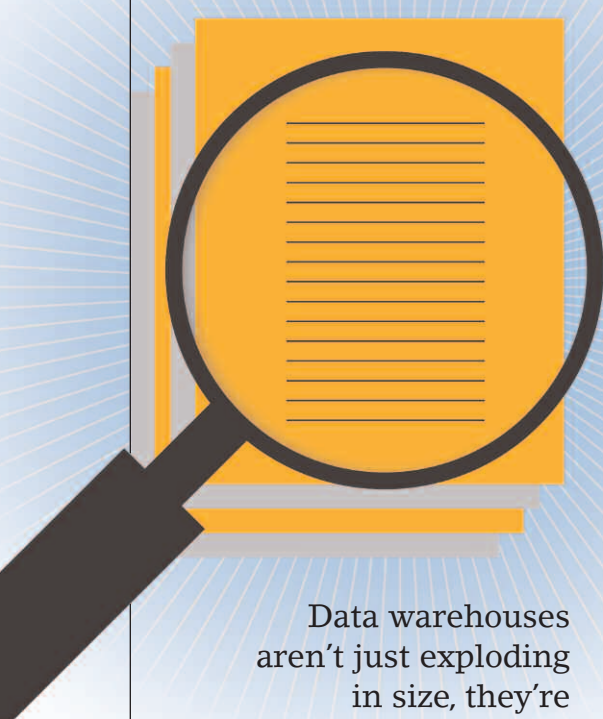


Data Strategy

Scaling The Data Warehouse



Data warehouses aren't just exploding in size, they're also supporting more users and increasingly complex queries, all in shorter time frames. Here's how to make sure yours is ready to scale.

By Richard Winter

LGR TELECOMMUNICATIONS has a 310-TB Oracle data warehouse that's used daily by 2,500 people at one of its telecom carrier clients. The warehouse powers an LGR service, called CDRlive, that gives its carrier customers access to call data records. It's updated round the clock, in near-real time, and is available for query 24 hours a day, 365 days a year.

"There are no batch jobs," says Hannes van Rooyen, chief architect at LGR, which supplies data warehouse software and services to the telecom industry. "Instead, as many as 13 billion records a day are added, and an equal number are dropped in an online update process that runs concurrently with user queries."

The data warehouse keeps more than a petabyte of disks spinning and has grown by a factor of 10 during the last four years. It's expected to at least double in the coming year.

Most companies still don't hold hundreds of terabytes of data, but they're up against the same data warehouse problems that face LGR—soaring data volume, more users, complicated queries, and fast-changing information. Throw in a growing number of vendor options and it's time for companies to re-evaluate their data warehouse strategies.

The new generation of data warehouses looks a lot like LGR's: growing at an extraordinary pace, in multiple dimensions, and supporting critical business processes that must react quickly to

ALSO IN THIS REPORT

Microsoft And Oracle Join The Scale-Out Crowd	3
EBay Turns Data Marts Into A Service .4	
Seven Gotchas Of Scalability	5
Teradata's Peta-Scale Appliance	8
A Closer Look At Teradata's Offerings .9	
Enterprise Search: Microsoft, Google, others Vie For Supremacy	10

events around the company. Whether your company has 250 GB or 250 TB of data, you're likely facing the same questions: Do we have the right architecture? Is it on the right platform? Is the warehouse about to run out of headroom? What will it take to service new users? How do we move from batch loading to continuous update? And with technology changing so rapidly, how do we know we're on the right system?

All the answers loop back to managing scalability. Getting control of scalability might mean embracing the highly parallel processing and scale-out architectures long offered by Teradata and IBM and elements of which are now emerging in new products from Oracle and Microsoft (see story, p. 3). Or it might just require more effective management of existing data warehouse practices, including quantifying requirements, measuring alternative solutions, and acting earlier on potential problems.

MULTIPLE DIMENSIONS OF SCALABILITY

The convergence of three key trends is driving the ever-expanding scalability challenges facing data warehouse managers. The first is well known: Data volumes are increasing rapidly. The largest data warehouses are tripling every two years, according to WinterCorp surveys.

That's about how fast LGR's data warehouse is growing; it will approach 3 PB in 2012. Hundreds of other data warehouses, including those run by retail, health care, and financial services companies, also will reach petabyte scale in the next few years and thousands will surpass 100 TB. In many cases, competitive pressures are driving businesses to capture more data in hopes they can better analyze, understand, acquire, and retain the most valuable customers.

Data warehouses also are getting more time

sensitive. The extraordinary velocity of the data in LGR's warehouse is a case in point: Billions of records pour in throughout the day, loaded into the database within minutes, and acted on almost immediately. If a mobile phone customer calls in because he had a bad experience, "we want to see exactly what happened, what calls were dropped, what tower was involved, and so forth, while they're still on the phone," van Rooyen says. "At the same time, you want the customer service person to know the customer's history." Problems are resolved faster, customers get better service, and "the business works better all around," he says.

