DIGITAL SPOTLIGHT | STORAGE

COMPUTERWORLD®

Peer Perspective. IT Leadership.
Business Results.
COMPUTERWORLD.COM
FEBRUARY 2013

Disaster Recovery on Double Duty

Virtualization and replication technologies are protecting against disaster while keeping business services humming.

Also inside: Big Data Overload | Who Holds the Encryption Keys?

COMPUTERWORLD

P.O. Box 9171
492 Old Connecticut Path
Framingham, MA 01701-9171
(508) 879-0700
Computerworld.com

» **EDITORIAL**

Editor in Chief
Scot Finnie

Executive Editor
Julia King (events)

Managing Editors
Johanna Ambrosio (technologies),
Ellen Fanning (features),
Sharon Machlis (online),
Ken Mingis (news),
Bob Rawson (production)

Assistant Managing Editor
Valerie Potter (features)

Director of Blogs
Joyce Carpenter

Art Director
April Montgomery

Features Editors
Barbara Krasnoff (reviews),
Tracy Mayor

News Editors
Mike Bucken, Marian Prokop

National Correspondents
Julia King, Robert L. Mitchell

Reporters
Sharon Gaudin, Matt Hamblen,
Gregg Keizer, Lucas Mearian,
Patrick Thibodeau,
Jaikumar Vijayan

Editorial Project Manager
Mari Keefe

Office Manager
Linda Gorgone

Contributing Editors
Jamie Eckle, Preston Gralla,
JR Raphael

» **CONTACTS**

Phone numbers, email addresses and reporters' beats are available online at *Computerworld.com* (see Contacts link at the bottom of the home page).

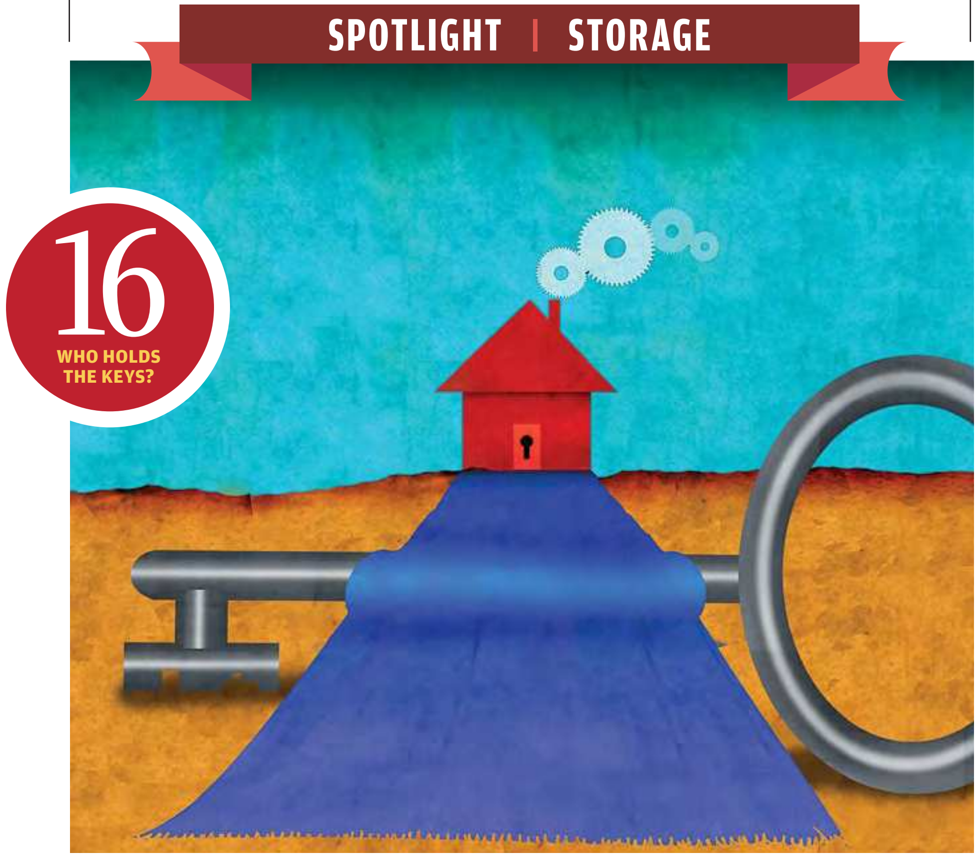
Letters to the Editor
Send to letters@computerworld.com. Include an address and phone number for immediate verification. Letters will be edited for brevity and clarity.

News tips
newstips@computerworld.com

Subscriptions and back issues
(888) 559-7327, cw@omeda.com

Reprints/permissions
The YGS Group, (800) 290-5460, ext. 100, computerworld@theygsgroup.com

THIS ISSUE | FEBRUARY 2013



COVER STORY

Disaster Recovery on Double Duty

6 Virtualization and replication technologies protect data from disaster, while keeping business services humming.

Big Data Overload

12 Complex requirements and never-ending demands for additional capacity are vexing storage administrators. Here's how to handle the data deluge.

Who Holds the Keys?

16 Encryption isn't bulletproof – especially if keys and digital rights are left out in the open. Learn how to lock down stored data.

HEADS UP | **4** Facebook plans to put user photos in "cold storage." | **Editor's note:** Why storage touches every facet of your business. **OPINION** | **11** **John Webster** says big data promises a style of computing that more closely mimics the functioning of the human mind. For IT, that means moving from provisioning of services to making a big impact on business results.

MICHAEL MORGENSTERN ILLUSTRATION

Fresh Insights

New Trends

Great Ideas

HeadsUp

THINKSTOCK

DATA STORAGE

Facebook to Put Photos in ‘Cold Storage’

FACEBOOK IS rethinking how it stores data to cope with the 7 petabytes of new photos that its users upload each month.

