUME 14 NO. 3

data management

IBM.COM/DMMAGAZINE

KNOWLEDGE. PERFORMANCE. RESULTS.

Five to watch

Data warehouse trends you need to know

COMMON GOOD
Public agencies reap benefits from analytics

HIGH SCORE
Mining models tested on InfoSphere Balanced Warehouse

PROGRAMMERS ONLY
How to reduce connects to DB2 for z/OS

FORMERLY IBM DATABASE MAGAZINE

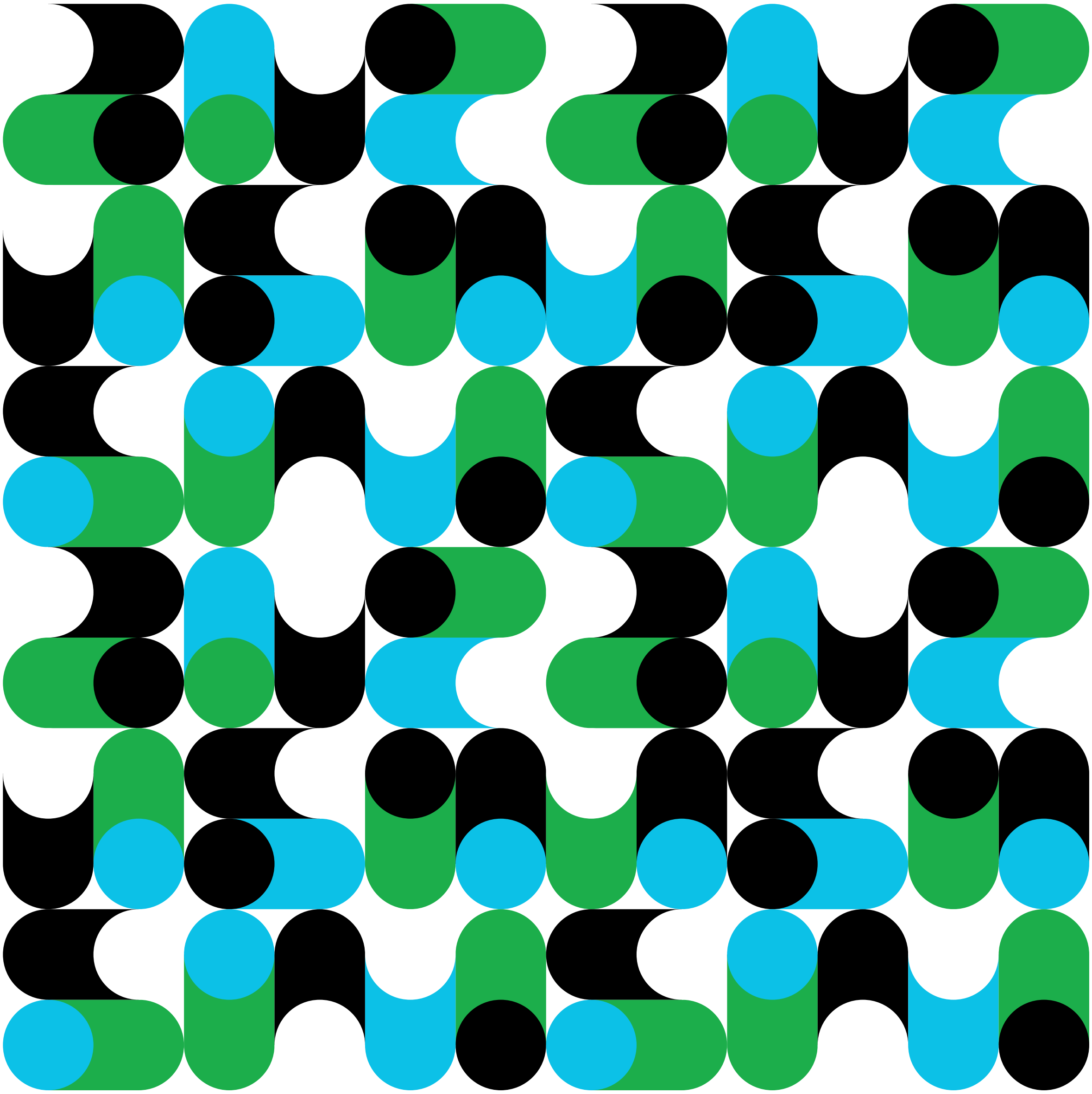
Smarter technology for a Smarter Planet:

Can an entire business be given a nervous system?

Today, instrumented devices connected by powerful service management systems are infusing intelligence into things like production equipment and supply chains, redefining the role of the infrastructure at the core of the enterprise. On a smarter planet, the datacenter is not simply the heart of IT—it's also the central nervous system of the entire business.

IBM is helping companies view their extended infrastructure not as a collection of disconnected pieces, but as an integrated system that connects the datacenter to all of the digital and physical assets of the business, creating a more dynamic infrastructure. From railway systems that can predict and schedule their own maintenance to power grids that match supply and demand, we're already helping customers improve service, increase flexibility and reduce operating costs by as much as 50%.

A smarter business needs smarter software, systems and services.
Let's build a smarter planet. ibm.com/infrastructure



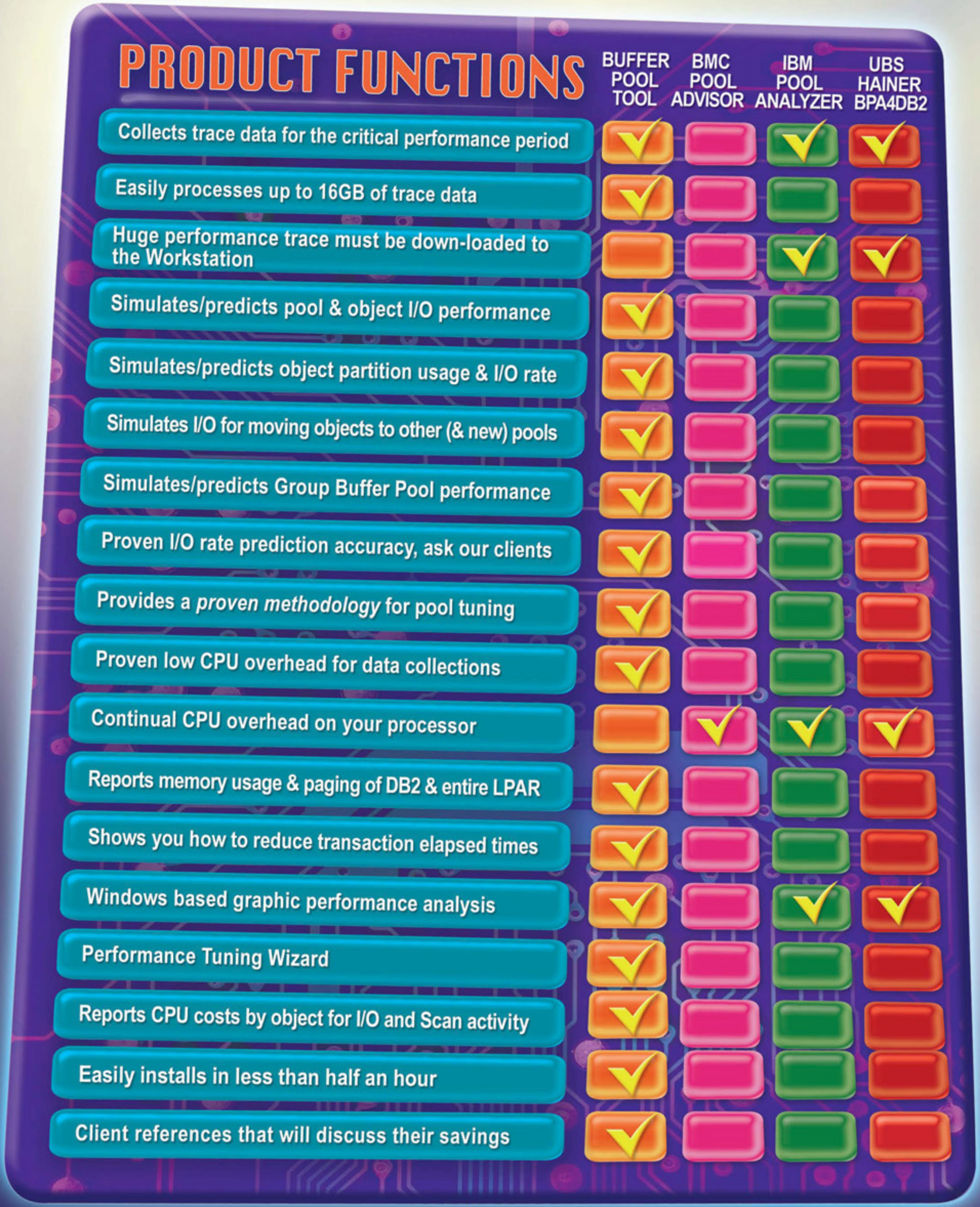
IBM
data
management
IBM.COM/DMMAGAZINE
KNOWLEDGE. PERFORMANCE. RESULTS.
ISSUE 3, 2009 / VOLUME 14 NO. 3

Features

Optimize Both Performance & DB2 Memory Utilization With

Buffer Pool Tool® For DB2

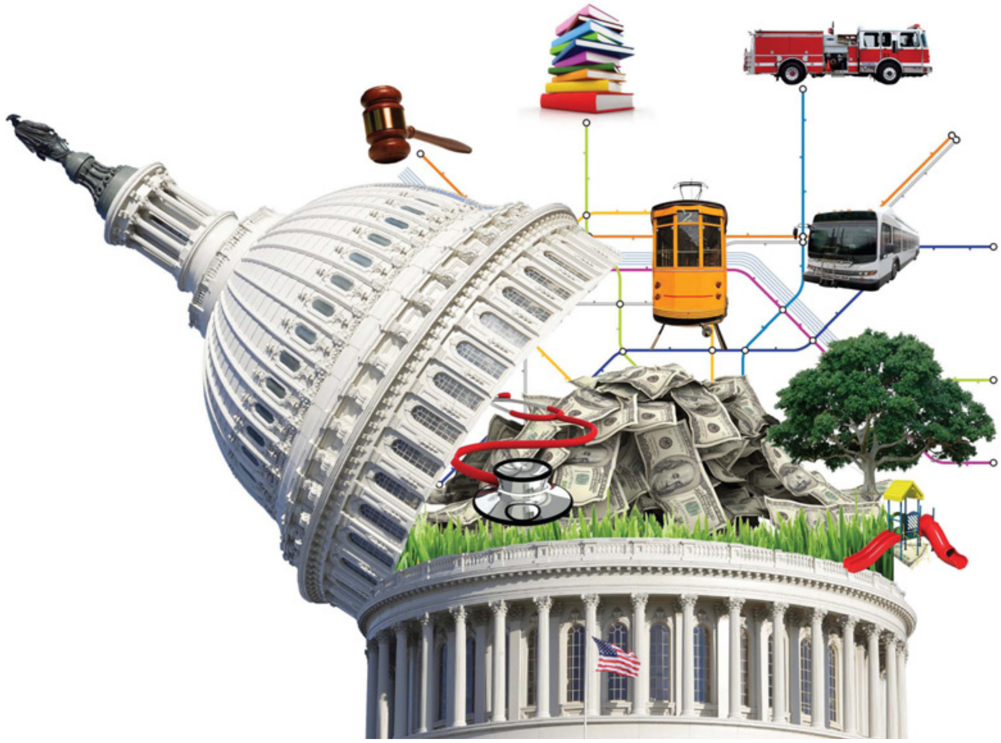
Reduce Your Processing Costs, Save $$ Now!



PRODUCT FUNCTIONS	BUFFER POOL TOOL	BMC POOL ADVISOR	IBM POOL ANALYZER	UBS HAINER BPA4DB2
Collects trace data for the critical performance period	✓		✓	✓
Easily processes up to 16GB of trace data	✓			
Huge performance trace must be down-loaded to the Workstation			✓	✓
Simulates/predicts pool & object I/O performance	✓			
Simulates/predicts object partition usage & I/O rate	✓			
Simulates I/O for moving objects to other (& new) pools	✓			
Simulates/predicts Group Buffer Pool performance	✓			
Proven I/O rate prediction accuracy, ask our clients	✓			
Provides a *proven methodology* for pool tuning	✓			
Proven low CPU overhead for data collections	✓			
Continual CPU overhead on your processor		✓	✓	✓
Reports memory usage & paging of DB2 & entire LPAR	✓			
Shows you how to reduce transaction elapsed times	✓			
Windows based graphic performance analysis	✓		✓	✓
Performance Tuning Wizard	✓			
Reports CPU costs by object for I/O and Scan activity	✓			
Easily installs in less than half an hour	✓			
Client references that will discuss their savings	✓			

RESPONSIVE
SYSTEMS

281 Hwy 79 Morganville, NJ 07751
Tel: 732 972.1261 Fax: 732 972.9416
Web: www.responsivesystems.com



Photograph by G.V. Gireesh

There are only 88 IBM Information Champions in the entire world. Meet two of the newest members of the data elite.

Departments

Ad Index

EM
where informa
Using EMC TimeFinder to Backup and Restore
DB2 for Linux, UNIX, and Windows Databases
Using EMC Tim
Recommended
reading

IBM
data
management
KNOWLEDGE. PERFORMANCE. RESULTS.

EDITOR
Cameron Crotty
editor@tdagroup.com

MANAGING EDITOR
Stephanie S. McLoughlin

ART DIRECTOR
Iva Frank

DESIGNERS
David Chan, Lalaine Gagni, Margie Preston

CONTRIBUTING WRITERS
Merv Adrian, Bonnie Baker, Jack Baker, David Beulke, Scott Bisang, Robert Catterall, John Edwards, Tam Harbert, Lester Knutsen, Stuart Litel, Leah MacMillan, Matthias Nicola, John B. Rollins, Roger E. Sanders, Chris Young

EDITORIAL BOARD OF DIRECTORS
Rick Myllenbeck (Chair), Scott M. Bisang, Debra J. Black, Cathy Elliott, Jeff Jones, Kimberly C. Madia, Nancy Miller, Bob Sawyer, Kathryn Zeidenstein

AD SALES EXECUTIVE
Randy Byers
advertise@tdagroup.com

AD COORDINATOR
Katherine Hartlove

SUBSCRIPTION SERVICES AND REPRINTS
To subscribe to the print or digital version of *IBM Data Management* magazine, change your address, or make other updates to your information, please go to **ibm.com**/dmmagazine. For instant access to the *IBM Data Management* magazine digital edition, visit ibmdmmagazinedigital.com/dmmagazine. For information about reprints, please send an e-mail to customerservice@tdagroup.com.

tda GROUP

PRESIDENT
Paul Gustafson

VICE PRESIDENT, STRATEGY AND PROGRAMS
Nicole Sommerfeld

VICE PRESIDENT, EDITORIAL DIRECTOR
Debra McDonald

VICE PRESIDENT, CONTENT SERVICES
Paul Carlstrom

Printed in the U.S.A.

IBM.COM/DMMAGAZINE

EDITOR'S NOTE / CAMERON CROTTY

For a certain crowd of data devotees, information fanatics, and statistics aficionados, late October in North America becomes something approaching paradise. That's because the continent's four most popular professional sports leagues—Major League Baseball, the National Football League, the National Basketball Association, and the National Hockey League—are either in season or finishing postseason play.

What do sports have to do with data management? Everything. Sports organizations (and fans!) have discovered that the flood of data produced by each game can be collected, mined, and analyzed—generating information and insight that can be used to win championships. Where a discussion of star athletes might once have started and ended with the number of points scored or prevented, today even casual fans compare teams and players by using complex formulas that describe everything from defensive range to situational performance.

Change "free-throw average" or "on-base percentage" to "seasonally adjusted sales growth" or "average revenue per unit," and it's easy to see that businesses have made the same discovery. For companies looking for an edge in the global marketplace, business intelligence (BI) and predictive analytics are no longer a luxury, they're a necessity.

In this issue of *IBM Data Management* magazine, we delve into the source from which all BI and analytics flow: the data warehouse. Whether you already have a data warehouse or are considering building one, start with our cover story, which examines the five business and technical trends that will shape your data warehouse strategy. Stop by the Data Manager column, where Merv Adrian brings us the story of a midsized company's experience starting up a data warehouse using pre-rolled data models. And don't miss our look at the data warehousing challenges faced by public sector organizations.

Let's get our hands dirty! Bonnie Baker is back this issue with ways to boost performance by eliminating unnecessary DB2 connections, while Matthias Nicola joins us from the IBM Silicon Valley Lab with a look at customizing XML storage in DB2.

Of course, there's another reason why late October is a big time for data professionals: the IBM Information On Demand 2009 Global Conference. This year, it's happening October 25–29 in Las Vegas. We hope that you're able to attend, but if not, it's never too early to start making plans for next year. And as always, we want to hear from you at editor@tdagroup.com.

Thanks for reading,

Cameron Crotty
Editor

EMC²
where information lives®

Cross the divide to the data center of the future on the industry's first Virtual Matrix Architecture.

Take your operations to unprecedented levels of scalability and efficiency with EMC® Symmetrix® V-Max™, purpose-built for the virtual data center.

Now you can break through the physical boundaries of traditional storage by making the shift:

- From managing individual products and technologies to providing value-add services within the data center
- From trading off between functionality and costs to providing the right levels of service at the right cost
- From managing physical devices to managing policies
- From providing high availability for selected operations to delivering 24x7xforever availability for all applications

The industry's first Virtual Matrix Architecture™ makes it all possible, enabling you to efficiently scale out your storage resources, manage them as a single entity, and easily provision resources as needed to meet any requirement.

Begin transforming your data center today. Learn more at www.EMC.com/vmax.

IBM Announces General Availability for IMS 11

Release highlights include enhanced ease of use and integration, manageability improvements, and support for 64-bit storage

On October 30, the latest release of IBM IMS—IMS 11—will become generally available. A hierarchical database system with few equals in database and transaction processing availability and speed, IMS 11 is now easier to use than ever before. It can be combined with a service-oriented architecture (SOA), enhancing on-demand business enablement, growth, and availability while helping to improve systems management.