High-velocity use of data—also called "operational business intelligence"—isn't a new concept. Teradata identified it several years ago as "tactical data warehousing," and IBM's Dynamic Data Warehousing pushes a similar notion of "right time" data. But the business pressure to provide such capability is rising.

Tactical data warehousing facilitates the moment-by-moment decisions employees must make. Many of these decisions are similar and repetitive: What should I offer this customer?



How do I treat this unexpected shipment that just turned up? Businesses that can make such decisions in a systematic way, informed by up-to-the-minute data, find they produce significantly better results.

Operational BI has big implications for data warehouse scalability. It results in larger user populations; more frequent, time-sensitive interactions; a need for fresher data; and support of business processes that can't tolerate downtime.

The third trend is rising complexity in data, queries, workloads, and analysis, all of which amplifies scale. When data warehouses are doing only simple things, such as predictable updating and straightforward reporting, they

can grow without creating fundamentally new problems. But when they have to respond inter-actively to complex and unpredictable queries—perhaps performing large, complex joins, aggregations, sorts, and calculations on trillions of records—the requirements have truly escalated.

Many modern data warehouses perform complex queries, analyses, and reports. They also operate on more complex schemas than in the past, with thousands of tables, hundreds of thousands of columns, and a complex web of data relationships.

EXTREME MULTIDIMENSIONAL GROWTH

There are few better illustrations of the multidimensional growth phenomenon than eBay. About 85% of the queries run on the company's

Microsoft, Oracle Join The Scale-Out Crowd

ORACLE AND MICRO-soft, in a bid to land mid-size data warehouse customers, are pitching new products aimed at a "scale-out" option—running large data warehouses on clusters of small, low-cost servers.

By bringing products for highly parallel architectures to midmarket users, two of the largest database vendors are acknowledging there are multiple dimensions to scalability. Companies can have as little as a terabyte of data but use complex queries or schemas, or have lots of people accessing the data. Such users often find they need a scale-out architecture.

Oracle last month rolled out the HP Oracle Exadata Storage Server and the HP Oracle Database Machine, both designed to raise performance for data warehouse queries. Oracle's products use the Exadata storage cell as a building block, relying on low-cost Hewlett-Packard hardware and intelligent Oracle software to off-load database processing to the stor-

age tier and increase disk I/O bandwidth. The performance version of an Exadata storage cell will store 1 TB of user data and deliver 1 GBps of raw I/O bandwidth.

The effective bandwidth in processing a query can actually be much greater than 1 GBps per cell because of compression and database operations, such as filtering and projections, performed within the storage cell. This lets Oracle data warehouses offer significantly higher performance, while requiring less space, power, and cooling; they also cost less compared with conventional storage arrays.

MICROSOFT BUYS IN

At Microsoft's Business Intelligence Conference last week, the company said it will integrate the technology it acquired as part of its purchase of data warehouse appliance vendor DATAlegro with Microsoft SQL Server. The first products are expected in 2010.

Before this move, Microsoft focused on growing data warehouses via scaling up; customers would buy larger SMP

servers when they needed a bigger warehouse. This approach has advantages in operational simplicity, but it imposes a ceiling on capacity. Microsoft still touts the scale-up option, but the DATAlegro technology adds a scale-out option.

The scale-out approach isn't new to large-scale data warehousing. Teradata has used it since 1984, IBM since the mid-'90s, and Oracle for nearly 10 years with RAC and now grid computing. HP Neoview and many data warehouse appliance startups emerging this decade are using it, too. In addition to reducing hardware costs, a good scale-out architecture promises modular capacity and potentially little or no disruption for upgrades.

With these latest announcements, Oracle and Microsoft are aiming to capture larger data warehouse deployments. Of course, like all highly parallel architectures, theirs will have limitations and bottlenecks. And they'll have to prove they're as good as those from vendors that started out with a highly parallel approach. —RICHARD WINTER

data warehouse are “exploratory in nature,” says Oliver Ratzesberger, eBay’s senior director of architecture and operations. They come from end users, with no opportunity for a database administrator to apply a tuning tool to them. “The queries hit the engine, and it has to handle them,” Ratzesberger says.

eBay’s data warehouse contains about 5 PB of disk storage distributed over primary and secondary systems, both running Teradata. The secondary system for disaster recovery is located about 1,000 miles from the primary one. Each system has a complete copy of the company’s core data, organized as an enterprise data warehouse. Both copies are updated every 15 minutes, round the clock, and are continuously active servicing queries.

There are more than 5,000 users and about 10 million queries each day. The daily update volume ranges from 10 billion to 15 billion records per day. Thousands of tables are involved, and queries range from simple lookups to

complex analyses that run for hours. The system is constantly managing a mixed workload with different service-level objectives for each of the various classes of work.

Given the scale of the system, the growth rates are even more remarkable: The number of users grew 25% last year, the number of queries doubled, and the size of the system has at least doubled each of the last four years.

eBay’s experience shows how data warehouses don’t just grow in quantity of stored data. They also expand in several dimensions at once, including data volume, number of users, query volume, data latency, and data and query complexity. Decisions on architecture and spending must take into account the likely growth of all these dimensions.