As the number of photos grows, Facebook needs to find cheaper, less power-hungry ways to store them all, according to Jay Parikh, the company’s vice president of infrastructure engineering.

Users upload about 300 million photos per day, and more on special occasions, Parikh told attendees of the Structure Europe conference in Amsterdam in October. “Halloween is one of our biggest photo upload days of the year. We will get somewhere between probably 1 and 2 billion photos uploaded just in a single day,” he said.

But people quickly lose interest in photos taken around holidays — after a few days or weeks, no one looks at them anymore. But Facebook’s deal with its users is that “we can’t delete the data; we have to keep it,” he said. That led to the idea of putting the photos into a sort of “cold storage,” Parikh said.

To do that, Facebook plans to build a new data center with different types of storage, server hardware and network gear. The new facility would consume less power and cost less than existing data centers — all without changing server response times, said Parikh.

Facebook plans to share features of its new data center infrastructure through the Open Compute Project, an initiative Facebook started to apply the open-source software collaboration model to the world of data center hardware.

– Loek Essers, IDG News Service

EDITOR’S NOTE

Storage Touches Everything in Your Business

You’re reading a bonus, digital-only issue of *Computerworld* focused on storage. Inside, you’ll find refreshers on three of the hottest storage topics: disaster recovery, big data and digital-rights management.

Our cover story, “Disaster Recovery on Double Duty” (page 6), is especially relevant in the wake of Hurricane Sandy. More and more IT shops are using technologies such as virtualization and replication to make disaster recovery just another (very important) service.

The nuts and bolts behind storing petabytes (and more) of data in a format that’s easy to access and analyze are more challenging than your average storage platform. “Big Data Overload” (page 12) tells you how to handle the data deluge.

Experts and IT leaders offer strategies for getting the most from the latest encryption and digital-rights management technologies in “Who Holds the Keys?” (page 16).

And in his column “Big Data Storage Must Support Analytics” (page 11), John Webster explains why big data computing closely mimics the functioning of the human mind. For IT, that means moving from provisioning of services to making a big impact on business results.

SCOT FINNIE IS *COMPUTERWORLD’S* EDITOR IN CHIEF. CONTACT HIM AT SFINNIE@COMPUTERWORLD.COM AND FOLLOW HIM ON TWITTER (@SCOTFINNIE).

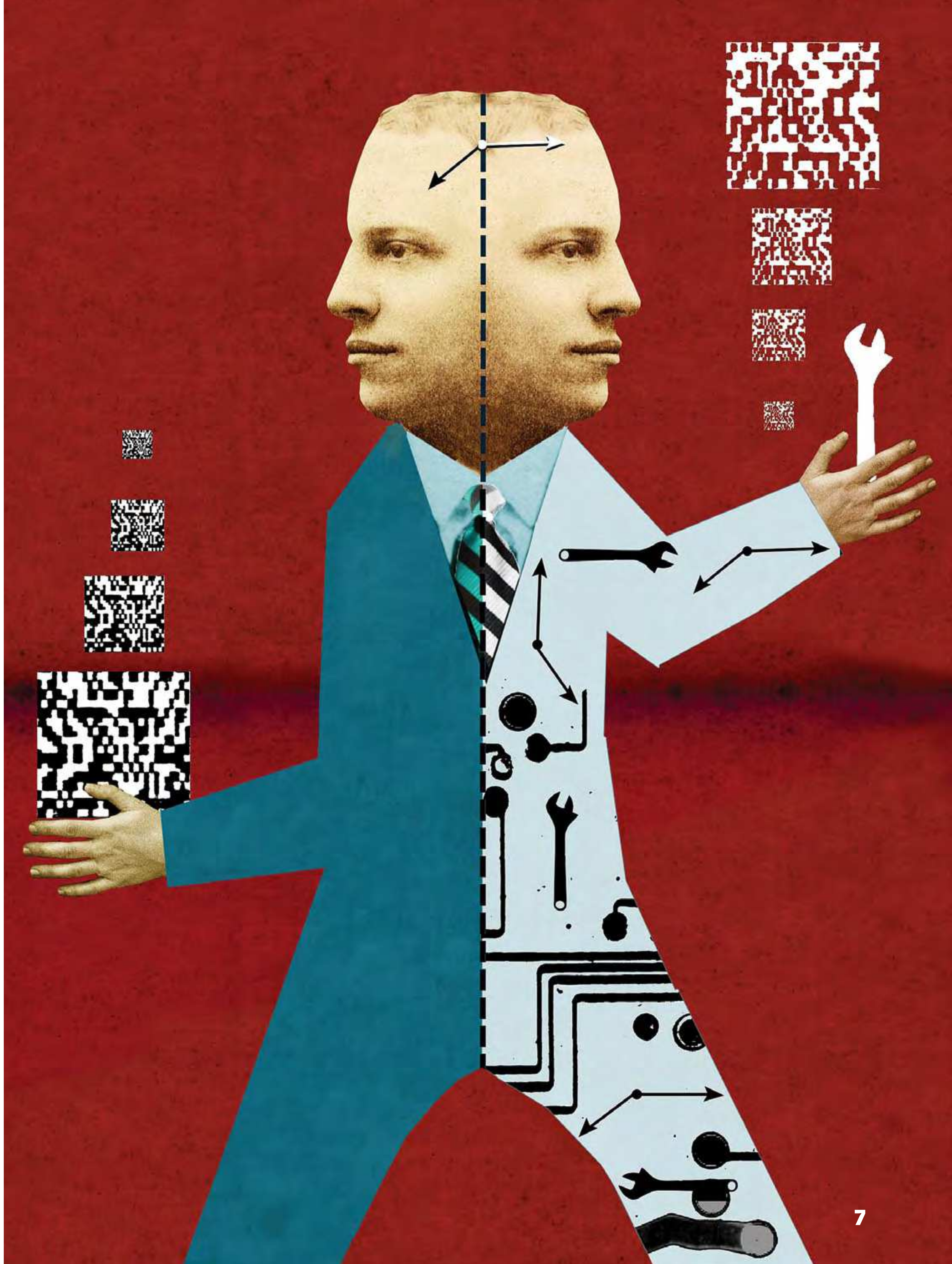
DISASTER RECOVERY ON Double Duty

Virtualization and replication technologies *protect data from disaster, while keeping business services humming.*

BY ROBERT L. SCHEIER

AT INGRAM MICRO, executive vice president and CIO Mario Leone doesn't think about how much he will spend on disaster recovery.