The new release includes integration enhancements and open-access improvements that enable greater flexibility, manageability enhancements that optimize staff productivity, and support for 64-bit storage that delivers increased performance and availability. IMS 11 also eases the complexity of creating database recovery points, reduces CPU time for online database reorganization processing, and enhances capacity optimization.

Key IMS 11 features include:

- Direct SQL access to IMS provides direct, distributed TCP/IP access to IMS data, both within an IMSplex or with platforms other than IBM System z. This enables cost efficiencies in application growth and helps improve system resilience. IMS 11 also includes new usability and serviceability enhancements, such as IMS Syntax Checker support for IMS Open Database.
- IMS Connect enhancements help improve IMS flexibility, availability, resilience, and security. IMS Connect is the TCP/IP gateway to IMS transactions and operations, and now to IMS data.
- Web services, Enterprise JavaBeans (EJB) components, and JavaServer Pages (JSPs) can be created from existing Message Format Service (MFS)–based IMS applications. New applications can be deployed to IBM WebSphere Process Server as part of your business choreography.

IMS ENTERPRISE SUITE COMPONENTS

- **IMS SOAP Gateway** maximizes reuse of IMS assets via standard interfaces with HTTP/SOAP transport to IBM and non-IBM components and IBM Rational tooling for XML conversion. IMS SOAP Gateway enhancements use industry standards and enhanced tooling to further extend the reach and participation of IMS assets in an SOA.
- **Connect APIs for Java and C** extend connectivity of distributed platforms to IMS and help simplify application development for stand-alone, user-written IMS Connect clients. C support is provided through the IMS Enterprise Suite V1.1 service process.
- **Java Message Service (JMS) API** expands Java application development in Java-dependent regions to offer synchronous callout support via the ICAL Data Language Interface (DL/I) call.
- **DLIModel utility plug-in** provides a simplified interface for transforming IMS database information into metadata for use in application development. DLIModel utility enhancements ease data transformation, data integration, and tooling installation.

- Enhanced user exits and new Type-2 commands help simplify operations and improve system availability. Support for IMSplex-wide recovery points taken while the system is online further facilitates database availability.
- 64-bit storage enhances system availability and overall system performance. IMS Fast Path Buffer Manager, the Application Control Block library, and local system queue area use 64-bit private storage, helping to simplify operations and improve storage utilization and performance.

"IMS 11 delivers the ultra-high performance and scalability that our customers need, combined with new features that make the system more accessible to a broader range of IT technologies and users," says Beverly Tyrrell, director of IMS development at IBM. "Today's business environment calls for the increased ease of use and improvements in resource management that IMS 11 is designed to deliver along with relief from many capacity constraints."

Also new for use with IMS 11 is the IMS Enterprise Suite, an innovative set of components that enhances connectivity, expands application development capabilities, extends standards and tools for an SOA, and eases installation. Planned to be generally available on November 6, 2009, the suite is a separate no-cost product for unlimited installs, designed to complement IMS 11 (see sidebar, "IMS Enterprise Suite Components").

Upward compatible from IMS SOAP Gateway Version 10 and IMS 10 DLIModel utility, IMS Enterprise Suite provides user-friendly standard interfaces, simplifies IMS metadata generation, and enables IMS business event data and monitoring. It also eases and expands IMS development (including Java and XML), administration, and access, and offers a visual representation of IMS databases and program definitions.

IMS 11 is complemented by a growing portfolio of tools, which offer day-one support for the new release. Recent announcements include IMS Cloning Tool V1.1 and new releases of IMS Audit Management Expert, IMS Connect Extensions, IMS Performance Analyzer, and IMS Problem Investigator.

MORE INFORMATION

ibm.com/software/data/ims

Moving DB2 to the Cloud

RightScale Cloud Management Platform to support IBM DB2 9.7

RightScale, a provider of cloud computing management solutions, announced in July that the RightScale Cloud Management Platform will now enable users to create, manage, and automate IBM DB2 Express-C 9.7 database software on the cloud. The release allows RightScale users to more easily build, test, and deploy applications on leading clouds, like Amazon Elastic Compute Cloud (EC2), using the most recent version of DB2.

"We are proud that RightScale customers will be among the first to take advantage of the latest advances in DB2 technology by deploying their database servers on a cloud in a fully managed RightScale cloud deployment," says Michael Crandell, RightScale CEO. "Both DB2 and the RightScale Cloud Management Platform have been developed with a focus on ease of use, and with this combination, organizations can get up on a cloud quickly with a cost-effective, enterprise-class data management solution."

DB2 Express-C 9.7 is available free of charge. This version of DB2 can be set up quickly, is easy to use, and includes self-managing capabilities designed to meet the needs of small and midsized businesses.

MORE INFORMATION

www.rightscale.com

IDUG OPENS UP MEMBERSHIP, REVAMPS WEB SITE

The International DB2 Users Group (IDUG) recently unveiled a new Web site and membership program that provides free premium membership.

The redesigned Web site includes more than 450 technical videos, a calendar of all IDUG and regional user group (RUG) events, user blogs, and the previous three years of conference presentations. Members have free access to all resources, except for the most recent conference proceedings.

MORE INFORMATION

www.idug.org

Information Management Communities

Network, learn, and share knowledge with your peers—online or in-person

Participating in an Information Management community is one of the best ways to enhance your personal and professional development, make you more efficient in your job and organization, and position yourself as a thought leader in your field.

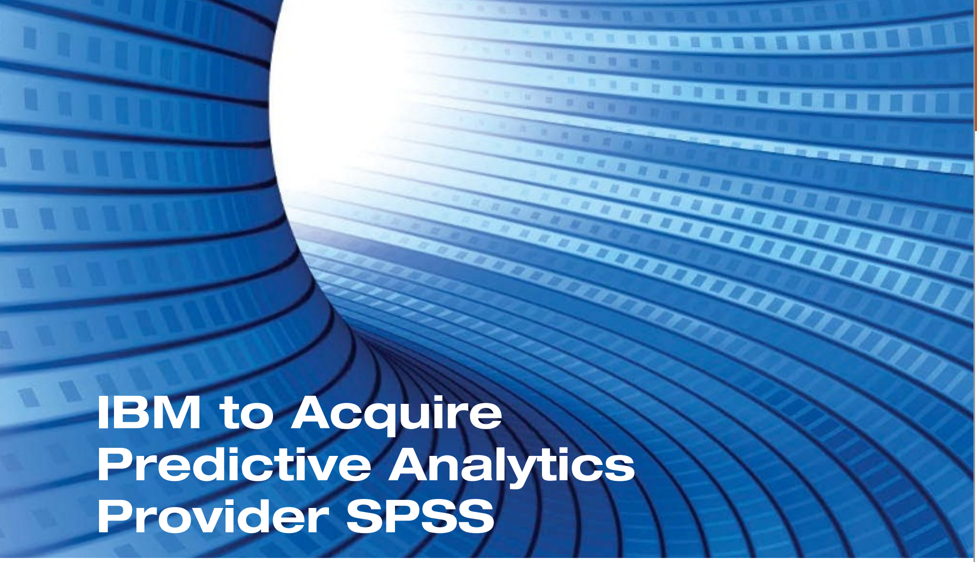
Members of Information Management communities include business and IT professionals, students preparing for their professional careers, IBM Business Partners, independent consultants, and IBM Information Management subject matter experts. Explore the many benefits by joining or participating in one or more of these communities:

- **User groups** are member-run communities that bring together thousands of like-minded professionals. Dedicated to specific technologies, user groups can empower you to expand your personal and professional development and improve efficiencies in your organization. They help you achieve these goals by delivering high-quality technical resources, online education and networking opportunities, face-to-face local events, and global conferences.
- **Online communities and forums** provide opportunities to share experiences and network with others who share similar interests. Participation in these communities provides many benefits, including peer-to-peer support and collaboration, interaction with subject matter experts, and access to best practices, insight, and innovation.

IBM developerWorks hosts many user forums covering a wide range of topics and interests at **ibm.com**/developerworks/data. For more communities and networking resources, check out the groups and forums below.

IBM User Groups	
International DB2 Users Group	www.idug.org
International Informix Users Group	www.iiug.org
ECM UserNet	**ibm.com**/software/data/content-management/usernet.html
Cognos International Users' Groups	**ibm.com**/software/data/cognos/usergroups
IMS Regional Users Groups	**ibm.com**/software/data/ims/usergroups.html
International U2 Users Group	www.u2ug.org

IBM User Forums	
Data Management	**ibm.com**/community/datamanagement
Enterprise Content Management	**ibm.com**/community/ecm
InfoSphere	**ibm.com**/community/infosphere
Business Intelligence and Performance Management (Cognos)	**ibm.com**/developerworks/data/products/cognos
ChannelDB2	www.channeldb2.com



IBM to Acquire Predictive Analytics Provider SPSS

Agreement increases IBM capabilities in advanced data capture, data mining, and statistical analysis

On July 28, IBM announced an agreement to acquire SPSS, a leading provider of predictive analytics software. The US$1.2 billion acquisition is expected to further expand the IBM Information On Demand software portfolio and the company's business analytics capabilities.

The ability to forecast future trends and spot shifts in consumer behavior can give businesses a competitive advantage in today's economy. Industry analysts foresee the market for business analytics software growing to US$25 billion this year, up 4 percent from 2008.

IBM is expanding its focus on business analytics technology and services to address enterprise requirements to cut costs, reduce risk, and increase profitability. Predictive analytics capabilities, which include advanced data capture, data mining, and statistical analysis, help organizations analyze trends and patterns found in historical and current data to predict potential outcomes and optimize business services, including product and service offerings for customers.

"With this acquisition, we are extending our capabilities around a new level of analytics that not only provides clients with greater insight, but with true foresight," says Ambuj Goyal, general manager for IBM Information Management. "Predictive analytics can help clients move beyond the 'sense and respond' mode—which can leave blind spots for strategic information in today's fast-paced environment—to 'predict and act' for improved business outcomes."

Data in Action Virtual Conference

Gartner Vice President Donald Feinberg discusses top challenges for database administrators

Held August 19, the Data in Action Virtual Conference highlighted effective data management solutions for smarter business outcomes. Listen to the replay to hear more, including remarks by Donald Feinberg, vice president and distinguished analyst at Gartner, about the state of the database management system software market and the data management challenges facing businesses today.

Feinberg addressed many topics, including the problem of exploding data volumes and new computing paradigms for handling it, data management cost-reduction techniques and technologies, and the perspective of the DBA now and in the future. The conference also featured a moderated panel of customers and partners who are leveraging IBM DB2 9.7, two interactive Webcasts, and eight exhibit booths.

› MORE INFORMATION

http://w.on24.com/clients/ibm/datainaction

More zIIP for DB2 Utilities Suite for z/OS

IBM engine gives DB2 utility sorts a boost

IBM DB2 Utilities Suite for z/OS for DB2 8 and DB2 9 now offers additional capabilities for leveraging IBM System z Integrated Information Processors (zIIPs), helping organizations reduce total cost of ownership by exploiting System z specialty engines.

This enhanced support opens zIIP eligibility to any DB2 utility sorting of fixed-length records in the memory object sorting path, except for REORG variable-length data sorts. When sorting fixed-length records using the memory object sorting technique, a portion of the workload will be redirected to a zIIP when one is available. This is different from the existing DB2 utility zIIP support (first introduced in DB2 8), which applied to index maintenance only.

Preliminary IBM lab testing indicates approximately 50 percent of sorts are eligible for zIIP offload for this kind of workload. Large, CPU-intensive sorts are expected to gain the greatest performance improvement. Actual results may vary depending on a number of factors, including the capacity and number of the eligible zIIP specialty engines and available system resources.

If you're going to the Information On Demand 2009 Global Conference in Las Vegas on Oct. 25–29, make sure to attend session #1608 (Maximizing the Performance of IBM DB2 9, DB2 Utilities, and IBM System z10) and session #1456 (DB2 Utilities—Update) to learn more about enhanced DB2 utility zIIP support.

› MORE INFORMATION

ibm.com/software/data/db2imstools/db2tools/db2utilsuite

Stuart Litel is president of the International Informix Users Group (IIUG; www.iiug.org/president), CTO of Kazer Technologies (www.kazer.com), an IBM Gold Consultant, member of the IBM Data Champion Inaugural 2008 class, and recipient of the 2008 IBM Data Professional of the Year award.

Uptime, Conference Time, Revenue Time

I'm no data warehouse expert, but Informix Warehouse looks pleasantly familiar

Years ago, back in the '90s before Informix became part of IBM, there was a version of Informix V8 that was also known as Informix XPS, which is still supported and used today. Although I did a short consulting assignment about 10 years ago on an XPS system, I never really got involved in the nuts and bolts of the system. However, two good warehouse expert friends, fellow *IBM Data Management* magazine columnist Lester Knutsen and IBMer Martie "The Brain" Lurie, both swear to me that XPS was about the best database warehouse product out there—and I will defer to Lester.

Now, IBM has Informix Warehouse, which is effectively the Informix version of IBM InfoSphere Warehouse. Informix Warehouse (**ibm.com**/software/data/informix/warehouse) actually incorporates some of the more commonly used features of Informix XPS directly into IBM Informix Dynamic Server (IDS) 11.5. It also adds several other IBM tools to enable you to build a mission-critical data warehouse on top of what anyone who spends any time with IDS knows is an extraordinarily fast, reliable database engine.

Making a point: Informix uptime

How fast and reliable is it? Just this morning I got an e-mail from one of the Informix architects at IBM asking me why the IIUG Web site is still using Informix V11.0 when the latest version is 11.5? It's because the IIUG server database has been up for almost two full years without a restart. Now, this is no knock on 11.5, and the upgrade would take only 15 minutes, but this old-school uptime hound just loves being able to say that our database has been up for that long. Yes, Informix really does stay up and running without a restart for well over a year on a server that receives more than a million hits a month!

2010 IIUG Conference location

Now that I have tried to fulfill the request from the editor to include something on data warehouses, let's get to some IIUG business. We have confirmed that the very successful annual IIUG Informix Conference will be at the Marriott Overland Park for the third year in a row, from April 25–28, 2010.

A call for papers—or, as I call it, "Submissions from Potential Presenters"—will go out later this year to encourage more users to submit their writings. The deadline will be November 15, 2009. Remember: if I can present, anyone can. And anyone selected to present gets a free conference registration pass. For more conference information, visit www.iiug.org/conf.

Speaking of conferences, I hope everyone will try to attend this month's IBM Information On Demand 2009 Global Conference, running October 25-29 in Las Vegas. If you are going to be there, make sure you say hello if we pass in the halls.

Go Informix and go huge

Finally, I have a great tip for ISVs. If you are not yet on the Informix bandwagon, meaning your product does not work with Informix, IBM has started a new program to lend a hand. Just port your existing or new product to IDS, and IBM will let you keep all the database sales for the following year (excluding maintenance).

Yes, you heard me right: port your existing ISV product to work with IDS, sell a billion dollars' worth of Informix databases, and you get to keep almost all of the database sales revenue, too.

Now, of course there are restrictions and I don't represent IBM, so you should contact Luis Pereira of IBM at lcpereira@us.ibm.com (tell them I sent you) for more information about the program. And special thanks to Robert Thomas, vice president of business development for IBM Information Management, for putting this all together.

Top Performance Features in DB2 9 for z/OS

Indexing, compression make your applications work faster

David Beulke (dave@davebeulke.com) is president of Pragmatic Solutions, Inc. (PSI), a training and consulting company specializing in designing and improving SQL, application, and system performance on DB2 for Linux, UNIX, and Windows, and z/OS. He has experience in the architecture, design, and performance tuning of large data warehouses and OLTP solutions and is a former president of IDUG.

IBM DB2 9 for z/OS is loaded with features that can boost the performance of your applications. In this column, I'll tell you about two that really stand out, but check out the online version of the column for a table of the top 50. Implementing any of these features can help reduce your applications' CPU burden and improve application response time.

After you convert to DB2 9 for z/OS, the first performance feature you should evaluate is one that probably will have the greatest performance impact: indexing on expressions. You can now put an index over a subset of a single column or derivatives of multiple columns, a formula, or a function used within your application. By creating a customized index on an expression, you can change the access path of an application using a tablespace scan, which is a combination of multiple indexes, to a single customized index—a much more efficient method. Because you define the index expression, you can design it to provide direct or superior access for any SQL statements within your applications. You can create an indexed expression on a table of any size, which means that you can use this feature to improve performance on all of your applications.

Almost any expression done in an SQL statement can be used to create an expression for an index, and the index expression can use multiple columns or expressions on multiple columns. DATE and other functions used in your application's SQL code are great candidates for indexes on expressions. For example, imagine grabbing the date from the insert timestamp of the row (DATE(IN_TIMESTAMP)). Instead of performing a tablespace scan to get the date portion of a timestamp, an application can access it as an index. This process is not only faster, but it also reduces the load on the CPU.

The second big-impact performance feature within DB2 9 for z/OS is index compression. Index compression offers the same benefits as DB2 data compression by storing more index entries per index page. The increased number of entries per page allows more keys to be available in the system and buffer, reducing I/O requirements and improving the number of index entries available in the buffer pools.

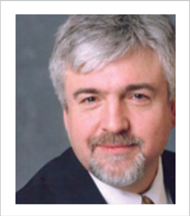
However, index compression works very differently from data compression: the index entries are compressed only at the physical disk page level, and a compression dictionary is not used. This allows index compression to be available without using a REORG or LOAD utility. The index entries are expanded in the buffer pools, allowing quicker memory traversing of the index structures for data retrieval. For example, compressed index pages may be only 4 K on disk but 8 K, 16 K, or 32 K in the buffer pool.

Index and data compression can save disk space, backup time, and recovery time over the life of your system, and can save CPU time by reducing processing I/O. Because compression does have overhead costs, it may not be appropriate for systems with insert/update loads—but it is ideal for large business intelligence or operational systems where data is mostly being read.

These are only two of many interesting features in DB2 9 for z/OS. For an extensive list of other performance features that can help your applications run faster and more efficiently and help save your company money, take a look at the table posted online at **ibm.com**/developerworks/data/library/dmmag/DMMag_2009_Issue3/IDUG/index.html. These new z/OS features and many other DB2-related topics are also covered at the revamped International DB2 Users Group (IDUG) Web site (www.idug.org), along with more than 450 videos at www.idug.org/db2-videos.html. Check it out!