FIVE-STEP PROGRAM

To be clear, don’t try to preach “multiple dimensions of growth” to business unit managers. They see scalability as simply the ability to buy

eBay Turns Data Marts Into A Service

E BAY HAS INSTITUTED a utility computing model to better manage the growth of its data warehouse that Oliver Ratzesberger, senior director of architecture and operations, refers to as “analytics as a service.”

This service lets authorized eBay employees access a virtual slice of the main data warehouse server where they can store and analyze their own data sets—either in isolation or in combination with core data in the enterprise data warehouse. eBay’s virtual private data marts have been quite successful—hundreds have been created, with 50 to 100 in operation at any one time.

They’ve eliminated the company’s need for new physical data marts that cost an estimated \$1 million apiece and require the full-time attention of several skilled employees to provision.

Virtual marts are often used only for a few days or weeks, so system resources are quickly reclaimed. Users typically introduce less than a terabyte of new data, which they often want to analyze in conjunction with the data in the enterprise data warehouse. If these projects were implemented as separate physical data marts, the required core data would probably be extracted to the data mart, swelling its size, requiring a way to keep replicated data

up to date, and multiplying cost and complexity in other ways.

eBay’s analytics as a service is a way for people to do “agile prototyping,” Ratzesberger says. “They can do experiments quickly and succeed or fail quickly and inexpensively.” This helps the company move faster to find and exploit opportunities in connection with Web site optimization, fraud detection, and revenue generation.

When an analytic environment is needed for more than 90 days, the data warehouse team explores whether the user’s data ought to be incorporated into the enterprise model.

—RICHARD WINTER

systems—including data warehouses—without unusual worries about growth. They expect data warehouse growth won't cause a disproportionate increase in costs, unreasonable disruptions in business activities, or big hits to performance. Oh, and never run out of headroom.

Sound daunting? Here's a five-step approach to deal with extraordinary data warehouse growth and meet business expectations for scalability:

1. Develop quantitative requirements. Use a systematic, measurement-based engineering process to document quantitative require-

ments. They should include working estimates of the size and macro structure of the database and workload, service-level objectives, and operating schedule. These key inputs provide much of the information required to develop a physical database and evaluate alternatives.

The database's macro structure covers the likely size and structure of the largest tables, the likely set of the most heavily used relationships, and the likely distribution of data values of the most significant columns. The macro structure of the workload covers the 10 to 25 query or transaction types expected to account

Don't Do This: 7 Gotchas Of Scalability

1. Wait until the system is built to test for scalability

There's always a temptation to wait too long before doing performance and scalability testing. The classic trap is waiting until the system is ready to go into production. Sure, the test is realistic because you can run the actual database and application, but if you discover something wrong, it's often too late to do anything about it. Test for scalability before you're committed.

2. Live with vague expectations

Database people think that if no requirements are established, no one can prove they failed. In reality, it's often worse in this situation; management assumes that the system will meet all expectations, so the system is never good enough. You're much better off setting realistic expectations that can be met.

3. Skip requirements Users often don't know what the requirements are. You have to help them visualize a new business process and the requirements for supporting it. Only then can you develop valid usage sce-

narios and engineering requirements.

If, for example, you currently mail a giant catalog to all customers quarterly, and you want instead to do 100 targeted mailings of specialty catalogs each going to about 2% of your customers, then hold two facilitated discussions with stakeholders. First, talk about what the new mailing process will be and how it will get carried out 25 times as often each year. Second, explore the information capabilities needed to support the process. Then work out usage scenarios and develop the necessary workloads, service levels, and other requirements. Don't skip identifying requirements, or you'll end up back at pitfall No. 2.

4. Skip risk analysis Once you develop requirements, identify, test, and manage the risks that emerge.

5. Accept flimsy "proofs" Beware of salespeople taking over the definition of the proof. Never let the vendor define the test to be performed. If you don't have the expertise in-house, get a consultant with experience defining

benchmark specifications for testing complex data management systems. Your test has to capture the key challenges of scale and performance.

6. Underestimate growth rates

Knowing this year's requirements isn't enough. Architectural and platform decisions will take awhile to implement and longer to change. Project requirements out two to three years, at least—better to have a projection that gets revised than shoot in the dark. And don't assume that the data and business growth rates are the same. Data and workloads tend to grow faster than the related business because data gets used more intensively as the business gains momentum.

7. Ignore any dimension of scalability

Data size is the dimension easiest to measure, but the workload, data complexity, query complexity, availability, and data latency dimensions are nearly as important. They can all drive configuration size and determine whether you're on the right platform. Take them all into account. —RICHARD WINTER

for the main performance challenges and their expected frequency.

When coming up with these estimates, the key is to get them in the ballpark—absolute accuracy is far less important than scale. Getting the scale right lets you understand whether you’re building a passenger car, an 18-wheeler, or a freight train. Don’t settle on this too soon: Document a set of numbers, talk through the estimates with stakeholders, and then use them in the management process as well as architectural and engineering decision making.