Instead, since 2010 the global electronics distributor has woven disaster recovery requirements into its broader business objectives and its service-level agreements (SLA) with its 15,000 users. The IT organization meets its service and disaster recovery commitments, and is reducing costs, through a "hybrid"



MICHAEL MORGENSTERN ILLUSTRATION

cloud made up of its own virtualized hardware located in colocation facilities in Chicago, Frankfurt and Singapore.

And rather than paying for dedicated recovery hardware that sits around waiting for a disaster, it uses virtualization to shift workloads from a failed server to one running a less critical workload. "We're always using that architecture for something," says Leone.

More and more IT shops are using technologies such as virtualization and replication to make disaster recovery just another service, sometimes using the same servers, network and storage that run order entry, email, application development or other services. This merges what historically were disaster recovery and business continuity efforts, protecting the business against not only rare disasters, but also human error or equipment failures.

Some store only data (and perhaps templates for virtual machines) off-site, creating (and paying for) the physical hardware to run them only when needed. "We can recover at our remote site much, much faster by just being able to fire up the system images of the VMs," says Justin Bell, systems administrator at Strand Associates, an engineering firm in Madison, Wis. Even if the server infrastructure at that site is less robust than the one at the primary site, "we could run in limited capacity, on much less hardware, until we got things back up at our primary site."

> **We can recover at our remote site** *much, much faster by just being able to fire up the system images of the VMs.*
>
> **JUSTIN BELL,** SYSTEMS ADMINISTRATOR, STRAND ASSOCIATES

Other organizations have done away with dedicated disaster recovery systems. Instead, they shift production work to test or development servers during outages and put less critical work on hold.

More Demands, More Risk

These changes are driven by ongoing pressure to cut costs while maintaining continual uptime, and by the flexibility provided by server, storage and network virtualization. Meanwhile, a recent spate of natural disasters, along with stricter regulatory requirements, has made disaster recovery the No. 1 subject of client inquiries at researcher Gartner, says analyst John Morency.

However, Forrester Research reports that enterprise disaster recovery/business continuity budgets are stuck at 6% of total IT capital and operating budgets and that concerns such as "consolidation, business intelligence and virtualization" are given higher priority when it comes to spending.

Meanwhile, the list of critical services that need protection keeps growing, with communication tools such as voice over IP and email gaining "critical" status alongside traditional business applications like order entry and ERP. Finally, it's necessary to ensure uptime not only after major disasters, but also in the event of localized failures, and many companies need the ability to quickly recover just one file rather than an entire system.

By separating virtual servers, networks and storage capacity from physical hardware, virtualization gives users many more choices in disaster recovery strategies. "When you recover a [virtual machine], it doesn't matter where we put it," says Kurtis Berger, IT manager at Provider Advantage NW, a healthcare software and services company in Beaverton, Ore. "At each of our data centers, all of our VM servers are pretty much the same. [Almost] any old box will handle the prescribed load, and it'll be good enough to recover some VMs onto."

Disaster recovery is also being transformed by fast, easy-to-use replication software that copies data between primary and recovery sites in near real time. One such offering, Double-Take software from Vision Solutions, allows users to sync data among servers and establish failover protection in about 20 minutes, says Joseph Pedano, senior vice president for data engineering at Evolve IP, a provider of cloud-based IT services in Wayne, Pa.

Martin Mazor, Ingram Micro's director of global information assurance, wouldn't discuss which products he uses, but he says replication allows his company to recover systems much more quickly than the full day it would take to ship tape off-site. Ingram Micro has also invested in tools that provide a single performance dashboard for all of its worldwide operations, and it has offered employees training in areas such as operational management and the handling of incidents and problems.

Evolve IP uses VMware virtualization technology, and Pedano says backup and recovery tools now feature improved VMware integration, making it easier to replicate and restore not just servers but also their associated databases and security systems.

To successfully restore a business service such as email or order entry, IT must recover not only the application server, but also associated components (such as an Active Directory server that contains user information or a database that holds inventory records), and it must do so in the proper order. Taking these dependencies into account is a major area of focus for vendors.

Recovery In the Cloud

- **Start with applications that already perform well in virtual or private cloud environments** but don't support your most important systems. This gives you time to try different approaches and vendors.
- **Be realistic about SLAs.** Most cloud providers won't take responsibility for your losses if you can't recover everything after a failure.
- **Understand the interdependencies** among the applications and services you host in the cloud and those you host in a traditional data center, so you can properly test recovery processes.

SOURCE: FORRESTER RESEARCH

Symantec, for example, recently announced that enhancements to its small and midsize business and enterprise backup products combine more granular backup and recovery of VMs with the ability to account for dependencies among VMs. The enhancements also make it easier to use multiple public or private cloud backup services, and to convert a physical server at a production site to a virtual server at a recovery site, says Dan Lamorena, director of product marketing for Symantec's storage and availability management group.

Continuity Software's RecoverGuard software is designed to automatically check all critical infrastructure components, such as the file system and virtualization components, and identify vulnerabilities that could cause downtime and data loss. It looks for vulnerabilities using a database of "signatures" similar to the ones antivirus tools use to identify malware. The database is updated by the vendor's researchers and its users, says CEO Gil Hecht.