Right-Sized Analytics

Rooms To Go finds a data warehouse–driven solution that fits its midsized business



Merv Adrian, an analyst and consultant, founded IT Market Strategy after three decades in the IT industry. During his tenure as senior vice president at Forrester Research, he was responsible for all of Forrester's technology research.

Ever wonder if all the brochures and flyers that spill out of your Sunday newspaper represent a careful, cost-effectively targeted marketing investment? Rooms To Go, a regional furniture retailer based in Seffner, Florida, wondered too.

With US$1.5 billion in annual revenue and 150 showrooms across the Southeast organized into three districts with multiple distribution centers, Rooms To Go is a leading independent furniture company in the United States. However, it has not been able to compete using the kind of information technology that drives marketing campaigns for the industry's giants. In fact, says CIO Russ Rosen, "We have been outsourcing our targeting efforts and mailing expensive brochures to too many people. Response rates being what they are, we were spending a lot of money on costly brochures with little linkage to results, and it was difficult to steer."

Rosen wanted to detect and predict the buying patterns of customers beyond simple metrics, such as which furniture pieces and categories are the best sellers by state and customer type. With better insights, Rooms To Go would be able to target repeat customers more effectively in its campaigns.

Sound familiar? Customer analytics fueled by data warehouses are one of the not-so-secret weapons of the world's largest retailers. The benefits, even if measured in only a point or two of improvement per store, can be massive. And as Rosen discovered, effective analytics are no longer out of reach or years away for small and midsized businesses.

Prebuilt data models offer best-practice expertise

Like many others in his position, Rosen quickly discovered that his team lacked both the critical expertise and the requisite time for building its own data warehouse from scratch. For example, each of the company's three districts is a separate distribution area and has individual data collection systems for both supply and sales. "Before I could do anything, I needed to roll the data into one place," says Rosen. "But we attempted a data modeling effort and realized that it was over our heads."

Rooms To Go solicited proposals from consultants; the bids were quite large, and the allotted budget was beginning to look inadequate for the task. The IBM InfoSphere team offered a different solution: InfoSphere Balanced Warehouse, including IBM retail business solution templates. The package included not only the hardware and software for database, analytics, and data warehouse design and administration, but also a pre-developed, ready-to-go retail model (see sidebar, "Solution components").

Such models have proliferated in recent years, as specialists codify years of experience in building systems within specific industries, pointed at specific tasks. Rosen believed that these best practices could shorten his development time, but was unsure whether they would be too different from his current systems to be easily adapted. “When we went over the business solution templates, I knew we were on the right track,” says Rosen. “The model defined which attributes would be needed for market basket analysis and for campaign promotion measurements. We weren’t modeling experts, so this seemed like it would save us a lot of time.”

The prebuilt model also addressed another of Rosen’s concerns. “We knew that the first time we analyzed our data would not be the last,” he says. “Our biggest fear was that we’d solve one problem and then have to rebuild as soon as we identified the next one. The promise of using a model, with the experience that had gone into it, was that we’d be able to build the next piece as a small increment without tearing everything down.”

Rosen initiated the project and quickly came to grips with one of the biggest challenges: mapping real data to business needs, in this case as specified by the data model.

“A big lesson learned for us was identifying the need for data we had not been collecting,” he explains. “I spent a few weeks with an IBM modeling consultant mapping the model to the data we actually had, and we figured out what we were missing. We had to build some new tables to collect information that had not been needed to support the original transactional requirements.”

The project took a pause to get all of the necessary data in place. “We actually went back and changed some of the transactional systems,” Rosen says. “As the sponsor of the project, I knew the delay meant much better information.” The changes to the transactional system were additions; because they didn’t alter any of the existing data structures, no line-of-business approval was needed.

Legacy systems put pressure on timelines

Another key lesson, and one that many smaller firms will encounter when deploying data warehouses, was the vital role of the extract, transform, and load (ETL) stage, and its potential cost in money and time. Rooms To Go runs on internally built transactional systems, using a Rocket Software UniVerse database with multi-value fields and distributed files—a specialized architecture that, while robust, is hardly representative of today’s typical data formats. The UniVerse data had to be converted before it would work with the new IBM DB2-based system.

As the project got underway in earnest, the ambitious three- to four-month time frame appeared threatened by the time the conversion might take. It’s not uncommon for this step to cause lengthy delays when legacy formats are involved: the skills to develop the data transfers may be in short supply, and commercial ETL products don’t handle every necessary task. Factoring in the time to convert complex data formats requires some careful assessment, and even well-planned projects can hit snags.

Rosen decided to bring in a consulting firm rather than hiring additional staff; the extra short-term cost was a better investment than adding in-house resources that would not be needed once the project was complete. Fortunately, the IBM solution included a “light” version of InfoSphere DataStage, which proved sufficient—with some tweaking—to move the data.

Once the data issues were resolved, Rosen turned his attention to delivery. “From start to finish, the first project took two months, and once it was done, we delivered our first executive dashboards in a week or two. Now we’re able to combine the market basket analysis with an understanding of customers’ linear decisions in repeated visits to the showrooms.”

The system meets not only Rooms To Go’s current IT needs, but has plenty of options for expansion—an important consideration for any business with an eye on growth. For example, the 3 TB IBM System Storage DS3200 has enough space to eliminate any scalability concerns for the near future. “We’ve loaded three years of data and we haven’t come close to using half the space yet,” says Rosen. “We’re not using DB2 Compression yet; we’ll investigate that, but we don’t need it so far.”

> “A big lesson learned for us was identifying the need for data we had not been collecting. I spent a few weeks with an IBM modeling consultant mapping the model to the data we actually had, and we figured out what we were missing.”
>
> —**Russ Rosen,** *CIO, Rooms To Go*

Rosen didn’t have to be told that few project variables are more important to success than payback. Even before the analyses begin to deliver results in campaign design, the internal systems are expected to deliver results in less than two years by eliminating the costs of the outsourced marketing systems. The entire package, including hardware and software, cost less than some of the design-and-develop–only proposals Rosen received. But the real payoff will come from Rooms To Go’s ability to develop campaigns that compare favorably with those of larger organizations—adding to the top line.

SOLUTION COMPONENTS

- IBM InfoSphere Balanced Warehouse
- IBM Retail Data Warehouse
- IBM InfoSphere DataStage
- IBM Cognos Business Intelligence
- IBM DB2 for Linux, UNIX, and Windows
- IBM System x3650 server with quad-core Intel Xeon processors running at 3.0 GHz with 4 GB RAM
- IBM System Storage DS3200 with 3 TB preconfigured storage

Change, and the Data Warehouse

What you need to know about 5 trends that are reshaping the data warehouse landscape

By John Edwards

Challenge,

Mike Randolph, vice president and senior technology manager for Bank of America, has seen data warehouse trends come and go. “Change has been constant over the years,” says Randolph, who supervises a 22-node, IBM DB2–driven warehouse that supports the bank’s credit card operations. “You either learn to adapt to the changes or get swallowed up by them.” ¶ These days, five major trends—exploding data growth; end-user demands for greater data analysis, granularity, and speed; requester and source proliferation; the growing popularity of prefabricated appliances/data models; and the challenge of working with unstructured data—are reshaping the data warehouse landscape, challenging adopters across all industries. But Randolph isn’t shrinking from the task. “You need to face change head on with the knowledge that any temporary disruptions created will be more than compensated for by better performance and the addition of new capabilities,” he says. ¶ For data warehouse managers like Randolph who are willing to embrace emerging trends, change and challenge present an opportunity to excel, says Warren Thornthwaite, a consultant with The Kimball Group, a data warehousing education and advisory organization. “Whether you’re dealing with things like growing data volume and the need for deeper data analysis or wondering how to handle unstructured data, you need to turn change into an opportunity,” he explains.

Exploding data growth

Data is expanding in at least two ways. The amount of information stored inside warehouses is snowballing as content accumulates over time. A 2008 study by a major market research firm revealed that enterprise data requirements are growing at an annual rate of 60 percent. Meanwhile, as more enterprise processes are instrumented and recorded, warehouse managers face a growing avalanche of data that must be organized and analyzed for possible warehouse use.

Data growth requires enterprises to create data warehouses that can be expanded quickly and efficiently, says Greg Lotko, vice president of warehouse solutions for IBM Information Management. "Look for an offering where modular building blocks allow enterprises to start with a warehouse of a certain size and then, as it grows, click in new modules of hardware and software together," he says.

But data warehouses can't be scaled upward infinitely. To prevent useless data from burdening systems, enterprises must also pay attention to the age and overall quality of their archived data, says George Goodall, an analyst at Info-Tech Research Group.

"Once information gets locked up in a database, organizations are very reticent to get rid of it," Goodall explains, noting that enterprises tend to err on the side of caution. Many opt to keep everything forever, either worried that the information may be needed to fulfill some type of regulatory mandate or simply assuming that at least some of the stuff may have future value. "Enterprises have to start paying attention to the effective life span of data," Goodall says. Information lifecycle management tools that help administrators rate and organize data can make this job easier.

Bank of America's Randolph feels that gaining the upper hand on mounting data is primarily a matter of creating strict—yet manageable—data retention guidelines. "Define retention periods and then stick to them," he says. "If exceptions are requested, make people justify why they need to go around whatever your standard for retention is—then really focus on keeping data only for the period of time that it's needed." Don't assume, for instance, that a compliance mandate requires permanent storage of a certain type of file or record—check the facts to learn what information is really needed and for how long.

Randolph says data modeling is the best way to manage the flow of information into a data warehouse. "You really have to focus on making sure that you're only bringing in data that adds value, as opposed to just saying, 'Hey, here's all this data, let's throw it in the warehouse and we'll figure out what to do with it later,'" he says. "It's simply a matter of thinking out and planning each data source."

Planning for growth

Data warehouse solutions must expand to match business growth, help organizations understand their data, and provide tools for removing data that is no longer in use. IBM InfoSphere Balanced Warehouse is available in configurations for businesses of most sizes, and can be expanded over time in a building-block approach. IBM InfoSphere Data Architect helps planners discover, model, and standardize data assets, while the IBM Optim software family offers a deep lineup of data management tools, including Optim Data Growth Solutions for automated archiving and storage of historical records.

ibm.com/software/data/infosphere/balanced-warehouse
ibm.com/software/data/studio/data-architect
ibm.com/software/data/optim

Picky end users

As data warehouses move deeper into the enterprise mainstream, end-user needs and expectations are driving demand for greater accuracy and more refined conclusions delivered in real time. "In just about anything in life, people always want more than they currently have," Goodall observes.

These increasing demands place new burdens on data warehouses and the people who manage them. Randolph says that carefully designed and configured data analysis tools can help managers satisfy increasingly picky end users without driving costs through the roof or sending performance levels crashing into the basement. "It's a mixture of building tools so that they have good response, and quicker response, but also being smarter on the front end where you're only populating

The right analytics at the right time

Balancing powerful tools with user-friendly interfaces can be a challenge. IBM Cognos solutions offer a wide palette of BI and performance management tools to help organizations efficiently deliver the information that users want. Also, the recently announced IBM Smart Analytics System helps organizations deliver a complete solution more quickly by providing broad analytic tools pre-integrated with a data warehouse foundation.

ibm.com/cognos
ibm.com/smart-analytics-system

the stuff that's really needed," he notes. Managers can, for example, provide end users with standardized analysis models that will help them achieve their desired goals quickly and easily.

Finding, creating, and fine-tuning data analysis tools to meet end users' growing expectations is becoming a major challenge for data warehouse managers, but so is tempering overly optimistic end-user expectations, says John Hagerty, a data warehouse analyst at AMR Research. "It's very important for IT, in combination with very visible business champions, to in essence paint the picture for people of what's really possible," he says. A few minutes spent with an end user, showing him or her how to effectively use a set of data analysis tools to perform various tasks, is often enough to diffuse complaints that the technology is slow, cumbersome, or ineffective.

Hagerty also suggests that managers regularly assess their tools to see if they are keeping pace with both system capabilities and end-user demands. "It's a continuing process," he adds. "You need to keep evaluating in order to ensure optimum performance."

The balancing act

Many data warehouses are at risk of becoming victims of their own success. As more departments and business partners learn how to exploit the technology to their own benefit, an unprecedented number of new requesters and sources threaten to slow performance to a crawl. For data warehouse managers, the challenge lies in maintaining access and stability in the face of growing system loads—without sacrificing speed and security.

Randolph says that the key to maintaining a successful balance between stability and speed is to use security and access control tools that don't adversely impact system performance. He suggests carefully scrutinizing specifications to find the products and services that impose the lowest infrastructure burden. "It's really a combination of having a strong gatekeeper, having an underlying infrastructure that adequately supports the data warehouse, and using a strong set of analysis tools," he says.

If, despite a manager's best efforts, a data warehouse is beginning to buckle under end-user pressure, it may be time to consider a new approach. "What we're telling our customer base is, spin off a logical datamart inside the data warehouse with Cubing Services," says Bill Wong, program director of data warehousing solutions, strategy, and market offerings at the IBM Toronto Laboratory.

Using IBM Cubing Services (see sidebar, "Performance Cubed"), organizations can create, edit, import, export, and deploy cube models over the relational warehouse schema. Cubing Services also provide optimization techniques to improve the performance of online analytical processing (OLAP) queries. "It's helping a lot of companies save on the real estate, the administration of extra servers, power, and things like that," Wong says.

The out-of-the-box warehouse

Like bespoke suits and hand-rolled cigars, the custom warehouse is becoming the exception rather than the rule. Today, a growing number of enterprises are turning to warehouse appliances and industry-specific data models that enable a data warehouse to be created in days or hours as opposed to weeks or months.

Goodall says that the "out-of-the-box" approach is highly appealing to organizations that want to build a data warehouse quickly, with less effort, and at a potentially lower cost. "These offerings have abstracted away a lot of the infrastructural complexity that one gets into with building a data warehouse," he explains. "They make a lot of the infrastructure side of things much easier as well; they make

Performance cubed

To help organizations manipulate their data more effectively, IBM InfoSphere Warehouse now offers direct support for optimized OLAP analytics with Cubing Services, a multidimensional analysis server that provides access for OLAP applications. With InfoSphere Warehouse, organizations can create, edit, import, export, and deploy industry-standard OLAP models over the relational warehouse schema. Included wizards also offer optimization recommendations to help improve the performance of OLAP applications and tools.

ibm.com/software/data/infosphere/warehouse/olap.html

Integrated solutions: Just add data

IBM has a wide array of options designed to give businesses a running start on data warehousing and analytics. IBM InfoSphere Balanced Warehouse solutions offer fully integrated, tested, and scalable components that combine easy-to-deploy warehouses with powerful reporting tools and BI capabilities. Another option is the IBM Smart Analytics System, which combines advanced, scalable analytic tools with a data warehouse on a storage and server platform.

ibm.com/software/data/infosphere/balanced-warehouse it very easy to scale up the scope, the complexity, and the size of the data warehouse."

As Goodall sees it, the signal challenge to prefabricated appliances and data models is that the one-size-fits-all approach should really be labeled "one size fits most." That's because product developers aim for the "average enterprise," not the organization that needs a data warehouse that reflects its exceptional or unique way of doing business. "If you're a leader, and you have gone out of your way to do something different from your competitors, then those industry-standard models can be a bit of a liability," Goodall observes.

On the other hand, despite its inherent limitations, prefabricated technology is certainly a time-saver that will help almost any enterprise get a running start on building its data warehouse. The infrastructure can then be further configured and tweaked to bring it in line with its adopter's specific and custom requirements.

Structuring unstructured data

As data warehouse technology matures and grows more sophisticated, an increasing number of enterprises would like to use their systems to tap into the hidden knowledge that's locked inside unstructured data.

Unstructured data—information that doesn't fit a standard data model—can arrive from many sources, including online surveys, Web forums, and e-mail. "Unstructured data means all the stuff that comes in on the questionnaires or document scans that you can now leverage directly and pair with traditional structured data," says IBM's Lotko. "Then you can derive new insights that you wouldn't have been able to create previously because you didn't have access to the information." Free-form text fields within customer relationship management (CRM) applications, for instance, can give enterprise decision makers the information they need to identify ongoing dissatisfaction trends as well as recurring issues that may be causing the problems.

AMR Research's Hagerty notes that an emerging family of business intelligence (BI) products and services are beginning to give data warehouse end users the ability to peer into and derive meaning from data contained in e-mail, call-center notes, chat transcripts, and Web pages. "Users get to see and track opinions, attitudes, sentiments, and other concepts that aren't easily represented in traditional data fields," he says.

Hagerty sees a bright future for unstructured data. "Once the technology catches up to the promise, unstructured data will become as ubiquitous as traditional BI or analytic technology," he predicts. But embracing unstructured data will require data warehouse managers to undergo a mind change: "One of the things a lot of data warehousing professionals have drilled into them is that things have to sit in rows and columns," he says. "Unstructured data will require these people to look at data in an entirely new light, understanding that text and even media can impart at least as much intelligence as numbers."

Tying it together

Recognizing emerging trends, while important, isn't enough to ensure a data warehouse's long-term viability, says IBM's Wong. He notes that it's equally important to act upon changes as they appear, perhaps by adding new solutions or by adapting established practices to new paradigms. "Warehouses that are not responsive or flexible—they'll die," he says.

Randolph agrees with the need for flexible and responsive systems. "To accomplish this, you've got to stay on top of things, become knowledgeable, and be open to considering new technologies and approaches," he says. "Then, you shouldn't be afraid to make changes, not for the sake of change itself, but always to keep your data warehouse on the leading edge." ✱

John Edwards (jedwards@gojohnedwards.com) is a technology writer located near Phoenix, Arizona.