2. Forecast long-term needs.

Within a few years, your data warehouse could be several times larger than it is today. To estimate long-term requirements, consider factors such as new applications, new or expanded subject areas, additional levels of data detail, as well as new users, tools, and data sources. The engineering requirements should define how a system will grow along each of the dimensions of scalability.

Don’t just extrapolate existing growth rates, since they don’t reflect changes in technologies and practices that might support major new opportunities. In retail, data scale increased dramatically first when point-of-sale and then when Web clickstream data were added to the data warehouse. In the supply chain, the next big leap in scale will come if there’s full deployment of RFID. Extrapolating from past trends might grossly understate the impact of future needs.

3. Identify the critical risks.

How To Manage Scale

- » **DEVELOP** quantitative requirements
- » **FORECAST** long-term needs
- » **IDENTIFY** critical risks
- » **MEASURE** potential solutions against the requirements and risks
- » **MANAGE** areas where the data warehouse doesn't meet requirements

documenting requirements—with vendors, user groups, reference companies, consultants—should raise the big risks: “We lose money if we can’t load that data in time,” or “We’re dead ducks if we go down on any of the big weekends.”

You’ll spot some yourself—like the engineering problems no one has a convincing solution for, or the recurring queries everyone knows are complex and time sensitive, but no one can say how much time they’ll take.

But not all engineering requirements are equally important; focus on the ones critical to

business objectives. In a fraud-detection application, it may be critical to get the data into the database within minutes or seconds of receipt—no matter what the circumstances. That may be fairly simple to do except during peak hours, yet those are the exact times it’s most critical to spot fraud because a lot of money is being spent. So ingesting data quickly during peak hours becomes a

critical factor. In other areas, response time may be important, such as with customer-facing queries. If a moderately complex query happens while a customer is talking with a call center agent, and thus has a desired 2-second window to complete, it could become a risk.

It may be easy to show that the requirements will be met initially, when data volumes are small and usage is light, but what happens in the second year, when volumes skyrocket? The trick is focusing on engineering with two characteristics: There’s no proof that the targets can be met, and missing the targets causes major pain to the business. These are the critical risks.

4. Measure solutions against targets. This step is key: Measure the solutions to a critical risk against the requirements today and as a company reaches the projections developed in the second step.

For this step to work, be realistic about scale and complexity. Don't cheat on the dimensions of scalability by, say, running a few simple queries, one at a time, on a 5% sample of the data. Instead, run a realistic simulation of the workload against a realistic, full-scale database, and take into account how the application is likely to evolve over the next three years.

5. Manage the gaps. Realistic analysis and testing often reveal that the intended data warehouse won't meet all the requirements. If so, address the issue with stakeholders before it becomes a problem. By measuring the alternatives, you can enter the discussion with real data on the options. Can users accept the 4-second response that is feasible under the current budget? Or would they increase the budget by 50% to get a 2-second response? Should we stay with the company standard platform, which has never been used with more than 10 TB of data, or take 90 days to evaluate other options, now that we understand we're likely to have 100 TB of data within 18 months?

A systematic engineering approach puts you in control, providing options with known outcomes and trade-offs as data warehouse requirements increase rapidly in six formidable dimensions. Where you have higher risks, you have analyzed, measured, and set up fallback plans. You can discuss the trade-offs and options with stakeholders and prepare them for the likely outcomes. This is a much better approach than the oft-used "forge ahead and hope" approach to managing data warehouse scalability (see story, p. 5).

ULTIMATE SCALABILITY

The new technology trend designed to deal with multidimensional data warehouse growth is toward highly parallel architectures. The HP Oracle Exadata Storage Server, announced last month, is designed to keep data flowing to and from more disks at once, increasing the pace at which I/O-intensive tasks can be performed. And Microsoft has just revealed that it will incorporate the DATAlegro technology acquired earlier this year into the next release of SQL Server, thereby increasing both I/O bandwidth and processor parallelism. Almost everyone is moving to exploit lower-cost hardware. Though big symmetric multiprocessor servers aren't about to disappear, there's an ever greater emphasis on scale-out architectures.

In the 1990s, conventional wisdom had it that massively parallel processing would never be more than a niche architecture, used for extreme requirements at the margins. But MPP has become reliable, manageable, and affordable—and suddenly it seems that nearly everyone is hungry for scalability. So highly parallel architectures—whether you call them MPP, cluster, or something else—have become part of the mainstream.

A lot of data warehouse practitioners are struggling with the changes brought on by rising data warehouse scale and rapidly evolving architectures. The most important thing to remember is that business problems aren't solved by buying new hardware or introducing new architectures. They're solved by determining the requirements of a solution and then implementing systems that meet those requirements.

To do that, follow these three recommendations in any data warehouse development project: Introduce a systematic management process to

deal with the scalability problems. Avoid the seven gotchas of scalability management. Emphasize quantitative requirements and use measurements or tests at every stage of the development life cycle. With a systematic approach, you will meet business expectations and have a scalable data warehouse with long-term business value.