Other products with those capabilities include VMware vCenter Site Recovery Manager, which also supports custom scripting and automation to ensure that VMs are brought up and reconnected in the proper order across multiple sites, says Gaetan Castelein, VMware's director of product marketing.

Making It Pay

Often, the only way to get funding for disaster recovery systems is to demonstrate that they deliver more than just "insurance," or that they can even pay for themselves.

For example, Strand uses FalconStor Software's Continuous Data Protector appliance to replicate about 50TB of data and 25 virtual servers between its remote offices and headquarters. This is not only easier and less expensive than using a colocation facility, but the higher bandwidth required for the replication also makes it easier for employees to videoconference and share complex engineering documents

Kevin Schmidt, senior network engineer, Thorntons

That bandwidth also allows Strand to "take snapshots every hour on the hour, so we can facilitate a file restore in about three to five minutes," says Bell. Given the expense the company would incur if an engineer had to repeat several hours of work, the ability to take snapshots helps justify the cost of disaster recovery even without a disaster, he says.

Thorntons Inc., a Louisville, Ky.-based convenience store operator, recoups much, if not all, of the cost of disaster recovery by using DataCore Software's SANsymphony storage virtualization software on XIOtech SANs. It purchased those SANs to support its newest servers, while moving its older Dell Compellent SANs and older servers to nearby space it already leased as a disaster recovery site. Senior network engineer Kevin Schmidt says that gives the company disaster recovery for its full application environment, not just its data, and it has improved performance and cut the time required to produce a profit and loss statement from 10 or 12 hours to less than five hours.

Another benefit is that virtualization allows the company to use the Dell Compellent storage, for which it paid $350,000 in 2007, as a recovery platform for its newer XIOtech storage.

Snapshots in Time

IMPROVING THE PERFORMANCE of replication systems and related technologies, such as snapshot tools, and tailoring them to shorten virtualized disaster recovery times is a key focus for vendors. Here are some examples.

- **Actifio's Protection and Availability Storage** (PAS) appliance allows users to execute a one-time transfer of data to a remote site and then send only changes to the data, with the changes themselves deduplicated, says Actifio CEO Ash Ashutosh. This not only reduces bandwidth requirements; it can also eliminate the need for backup software, he says.

  The distributed object file system within PAS contains information about each block of stored data that makes it easier to find and reuse the data for purposes other than disaster recovery, such as test and development, regulatory compliance or legal searches, he says.

- **FalconStor's CDP** aims to speed recovery by ensuring the most recent snapshot is always the most complete. This eliminates the need to factor in the incremental changes since the initial backup before recovering the data. And it can save hours when recovering tens of terabytes of data, says Fadi Albatal, vice president of marketing and product management.

- **Asigra's Cloud Backup** eliminates the need for dedicated physical recovery hardware by automatically backing up VMs to virtualized environments and scaling up the VMs in the recovery environment so they can meet production needs. By automatically creating new servers and provisioning storage, it can reduce restore times from hours to minutes, says Eran Farajun, an executive vice president at Asigra.

- **Egenera's PAN (Processing Area Network) Manager** software virtualizes connections between physical or virtual hosts to a customer's network or storage resources, which speeds restoration by making it easier to create not just VMs but also the network and storage connections needed to make them work, says John Humphreys, vice president of marketing.

  PAN can also automatically detect failures in production servers and move them to the recovery environment, with the new server looking "just like it did before, with the same MAC address and same resources," says Scott Geng, senior vice president of engineering.

- **Dell's Compellent Live Volume** enables a physical server or VMs to share a virtual storage volume among Dell's Compellent Storage Center SANs in a semi-synchronous configuration that enables always-available failover volumes or LUNs, and makes it possible to move data closer to users for performance reasons, says Brett Roscoe, general manager and executive director of data management at Dell.

  Jason Buffington, an analyst at Enterprise Strategy Group, says applications like Microsoft Exchange, Microsoft SQL Server and some network-attached storage platforms offer capabilities such as replication and failover at little or no extra cost.

– ROBERT L. SCHEIER

Cloud Disaster Recovery? Not So Fast

Some providers say cloud-based disaster recovery will bring the benefit of true disaster recovery, rather than just backup, to small and midsize businesses that until now couldn't afford it.

Pat O'Day, co-founder and CTO of Bluelock, a provider of public cloud virtual data centers, says customers are increasingly satisfied with cloud security. Many security experts say even public cloud environments in which multiple customers share hardware can be made secure with the proper processes.

But a fall 2011 Forrester Research survey showed only 11% of large enterprises and 9% of small to midsize businesses had adopted recovery as a service, with 35% of large enterprises and 41% of SMBs saying they were interested in it but had no plans.

Berger of Provider Advantage NW says cloud providers only promise "not to go into your servers" when he questions them about security. "To me, that's not enough," he says, adding that the disaster recovery prices he's hearing — $500 per month per server — are "more than I can justify."

Kurtis Berger, IT manager, Provider Advantage NW

He instead uses Acronis Backup & Recovery to back up approximately 60 VMs at two data centers. The facilities are only a half-hour apart, so this setup would not meet some definitions of a disaster recovery system, but he says it covers most of his needs because the applications aren't mission critical.

Hecht downplays resistance to cloud-based disaster recovery, saying the smallest companies typically host their entire infrastructures in the cloud, and thus get some level of disaster recovery simply by having applications and data off-site.

Smaller companies that do choose the cloud typically don't do it for the savings, he says, but because "it's just so much simpler to have a system you set up and forget."

While midsize organizations have some incentive to consider disaster recovery in the cloud, few of them use the cloud for mission-critical systems that require true disaster recovery — and what they get in the cloud is closer to dedicated hosting (with the customer's data and systems running on separate hardware) rather than a multitenant, elastic, "pay-as-you-go" public cloud, Hecht says.