> "Warehouses that are not responsive or flexible—they'll die."
>
> —Bill Wong
> *Program Director*
> *Data Warehousing Solutions, Strategy, and Market Offerings*
> *IBM Toronto Laboratory*

Structurally sound

Text analysis is just one example of the extensive unstructured data analysis capabilities available in IBM InfoSphere Warehouse. InfoSphere Warehouse uses the Unstructured Information Management Architecture (UIMA), an open, scalable, extensible platform for creating, integrating, and deploying text-analysis solutions. InfoSphere Warehouse provides operators and tooling for dictionary-based and regular expression–based named entity recognition, and UIMA-based components can be imported and used within InfoSphere Warehouse, helping organizations dig deeply into their unstructured data.

ibm.com/developerworks/data/library/techarticle/dm-0906textanalysis/index.html
ibm.com/software/data/infosphere/warehouse/unstructured-data-analysis.html

When it comes to
performance
Rely on
Informix
WWW.IIUG.ORG
LINE-X
Informix SOFTWARE
74
IIUG Conference 2010
April 25-28, 2010
Overland Park, KS, USA
iiug.org/conf

PUBLIC SERVICE GETS SMARTER

BY TAM HARBERT

DATA WAREHOUSING AND BUSINESS INTELLIGENCE HELP THE PUBLIC SECTOR SHARE AND ANALYZE VALUABLE DATA STORES

When it comes to using data warehouses and business intelligence (BI), the public sector has historically lagged the private sector. Part of the reason may be that public organizations face challenges that are greater than—or at least different from—the private sector's hurdles when implementing these projects.

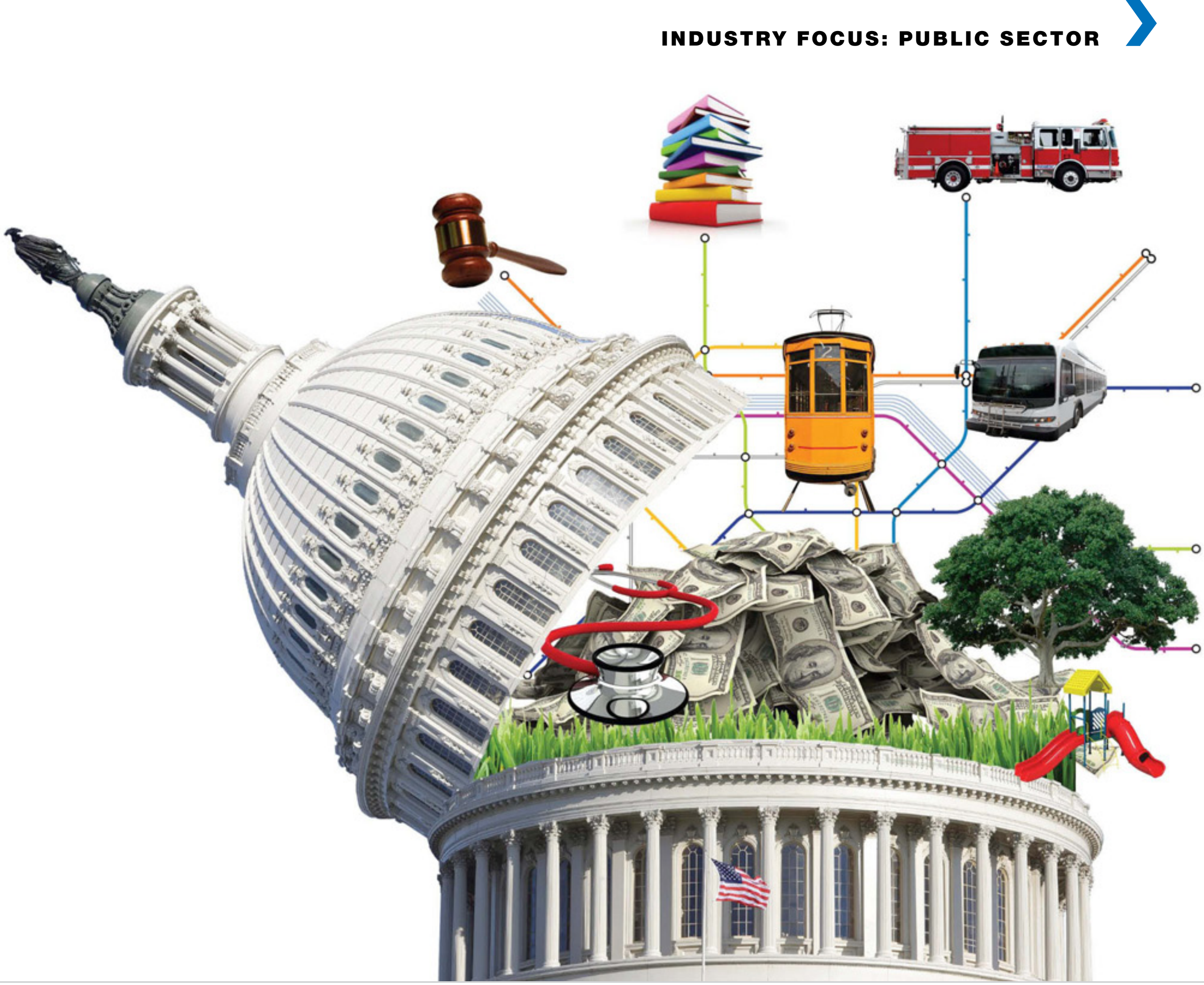
First, finding the money to pay for improving IT is seldom easy and can be even more difficult for public-service groups amidst today's budget shortfalls. Second, public institutions often run into substantial barriers when it comes to sharing data. And third, a public organization's IT strategy can shift rather arbitrarily due to frequent leadership turnover.

For a sector that's facing new demands for transparency and performance, the benefits of analytic software powered by data warehouses are proving irresistible, especially given that recent legislation requires public institutions to better manage and track the use and effectiveness of federal funding. As public organizations adopt these technologies, they're finding new solutions to their long-standing challenges.

Data reporting regulations ring the BI bell

Any discussion of data warehousing and BI in the public sector has to start with money—or too often, the lack of it. Against that backdrop, the American Recovery and Reinvestment Act (ARRA), passed by the U.S. Congress in February 2009, promises to shower US$787 billion across the country to help promote job growth and economic activity. Twenty-nine federal agencies are tasked with distributing the funds. State and local governments will receive more than US$541 billion in discretionary and direct spending and tax cuts; of that total, US$204 billion is specifically tagged as discretionary spending, according to INPUT, a market research firm that tracks government spending. Some US$100 billion of ARRA funds is going to public schools and colleges, including US$41 billion in grants to local school districts.

The Mobile County Public School System (MCPSS) in Alabama is one of them. The largest school district in the state, MCPSS has 63,000 students that generate 1.2 million attendance records and 3.5 million grade records a year. This and other data reside in three different databases (human resources, school and student information, and federal programs) in three different physical locations. Gathering the right information to create timely reports was difficult and time-consuming. For example, federal programs want test scores coupled with both student attendance and teacher qualifications, all of which were in different databases.



"We were spending just incredible amounts of time building these reports, pulling the data sources together to build queries, to build analysis tools," says David Akridge, CIO of MCPSS. "We had to pull all this together in multiple data sets, push it into an Excel spreadsheet or something like that that could be given to them," he explains. "If they wanted to change something, we'd have to go back and rebuild it."

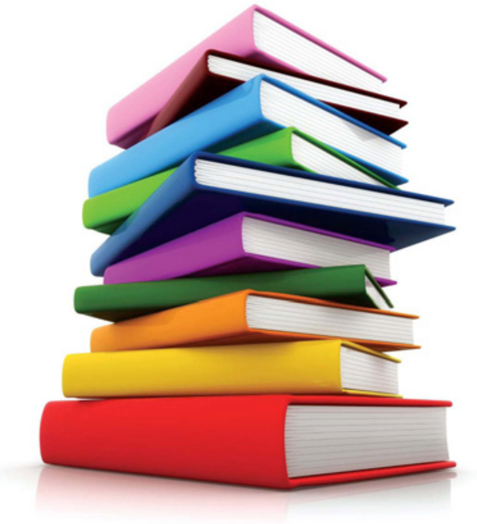
Akridge saw the stimulus money program as an opportunity. He convinced the board of education to revive a former attempt to build a data warehouse to help the district pull relevant information together more quickly and efficiently. The board gave its approval and, after extensive review of competitive offerings, the district hired IBM and its local business partner, DecisionEd Group, to develop a data warehouse based on IBM analytics and BI technology. By August 2009, the district had launched the new system and was rolling it out to teachers and administrators.

MCPSS is not unique. Public-sector organizations across the country, from schools to state and local governments, are using stimulus money to embrace data warehousing and BI. "Business intelligence and performance management have always been strong in the public sector, and the stimulus is just making it stronger," says Robert Dolan, IBM global government and education industry executive, BI and performance management. "We're seeing a lot of interest from government organizations that probably didn't think they needed the technology. There's a buzz around it right now."

In fact, the ARRA virtually mandates the use of BI in government by including strict requirements for accountability and transparency in the use of taxpayer dollars (see sidebar, "Stimulating intelligence"). Organizations that receive stimulus funds are required to publish accounting, allocation, and results data for the money received. The law also mandated the creation of a Web site, Recovery.gov, which is intended to provide increasingly detailed information to citizens on how stimulus funds are being used.

Limits to sharing data

But public-sector organizations face unique challenges when implementing data warehouse and BI projects. For example, when a corporation wants to share data, it usually has the option of standardizing on a certain database platform and sharing data freely among its business divisions. That's a grand oversimplification, and the reality is usually fraught with internal politics and technical challenges, but



Mobile County
Public School System

Project:
Build a data warehouse to improve information delivery and enable more effective student management

Cost: US$1.2 million

Challenge: Information was in three different physical databases. Rendering reports was a time-consuming, complex process that provided limited insight into student performance. School administrators had to wait for quarterly reports, which arrived too late in some cases to flag at-risk students before they got into trouble or dropped out of school.

Solution: IBM Cognos Business Intelligence software (**ibm.com**/cognos), deployed in a relational database environment. Cognos BI produces customizable dashboards that give administrators and teachers up-to-date reports and measures, enabling them to effectively monitor each student based on a number of factors, such as attendance and grades. The system can proactively alert teachers and administrators if a confluence of these factors crosses a threshold, indicating that a student may be at risk.

the path to data sharing in the private sector is—at least theoretically—relatively straightforward.

In the public sector, it can be a lot more complicated. Start with the fact that there may be legal restrictions on sharing data. Even state agencies that send their reports to the same person—the governor—may have different rules and legal restrictions on distributing data, says Geoff DePriest, senior manager in the performance business unit of Crowe Horwath LLP. If one agency is getting funding from the U.S. Department of Labor and another is getting funding from the U.S. Department of Housing and Urban Development, for example, they may have to abide by different rules, he explains.

Even without legal restrictions, effective data sharing is often thwarted because of another obstacle familiar to private-sector businesses: various agencies structure the same data differently. For example, when DePriest worked for the State of Indiana, the Department of Workforce Development architected its data on a person-by-person basis, while Family Social Services structured the data on a case-by-case basis.

One agency that is trying to surmount such data-sharing challenges is the U.S. Census Bureau. In 2007, the Census Bureau granted a contract to IBM to provide data tabulation and dissemination services to support the 2010 Census and other key surveys. The Data Access and Dissemination System (DADS) division of the Census Bureau is responsible for disseminating data from five major surveys, including the decennial (10-year) census that will be conducted in 2010. But each of the five surveys collects data differently, and they use different database systems on the back end to summarize the data. “One shop may be using SAS, another produces database tables, and in some cases we simply get flat files,” says Jeff Sisson, DADS program manager. Because of this, the IT staff spends a lot of time custom-coding data for the system.

With IBM’s help, DADS is building a data-loading system that will be flexible enough to handle a variety of data formats. By using standard technology and designing the system around metadata, DADS expects to increase efficiency and save back-end costs.

In addition, IBM is coordinating work on American FactFinder (www.factfinder.census.gov), the Census Bureau Web site that publishes data for public consumption, to make it easy for the average citizen to search for and work with data on the Web site.

On the current version of American FactFinder, for instance, if citizens want to see all married households in the United States, they must first specify which of the five major surveys to search. “We’re going to make a significant leap forward when we go to the new dissemination system,” says Sisson. “We’re changing the main paradigm in terms of search and navigation.”

Rather than being survey-driven, the revamped site’s search function will be topic-driven. “If I want to find information on married households, I just enter that as a search term and it will bring up everything on married households, regardless of what survey the data is from. It will be a much more powerful tool for Joe Public,” adds Sisson.

The system will also incorporate more sophisticated search and navigation tools, as well as enhanced mapping and charting capabilities. The goal is to accommodate a variety of users with a wide range of skill levels.

The new American FactFinder Web site should be ready for the public by January 2011, says Sisson. And, by law, the data from the 2010 census has to be loaded into the system by March 31, 2011, and distributed to state governments—a challenge that the data-loading system improvements will help address. "That's all the data that states use to reapportion their congressional districts," he says. "We can't be late."

New election, new boss

Another challenge in the public sector is frequent (and practically guaranteed) leadership turnover. In state governments, for example, priorities can change with every four-year election cycle. "You can't assume that the next round of leadership is going to want to share and leverage data in the exact same way," notes DePriest.

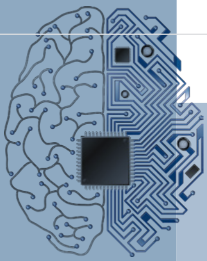
It happens in school districts, too. MCPSS had been trying to launch a data warehouse project for four years, says Akridge. It put out an RFP back in 2006 and IBM won that bid. Then the district got a new CIO who decided to switch to another company. That project was unsuccessful. Meanwhile, Akridge was appointed CIO and saw an opportunity to try again.

"We've been through so much with this," he says. "We went a long time being very disappointed." Akridge is thrilled that the project is now going so quickly. "IBM Cognos and DecisionEd have done in four weeks what other companies we've worked with couldn't do in a year. Through invaluable information insights delivered by IBM technology, we're moving closer to fulfilling our mission of graduating citizens who are prepared with the skills they need for the 21st century."

As electronic reporting requirements tighten and citizen demands for transparency and access—not to mention the number of data-gathering applications and systems—rise, warehousing, performance management, and BI technologies are likely to become central pieces of the public-sector data management puzzle. An organization may start using these technologies just to track and manage recovery funds or demonstrate compliance with government regulations, but ultimately it can use the systems to make decisions over the long term that improve productivity, reduce costs, and deliver better service. ✱

Tam Harbert is a Washington, D.C.–based journalist who covers technology, business, and public policy.

Stimulating intelligence



Four parts of the American Recovery and Reinvestment Act (ARRA) are likely to require BI, according to Ramon C. Barquin, president of the Barquin International consultancy and co-founder of The Data Warehousing Institute.

1. Upgrade general government IT. Significant mandates require funds be used specifically to develop, enhance, or modernize IT systems throughout the federal and state government system.

2. Develop health IT systems. "Business intelligence will, of course, be one of the pillars of health IT since the massive amounts of data from electronic health records will be the prime object of significant analysis," according to Barquin.

3. Drive education improvements. The law sets out specific goals and reforms, and requires grantees to measure and track progress toward those goals. "A boatload of metrics is mandated from grantees, whether they be states, local governments, contractors, or institutions," Barquin notes.

4. Require transparency and accountability reporting. The law requires all recipients to track how funds are used and how many jobs are created, for example. "Without a robust BI toolkit, recipients of ARRA funds will probably not be able to take full advantage of the funding and will not be able to comply with the reporting requirements," says Barquin.

U.S. Census Bureau

Project: Improve data tabulation and dissemination for the 2010 Census and other key Census Bureau surveys

Cost: US$89.5 million

Challenge: Expedite data tabulation assessment for five major surveys of the Census Bureau, including the 2010 Census, the American Community Survey, the economic census, annual economic surveys, and the population estimates program. Increase flexibility in analysis of the data and improve the usability of the information on the American FactFinder Web site.

Solution: IBM Global Business Services is integrating a variety of technologies including a pre-existing data warehouse, IBM WebSphere software, IBM Tivoli Workload Scheduler, Space-Time Research software, an ESRI mapping and charting engine, and Endeca search and navigation solutions.

6 Keys to Real-Time Analytics

By Leah MacMillan

Are you getting all you can out of your analytics initiatives?

The need for information in the 21st century continues to intensify—and shows no sign of abating. Today's decision makers need to make sense of a tremendous volume and variety of information, leading more enterprises to deploy analytics that not only help them sense and respond to key business issues, but also help them make predictions and act based on real-time data.

IBM Smart Analytics System: The combo platter

Released in September, the IBM Smart Analytics System is designed to minimize the time, expense, and technical skills requirements that hinder broad adoption of advanced analytics systems. Organizations of all sizes can use the platform to rapidly deploy and operate advanced analytics for solving complex business problems.

The system combines an analytics platform, trusted information platform, and system platform. The analytics platform provides cubing services, data mining, text analytics, intuitive business intelligence (BI) reporting, analysis, dashboards, and scorecards. The trusted information platform offers high-performance data warehouse management and storage optimization. The system platform provides scalable server and storage resources. The Smart Analytics System also includes installation services and a single point of support.

The new system requires a small amount of storage to do its work, saving both floor space and energy. It is designed to uncover insights and hidden relationships among massive amounts of data—not just structured information found in databases, but unstructured and incompatible data from such diverse sources as videos, e-mail, Web sites, podcasts, blogs, wikis, archival data, and more.

These intelligent data mining features, combined with speedy analytics and other business-critical capabilities, help make the Smart Analytics System a powerful warehouse-based option for developers facing increased demands for faster and more accurate information access.

Business analytics derive their value through the ability to extract specific and changing data from a wide variety of heterogeneous sources for smarter decision-making. Potential sources extend far beyond the classic IT portfolio of enterprise resource planning (ERP) transaction systems, databases, and data warehouses to include information from external sources, such as customer surveys, market research, and buyer behavior trends. Analytics applications transform this information in real time (or near real time) to deliver fresh insights.

For example, business analytics can help organizations monitor blood supplies, report on carbon footprints, or increase visibility across their supply chains. Retail outlets can determine the best end-cap product positioning based on customer preferences for Coke or Pepsi. Police services are using analytics to put crime information into the hands of patrol officers so they can quickly identify problems and associate trends and locations of crimes.

How can data management professionals ensure that their performance management and analytics initiatives are set up for success? Here are six best practices that can help you overcome the twin challenges of increasing user demands and more complex data sourcing requirements.

1. Cast a wide net

Making decisions and developing processes based on only part of the picture can negatively impact business performance. The first step, therefore, is to make sure that your analytics implementation has direct access to all relevant, available data no matter where it resides. The analytics system should also serve as the authoritative source for all historical and transactional data, so you can properly glean insights on trends and make decisions that will impact future performance. One-off dashboards, custom-developed programs, or stand-alone spreadsheets that don't connect back to the trusted pool of data are generally not reliable, sustainable, or scalable. Each solution adds its own layer of query and reporting complexity and introduces associated reconciliation and usability challenges.