Richard Winter is the president and founder of WinterCorp, a consulting firm focused on large-scale data management. In addition to advising companies in industries including retail, health care, financial services, and distribution, WinterCorp provides consulting services to vendors including Hewlett-Packard, IBM, Microsoft, Netezza, Oracle, and Teradata.

Performance Upgrade



By Doug Henschen

Teradata Rolls Out Peta-Scale Appliance

NOT EVERY COMPANY NEEDS AN APPLIANCE that can scale up to 50 petabytes, but who would say no to database upgrades that claim to improve performance by as much as 30%? Promising something for everyone, a new peta-scale appliance and the Teradata 13 database upgrade were the headliners at this week's Partners Teradata User Group Meeting in Las Vegas. Teradata also demonstrated a prototype appliance of the future using solid-state disk (SSD) drives that promise both faster performance and lower power consumption.

Teradata's new Extreme Data Appliance 1550 is designed for high-data-scale applications that are characterized by focused queries, departmental scale, and not-so-time-sensitive querying. Examples include Web site clickstream analysis, multiyear regulatory compliance, manufacturing processing and testing, RFID-product movement, and cell phone network usage. Many of these apps were heretofore viewed as impractical, or they were relegated to server farms and flat-file processing. The 1550 combines 1-TB hard drives with Intel quad-core nodes and built-in data protection software to offer extreme storage density, starting at a list price of \$16,500 per terabyte.

"This gives many people an option to get [vast sets of data] into a relational format and start doing analytics on it for discovery and huge data sifting," says Scott Gnau, Teradata's chief development officer. "It's for applications in which performance isn't extremely important, but ease of use and ease of integration will be important."

Teradata 13, the latest upgrade of the vendor's core database, is said to boost performance in several ways. New data extract, load, and transform tools, for instance, are said to speed loading, while workload optimization has accelerated online analytical processing query performance by as much as 30%. Perhaps the biggest news in the database upgrade is Teradata Virtual Storage, which enables customers to manage heterogeneous drive types as a virtual storage pool.

"In parallel processing environments, you always need the system to be balanced, because the slowest thread will dictate how fast the overall warehouse performs," Gnau explains. "With Virtual Storage, you can mix drive sizes, say 73-GB and 300-GB drives, in the same storage pool, and the software will automatically move the data, based on service-level requirements and the frequency of access, among the [fast and slow] areas of available storage."

Thus, the new Virtual Storage capability lets Teradata customers mix generations of devices without crimping overall performance. Low-demand data will be moved to the inner tracks of the slower, fatter drives while the high-demand data will be migrated to the faster, outer tracks of the higher performance drives.

Teradata's prototype SSD-based appliance appearing in Las Vegas is strictly for demonstration purposes. SSDs are still far too expensive to go into a production data warehousing device, but Gnau says they may be more practical by 2010 or 2012. The concept device looks like Teradata's 2550 appliance, but inside are 128-GB SSDs rather than spinning disks. Currently making their way into laptops and ultra-portable devices, SSDs promise five to 10 times the performance and an order-of-magnitude lower power consumption than conventional drives with motors and moving parts.

Closer Look



By Mark Madsen

Teradata Adds To A Growing Portfolio

T

HE TERADATA 1550 IS POSITIONED to deal with the very large data volume problem, not so much typical data warehouse usage. When you look at data usage, there are two types of large data problems. The classic data warehouse model involves analyzing subsets of the total data and occasionally scanning all the data. The other model is the need to analyze very large data sets that would normally be impractical, like looking at a year of Web traffic or call details.

The Teradata product line now covers the entire platform range, from smaller projects (subject-area marts, smaller warehouses) to real-time large enterprise data warehouses. A plus is that its products run

the same database across the entire line. Teradata has been very open about its product capabilities and pricing.

Teradata has a strong competitive position in the market. Most of the appliance vendors are still in venture-funded startup mode. In an uncertain financial market, this means they have to work harder to preserve cash since the likelihood of closing new funding rounds is low, as is an IPO or acquisition. DATAlegro was lucky to get out of the market when it did. If investment funds stay tight for the next few quarters, some companies with lower cash positions could run into trouble.

Teradata also announced the next major release of its database, improving performance and manageability (the types of things you'd expect in any major release). It added new features like automatic sensing of data temperature so data placement can be optimized, geospatial capabilities, and improved workload management features.

It talked about solid-state disks to augment performance as part of the virtual storage announce-

ment, allowing SSD and spinning disks to be used together with software that moves data to and from SSD based on performance rules.

It showed a prototype storage array that uses 128-GB solid state drives, and Toshiba recently announced that it has been able to build 256-GB SSDs with a read rate of 120 MB per second. Solid state disks offer incredible performance under random I/O workloads, but Teradata's experience is that SSDs don't perform that much better than spinning disks when it comes to database scans. This makes sense since you can only access the data as fast as the drive controllers and channels can move the data. There's significant work ahead to change how I/O subsystems perform before we can boost channel performance to match SSD read rates.