Most large organizations are big enough to provide disaster recovery themselves, he says, and even if they weren't, "there's no good solution" for protecting sensitive applications in the cloud.

Cloud disaster recovery is also not suited for applications on older platforms that most cloud providers don't support, or large databases that don't perform well in the cloud, says Morency. Users also need to watch for hidden costs in software licenses, he adds, noting that some cloud vendors charge for software sitting unused on remote VMs or disaster recovery systems.

Gartner and Forrester also warn that most cloud disaster recovery providers will refund only a portion of a customer's fee if disaster recovery falls short — nowhere near enough to make up for the potential revenue loss that such an event could cause.

The cost of the bandwidth required to quickly recover an organization's VMs and data from the cloud is often an unwelcome surprise, says Alan Arnold, executive vice president and CTO at Vision Solution Management, which provides high-availability and disaster recovery software and services. Some customers and providers opt to physically ship portable hard drives via overnight courier, says Arnold, recalling that one user joked that "FedEx is still the largest bandwidth network out there."

With IT so central to the business and budgets so tight, it's essential to get input from top business managers to assess which applications deserve the highest levels of protection. Ingram Micro, for example, conducted a business impact analysis that put various applications in different tiers, with voice, email, ERP and ordering among the top priorities. The company thought of it "just like an insurance policy," says Mazor. "It helped us think of how much insurance we're going to buy." ◆

Scheier *is a veteran technology writer. You can contact him at bob@scheierassociates.com.*

New Approaches To Data Recovery

SOME IT shops are expanding disaster recovery to include not only servers but also user devices. They're using portions of backup sites to store images of virtual desktops, laptops or even tablets so users can have access to their data and applications while they await replacement devices, says Eran Farajun, executive vice president at Asigra, which is also giving customers the ability to back up and restore data from consumer devices such as smartphones and tablets.

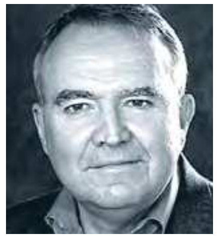
Jason Buffington, an analyst at Enterprise Strategy Group, says many companies now require branch offices to adopt the same protection standards as headquarters. He says products designed to help with such efforts include Riverbed Technology's Steelhead EX+ Granite appliances, which optimize the performance of WANs to speed backup and replication from branch offices to central data centers.

And many organizations are reducing or ending their use of tape for disaster recovery, although some still use it for long-term archiving.

"For us, tape is dead," says Kurtis Berger, IT manager at Provider Advantage NW. "It was the second tape drive that failed that finally pushed us toward a hard-drive-only solution. Hard drives are faster, and so cheap. We just couldn't find any reason to entertain the idea of tape anymore."

"Tape has been a love-hate relationship – mostly hate," says Jason Axne, system administrator at conveyer belt manufacturer Wire Belt Company of America. He cited tape's unreliability, the lengthy recovery periods for even single files or email in-boxes, and the time required to manage backups. Using Actifio PAS and disk-based storage, he says, "I don't spend any time during the day managing our backups . . . because it just works."

– ROBERT L. SCHEIER



JOHN WEBSTER

Big Data Storage Must Support Analytics

PERHAPS YOU'VE heard that the next new thing in IT is "big data" and concluded that the hype-cycle machine is turning out another attention-getter. I'm not big on predicting paradigm shifts, so I won't in this case. But I will say that if you're an IT professional, you ignore big data at your peril. I believe this one is all it's cracked up to be and more.

First, a word of caution. As with the cloud (the last new thing), we are now in the definition stage. New and often conflicting definitions abound as vendors attach their own meanings to the term *big data*.

The most common source of confusion results from the conflation of big data storage with big data analytics. Big data analytics is the big deal. Big data storage is really nothing more than storage that handles a lot of data for applications like high-definition video streaming.

One large storage vendor that has yet to make a big-data statement told me that his company was considering "Huge Data" as a moniker for its big data storage entry. Seriously. Someday soon, big data storage will begin to support big data analytics. Right now, though, I think it's key to first figure out if the vendor is pitching storage or analytics.

The definition of big data analytics is also getting pulled in somewhat conflicting directions. One can start with an understanding of data warehousing and add capabilities that the classic data warehouse doesn't offer.

For starters, big data analytics encompasses unstructured and structured data. It's widely believed that 80% of all data is unstructured. Big data analytics means that unstructured data — the bulk of what's out there — can now be mined.

The classic data warehouse user sets up queries and gets results anywhere from a day to a week later, whereas the goal for many big data analytics processes is to deliver results in real time.

> "The ability to produce new types of information in real time is decidedly new and powerful.

Finally, data warehousing works with a limited number of data sources. Big data analytics has the power to combine disparate sources — like a supply chain tracking system that commingles RFID, GPS and product shipment data — to deliver information previously unattainable.

I could say that any definition of big data analytics must combine all three of these attributes, but that would be misleading. What isn't helpful is relabeling something like a traditional data warehousing product as "big data" simply because it handles bigger data volumes.

Rather than quibbling over definitions at this stage, what we really should be after as IT professionals is understanding and hopefully leveraging what is new. The ability to encompass unstructured data into the business analytics process is new. The ability to converge multiple data sources — structured and unstructured — is new. And the ability to produce new types of information in real time is decidedly new and powerful.

Here's why I think that big data is worth the attention. Yes, it has the potential to deliver new types of information to both business users and consumers in real time. Beyond that, however, lies the promise of a style of computing that more closely mimics the functioning of the human mind as it takes in data from many different sources, forms thoughts and makes decisions in real time. For IT, that means moving from the provisioning of services to making a big impact on business results. ◆

Webster is a senior partner at Evaluator Group, a storage research firm. You can contact him at john@evaluatorgroup.com.