Analytics solutions need a rich variety of information to yield meaningful insights. With so much data fragmented across any number of systems, you need a broad reach to ensure you can connect to any and all transactional systems, warehouses (relational and online analytical processing [OLAP]), flat files, and legacy systems, as well as XML, Java Database Connectivity (JDBC), Lightweight Directory Access Protocol (LDAP), and Web Services Description Language (WSDL) sources.

Casting a wide net helps you break down the data silos that hamper analysis and allows you to deliver a timely and complete enterprise view of relevant information. Plus, when new data sources become accessible, all analytics capabilities can access that data immediately.

2. Plan a caching strategy

Performance optimization is a critical part of fast reporting and interactive analysis. Switching between different back-end systems to access data is a familiar requirement, but it can seriously hinder performance if done on the fly.

Instead, create a caching strategy to both improve system performance and minimize any negative impact on the performance of source systems caused by repeated requests for data. Common techniques include enterprise information integration (EII); virtual caching; OLAP caching; caching to disk or local database; event-driven, scheduled, and manual refreshes; and advanced hybrid memory/disk utilization options.

3. Adopt a common, multilingual business model

Once the IT team has accessed and integrated the data needed to provide a complete view of the organization, modelers must convert it into information that is meaningful to business users. They must also ensure that the right information reaches the right users at the right time and is delivered in the right way.

The key to delivering this information in terms that business users understand is a common metadata business model that applies consistent business rules, dimensions, and calculations to all data regardless of its source. This makes it easier for a business to accurately report and analyze information such as sales invoices, general ledger charges, and order receipts.

A common business model provides the single view of the organization necessary for reliable, cross-enterprise reporting for all roles, locations, and languages. This approach not only supports a level of information consistency that leads to confident decisions, but reduces the cost of maintaining the modeling environment. It also reduces report proliferation by allowing a single report to be produced for all geographies.

4. Model once, package for many

Large data warehouses can overwhelm those trying to produce reports and analyses because there are simply too many data objects to choose from. Instead, build one model and publish sections of it that address the needs of different business users or communities. Whenever possible, create reusable objects and build multi-tier models that separate physical models from business models. This will decrease the downstream effect of changes and enable you to evolve your models more easily, as well as add or change data sources and sourcing strategies.

By publishing sections of a single common business model, you avoid the pitfalls of duplication and divergence. This strategy helps decrease model proliferation, supports consistency across the enterprise, and reduces the time required to deliver different models for different user groups—and it ensures that each user community receives only the specific information it requires.

5. Establish role-based security

Similarly, just because you have a single common business model fueling your analytics engine doesn't mean you want every user to see every analysis or report. Assign role-based user access to avoid the pain and expense of generating separate models or reports. The single model also restricts authorized users to only their view of the data, which may also help you comply with data governance and privacy regulations.

6. Develop models collaboratively

It's not easy to quickly build, deploy, and maintain an effective model, so organizations typically employ teams of data modelers to accomplish this task. To maximize productivity, craft processes and deploy tools that enable modeling teams to work collaboratively. For example, data modelers will need the ability to work on different parts of the model simultaneously, without jeopardizing one another's changes or creating "downstream ripples" before aggregating the segments into a single view.

Leah MacMillan is director of Business Intelligence and Performance Management Product Marketing at IBM.

RESOURCES

IBM Smart Analytics System: ibm.com/software/data/infosphere/smart-analytics-system

IBM Information Management: ibm.com/software/data

High-Performance Data Mining

Parallelized scoring performance with an SAS PMML model in InfoSphere Balanced Warehouse

By Jack Baker
and John B. Rollins, Ph.D., P.E.

When it comes to predictive analytics and business intelligence (BI), organizations with large information warehouses usually face a choice: create and implement data mining models directly within the database environment, or create them in a separate analytic environment, such as a data mining workbench.

Deploying mining models in a database environment can generate business results faster by improving scoring performance through parallelization and reducing overall software licensing costs. In addition, significant savings in software licensing may be realized by limiting high-cost data mining software to a small development environment and then porting the data mining models to the large-scale production environment. On the other hand, some organizations develop and deploy data mining models in an analytic environment, investing significant resources and experience in their mining models and process.

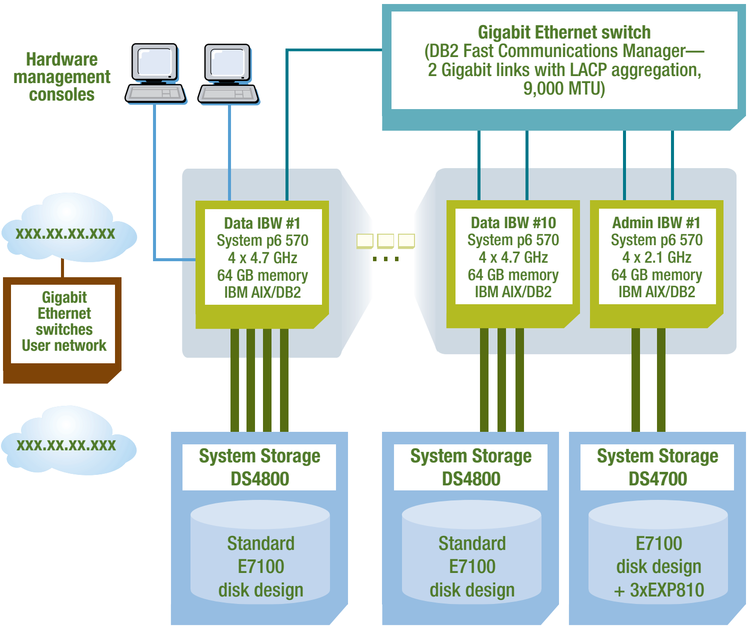


Figure 1: *The server environment for the test consisted of an InfoSphere Balanced Warehouse E7100 system*

For many organizations, the best choice may be a hybrid scenario that enables data mining models to be developed in an analytic environment and then deployed in a database environment optimized for high-speed, high-volume scoring processes. This approach is made possible by Predictive Model Markup Language (PMML), which defines a format for expressing data mining models. Data mining models that are created with PMML can be easily imported into a database, making them available to a scoring process within the database environment.

IBM InfoSphere Balanced Warehouse (IBW) is an excellent platform for organizations to deploy externally created PMML data mining models to create a high-speed, high-volume BI and predictive analytics environment. To demonstrate the capabilities of IBW, IBM conducted a scoring performance study with the following objectives:

- Demonstrate that a PMML data mining model can successfully be used for scoring in a high-speed, high-volume IBW environment.
- Assess the scaling performance of scoring in an IBW environment across a range of hardware configurations and data volumes.
- Develop best-practices recommendations for configuring an IBW data mining environment.

Setting up the test environment

The study consisted of four steps:

- Prepare the server environment.
- Establish a data mining model in IBW by obtaining a SAS logistic regression model in PMML format and importing it into the database.
- Build a set of five data tables in the database, with each table defined on four different partitioning schemes.
- Use SQL scripts to apply the data mining model to each of the five data tables for each of the four partitioning schemes and to report the execution time.

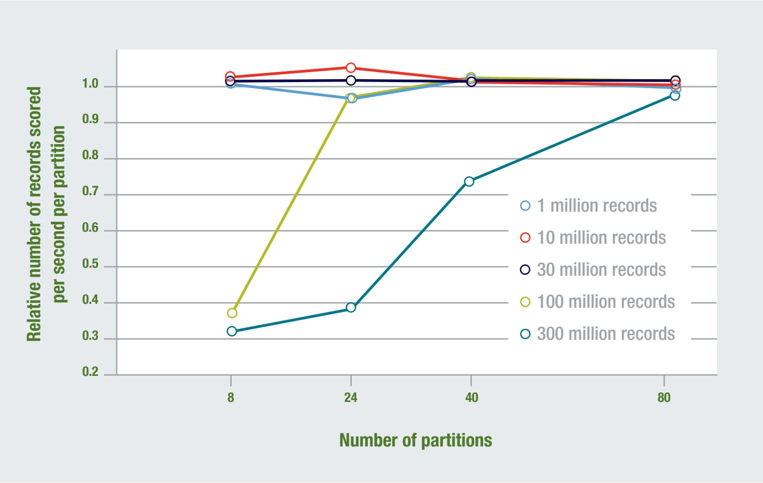


Figure 2: *Scoring performance: scoring rate per partition vs. number of partitions (relative to 1 million records on 8 partitions)*

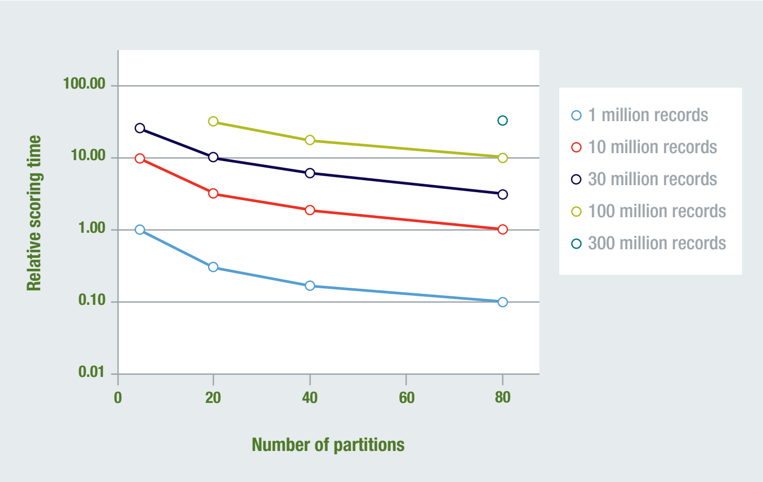


Figure 3: *Scoring performance: scoring time vs. number of partitions (relative to 1 million records on 8 partitions)*

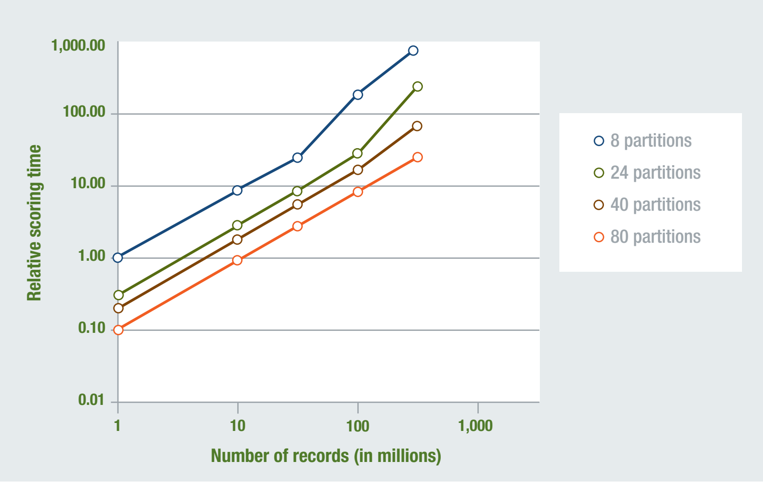


Figure 4: *Scoring performance: scoring time vs. number of records (relative to 1 million records on 8 partitions)*

Server environment

An IBW environment was configured on an IBW E7100 consisting of a cluster of 11 IBM System p6 570 servers running IBM AIX (see Figure 1). The E7100 cluster consisted of one administrator server and 10 data servers. Each server contained four processors and 64 GB of memory. The servers were connected over a Gigabit Ethernet switch. IBM System Storage DS4800 and DS4700 units were used for storage.

Data mining model

We obtained a logistic regression model created using SAS Enterprise Miner 5. This data mining model was exported from SAS Enterprise Miner in PMML format and then imported into a DB2 table in the IBW environment, making it available for incorporation into a DB2 scoring process.[1]

Data creation and partitioning

The data for the study was extracted from a very large database provided by an IBM customer. The extract contained two random samples of 1 million and 10 million account records, respectively. These two samples were used to construct five tables for the scoring runs, comprising 1 million rows, 10 million rows, 30 million rows, 100 million rows, and 300 million rows.

These five tables were created across four different partitioning schemes to demonstrate the "scaling up" (more records) and "scaling out" (more partitions) of data mining queries in an IBW environment. The partitioning schemes consisted of an administrative server and from 1 to 10 data servers, with each scheme having 8 partitions on each server. The four schemes were set up in the following configurations: 1 data server and 8 partitions, 3 data servers and 24 partitions, 5 data servers and 40 partitions, and 10 data servers and 80 partitions.

Scoring performance

We assessed scoring performance for each combination of partitions and rows. For each run, the DB2 buffer pools were warmed to a consistent state to help ensure that the recorded times would be consistent and comparable across runs. For each case, the number of records scored per second and the number of records scored per second per partition are reported relative to the base case of 1 million records and eight partitions. To view a table of scoring execution times and performance metrics, go to **ibm.com**/developerworks/data/library/dmmag/DMMag_2009_Issue3/IBW/index.html.

Figures 2–4 illustrate the scoring performance. In Figure 2, we see that the relative scoring performance measured as the number of records scored per second per partition relative to the base case (i.e., the relative scoring rate per partition) remains constant as the number of partitions increases. For the cases of 100 million records on 8 partitions and of 300 million records on 8, 24, and 40 partitions, performance was limited by available physical memory.

Investigation of this performance limitation showed it to be the result of buffer pool thrashing. In our tests, the source and target tables were in the same tablespace and buffer pool. Once the number of records became large enough, reads and writes began competing for the same buffer pool resource and thus slowed down the overall scoring rate.

In Figure 3, we see that changing the number of partitions changes the relative scoring time performance (number of records scored per second relative to the base case) by the same factor. For the case of 1 million records, we see that tripling the number of partitions from 8 to 24 reduces the relative scoring time threefold (from 1 to 0.3).

$$\text{Time (seconds)} = \frac{\text{\# of rows to process}}{(\text{\# of rows processed per second per partition}) \times (\text{\# of partitions})}$$

$$\text{\# of partitions} = \frac{\text{\# of rows to process}}{(\text{\# of rows processed per second per partition}) / (\text{Desired time in seconds})}$$

***Figure 5:** Formulae to calculate configuration metrics for a data mining model*

In Figure 4, we see that the relative scoring time increases linearly as the number of records increases. Deviations from linear performance in the cases of 8, 24, and 40 partitions reflect the limitation of available physical memory.

Findings and conclusions

The results of this study indicate that scoring performance using PMML data mining models in an IBW environment scales linearly with data volumes and hardware configuration. Furthermore, performance is constrained by available physical memory. Specifically:

- Performance is constant on a per-partition basis regardless of data volume and configuration size (see Figure 2).
- Changing the number of partitions changes the relative scoring time performance by the same factor (see Figure 3).
- Performance scales linearly with increasing data volumes (see Figure 4).
- Exhaustion of the available physical memory on the servers indicates a need to increase the total number of servers in the cluster.[2]

- The onset of buffer thrashing indicates a need either to add additional memory to each server or to add more servers to the cluster. Until this memory threshold is reached, performance remains linear.

IBW's capability to import a PMML data mining model means that analytic and IBW environments can be used synergistically to create and use data mining models for high-speed, high-volume scoring through operational business applications or automated processes. Organizations using this strategy can leverage their existing investment in analytic expertise and tools with an IBW environment to better support the decision-making process.

Best-practices recommendations

Our findings and conclusions lead to three best-practices recommendations for configuring an IBW data mining environment:

1. To eliminate logging overhead, set an output table for scoring results to Not Logged Initially.
2. To facilitate buffer pool tuning and to separate disk activity, source tables and scoring results tables should be placed in different tablespaces and different buffer pools.
3. To calculate configuration metrics for a particular data mining model, the formulae displayed in Figure 5 can be used, where the model's performance (number of rows processed per second per partition) has been determined by scoring a randomly selected subset of the data to be scored.

Jack Baker is a benchmark technical lead on the IBM Advanced Warehousing team. He works with customers to demonstrate InfoSphere Balanced Warehouse in their environment and consults with customers on database performance issues.

John B. Rollins, Ph.D., P.E. is an executive analytics architect on the IBM Advanced Warehousing team, where he serves as a senior technical leader for predictive analytics in support of customers, software sales, and integration of analytic technologies.

RESOURCES

InfoSphere Balanced Warehouse: ibm.com/software/data/infosphere/balanced-warehouse

PMML Version 3.0: www.dmg.org/pmml-v3-0.html

[1] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E., and Chodagam, J. *Dynamic Warehousing: Data Mining Made Easy.* IBM Redbook SG24-7418-00 (2007).

[2] In an IBW configuration having servers with standard preconfigured memory, the capacity-planning decision would focus on the number of servers, not on the amount of memory per server.

Sharing Knowledge, Driven by Passion

By Scott Bisang

Meet the newest IBM Information Champions from India: Raghuveer Babu and Pradeep Kumar

In his first year at the K.L.N. College of Information Technology, Raghuveer Babu undertook a project that required him to consider the total capacity of the human brain. Calculating how much data the brain can truly hold is not exactly the same task as calculating the storage capacity of a hardware system. But it did force Babu, then a student majoring in computer science and engineering, to think about how much data actually exists—and more important, how to quickly distinguish the meaningful data from less valuable data.

The project piqued Babu's interest in databases as tools to organize and efficiently access data. "Databases are everywhere—they play a vital role in every project," he explains. During their years at K.L.N., an IT and engineering college located in Madurai, about three hours from the Indian Ocean in Tamil Nadu, a state in the southern part of India, Babu and classmate Pradeep Kumar not only learned about databases, but also taught classmates about them—especially IBM DB2. Their volunteer work surrounding DB2 recently earned Babu and Kumar the designation of IBM Information Champions.

Advocates and educators

The Information Champion program recognizes dedicated IBM product advocates who share their opinions and years of experience with others in the same field through technical communities, books, Web sites and blogs. (See sidebar, "IBM Information Champions: Making valuable contributions.")

"It feels really great to be recognized," says Babu. "I'm very happy that IBM has such a wonderful program to recognize the efforts taken by people for communities all around the world."

IBM Information Champions Pradeep Kumar (left) and Raghuveer Babu helped hundreds of their peers build DB2 skills from scratch.