Based on the price of SSDs and the engineering challenges to improve both I/O subsystem hardware and software performance, don't expect vendors to roll out racks of SSDs any time soon. In a few years it will be more common to see blends of SSD and spinning drives to boost performance.

Enterprise Search



By Andrew Conry-Murray

Microsoft, Google, Specialized Players Vie For Supremacy

ENTERPRISE SEARCH TOOLS ARE EVOLVING to meet significantly different business requirements. IT and legal may need to scoop up documents, files, and e-mail relevant to forthcoming litigation. Security and compliance officers want to search laptops to make sure credit card numbers aren't hitting the road. Meanwhile, lines of business are clamoring for better ways to extract value from reams of enterprise data. Cracking open different repositories could help



salespeople better use information gathered about customers.

Companies approaching enterprise search must match their requirements to the capabilities of competing search platforms from Google, Microsoft, and a growing field of specialized vendors. Yet even if CIOs scope out requirements perfectly, they may find themselves running multiple search products for different business units to address diverse needs, and piling on the storage and server resources.

AND THAT'S OK.

Take National Instruments, a maker of computer-based measurement and automation products for manufacturers and scientists. The company has seen its search infrastructure—cover-

ing information from customers outside the firewall and employees inside it—grow from 10 servers to 25 in about three years. Eight of those are production servers, with the rest dedicated to testing and development, security, and processing. Of particular note is the wildfire growth of National Instruments employees' use of search. John Graff, VP of marketing and customer operations, says CPU requirements to index data and respond to employee queries are growing 152% year over year.

But National doesn't begrudge the increase in resources. "As IT comes back to me to say 'We need more,' it's an easy sign-off because the value is so clear," Graff says.

In this business climate, what kind of technology draws that kind of support? One that solves

Impact Assessment: Enterprise Search

● **Benefit**

● **Risk**

IT organization

IT is on the front lines of e-discovery, and a search tool is indispensable. These products also can help with security and compliance and provide insight into what data the organization has, its business value, and whether it can be disposed of.



IT may run into operational problems balancing bandwidth, latency, and OS consumption when indexing remote users across a WAN. These tools must be used in consultation with legal and HR to ensure IT scopes the search effort properly.

Business organization

Companies are sitting on incredible volumes of information. Enterprise search can help solve business problems by getting relevant data into users' hands more quickly.



Search for search's sake yields little value; have a clear vision of a business need that search can meet. Underestimate the complexities of a search infrastructure and you may fail to provide the necessary resources to support a robust system.

Business competitiveness

Enterprise search can cut through information silos. Discovery- and compliance-oriented search help keep litigation costs low, underpin retention and disposition policies, and meet regulations around privacy and security.



Search deployments that aren't backed by explicit business plans will waste human and technological resources that could be applied to other efforts.



Bottom Line

Enterprises can let data volumes grow unchecked and sink under their own weight, or they can construct information management policies and systems designed to extract value from hoards of data. Enterprise search, both compliance- and business-focused, plays a central role in information management.



problems. Still, purchasing decisions are complex. There's not only no clear market leader, but the category is diverging into two distinct paths.

LOOK PAST THE OBVIOUS

While Google is synonymous with Web search, it's only one of many players in this market—and by no means dominant. Autonomy, Microsoft via its Fast Search & Transfer acquisition, Recommind, and others more than hold their own against the Big G. Endeca and IBM offer search products aimed at specific business problems. And companies such as Guidance Software, Kazeon, and StoredIQ Software are winning customers faced with e-discovery burdens.

InformationWeek separates the enterprise search market into two major categories: compliance search and business search. Vendors in the first category aim at IT and corporate officers, such as legal counsel, human resources professionals, or compliance officials. These constituencies aren't trying to find one relevant result out of a 1,000, but 5,000 relevant results out of 1 million. This category is dominated by e-discovery, and search products not only must find information, but also manage it, whether by moving it to a new repository or applying controls to ensure files aren't changed or deleted. Vendors in the second category aim at employees, whether a business unit looking to extract more value from the information in various repositories, or a broader audience that needs help finding mislaid documents.

DISCOVERY CHANNEL

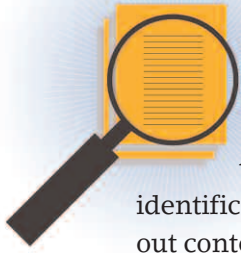
Autonomy, Guidance, Kazeon, and StoredIQ offer compliance search technology, with e-discovery and information management as the major drivers. In December 2006, the Federal

Rules of Civil Procedures, which govern the processes and requirements of parties in federal civil suits, clarified the rules regarding electronically stored information. These rules have had a significant impact on the breadth of data that companies are expected to find and produce in litigation.