Big Data OVERLOAD

Complex requirements and relentless demands for capacity vex storage administrators. *Here's how to handle the data deluge.* **BY STACY COLLETT**

IT USED TO BE ONLY for scientists, Internet giants and the mega-social-media set — Amazon, Twitter, Facebook, Shutterfly. But now, more and more enterprises of all kinds are aiming to gain a competitive edge by tapping into big data in hopes of unearthing the valuable information it can hold. Today, companies such as Walmart, Campbell Soup, Pfizer, Merck and convenience store chain Wawa have big plans for their big data.

Some are venturing into big data analytics to respond to customers faster, keep better track of customer information or get new products to market quicker.

"Any business in this Internet Age, if they don't do it, their competition is going to do it," says Ashish Nadkarni, a storage analyst at IDC.

Organizations of all sizes are being inundated by data, from both internal and external sources. Much of that data is streaming in real time — and much of it is rendered obsolete in minutes, hours or a few days.

The resulting growth of storage needs is especially troubling for large enterprises, where the amount of structured and unstructured data requiring storage grew an average of 44% from 2010 to 2011, according to Aberdeen Group. At companies of all sizes, data storage requirements are doubling every 2.5 years. What's more, different tools are required to optimize the storage of video, spreadsheets, formatted databases and fully unstructured data.

"The challenge is to try to keep your spending on storage from being linear with your rising storage requirements," says Dick Csaplar, a virtualization and storage analyst at Aberdeen. Technologies that can help mainstream users of big data avoid that fate include storage virtualization, deduplication and storage tiering. For heavy-hitters, such as scientists, social media websites and simulation developers, object-oriented and relational database storage are the best options.

But the nuts and bolts of systems designed to hold petabytes (and more) of data in an easily accessible format are more complex than the inner workings of everyday storage platforms. Here's some expert advice on managing and storing big data.

WHAT KIND OF DATA ARE YOU ANALYZING?

The type of storage required depends on the type and amount of data you're analyzing. All data has a shelf life. A stock quote, for example, is only relevant for a minute or two before the price changes. A baseball score is sought after for about 24 hours, or until the next game. This type of data needs to reside in primary storage when it is most in demand and can then be moved to cheaper storage. A look at trends over multiple years supports the idea that data stored for long periods of time usually doesn't need to be on easily accessible primary drives.

HOW MUCH STORAGE DO YOU REALLY NEED?

The amount and type of storage you need for big data depends on both the amount of data you need to store and how long that data will remain useful.

There are three types of data involved in big data analytics, Nadkarni says. "It could be streaming data from multiple sources being sent to you literally every second, and your time slice is a few minutes before that data becomes old," he says. This kind of data includes updates on weather, traffic, trending topics on social networks and tweets about events around the world.

Big data can also include data at rest or data generated and controlled by the business for moderate use.

Streaming data requires only fast capture and analytics capabilities, Nadkarni says. "Once you've analyzed it, you don't need it anymore." But for data at rest or business-controlled data, "it is incumbent upon you to store it," he says.

WHAT TYPE OF STORAGE TOOLS WORK BEST?

For enterprises just starting to grapple with big data storage and analysis, industry watchers advocate storage virtualization to get all storage under one umbrella, deduplication to compress data and a tiered storage approach to ensure that the most valuable data is kept in the most easily accessible systems.

Storage virtualization provides an abstraction layer of software that hides physical devices from the user and allows all devices to be managed as a single pool. While server virtualization is a well-established component of today's IT infrastructures, storage virtualization has yet to catch on.

In a February 2012 Aberdeen survey of 106 large companies, only 20% of the respondents said that they have a single storage management application. The average was three management applications for 3.2 storage devices.

However, many storage vendors are reluctant to have their devices managed by another vendor's product. Storage virtualization is "much more complex [and] takes more time, so it hasn't caught on like server virtualization," Csaplar says. Instead, many storage administrators are looking at cloud-type implementations for third- or fourth-tier storage to move data more easily across different infrastructures and reduce storage costs. "Some companies have done it and gotten good results, but it's not a slam dunk," he adds.

Csaplar expects to see an increase in utilization of cloud-based storage and other cloud-based computing resources in the near future as network connectivity improves, costs decline and the ability to encrypt and decrypt data in flight improves. "With the cloud, you get a monthly bill paid out of the operational budget, not a separate capital budget," he says.

DEDUPLICATION AND COMPRESSION

Administrators can shrink the amount of storage needed with deduplication, which eliminates redundant data by using data compression tools that identify short repeated identical strings in individual files and store only a single copy of each.

How much can storage needs be reduced? In the Aberdeen survey, 13% of the respondents said they had reduced data by 50%, but a more likely figure for most enterprises would be a 30% to 50% reduction of highly repetitive, structured data, Csaplar says.

STORAGE TIERING

Once the business decides what data it wants to analyze, storage administrators can put the newest and most important data on the fastest and most reliable storage medium. As the data grows older, it can be moved to slower, cheaper storage. Systems that automate the storage tiering process are gaining ground, but they're still not widely used.

When developing storage levels, administrators must consider the storage technology, the speed of the device and the form of RAID needed to protect the data.

The standard answer to failover is replication, usually in the form of RAID arrays. But at massive scales, RAID can create more problems than it solves, says Neil Day, a senior vice president and CTO at Shutterfly, an online photo site that allows users to store an unlimited number of images at the original resolution. Storage has exceeded 30 petabytes of data.

> **The challenge** *is to try to keep your spending on storage from being linear with your rising storage requirements.*
>
> **DICK CSAPLAR,** ANALYST, ABERDEEN GROUP

In a traditional RAID data storage scheme, copies of each piece of data are mirrored and stored on the various disks of the array, ensuring integrity and availability. But that means a single piece of data stored and mirrored can inflate to require more than five times its size in storage. As the drives used in RAID arrays get larger — 3-terabyte drives are very attractive from a density and power consumption perspective — the time it takes to get a replacement for a failed drive back to full parity grows longer and longer.