Babu and Kumar became interested in DB2 through The Great Mind Challenge (TGMC), a technical contest aimed at students in India and sponsored by IBM. The pair learned about DB2 while recruiting classmates (enough to create more than 80 teams) to enter the contest. As a result of their efforts—which included visiting every class to meet with students—K.L.N. won an award in 2008 for having the most participating students in all of India.

That dedication, along with a passion for working with databases, sparked Babu's and Kumar's drive to teach DB2 to other students. They started the KLNCIT DB2 User Group, the largest IBM university user group in Asia with more than 400 registered members. They set up social networks for the group, recruited members, and encouraged 120 fellow students to become certified on DB2. They have also worked with three separate colleges to plan a full developer conference (DB2 DevCon). And although Babu and Kumar have now finished their engineering courses and are currently working as developers for a private software company, they remain involved in DB2 education efforts.

"I find DB2 very easy and flexible to work with, while providing good security and performance compared to other databases," says Babu. "DB2 can also store hierarchical data, like XML, in a relational database model. In my opinion, DB2 makes projects easier."

Helping students and broadening horizons

The pair present regular sessions to the university user group on IBM DB2 Express-C, a full-function version of DB2 that can be downloaded and deployed at no charge, making it an ideal option for students who are learning about data servers. The duo's free training sessions include:

- An introductory program on databases and related technologies
- Comparisons of DB2 features with those from other databases
- An introduction to DB2 9.5 and its innovative features
- Preparation for DB2 9 Family Fundamentals certification (exam 730), including practice sessions
- Hands-on labs

"In our introductory class, we explain the importance of databases in the real world and how they work in IT infrastructures," says Kumar. "This grabs their attention and gets them excited about databases and related technologies. From there, the students are constantly approaching us, wanting to learn about new features."

Photograph by G.V. Gireesh

After their successful first attempt at recruiting students to become certified on DB2 earlier this year, Babu and Kumar plan to conduct two more certification training sessions in 2009, plus training sessions for students who originally registered in TGMC.

Both men were named DB2 Student Ambassadors in 2008, a designation they held during the rest of their time at K.L.N. A DB2 Student Ambassador is someone enrolled in a particular college who organizes events, advocates the use of DB2 in class projects, and arranges technical presentations by local IBM employees. Some ambassadors—including Babu and Kumar—also deliver technical presentations.

"Pradeep and Raghuveer have been the most active DB2 Student Ambassadors in India," says Raul Chong, IBM DB2 on Campus program manager. "They've not only organized several events, including the offering of certification exams and preparation classes, but they've also helped by recruiting and interviewing additional ambassadors in other universities in their state."

Developing skills for the real world

Student ambassadors help increase awareness of DB2, but they also provide significant benefits to their university peers.

IBM Information Champions: Making valuable contributions

Less than a year after its launch, the IBM Data Champion program has been expanded and renamed, becoming the IBM Information Champion program. The change was made to recognize the multitude of community members who contribute to the other segments of the IBM Information Management software division, such as Cognos, Enterprise Content Management, and InfoSphere, says Amit Patel, who leads the program for IBM.

An Information Champion is a dedicated individual who willingly shares his or her opinions and years of experience with IBM Information Management products. Information Champions often run user groups, manage community Web sites, write blogs, author technical journal articles, and speak at conferences, among many other volunteer activities. Currently, 88 individuals from 23 countries hold the Information Champion designation.

Information Champions receive a featured profile on the IBM Web site, invitations to events, and a plaque honoring their status, along with a virtual badge for e-mail signatures and community-site recognition, and special contacts within product development teams to facilitate communication with the IBM labs. They also receive additional technical resources and benefits, like early access to Information Management software releases.

"Individuals who give their time and share their expertise are the lifeblood of a software community, enriching the experience for all," says Paula Wiles Sigmon, program director for Information Management communities at IBM. "We are glad to extend the Information Champion program to recognize the varied and vital contributions of exceptional members across the entire Information Management community."

For example, at most universities, a single professor or dean will dictate what type of software is used, which can leave students unprepared to work in the real world with other database management systems. Babu and Kumar, along with many other student ambassadors, have worked to make sure students learn a wider variety of technologies, better preparing them for life after graduation.

Through their efforts, student ambassadors reap additional rewards, such as developing presentation and marketing skills that will be invaluable in their careers. They also become more visible to professors and potential employers. "I tell students that you can be the best programmer in the world, but if nobody knows who you are, you won't find a job," says Chong.

Babu and Kumar did enjoy those perks, but that's not what drives them to continue to spend their free time working with the user groups. "My interest in technology and desire to teach others motivate me do this work," says Kumar. Babu puts it even more simply: "Passion."

And that's what being an Information Champion really boils down to: passion. Because no matter what benefits you gain from writing a technical article or teaching others about software, you must be passionate about the topic to spend so much of your own free time on it.

"Raghuveer and Pradeep have demonstrated exemplary leadership in the community so early in their careers," says Amit Patel, team lead for IBM Information Management community programs and head of the IBM Information Champion program. "They voluntarily helped hundreds of their peers build DB2 skills from scratch. It is their passion for technology and dedication to the success of their peers that makes them Information Champions." ✱

__Scott Bisang__ is a marketing communications specialist at IBM. He previously worked as a journalist and has written for several newspapers and magazines.

RESOURCES

IBM Information Champion program: ibm.com/software/data/champion

DB2 on Campus and DB2 Student Ambassador Programs: ibm.com/software/data/db2/express/students_programs.html

DB2 on Campus Facebook group: www.facebook.com/group.php?gid= 3000790461

QUEST SOFTWARE
Administration & Space Management
Workload Analysis & Trending
Performance Diagnostics
SQL Tuning

System z Rocks (Again)

The mainframe isn't dead—in fact, it's turning heads as a data warehouse platform

Robert Catterall (rcatterall@catterallconsulting.com) is president of Catterall Consulting, a provider of DB2 consulting and training services.

When IBM announced DB2 in 1983, the product was mainframe-based and positioned as a new technology foundation for decision-support applications (the term "data warehouse" had not been coined yet). Organizations jumped on DB2, but often used it to run online transaction processing (OLTP) workloads. In response, IBM added lots of performance-enhancing features to DB2, and OLTP became the primary workload running on DB2 systems all over the world.

As time went by, the predominance of OLTP fed a perception that mainframes were not a great fit for business intelligence (BI) applications. Now, however, IBM System z and DB2 are a popular combination for BI. How is it that DB2 and the mainframe—the combination born for BI—are once more seen as a primo platform for data warehousing?

The mainframe market: Very healthy, thank you

The "mainframes are dead" pronouncements reached a crescendo in the late 1990s, and some folks continue to stick to that line. What a load of bunk. The market for IBM mainframe servers remains very robust (see sidebar, "Mainframes still in the mainstream").

In fact, the System z server line is thriving because it solves major challenges that almost every company is facing, starting with the drive toward "greener" computing. System z servers are highly space- and energy-efficient, delivering a lot of computing power per unit of floor space occupied and unit of energy consumed.

Another mainframe-boosting trend is the desire for continuous availability and consistently good performance of online applications, underscored by users' trust that sensitive data will be well protected. System z has long been the benchmark computing platform when it comes to availability, scalability, workload management, and security (advantages that are even more pronounced when several IBM z/OS systems work together in a shared-data, Parallel Sysplex cluster).

Just because mainframes are hot technology doesn't mean that z-based data warehousing should be on the rise—but that's what's happening. Why? For one thing, there is an increasing interest in getting more BI work done closer to the warehouse source data, and a great deal of that source data is on mainframe systems. Getting the data warehouse close to the data supply facilitates frequent, near-real-time updating of warehouse data, a huge plus for many organizations.

Further, many companies now see their data warehouse systems to be just as mission critical as their run-the-business OLTP applications. Tolerance for unscheduled outages is low, especially when failures trigger financial penalties built into service-level agreements. That often leads organizations to build their warehouse on a System z foundation.

DB2 for z/OS steps up

While DB2 debuted on the mainframe platform as a DBMS designed for decision support, DB2 for Linux, UNIX, and Windows (LUW) delivered BI-friendly features that were lacking in DB2 for z/OS. This gave some folks the impression that IBM was pushing distributed system servers as the DB2 platform of choice for data warehouse applications. In recent years, DB2 for z/OS got its BI groove back, thanks to enhancements in DB2 8 and 9, including:

- **64-bit addressing (DB2 8):** Data warehouse applications are often I/O-intensive, and it helps to be able to allocate really big buffer pools.
- **Materialized query tables (DB2 8):** By providing prebuilt intermediate result sets that would otherwise have to be materialized at query execution time (often related to aggregation functions or table join operations), MQTs can dramatically reduce query run times.
- **In-memory work files (DB2 8, extended in DB2 9):** When an intermediate result set is materialized at query execution time and is subsequently re-accessed by DB2, that re-access is accelerated through the use of in-memory work files.
- **Indexes on expressions (DB2 9):** BI queries are often complex and regularly involve predicates that contain column expressions. The ability to create indexes on column expressions can deliver orders-of-magnitude performance improvements.
- **Index compression (DB2 9):** In data warehouse environments, the disk space used for indexes can exceed the space used for tables. Index compression can help reduce index disk space consumption significantly.
- **Richer SQL:** Includes common table expressions and recursive SQL (DB2 9); INTERSECT and EXCEPT for result set comparison (DB2 9); online analytical processing (OLAP) functions such as RANK, DENSE_RANK, and ROW_NUMBER (DB2 9); TRUNCATE for quickly emptying data from a table (DB2 9).

Add some financial incentives

A BI application can be a particularly cost-effective mainframe workload, thanks to a couple of financial incentives provided by IBM:

- **zIIP engines:** System z Integrated Information Processors (zIIPs) are specialized mainframe CPUs that help lower the cost of computing: they cost less than general-purpose processors, and they do not factor into mainframe software pricing. DB2 query parallelism—a big performance booster for data warehouse queries—is a zIIP-eligible system activity. zIIPs can also handle some of the work associated with queries that get to DB2 by way of the Distributed Data Facility, using the Distributed Relational Database Architecture (DRDA) protocol (often through IBM DB2 Connect)—something that's very common in DB2-based data warehouse environments.
- **DB2 for z/OS Value Unit Edition pricing:** For certain types of application workloads (and data warehousing is one of the eligible types), DB2 for z/OS can be acquired for a one-time charge.

A green, super-scalable, super-available, and super-secure server platform. Advanced BI technology. Budget-friendly financial incentives. DB2 and System z offer a compelling solution for organizations seeking a rock-solid foundation on which to build mission-critical data warehouses. How to get from A to B, then? That's where IBM InfoSphere can help.

InfoSphere: Design, populate, accelerate

A data warehouse does your organization good when it is populated and providing actionable information to users. IBM InfoSphere Warehouse on System z provides an integrated toolset that helps you get there faster, particularly when the warehouse source data is managed by DB2 for z/OS. The toolset includes a Design Studio that helps to model data for OLAP access (including physical database design and data movement flows); the SQL Warehousing Tool (SQW), which delivers a SQL-based data movement and transformation capability; Cubing Services for optimizing multidimensional reporting and analysis, including a caching capability that can greatly improve performance for queries expressed in the industry-standard Multidimensional Expressions (MDX) query language and that supports popular end-user tools such as IBM Cognos 8 Business Intelligence and Microsoft Excel; and an administration console to manage the run-time environment.

So, get with the in crowd. The mainframe platform that runs your mission-critical OLTP applications is also a great choice for data warehousing. After all, BI is in the DNA of DB2 for z/OS. ✱

MAINFRAMES STILL IN THE MAINSTREAM

Far from being extinct, mainframe servers maintain a solid—and growing—enterprise presence.

- During the fourth quarter of 2008, IBM led the market for high-end servers (those costing more than US$250,000), with 63.5 percent of factory revenue share.[1] System z's share of this mainframe market has nearly doubled since 2000.[2]
- This growth is not just a matter of long-established mainframe applications getting bigger. Installed capacity of *new* workloads on IBM mainframes grew significantly in the first half of 2008 versus the same period in 2007.[3]
- System z growth extends beyond traditional large-scale-computing markets. IBM mainframe revenue in emerging markets such as Brazil, Singapore, Thailand, and the Philippines grew 21 percent in the first six months of 2008 compared to the first half of 2007, according to IBM.[3]

[1] "IBM Tops in Server Hardware in 2008." Feb. 25, 2009. **ibm.com**/press/us/en/pressrelease/26777.wss

[2] Robb, Drew. "Server Snapshots: IBM z10 EC." March 27, 2008. serverwatch.com/hreviews/article.php/3737101/server-snapshots-IBM-z10_EC.htm

[3] "Clients Across Major Industries, Mature and Emerging Markets Choose IBM Mainframes to Run Their Most Sophisticated Business Transactions." Nov. 11, 2008. **ibm.com**/press/us/en/pressrelease/25973.wss

RESOURCES

DB2 for z/OS: ibm.com/db2/zos

Data warehousing and business intelligence on System z:
ibm.com/software/data/businessintelligence/systemz

Customizing XML Storage in DB2

Tailor XML storage to the needs of your application

Matthias Nicola (www.matthiasnicola.de) is a senior engineer for DB2 pureXML at the IBM Silicon Valley Lab. He also works closely with customers and business partners, assisting them in the design, implementation, and optimization of XML solutions.

DB2 pureXML makes storing and querying XML data easy: just define a column of type XML and then insert or load XML documents into that column. But what if you need to go beyond the basics? With IBM DB2 for z/OS and DB2 for Linux, UNIX, and Windows (LUW), that's no problem. Let's review some best practices for when and how to customize your XML storage.

To get started, we'll use the XML document in Figure 1, which represents an order with an ID, date, customer ID, and multiple items. Note that items can vary in the elements that describe them, such as size or color. Let's assume that we need to manage many such documents in DB2.

```
<order OrderID="9001" OrderDate="2009-10-18">
  <customerID>26914</customerID>
  <item id="LK-486">
    <name>Magic Potion</name>
    <size>300ml</size>
    <price>19.99</price>
  </item>
  <item id="VF-145">
    <name>Crystal Ball, Deluxe</name>
    <color>crystal clear</color>
    <price>295.00</price>
  </item>
</order>
```

***Figure 1:** A sample XML document*

How to split and reassemble XML documents

The first tip in my article "15 best practices for pureXML performance in DB2" (see sidebar, "Resources") is that you should choose your document granularity wisely. Essentially, this means that the XML documents you store in DB2 should match the logical business objects of your application and the predominant granularity of access.

In our example, let's assume that orders are subject to frequent changes and that reading, adding, or deleting individual items within an order are the most critical operations that require optimal performance. In this case you could consider splitting order documents upon insert and storing each item as a separate document in a separate row. The advantage of this storage approach (compared to intact storage of the original order documents) is that it allows easier and faster manipulation of the stored data:

- You can *retrieve* an item with a single row read, and without extracting the item from an order document.
- You can *remove* an item from an order simply by deleting a row from the items table. Manipulating an entire order document is not required.
- You can *add* an item to an order by inserting another item that has the appropriate order ID, customer ID, date, and sequence number. Again, manipulating an order document is not required.

This ease of adding and removing items in an order is particularly valuable in DB2 9 for z/OS, which doesn't support inserting or deleting elements within an existing XML document.

Figure 2 shows a possible table definition and INSERT statement for splitting an order document in DB2 for z/OS and DB2 for LUW. Relational columns store the order ID, customer ID, order date, and a sequence number for the items. Item information is stored in XML format because items may have different elements and attributes (see Figure 3).

The INSERT statement contains a fullselect with an XMLTABLE function. This function extracts values from the incoming XML document for insertion into the columns of the items table, and it splits the incoming XML document and produces separate item documents.

The XMLTABLE function contains a parameter marker through which the application can pass an order document. With the XPath

```
CREATE TABLE items(ordID INTEGER, custID INTEGER,
                   odate DATE, seqNo INTEGER, item XML);

INSERT INTO items(ordID, custID, odate, seqno, item)
 SELECT T.ordID, T.custID, T.odate, T.seqno, XMLDOCUMENT( T.item)
 FROM
  XMLTABLE('$d/order/item' PASSING cast(? AS XML) "d"
   COLUMNS
    ordID        INTEGER    PATH     '../@OrderID',
    custID       INTEGER    PATH     '../customerID',
    odate        DATE       PATH     '../@OrderDate',
    seqNo        FOR ORDINALITY,
    item         XML        PATH     '.') AS T;
```

Figure 2: *Table and INSERT statement to store each item in a separate row*

```
ORDID  CUSTID  ODATE        SEQNO    ITEM
------- ------- ------------ -------- -----------------------------------
 9001  26914   10/18/2009          1  <item id="LK-486">
                                          <name>Magic Potion</name>
                                          <size>300ml</size>
                                          <price>19.99</price>
                                      </item>
 9001   26914  10/18/2009          2  <item id="VF-145">
                                          <name>Crystal Ball, Deluxe</name>
                                          <color>crystal clear</color>
                                           <price>295.00</price>
                                      </item>
2 record(s) selected.
```

Figure 3: *Contents of the items table after inserting the sample document*

expression $d/order/item, the XMLTABLE function produces one row for each item element in the input document, then extracts the order ID, customer ID, and order date. The special column definition FOR ORDINALITY numbers the generated rows. (For details on the XMLTABLE function, see “XMLTABLE by example, Part 1” in the Resources sidebar.) The XMLDOCUMENT function ensures that each item fragment can be inserted as a stand-alone XML document.

Figure 3 shows the data in the items table after our sample document has been inserted using the INSERT statement in Figure 2.

Figure 4 shows how you can reconstruct the original order document in Figure 1, if needed. The functions XMLELEMENT and XMLATTRIBUTES construct the top of the document with values from the relational columns of the items table. The function XMLAGG combines all items for a given order in the constructed document. Note that XMLAGG contains an optional ORDER BY clause on the column seqno.

```
SELECT XMLELEMENT(name "order",
          XMLATTRIBUTES(ordID AS "OrderID", odate as "OrderDate"),
          XMLELEMENT(name "customerID", custID),
          XMLAGG(item ORDER BY seqno) )
FROM items
WHERE ordID = 9001
GROUP BY ordID, odate, custID;
```

Figure 4: *Reassembling the original order document*

This ensures that the items appear in the same sequence as in the original document.