While case law around these updated rules is evolving, the upshot is clear: Courts won't accept "I can't find it" as an excuse for not producing information relevant to a lawsuit. In January 2008, Qualcomm was slapped with an \$8.5 million penalty because it mishandled the e-discovery process and failed to produce e-mail relevant to a lawsuit with Broadcom. And without search in place, e-discovery costs can fast eclipse the amount that a company may stand to lose in a lawsuit. Even employing precise keyword searches, Verizon places the price of processing, reviewing, culling, and producing 1 GB of data at between \$5,000 and \$7,000, according to a study by the University of Denver's Institute for the Advancement of the American Legal System. Multiply that by the size of your data stores, and the cost of a few new servers seems downright reasonable.

Mike Brooks, CIO and senior VP of CVR Energy, a \$3 billion-a-year refinery, uses Autonomy's Idol as part of his e-discovery program. Idol is a search and indexing technology that underpins all of Autonomy's software products. Brooks runs discovery searches using Idol, and then uses a homegrown software tool to move relevant information to a secure repository. "We are trying to make sure there's nothing in our enterprise we don't know about, to avoid surprises," Brooks says.

E-discovery phases can be mapped out using the Electronic Discovery Reference Model, an independent framework that has been adopted



by the vendor and legal communities. The search products described here focus on the initial phases of the discovery process, including identification, collection, and preservation. In the identification phase, search products must seek out content relevant to a lawsuit; they connect to various repositories, crawl the content, and create a searchable index. Users—in this case IT, HR, and legal counsel—run queries and get back a list of matches. And like business search tools, these products offer capabilities that go beyond simple keyword and Boolean search, such as support for multiple languages, natural language processing, and pattern recognition to extract additional layers of meaning from the information being indexed.

The features needed for the collection and preservation phases separate discovery-focused search products from their business kin. For instance, to do collection effectively, these products must be able to move content from one repository to another while preserving metadata, such as time stamps, to demonstrate that information wasn't altered during the discovery period. StoredIQ addresses this by logging the original metadata and then adjusting the necessary fields when files are moved or copied to a new location. Autonomy, Guidance Software, and Kazeon say their products can move files without changing metadata at all.

Preservation requires relevant data to be maintained in an unaltered state. In the discovery process, employees involved in litigation, called custodians, are issued a preservation notice by legal counsel instructing them not to destroy or tamper with files, e-mail, and other information related to the case.

Human nature being what it is, custodians may be inclined to do exactly the opposite, so these

search products have to provide a legal hold, in which information is preserved for the duration of a case. Discovery search products enforce these holds either by moving relevant information to a secure server or archive, or by altering write, open, or delete permissions.

Laptops and desktops present problems for collection and preservation that don't exist with business search, and vendors have approached those challenges differently. Autonomy's Zantaz Introspect, Kazeon, and StoredIQ can access and search PCs over the network. They can collect information and enforce legal holds without the use of an agent.

Guidance requires a small piece of software, which it calls a servlet, to be installed on systems to be searched, though the company says the servlet doesn't install DLLs or interact with the host operating system. Autonomy also includes an agent with its Aungate Legal Hold software to lock down relevant information on laptops, PCs, and servers. StoredIQ says it plans to release an agent for laptops and desktops at the end of the year.

BEYOND DISCOVERY

Deploying a search product is often a tactical response to an e-discovery emergency. But there are long-term strategic benefits to these products that involve understanding and managing corporate information, particularly unstructured data.

In addition to discovery, CVR's Brooks uses Idol to index work orders that are generated as part of plant maintenance operations. While these work orders contain pre-defined codes that identify common operations, employees also include detailed comments about problems and solutions that provide context about maintenance issues that can't be gleaned from codes.



“When Idol goes through the data, it groups together like topics, so when I’m running through a set of work orders, I can look at what the issues have been,” Brooks says. If he sees large clusters of work orders around a specific topic, it allows him to identify reoccurring problems.

IT search also can shine a light into hidden corners of an organization, such as laptops and desktops. IT often has little visibility into the kinds of information stored there. Popular repositories such as SharePoint, which can be deployed without IT’s input or even awareness, also are prime candidates for compliance search. “We see people running our technology once a week for audits, like finding personally identifiable information, source code, intellectual property,” says John Patzakis, chief strategy officer at Guidance. Kazeon CEO Sudhakar Muddu says 50% of his company’s business is e-discovery, with the rest in support of governance, security, and data management.

This process of looking at search and indexing to serve e-discovery needs also can help companies manage—or create—a retention and disposition strategy. “Forward-looking companies are being driven by chief risk officers to get a handle on data,” says Craig Carpenter, general counsel and VP of marketing for Recommind. “They want to have it organized and start retiring data they don’t need.” Getting rid of data may go against the natural instincts of technology professionals, but as organizations add terabytes of information to the infrastructure every year, those instincts may be swamped by necessity.

THE ‘17 DATABASES’ PROBLEM

Business enterprise search is evolving from its

original use case, which can be described as “search for search’s sake.” The goal then was to generate a general index of information repositories and provide a front end for employees to browse through it, the way they would the Web—with simple queries that coughed up a long list of results. Today, companies approach business search to get better insight into specific domains and address business problems. “Customers aren’t looking to buy search,” says Craig Reinhardt, director of enterprise content management at IBM. “They want better business results. We look at search as a critical ingredient that needs to be integrated with other applications.”