Shutterfly eventually adopted erasure code technology, where a piece of data can be broken into chunks, each useless on its own, and dispersed to different disk drives or servers. At any time, the data can be fully reassembled with a fraction of the chunks, even if multiple chunks have been lost due to drive failures. In other words, you don't need to create multiple copies of data; a single instance can ensure data integrity and

MICHAEL MORGENSTERN ILLUSTRATION

Hadoop Isn't the Only Option Anymore

THE CONCEPT of "big data" has widened. The term once applied to intricate data that had to be available instantaneously for highly repetitive queries by power users such as scientists and social media websites. Now it includes the multiple petabytes of structured and unstructured data that most enterprises must store.

While the open-source systems Hadoop and Cassandra are the go-to big data options for hard-core data crunchers, some commercial vendors are ramping up their storage systems to handle multiple petabytes of data and offer quick, easy ways to analyze it.

"Big data used to be a tool that only the biggest [enterprises] could use, but now it's hard to find someone who's not using something to get insight from data," says Ed Walsh, vice president of marketing strategy for storage products at IBM. To do that, "you'd better have efficient storage, or the cost can get to you," he says. "You also have to have decent performance from these apps, which are very dynamic. And you'd better be able to back it up."

For its part, IBM for several years has been building a portfolio of high-performance storage and analytics products and technologies, including Hadoop. But in June of last year, it announced a "formal approach" to the way it markets its storage and analytics products, called IBM Smarter Storage. The company also announced its first offerings that incorporate software from its acquisition of Platform Computing earlier last year – all intended to help a broader set of enterprise customers.

"We did that because we do have a very complete portfolio, and sometimes it comes off as being too complex, so part of it is helping people see the holistic view," Walsh says. "It helps people understand what they're trying to do" with their data.

– STACY COLLETT

availability. Because erasure codes are software-based, the technology can be used with commodity hardware, bringing down the cost of scaling even more.

One of the early vendors of erasure-code-based software is Cleversafe, which has added location information to create what it calls dispersal coding, allowing users to store chunks — or slices, as it calls them — in geographically separate places, like multiple data centers.

■ MEGA-BIG-DATA USERS

Like Shutterfly, enterprises with massive storage needs must look beyond block storage, Nadkarni says. "When you're talking about massive data sets in the petabyte range, you have to look at either object-based storage or a distributed file system," he says. "Think about [commercial offerings like] EMC's Isilon scale-out storage or the Dell Fluid File System . . . and open-source solutions, as well. They're much cheaper to store data, and from a performance perspective, they can offer you a much better price/performance ratio. And ultimately, they're scalable."

Users of commercial software often have data that is partially disposable or has very little post-process requirements, he adds.

■ FEWER ADMINISTRATORS REQUIRED

When deployed correctly, storage virtualization, deduplication, storage tiering and erasure code technologies should reduce your need for administrators, because the tools enable you to manage data through a single pane of glass. In Shutterfly's case, the automated storage infrastructure allowed the company to slow the growth of its maintenance staff. As the company's daily maintenance workload declines, administrators can spend more time on proactive projects.

In some cases, big data projects are done by special teams, not traditional IT staff, Nadkarni says. "They're owned and operated by business units because the IT infrastructure is not agile enough to support big data environments or may not have the skill set for it."

"You might have a situation where storage administrators aren't involved," he adds. "Or they might just have a small role where [they provision] some storage and everything else is done by the systems folks."

■ COMING SOON

One trend Nadkarni sees catching on is the concept of moving the compute layer to the data. "You look at solutions from Cleversafe and solutions from other storage providers who are building compute capabilities in the storage layer," he says. "It is no longer feasible to move data to where the compute layer sits. It's practically impossible, especially if you only have a few minutes to analyze the data before it gets stale. So why don't I let the compute layer sit where the data sits?"

Cleversafe offers a high-end, Hadoop-based solution for big data heavy-hitters like Shutterfly, "but they're trying to make it more all-purpose," Nadkarni says. "Cleversafe breaks the model of procuring [compute power] from one vendor and app storage from another vendor." To be successful with mainstream enterprises, "business units will have to start thinking in different ways. I'm confident that it will eventually catch on, because the efficiencies in the current model just don't lend themselves to be favorable for big data."

He adds, "Big data is a way for people to maintain their competitive edge. In order to make the most out of their data, they're going to have to change processes and the way they function as a company — they're going to have to be very quick to derive the value from this data."

But before diving into a new big data storage infrastructure, "people have to do their homework," Csaplar says. "Research it and talk to people who've done it before. It's not cutting-edge, so talk to someone who has already done it so you don't make the same mistakes they've made." ◆

Collett *is a* Computerworld *contributing writer. You can contact her at stcollett@comcast.net.*

WHO HOLDS THE Keys?

Encryption isn't bulletproof if keys and digital rights are left out in the open. *Here's how to lock down stored data.* **BY STACY COLLETT**

ENCRYPTION can make up for a litany of security snafus — from a bad firewall to an unrelenting hacker to a lost laptop. Once data is encrypted, criminals can't use or sell it. Plus, if encrypted data goes missing, companies are protected from disclosure requirements in most states. No wonder 38% of companies surveyed by Forrester Research have already adopted full-disk encryption technology. But data protection doesn't stop there. Encryption keys and digital rights also must be well orchestrated and secured, or else encryption protection goes out the window.

For instance, encryption keys kept in a predictable place are like house keys left under a welcome mat: They're easy prey for intruders.

Last year, hacking group Anonymous broke into SpecialForces.com, a provider of law enforcement equipment, and stole thousands of customers' data and credit card numbers. The data was encrypted, so the crisis appeared to have been averted. But the hackers didn't stop there. They broke into the company's servers and stole the encryption keys. The group then leaked roughly 14,000 passwords and 8,000 credit card numbers of customers on its website.