Use generated columns to your advantage

The new IBM DB2 pureXML features in DB2 9.7 for LUW allow you to use XML columns together with the Database Partitioning Feature (DPF), range-partitioned tables, and multidimensional clustering (MDC) tables. However, the partitioning or clustering keys must consist of relational columns. You just saw how to use INSERT and XMLTABLE to extract values from XML documents into relational columns. You can certainly use those relational columns to partition or cluster your table. If you prefer that your application uses simpler INSERT statements (such as INSERT INTO orders(order) VALUES(?);) and is unaware of the extraction, consider using a generated column.

DB2 9.7 supports XML parameters in user-defined functions (UDFs), enabling you to define generated columns that are automatically populated with values from inserted XML documents. Figure 5 shows a UDF that takes an XML document, such as the sample order in Figure 1, as input. This UDF uses the functions XMLCAST and XMLQUERY to extract the OrderDate attribute from the input document.

```
CREATE FUNCTION extractDate(doc XML)
   RETURNS DATE
   LANGUAGE SQL CONTAINS SQL
   NO EXTERNAL ACTION DETERMINISTIC
   RETURN XMLCAST(XMLQUERY('$d/order/@OrderDate'
            PASSING doc AS "d") AS DATE);
```

Figure 5: *Scalar UDF with parameter of type XML*

You can use this UDF in SELECT queries and other SQL statements, but also to define a generated column. For the following example, let's assume that inserting and retrieving complete orders are the most critical operations. In this case, storing the order documents intact is a good choice. Figure 6 defines a table that stores orders in an XML column and automatically extracts the order date into a generated relational column (odate). An INSERT statement can now simply insert an XML document into the order column and does not need to be concerned with extracting values into relational columns.

```
CREATE TABLE orders(
  order XML,
  odate DATE GENERATED ALWAYS AS (extractDate(order)));
```

Figure 6: *Table definition with a generated column*

If you store many orders on an ongoing basis, you may need to archive (roll off) old orders. This is best accomplished with range partitioning. The table in Figure 7 is partitioned by values in the column odate, which is generated from the XML column. Similarly, you can use generated columns as the distribution key in a partitioned database (with DPF) or as the clustering key for MDC tables.

```
CREATE TABLE order2(
   order XML,
   odate DATE GENERATED ALWAYS AS (extractDate(order)) NOT NULL)
 PARTITION BY RANGE (odate)
 (PART    q109    STARTING('01-01-2009')    ENDING ('03-31-2009') INCLUSIVE,
  PART    q209    ENDING ('06-30-2009')    INCLUSIVE,
  PART    q309    ENDING ('09-30-2009')    INCLUSIVE,
  PART    q409    ENDING ('12-31-2009')    INCLUSIVE);
```

***Figure 7:** Range-partitioned table based on a generated column*

Keep XML storage under control

Customizing your XML storage has various benefits. Splitting large XML documents into smaller documents can allow for simpler and more efficient manipulation of XML data. Using UDFs to define generated columns simplifies the extraction of XML values into relational columns. These columns then help you manage XML in partitioned databases, range-partitioned tables, or MDC tables. You can find other useful best practices for XML data in *The DB2 pureXML Cookbook* and the articles listed in the Resources sidebar. ✱

RESOURCES

15 best practices for pureXML performance in DB2: **ibm.com**/developerworks/db2/library/techarticle/dm-0610nicola

The DB2 pureXML Cookbook (IBM Press, 2009): ibm.com/developerworks/wikis/display/db2xml/DB2+pureXML+Cookbook

XMLTABLE by example, Part 1: ibm.com/developerworks/db2/library/techarticle/dm-0708nicola

Enhance business insight and scalability of XML data with new DB2 9.7 pureXML features: ibm.com/developerworks/data/library/techarticle/dm-0904db297purexml

Exploit XML indexes for XML query performance in DB2 9: **ibm.com**/developerworks/data/library/techarticle/dm-0611nicola

Updating XML in DB2 9.5: ibm.com/developerworks/data/library/techarticle/dm-0710nicola

DB2 pureXML wiki: ibm.com/developerworks/wikis/display/db2xml/Home

Lowering the Cost of Data

It's no secret that businesses can gain strategic advantages from turning data into insights faster than their competitors, but the exponential growth of data threatens any effort to reduce costs and lower the data center's environmental footprint.

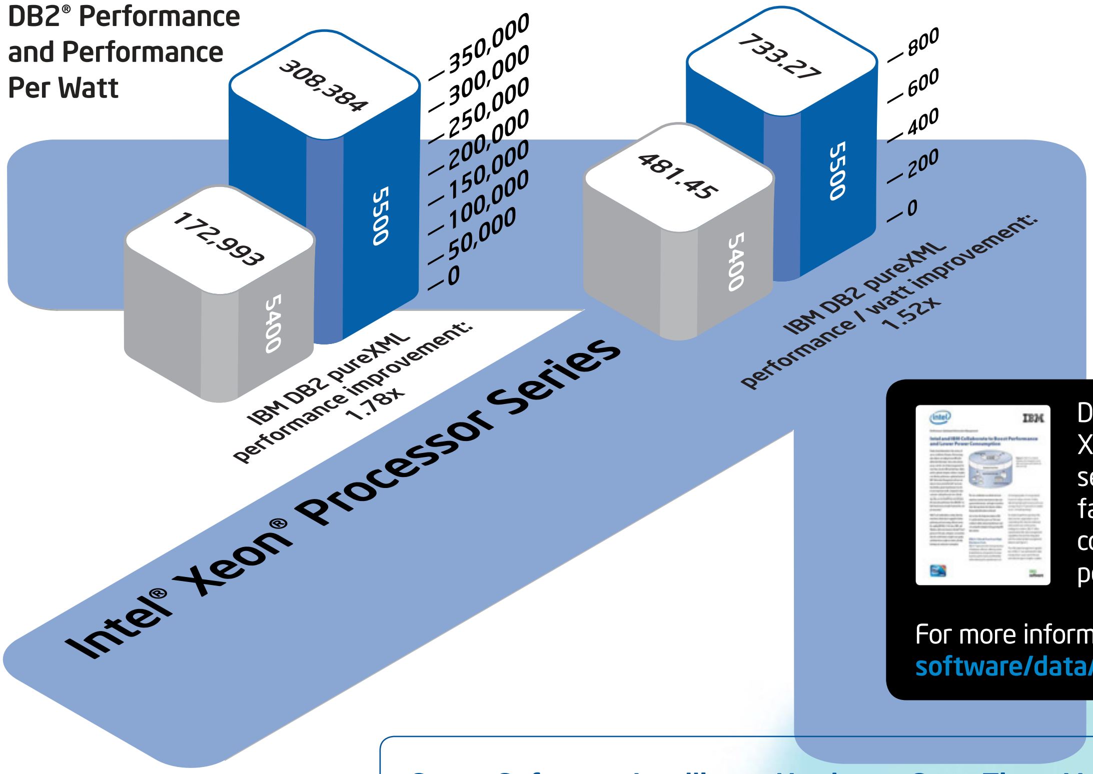
IBM is one of many leading companies helping customers optimize these trade-offs. IBM's next-generation database software, DB2® 9.7, offers sophisticated features designed to increase business performance and flexibility and reduce the operational costs of managing data. IBM's deep compression technology yields compression rates of up to 83 percent, producing a positive impact on storage-related costs.

DB2 is fully optimized for the Intel® Xeon® processor 5500 series and delivers 78 percent more performance and 52 percent better performance per watt than on the Intel Xeon processor 5400 series.[1] That's the largest single-generation improvement since IBM and Intel began collaborating in 1996 to optimize DB2 performance on Intel-based servers. It produces faster reports and responses at a lower cost and with a smaller environmental footprint.

IBM developers say it's easy to get the performance. "Not only can you achieve superb performance results by combining the DB2 product with the Intel® processor, but we were able to do that with an absolute minimum amount of tuning," said Berni Schiefer, distinguished engineer at IBM. "Through an out-of-the-box experience, anyone can achieve those results."

[1] Intel TPoX performance comparison between Intel® Xeon® processor 5570 versus Intel Xeon processor 5460 platforms

DB2® Performance and Performance Per Watt



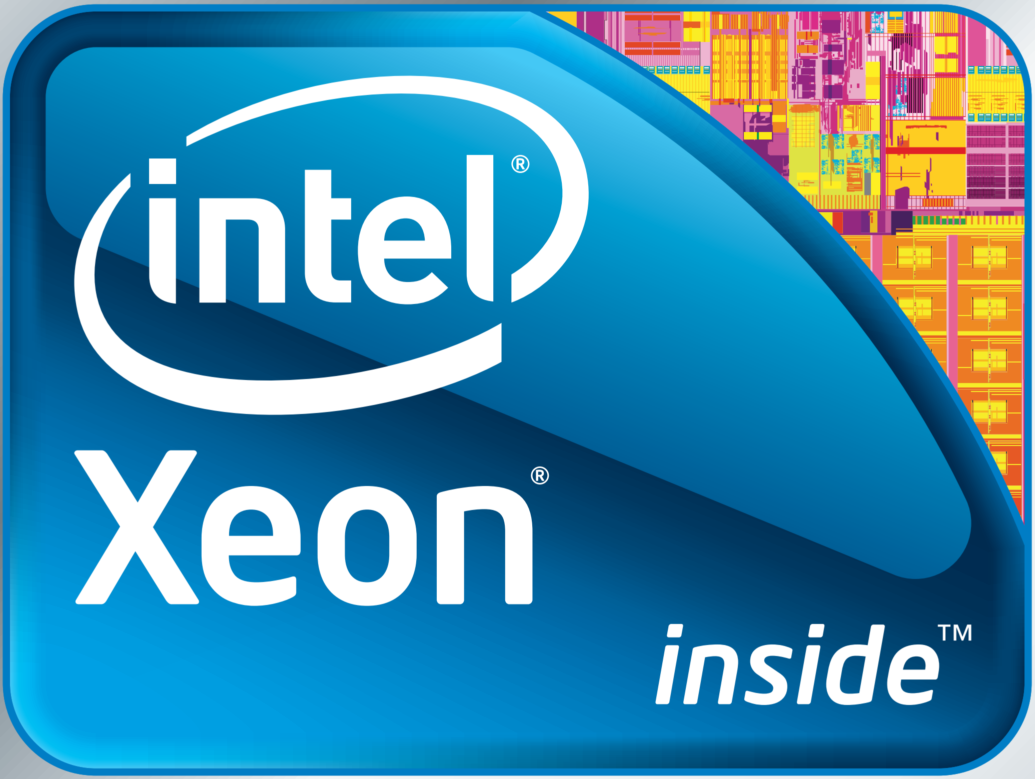
DB2® uses the Intel® Xeon® processor 5500 series to produce faster reports at lower cost—without elaborate performance tuning.

For more information, go to **ibm.com/software/data/db2/intel**

Smart Software, Intelligent Hardware Save Time, Money, Energy

The Intel® Xeon® processor 5500 series includes intelligent performance that can increase frequency on demanding workloads when conditions allow and turn off processors to save energy when they're not being used. IBM DB2® 9.7 incorporates intelligence that automates many time-consuming database administration tasks. For example, DB2 9.7's self-tuning memory manager allocates system memory for top performance depending on the type of workload. In a head-to-head comparison between DB2 9.7's self-tuning memory manager and some of IBM's best performance engineers, the self-tuning memory manager tuned the memory better than the performance engineers.

WhereScape's leaders are impressed with the combination of the two technologies. "When you consider what's going on now with Intel's intelligent performance and what IBM is up to with DB2 9.7, this is not business as usual," says Mark Budzinski, vice president and general manager for WhereScape USA, which builds data warehouses. "This is really game-changing technology."

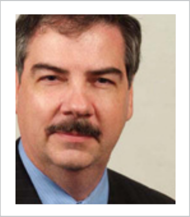


Powerful.
Intelligent.

Learn more ▶

Using Materialized Query Tables

Boost database response time for complex queries in IBM DB2 for Linux, UNIX, and Windows



Roger E. Sanders *(roger_e_sanders@yahoo.com) is a consultant corporate systems engineer at EMC Corporation. He is the author of 18 books on DB2 for Linux, UNIX, and Windows and teaches classes at many DB2 conferences. He is currently working on a new book that outlines how to write technical magazine articles and books and get them published.*

Special thanks to IBM Consulting Learning Specialist–Data Management Melanie Stopfer, IBM Senior DB2 Technical Evangelist Dwaine Snow, and IBM Senior Competitive Specialist Reed Meseck for providing information used to develop this article.

For database administrators, how well database applications perform is of the utmost importance. One way to significantly speed up the response time of decision support queries is to use materialized query tables (MQTs).

In this column, I'll explain what MQTs are and I'll show you how to create and populate a system-maintained and a user-maintained MQT. I'll also show you some situations where using MQTs can be beneficial.

What are MQTs?

The definition of an MQT is based upon the results of a query. Think of an MQT as a kind of materialized view, because the data for an MQT comes from one or more base tables. The difference lies in how MQT data is generated and where it is stored. Usually, data for a view is generated by executing the query upon which the view is based each time the view is referenced; the data resides in the underlying base tables that are referenced by the view. MQT data is generated by executing the query upon which the MQT is based at regular intervals or at a specific point in time (which you control); the data resides in the MQT itself. Like any other table, an MQT can have indexes, and the RUNSTATS utility can be used to generate and store statistics about MQTs.

MQTs provide a powerful way to improve response time for complex queries, especially queries that perform one or more of the following operations:

- Aggregate data over one or more dimensions
- Join and aggregate data over a group of tables
- Perform repeated calculations
- Perform resource-intensive scans
- Access a common subset of data—that is, retrieve data from a "hot" horizontal or vertical database partition
- Retrieve data from a table, or part of a table, in a partitioned database environment

Knowledge of MQTs is tightly integrated into the IBM DB2 SQL and XQuery compilers. During the query rewrite phase, the DB2 optimizer matches queries with existing MQTs and determines whether to substitute an MQT for a query that accesses base tables. (If an MQT is used, the Explain facility can provide information about which MQT was selected.) The larger the base tables, the greater the potential response time improvements, because an MQT grows more slowly than its underlying base tables.

Creating an MQT

MQTs are created by executing a special form of the CREATE TABLE SQL statement:

```
CREATE TABLE [TableName] AS
([SELECTStatement])
DATA INITIALLY DEFERRED
REFRESH [DEFERRED | IMMEDIATE]
[ENABLE | DISABLE] QUERY OPTIMIZATION
MAINTAINED BY [SYSTEM | USER]
```

where:

- *TableName* identifies the name to be assigned to the MQT you're creating
- *SELECTStatement* identifies the SQL query that is to be used to populate the MQT you're creating

Two different types of MQTs can be created: system-maintained MQTs and user-maintained MQTs. Insert, update, and delete operations cannot be performed against system-maintained MQTs. However, a REFRESH IMMEDIATE system-maintained MQT is updated automatically, as changes are made to all underlying tables upon which the MQT is based. The REFRESH keyword lets you control how the data in the MQT is to be maintained: REFRESH IMMEDIATE indicates that changes made to underlying tables are cascaded to the MQT as they happen, and REFRESH DEFERRED means that the data in the MQT will be refreshed only when the REFRESH TABLE statement is executed.

User-maintained MQTs allow insert, update, or delete operations to be executed against them and can be populated with import and load operations. However, they cannot be populated by executing the REFRESH TABLE statement, nor can they be created with the REFRESH IMMEDIATE option specified. Essentially, a user-defined MQT is a summary table that the DBA is responsible for populating, but one that the DB2 optimizer can utilize to improve query performance. (If you don't want the DB2 optimizer to utilize an MQT, simply specify the DISABLE QUERY OPTIMIZATION option with the CREATE TABLE statement used to construct the MQT.)

When to create MQTs

How do you decide if having an MQT would be beneficial, or determine which MQTs should exist? IBM DB2 Design Advisor can help. Using current database statistics, the DB2 optimizer, snapshot monitor information, and/or a specific query or set of SQL statements (known as a workload) that you provide, the Design Advisor recommends indexes, MQTs, or multidimensional clustering (MDC) tables that would improve performance. The indexes, MQTs, or MDC tables; the statistics derived for them; and the Data Definition Language (DDL) statements required to create them can be written to a user-created table named ADVISE_INDEX.

The Design Advisor is invoked by executing the db2advis command. A GUI version of the Design Advisor, known as the Design Advisor Wizard, is also available; activate it by selecting the appropriate action from the Databases menu in the DB2 Control Center.

MQTs and subdomains

While it is possible to define MQTs for each and every query predicate used (which, by the way, is *not* a good idea), a decision support and data warehouse environment often contains a very small set of common query subpredicates and qualifiers that are executed over and over. A powerful, yet regularly overlooked use of MQTs is to optimize access to frequently used subdomains of data that resolve such subpredicates.

In this scenario, MQTs do not contain summarized data, but instead are used to help the DB2 optimizer quickly identify and isolate qualifying rows; the MQTs act as a quick prequalification of rows that are involved in several more complex queries. For example, a reporting system might have a dozen or more reports that use the subdomain of data representing YESTERDAY. One report might look at yesterday's overall sales, another looks at yesterday's sales by region, while a third looks at yesterday's sales by product. Each of these three queries likely shares the common subpredicate WHERE DATE = YESTERDAY, which immediately limits the record set.

If you created three separate MQTs to answer each of these queries, they would require a significant amount of disk space to maintain. If, however, you constructed an MQT that simply delimits the domain, i.e., YESTERDAY; included attributes that are not likely to change, such as "Region" and "Product_ID"; and then created an appropriate index over this single MQT, all queries about YESTERDAY could be satisfied by this single MQT. Back JOINs and row fetches would most likely be required, but these operations would use a far smaller subset of the data.

The idea here is simple: make it as fast and easy as possible for DB2 to cut the data involved in multiple queries down to size without having to read several indexes, while also avoiding the creation and subsequent management of many similar MQTs. Find the common patterns in your queries, think about them in terms of the domains they express, and determine which attributes are used most often and which attributes will most quickly reduce the size of the data. Then, create an appropriate MQT for the queries being executed, create appropriate indexes on the base table(s) and the MQT, and keep the statistics current. Finally, let the DB2 optimizer choose whether to use the base table or the MQT; do not explicitly reference the MQT in the SQL.