Reinhardt points to customers, such as those in law enforcement, that use IBM’s OmniFind Enterprise search platform to find patterns in criminal records, or to manufacturers that use the search software to analyze customer comments on blogs and wikis.

Microsoft sees significant opportunity in this business-oriented approach to search, which was a major driver for its January acquisition of Fast Search & Transfer. National Instruments’ situation illustrates why Microsoft’s view makes sense. In early 2005, National was looking for a way to streamline access for different lines of business to all the content it gathers on its customers.

“We made a list of all the databases and repositories where we had customer information,” VP Graff says, adding that he quickly found certain groups were tapped into different systems—for instance, the sales group used the CRM application, while engineering tracked the company’s tech support Web pages—but no one group had access to everything. “We called it the ‘17 databases’ problem,” he says.



The company had deployed the Fast Enterprise Search Platform to enhance the search capabilities of its customer-facing Web site, and Graff thought search might be a good way to unlock the silos that contained information about its customers. These silos included the CRM system, corporate file servers, Lotus Notes, Oracle databases, and an internal wiki.

"It's been a huge hit," Graff says. "Our salespeople use it to do research on customers prior to visits. Marketing and engineering management use it to get feedback on what customers are doing with products. Even our CEO uses it."

A key to success is the search interface, which employees access through an intranet. National customized the user interface to let people select facets of a search. One surprisingly popular choice is age. "You can look at the creation date an order was processed, or the date a tech support query happened," he says. "Users can bring the freshest information to the top of the pile."

Companies also are asking search to provide more context, based on a variety of factors, such as the person conducting the search. For example, Google's latest version of its Search Appliance leverages Active Directory and LDAP-based directories to personalize search results based on the searcher's organizational role.

"You can create a policy group for the sales department that gives higher priority to documents talking about pricing," says Nitin Mangtani, lead product manager for Google Enterprise Search. "For engineers, you can give higher importance to engineering documents."

Another example of context is Recommind's MindServer search platform, which can be aug-

mented with modules, such as Expertise Location. MindServer uses information gleaned from indexed data and other sources, such as HR portals, to associate users with expertise in different content areas based on that user's work product. It can serve as an extended company directory to help employees locate colleagues with specialized knowledge.

MICROSOFT ALSO IS PURSUING EXPERT SEARCH.

"We refer to it as 'people search,'" says Jared Spataro, director of enterprise search at Microsoft. "It will relate concepts to people in the organization who might know something about something. It's something we hear a lot about from our customers." Spataro says the company wants to lead in this area, but it hasn't yet announced specifics.

KEY ISSUES

Regardless of the type of search you're interested in, there are technological issues that must be addressed, including indexing speed, index size, and security. In a discovery effort, time is of the essence. Initial results may need to be available to counsel within weeks. That may sound like a long time, but not when faced with repositories that hold multiple terabytes of information that have to be indexed before anything else can happen.

Indexing times are fluid. How quickly an engine can create an index depends on the content. A file share full of PowerPoint slides with 25 words per page will be indexed in a blink. Text-heavy documents take longer, as do PST files that have to be cracked open or files that may have multilevel attachments.

Some search products will federate with an index that has been created by the repository's native search feature, such as a Documentum repository or an e-mail archive. This speeds



indexing time and saves on storage space. Through federation, the third-party search engine essentially brings the query to the application's native search field, and then incorporates the results into its own user interface.

Note that most e-discovery search vendors prefer to index content themselves, whether or not the targeted repository has native search capability.

Customers also have to take the search infrastructure into account. Google and StoredIQ deliver via an appliances, while the other search products are pure software deployed on servers. IT must provide sufficient processing capacity to handle volumes of queries. This may not be an issue with compliance search, which isn't intended to address simultaneous search requests from a large audience of users.

Companies must also provide storage for the index (except for Google and StoredIQ). Vendors usually estimate the index size as a percentage of the content being cataloged. For example, if the index is 10% of the content, a 100-TB body of data will yield a 10-TB index. The primary factor is how detailed you want the index to be. For instance, the Fast search engine can produce an index that runs about 20% of the size of the content, but most organizations will enrich the index through ad-

vanced linguistics to provide more detailed search results. Microsoft's Spataro says customers opting for a rich index should expect it to run two or three times the size of the actual content store.

Another issue is how the search engine links to content repositories. Most search products include out-of-the-box connectors for popular platforms, such as Exchange, Notes, SharePoint and Documentum, as well as general-purpose connectors for file and Web servers. However, IT may need to tweak connectors or build one-off integrations if a critical application or repository isn't supported.

CIOs also need to make sure users don't get access to search results that violate corporate access controls. Most search engines can match user identities to permissions associated with groups in the company's directory system.

Bottom line, enterprises should approach search as a strategic technology that will help solve specific business problems. To that end, companies must understand their own requirements when evaluating search platforms—IT should involve business units, legal, and HR, and start with the business case to see where search will provide value. Done right, this is one technology that will pay off not just in dollars, but in productivity and peace of mind.