"Most of the standardized encryption methods or algorithms specified by NIST are really good; it's just a fact of how you implement them and how you do key management," says John Kindervag, an analyst at Forrester Research.

While many companies have deployed full-disk encryption to meet regulatory compliance mandates or to avoid public disclosure requirements under state privacy laws if data is lost or stolen, an alarming number of companies still don't take precautions.

More than half of the 500 IT professionals surveyed by Ponemon Institute and Experian Information Solutions earlier in the year said their lost or stolen data wasn't encrypted. Lost data most often included email (70%), credit card or bank payment information (45%) and Social Security numbers (33%). If the organization was able to determine the cause of the breach, most often it was a negligent insider (34%). Some 19% said outsourcing data to a third party was to blame, and 16% said a malicious insider was the main cause.

"Any device that leaves your organization needs to be protected, and with more than just a password," says Gartner analyst Eric Ouellet. "We know you can jailbreak these things very easily." Data at rest must be protected, too, he adds. "Even mislabeling a tape [in storage] or not being able to find it is a disclosure event," unless the data is encrypted.

Semiconductor production equipment maker Applied Materials faces strict customer and legal requirements to protect information. The company, which operates in 25 countries, began rolling out full-disk and message encryption in late 2010 as part of a tech refresh of its 13,000 laptops. Today, 78% of laptops are encrypted, with only a few holdouts.

"The change has been positive all over the world," says Matthew Archibald, who serves as both chief information security officer and chief privacy officer at the Santa Clara, Calif.-based company. "On the engineering side, they believe anything slows [the system] down, so you have to go through different sets of communication and show them that it doesn't impact them in any way."

At Intel, 85% of laptops have full-disk encryption, but CISO Malcolm Harkins is already assessing the next big thing — self-encrypting hard drives, which will address encryption gaps when laptops are in standby, sleep or hibernate modes.

"As you're moving to products that are always on/instant on, if you've got a nine-hour battery life and it's always on standby, the data is not encrypted," Harkins says. "I also want to improve the user experience," he adds, referring to the passcodes and reboot times that users currently endure for encryption. "If I can do that, as well as potentially lower my cost of control, self-encrypting drives might be the answer."

> **I also want to improve the user experience.** *If I can do that, as well as potentially lower my cost of control, self-encrypting drives might be the answer.*
>
> **MALCOLM HARKINS,** CISO, INTEL

MICHAEL MORGENSTERN ILLUSTRATION

Key Management

While encrypting data is important, the keys that control the encryption and decryption processes are even more important because, well, data is useless without a key. And with so many programs and devices requiring encryption and individual key management, it's easy to see why keys can be mismanaged or why dangerous shortcuts are taken to manage them.

Today, most encryption solutions have their own built-in key managers that also create backups, "so at least you have some consistency," Ouellet says. "The key manager that comes with those solutions is probably good enough." But as the number of encryption solutions and keys grows in a company, centralized key management may be the answer.

A quarter of companies surveyed by Forrester have adopted centralized key management in some form, he adds, but that number will grow as interoperability standards take hold.

Open standards organization Oasis has developed a key management interoperability protocol (KMIP) as a standard within cryptographic systems. "This standard has been growing over the last few years and is replacing older standards," Ouellet explains. "The only catch is that while most organizations that provide cryptography want to support KMIP, they'll do it as a means to manage others' keys. They're not allowing others to manage their keys. It's kind of a chicken-and-egg thing," which will hold back adoption "unless the vendors start opening themselves up."

$184 MILLION TO $330 MILLION:

The average loss in a brand's value after a data breach.

SOURCE: EXPERIAN INFORMATION SOLUTIONS

Holder of the Keys

Analysts say to leave key management to the professionals. Kindervag advises IT staff to deploy an enterprise-quality key management program that understands key management in your company. "Don't try to build your own," he cautions. "Don't email keys back and forth to each other, and don't leverage things like Active Directory to store your keys."

Do keep the key management function in a segment of your network that is completely separate from the encrypted data, and protect it with features such as Layer 7 firewalls, IPS devices and strong access control, he adds. Only a few people who are designated to manage keys should have access to that segment of the network, and they should constantly monitor what is happening on the key management servers, such as who is seeking access.

In the near future, key management will be available in the cloud with a service provider that specializes in enterprise key management. "Traditional PKI vendors are moving in that direction," Kindervag says, and credit card payment processors are capable of expanding their key management technologies into intellectual property and custodial data areas.

Cloud key management is also "a big trend right now" for smaller organizations that don't feel comfortable owning and managing keys, Ouellet says. Cloud providers can create private virtualized environments for small businesses and manage the technology side.

The key to successful deployment of encryption, key management and digital rights is to make it easy.

"Spend quality time with self-installing packages," says Applied Materials' Archibald. "We have automated distribution of the software, and it's just a matter of having it enabled for the user. There are only two or three things an individual needs to do — set their pass phrase, sync that to their Windows login and reboot their machine." ◆

Collett *is a* Computerworld *contributing writer. You can contact her at stcollett@comcast.net.*

Proceed With Caution

WHILE ASSIGNING RIGHTS for viewing and editing documents seems like a good idea, it's not something that Gartner's Eric Ouellet recommends for organizations that need to keep documents for a long time.

"There are no standards for EDRM [enterprise digital rights management]," he explains. If a vendor changes the cryptography or the way it applies the technology, users must upgrade or retrofit all existing documents or run the risk of having orphaned documents that no one can open. One of Gartner's clients had to upgrade twice over the past eight years, he adds.

"If documents are only going to live for 12 to 18 months, that's a risk window that you can manage," he says. "But if the documents need to live for four to five years or more, then you have to start building alternate systems," such as ones for keeping copies in plain text that are accessible to only one or two people in the organization.

– STACY COLLETT