REFRESH IMMEDIATE VERSUS REFRESH DEFERRED

REFRESH IMMEDIATE MQTs affect query performance in the same manner as indexes. Specifically, they:

- Speed up the performance of relevant SELECT statements while returning current data
- Are automatically chosen by the optimizer whenever it makes sense
- Can degrade the performance of insert, update, and delete operations
- Cannot be updated directly
- May occupy considerable disk space
- May have exclusive locks held during updates of their base tables

On the other hand, REFRESH DEFERRED MQTs have no effect on the performance of insert, update, and delete operations. As a rule of thumb, use REFRESH IMMEDIATE MQTs in moderation to optimize frequently run queries in which current data is important.

Reducing Conversations with DB2 for z/OS: Part 1

As Prem Mehra said, "There is no better-performing SQL than the SQL that is not executed."

Bonnie Baker
(bkbaker@bonniebaker.com) specializes in teaching on-site classes for corporations, agencies, and DB2 user groups. She is an IBM DB2 Gold Consultant, an IBM Information Champion, a five-time winner of the IDUG Best Speaker award, and a member of the IDUG Speakers' Hall of Fame. She is best known for her ability to demystify complex concepts through analogies and war stories.

Long ago, a learned colleague (Prem Mehra) introduced me to the concept of unnecessary SQL and his oft-quoted truism: "There is no better-performing SQL than the SQL that is not executed." In this three-part series, we are going to look at SQL that is totally unnecessary and should be eliminated, SQL that is executed far more times than necessary, and SQL that should be replaced with newer, better-performing SQL. The goal? To reduce connects to DB2 and, if possible, to eliminate some connections completely. The other goal? To learn about the latest solutions in IBM DB2 8 and DB2 9 for these old problems. So with that in mind, welcome to part 1 of my series on eliminating or reducing connects to DB2.

COUNT(*): Problem 1

There are many fundamental performance rules that should be obeyed when writing programs. One is to eliminate all unnecessary SELECT COUNTs in a program. Another is to reduce the number of executions of COUNTs if they cannot be totally eliminated. Let's look at a few examples:

```
Select count(*) into :hvcount
  from employee_master
Where status_flag = 'T'
  and hiredate <= current date - 90 days
```

Followed by:

```
Update employee_master
  Set status_flag = 'P'
Where status_flag = 'T'
  and hiredate <= current date - 90 days
```

Whenever I see SQL like this, my internal warning buzzer goes off. Why is there a COUNT before the UPDATE? I check further. Is the host variable ever examined? If it's 0, what is the action? What if the COUNT is greater than 0? Is the next action dependent upon the content of :hvcount?

Often the COUNT is totally unnecessary and can be completely eliminated. In this case, the programmer may want to know how many temporary (T) employees will be updated to permanent (P) status when/if the UPDATE statement is executed.

Two things are wrong with this approach. First, unless you are using ISOLATION RR or have exclusive maintenance access to the table, the COUNT of the number of qualified rows read and the actual number of rows that will be updated might differ. Rows could be deleted, inserted, or updated between the COUNT and your actual UPDATE.

However, the other—and far more important—problem is that the COUNT is unnecessary because DB2 counts the rows as they are updated and returns the count in SQLERRD(3) of the SQL Communications Area (SQLCA). As you test your programs, check to see if this SQLCA field contains the COUNT information that you need.

As far as I know, DB2 counts your rows as you do maintenance whenever it can. However, I do know of at least two exceptions:

1. When you DELETE all of the rows from a table (i.e., no WHERE clause) in a segmented tablespace
2. When you DELETE all of the rows from a table in a universal tablespace (new in DB2 9)

When doing a mass DELETE from a table that is in either of these two types of tablespaces, DB2 has no need to individually address the rows and therefore does not count them—with three exceptions:

1. When DATA CAPTURE is on
2. When a VALIDPROC is involved
3. When row-level security is being used

However, mass DELETEs are most commonly used for work tables or temporary tables, which are highly unlikely to employ any of these three exception categories.

A new feature of DB2 9 allows you to truncate a table, thereby "reinitializing" it (either by resetting the High-Used Relative Byte Address [HURBA] or by deleting and redefining the underlying VSAM data set—much like a LOAD with an empty INPUT data set). This, like the mass DELETE, does not count the rows for you.

COUNT(*): Problem 2

Let's look at another real-life example of an improper/inadvisable/inappropriate use of COUNT:

```
Select count(*) into :hvcount
  from employee_master
Where status_flag = 'T'
  and hiredate <= current date - 90 days
```

Followed by coded logic:

```
If :hvcount = 0, then get next, else if :hvcount =
1, then
```

```
Select col1, col2... into :hvcol1, :hvcol2
  from employee_master
where status_flag = 'T'
  and hiredate <= current date - 90 days
```

```
Else if :hvcount > 1, then
```

```
Declare mycursor cursor for
  Select col1, col2... into :hvcol1, :hvcol2...
    from employee_master
  where status_flag = 'T'
    and hiredate <= current date - 90 days
```

```
Open mycursor ...
```

```
Fetch mycursor into :hvcol1, :hvcol2...
```

Let me reassure you: I am not manufacturing these examples. Here the programmer has very honorable intentions: to use a lower-CPU, shorter path-length singleton Searched SELECT if only one row qualifies, and to use a higher-CPU, longer path-length CURSOR if and only if >1 row qualifies.

But reading the row(s) to count them and then (again) reading the row(s) to process them is *not* the answer. Double dipping is not good, even if the pages are already in the buffer pool. How many times have I heard that? Just because the pages may be in the buffer pool does not mean the second read is free. Now that DB2 is one serious, flying, crunching maintenance machine, eliminating connects is one of the few places we can see huge reductions in CPU and GET PAGE overhead.

So, what is a better approach for this problem? The programmer is trying to avoid a higher overhead cursor. Therefore, think: How often is there only one row? Ninety-eight percent of the time, you say? Then do the singleton SELECT first. Then, if and only if you receive an -811 ("multiple rows returned to a Singleton Select") error message, open a cursor. Conversely, if the majority of the time more than one row will be found, then open a cursor immediately without worrying about the small payback from a singleton SELECT. In other words, do what is most common first, and don't do the COUNT.

COUNT(*): Problem 3

Let's look at yet another real-life example:

```
Select count(*) into :hvcount
  from employee_master
Where status_flag = 'T'
  and hiredate <= current date - 90 days
```

This COUNT is followed by this logic: divide :hvcount by 10 (the number of rows that fit on a screen) to see how many screens of data there are. Why? So that each screen, including the initial one, can display "Page 1 of n" for the user.

Here we have an "uh-oh, can't be eliminated" requirement. But is there a better way to find out how many 10-row screens will qualify if our user actually, and maybe unrealistically, hits PF8 until there are no more rows/screens?

Yes. First of all, ask, "How many screens are there usually?" If the answer is one screen or less than a full screen, then don't count the rows before you read them. Instead, open the cursor, fetch the rows, and count the rows as you read them. If you reach an End of File before you hit row 11, you know your answer without connecting to do the COUNT. And you have eliminated all of the GET PAGEs (maybe READ I/Os) incurred when you made DB2 do the COUNT.

What if there are usually between 11 and 30 rows that qualify? Then read those rows, too, counting as you FETCH. If you hit +100 before or as you read row 31, then save those rows in some area that hangs around between screen displays (e.g., in CICS, the COMMAREA). Put "Page 1 of 3" on page one and eliminate the instructions that would have been needed for the PF8s to display "Page 2 of 3" and "Page 3 of 3".

If more than three pages of rows qualify, then store what the user will realistically look at in the COMMAREA (or wherever). Then, and only then, do the COUNT. You have deferred the COUNT until the last possible moment. You have avoided the double read in most situations. Most important, you have a good program with great logic that others can replicate to write their programs.

We get it. What's next?

For those of you who work with DB2 daily and know a thousand examples and exceptions to my suggestions, I thank you for being in the trenches of the "It depends!" world of reality—and for sharing your insights and wisdom.

For the rest of you, the next column will include more examples of unnecessary SQL and the whys, along with better and newer solutions for replacing the common code that is out there. Stay tuned for part 2.

RESOURCES

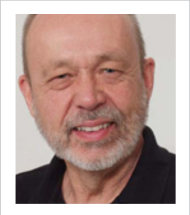
DB2 for z/OS: ibm.com/db2/zos

The Mystery of DB2 Sorts:
www.dbmag.intelligententerprise.com/story/showArticle.jhtml?articleID=211300275

Programmers Only archive:
www.dbmag.intelligententerprise.com/columns/programmersonly.jhtml

Fastest Informix DBA: How Did They Do It?

Here are the techniques that worked for the Fastest Informix DBA contest winners



Lester Knutsen (lester@advancedatatools.com) is president of Advanced DataTools Corporation, an IBM Informix consulting and training partner specializing in data warehouse development, database design, performance tuning, and Informix training and support. He is president of the Washington, D.C. Area Informix User Group, a founding member of IIUG, an IBM Gold Consultant, and an IBM Data Champion.

At the IIUG Informix Conference in April, we ran a Fastest Informix DBA contest. I took a simple customer billing process and added some bad SQL and a default ONCONFIG file with some bad configuration options—recreating the sort of challenges we see every day. The unchanged benchmark took about 30 minutes to run, and I challenged participants to make it run faster. The fastest DBAs tuned it to run in fewer than 4 minutes.

In my last column, I talked about the challenge and listed the winners; this time, we'll look at what worked and how they did it.

First, study the problem

The DBAs who succeeded spent as much time as they could studying the problem. We had a document that described the benchmark, including all the code and the expected results. We also had a video where I described the problem they needed to solve, and I showed them how to run the benchmark. Those who spent more time studying all the material did better, because they were able to better focus their efforts. For example, I purposely designed the schema so that there was a very high buffer turnover rate, and the fastest DBAs looked at the data and the schema and figured this out. Also, the benchmark system had only one disk, so a lot of tuning to take advantage of Informix parallel disk reads and writes would be unlikely to help much. The first thing that all the successful DBAs did was study the problem, analyze the facts, and then come up with a plan.

Informix configuration: ONCONFIG file changes

One challenge involved deciding what changes to make to the Informix Dynamic Server (IDS) configuration file, the ONCONFIG file. These changes were specifically targeted for the benchmark environment and may not help in all situations, but they do give you a good idea of what to look for in your own ONCONFIG file.

BUFFERPOOL—All of the fastest DBAs increased the number of buffers that the server used. The server had 3 GB of RAM. The fastest DBA used almost half of the RAM for buffers, created a dbspace with a 16 KB page size (larger than the default 2 KB page size), and allocated one-third of the memory for buffers for this 16 KB page size. This solved the problem of the record length being too large to fit on a default 2 KB page size, kept the records together, and put the most number of records in memory. Overall, I suspect that BUFFERPOOL adjustments made the biggest difference in performance.

The fastest DBAs were also careful not to make the BUFFERPOOL too large, as this would have caused the OS to start swapping to disk and would have slowed down the entire system. When you add buffers, you also need to consider the number of Least Recently Used (LRU) queues that manage all these additional pages in the BUFFERPOOL. The fastest DBAs increased the LRU queues to more effectively handle the additional memory.

SHMVIRTSIZE—This is the amount of memory Informix will allocate to workspace and virtual memory. All of the faster DBAs increased it, and the fastest increased this the most. There were a couple of very large "group by" statements in the SQL, and increasing virtual memory along with some other changes to the Parallel Database Query (PDQ) process allowed more of this work to take place in memory.

DS_TOTAL_MEMORY—This is the amount of memory from the SHMVIRTSIZE memory that will be used for PDQ operations. The default is very small, and increasing this may have helped with sorts and index builds.

DS_NONPDQ_QUERY_MEM—This is the amount of memory allocated to sorting when a sort is performed that does not use PDQ. With only one disk drive on the benchmark system, not much could be done with PDQ. Increasing this parameter helped with sorts and index builds.

LOCKS—This is the number of LOCKS and memory available for these LOCKS. If this configuration parameter is not big enough, Informix will dynamically increase it, but increasing it on the fly is very inefficient. The fastest DBAs set the number of LOCKS so the server did not need to dynamically increase this parameter.

RESIDENT—Setting this keeps Informix in memory and tells the OS not to swap the database server out to disk. All of the faster DBAs set this to keep Informix in memory.

CPU VPs—The benchmark machine was a four-core machine and could support four Informix CPU virtual processors (VPs). All of the faster DBAs set the number of CPU VPs to between three and four to take advantage of all CPUs on the machine.

DBSPACETEMP—I created a temp dbspace in the base configuration but did not define it in the ONCONFIG file, so it was not used. Instead, the rootdbs was used for sorts and temp files. Again, all of the faster DBAs changed the ONCONFIG file to identify and define this parameter. Several even created two or three additional temporary dbspaces, so Informix could read and write to tempdbs in parallel.

Some of the fastest contest participants changed other ONCONFIG parameters, including PHYSBUFF and LOGBUFF, DIRECT_IO, VP_MEMORY_CACHE_KB, and CLEANERS. It's hard to know exactly how much these contributed, but these are the places where the very fastest DBAs found extra speed. I was also interested in which parameters did not get changed. It may have been because of time, but no one changed the read-ahead parameters RA_PAGES and RA_THRESHOLD, and no one changed the index-cleaning parameters or changed the BTSCANNER.

Informix schema changes

I purposely designed the database with two tables that had very large columns, a CHAR(2000) column in the customer table and a CHAR(1000) column in the bills table. However, most of this was wasted space. In the customer table, only about the first 100 characters were used, and in the bills table this field was never used. Not only did this waste space, but it caused the tables to overflow a 2 KB page and created most of the buffer thrashing and very high buffer turnover rates. There are a couple of solutions to this, one of which is to alter the table and turn these columns into LVARCHAR columns. This change reduced the number of buffers that were read and written during the benchmark and may have had one of the biggest impacts on overall performance.

Another schema change that a few DBAs made was to move the index creation on the bills table to after all the data was inserted, instead of before that data was inserted. This made for a faster load of the table without an index, and when the index was built, it was more compact and optimized. Also, in IDS 11.50, building an index performs an automatic UPDATE STATISTICS HIGH on the table, which would provide the Informix query optimizer with better information about the table. A few of the DBAs added additional indexes on the customer table, which may have helped the performance of the last query in the benchmark.

SQL optimization

The benchmark process comprised two INSERT statements into a bills table and three UPDATE statements. The UPDATE statements contained subqueries. At the end of the benchmark process, two SELECT statements with group by clauses were executed to check the numbers. The resulting numbers from the last two statements had to match the expected results; this is how we verified that the benchmark was completed and correct. I added redundant and unnecessary code to the SQL statements to make it challenging.

With some careful planning, I think the whole process could have been done as one INSERT statement and one or two UPDATE statements, but no one managed that. However, several of the faster DBAs did identify the unnecessary code in the SQL statements and removed that code from the benchmark process.

More to come at IOD 2009

The contest was a lot of fun to run and monitor, and it is exciting to see the ingenuity and creativity that all the Informix DBAs who participated put into it. Congratulations to the winners, who were announced in the last issue (**ibm.com**/developerworks/data/dmmag/archive.html) and are listed on our Web site at www.advancedatatools.com/Informix/index.html.

We sponsored another version of this contest, the Fastest Informix DBA Contest II, from June 18 to September 30, 2009. At the IBM Information On Demand 2009 Global Conference in October we will hold a Webcast and a Birds of a Feather session on the contest—visit the Advanced Data Tools Web site (above) for more details. Hope to see you there. ✱

Smarter is...

Boosting the IQ of Galway Bay

Marine Institute Ireland monitors conditions in real time with IBM InfoSphere Streams software

Chris Young *is a technology writer based in the Pacific Northwest.*

Is water quality changing? Are fish stocks dwindling? Are there enough waves in the bay to generate electricity?

Ireland's Galway Bay, like many waterways around the world, faces pressing questions about pollution, flooding, fish populations, green energy generation, and climate change threats. Gathering and processing environmental information is the key to answering these queries, and the Marine Institute Ireland (MII) SmartBay information project is netting impressive results with the help of IBM InfoSphere Streams software.

Hazards like a pollution spill can cause damage more quickly in Galway's confined waters than in the open sea, so scientists and environmental agencies need to react to any signs of distress without delay. That takes real-time data—and lots of it. "Monitoring water quality and marine life in the bay requires frequent sampling to stay ahead of problems," says Dr. Harry Kolar, IBM chief IT architect for the SmartBay project. "You also need a way to rapidly analyze and deliver that information where it's needed."

InfoSphere Streams (**ibm.com**/software/data/infosphere/streams) is proving to be just the right solution. MII collaborated with IBM to implement the SmartBay middleware and user interface, and IBM is now tackling the unique analytics issues presented by marine environments. "In traditional processing, systems often run queries against static data," explains Kolar. "InfoSphere Streams is designed to continuously evaluate ever-changing data like that found in Galway Bay. Information provided can range from how much wave energy is present for sustainable power generation, to whether chemical levels are increasing."

The data doesn't have to go into a back-end system for analysis, saving both time and money. "InfoSphere Streams allows us to perform real-time analysis where the data is acquired," says Kolar. "For example, we're experimenting with sensors and processors deployed together on buoys out in the bay. The buoys carry sensor arrays that measure over 35 different parameters, from wind to water chemistry, that are vital for reporting and predicting weather, sea conditions, and water quality." Agencies can acquire information directly from the waterborne buoys, eliminating the delay of sending data to a shoreside data center.

Meanwhile, MII and IBM are working on the project's next phase: the buoys will be a beta test bed for distributed platforms capable of performing analytics and intelligently responding to changes. "If a Streams node sees something shifting from from a baseline, it will have enough intelligence to increase the sampling frequency and alert key people," says Kolar.

The successes at MII are opening up possibilities for similar systems at other sites. MII, for one, hopes to use lessons learned from SmartBay to extend the monitoring systems over Ireland's continental shelf and down to the abyssal ocean plain more than two miles beneath the ocean surface. "SmartBay shows we can meet many environmental challenges using large-scale data collection and distributed intelligence," Kolar says. "With the help of technologies like IBM InfoSphere Streams, it's already becoming a smarter planet."

TDA GROUP
800 El Camino Real, Suite 380
Mountain View, CA 94040