VIRTUALIZATION
CLOUD
APPLICATION DEVELOPMENT
HEALTH IT
NETWORKING
STORAGE ARCHITECTURE
DATA CENTER MANAGEMENT
BI/APPLICATIONS
DISASTER RECOVERY/COMPLIANCE
SECURITY

Handbook

In the Cloud, Big Data's a Big Deal

Big data analytics in the cloud may solve some problems, such as reducing data center infrastructure costs. But does it introduce new ones as well?

TechTarget

Big Data: More Manageable in the Cloud?

When cloud computing came on the scene, it marched in step with other trends that have begun to affect data center infrastructure. Big data is one such trend.

Still murky, some describe the concept of big data as data so voluminous and, often, unstructured that a single department can no longer effectively manage it on its own. Others characterize it as data that requires a series of complex number-crunching processes to enable real-time decision making.

However you characterize the concept, cloud computing has eased some challenges associated with big data by offering scalable resources that companies can access in the cloud but that could reduce infrastructure costs.

Still, big data is widely recognized as complex and difficult. As Alex Barrett notes, even if you farm out the infrastructure required to support big data, companies are struggling with having the in-house skills required to work with it. Other companies find that to get the scale and performance they seek, they still need complex infrastructure such as Hadoop or Cassandra to get the job done.

As Beth Pariseau observes, large companies with extensive virtualization landscapes may find that their environments don't work well with big data approaches. And finally, Pariseau talks about some important big data concepts as well as the impact of the shift on traditional data center organization in her interview with Medio's Brian Lent and Ivan Sucharski. ■

Lauren Horwitz
Executive Editor
Data Center and Virtualization Media Group

When Cloud Computing Meets Big Data

For years, IT organizations have generated, collected and stored vast amounts of data. Now, IT is being asked not just to store it but to provide the infrastructure to perform analytics on this data. The trouble is that the task is a resource-intensive proposition. For organizations that don't have idle servers to throw at big data workloads, could tapping into the public cloud be a valid alternative to maintaining costly internal infrastructure?

When it comes to processing big data, the public cloud has a lot going for it. Public clouds are pay per use, so they are a good fit with finite big data workloads. Further, many big data analytics jobs can be easily "parallelized"—that is chopped up into smaller, more discrete tasks—which maps nicely to public cloud. And many public cloud providers offer templates for popular big data platforms such as Hadoop, making it easier for administrators to set up the requisite infrastructure.

For some big data analytics workloads, going to the cloud is the only valid scalable solution. Medio, a real-time analytics firm, rearchitected its software for multi-tenancy, tapping into the scale of Amazon Web Services to complement its own data center.

"One billion data events hit our data borg every day; we have 15 million monthly unique visitors," said Rob Lilleness, Medio's CEO. In Hadoop running on Amazon Web Services, Medio encountered a platform that could handle the scale that the company sought.

Even for modest workloads, public cloud's pay-per-use model is appealing. "The cloud is good at spin-up and spin-down," said Frances Guida, Hewlett-Packard Co.'s manager for cloud solutions in its enterprise group. That's a nice fit with big data. "A lot of analytics aren't predictable, and when you answer the question, you don't necessarily need the infrastructure again," she said.

Nor does big data in the cloud need to be an all-or-nothing proposition; some organizations see value in taking a hybrid approach. Archimedes Inc., a medical simulation firm in San Francisco, manages a private Hadoop cluster for data processing with the help of Univa Grid Engine software, but runs its front end on Amazon's cloud. "We could have run [Hadoop] on AWS as well, but when we calculated the cost, we figured out that if we could keep the hardware busy 30% to 40% of the time, it was cheaper to run it in-house," said Katrina Montinola, vice president of engineering.

While using the cloud for big data may seem like an obvious solution, there are caveats: security, for one, but also physical constraints concerning data movement.

But while using the cloud for big data problems may seem like an obvious solution, there are a lot of caveats: security, for one, but also physical constraints concerning data movement and latency. Even more daunting is the lack of trained professionals who know how to pose business questions of the data in a meaningful way. While the latter problems can be addressed with time, money and technology, data science skills are harder to come by—and are certainly not the domain of your average IT administrator.

SAAS TO THE RESCUE

Arguably, cloud's biggest contribution to solving the big data problem is the number of analytics vendors that have adopted the Software as a Service (SaaS) model. IT departments not only don't need to buy infrastructure but also don't need to set anything up.

That's been a powerful selling point for Emcien Corp., which offers its pattern detection software as a service running on Amazon Elastic Compute Cloud and counts large retailers, telecommunications providers and intelligence agencies among its customers. "The IT user doesn't need to buy all that hardware. All you need is a Web browser and you're in business," said Radhika Subramanian, Emcien CEO.

Big data platforms are notoriously complex, said Ryan Sousa, senior vice president of engineering at Medio. "To get to the scale and cost-effectiveness from analytics, you need Hadoop and Cassandra and other foundational components to get to the necessary size, throughput and performance," he said. "Building out that framework can be really challenging, and it's rarely in-house IT's core business," he said.

Even longtime business analytics users are considering a possible switch from on-premises to cloud. Janet Grimsley is vice president and information specialist at The Fauquier Bank in Warrenton, Va., and uses on-premises information optimization and visualization tools from Datawatch Corp., which recently began offering its tools as a service. "We would entertain using it as a service," she said. "It's just a cost issue for us."

As such, the emergence of the SaaS model could close a longstanding rift between business users and IT, said Rod Smith, vice president of emerging Internet technologies at IBM, which now offers a SaaS version of its Social Media Analytics tool. Big data analytics allows line-of-business users to look for insight and follow hunches, but historically, "IT can't plan for a hunch," Smith said. Now, with SaaS-based analytics tools, "line-of-business [users can say] 'I can move as quickly as I'm willing to spend money.'" IT, in turn, can "help line-of-business follow its hunches."

BIG DATA, BIG PROBLEMS

But SaaS-based analytics can take you only so far. Emcien, for instance, used to run its software exclusively in the cloud, but recently, has started to offer a virtual appliance version of its software that customers can run in-house. "The cloud is a fantastic resource for us because of its scalability and because it is extremely

"The emergence of the SaaS model could close a longstanding rift between business users and IT."

*—**ROD SMITH,** vice president of emerging Internet technologies, IBM*

economical," said Subramanian. However, as data sets grow larger, some customers have balked at the prospect of moving those sets to and from the cloud. The company routinely does evaluations on sample data sets of 60 TB, and production data sets soar into the hundreds of terabytes or even petabytes, she said.

Even if the data sets aren't that large, convincing organizations to put them in a cloud can be tough, conceded Medio's Lilleness. The first two generations of Medio's product ran on-premises because "people wanted their customer data inside their own data center—they had a greater degree of comfort with that." And when Medio relaunched in the cloud, "people were really resistant to moving their data there."

But even traditional on-premises technologies are being retooled for the cloud. HP's CloudSystem converged infrastructure platform, for example, features the same cloud management layer as the company's public cloud offerings, making it simpler for organizations to "burst" to the cloud to meet peak demand, said Margaret Dawson, HP vice president of product marketing and cloud evangelist. At the same time, the company plans to announce deeper integration between big data tools such as its Vertica database and its cloud offerings.

Even on the security front, resistance is softening, said Lilleness, and traditional offerings from business intelligence and data warehousing vendors like IBM Netezza, Oracle and Teradata are seeing declining adoption, he claimed.

"It's too expensive. Customers can't afford to have enough capacity to analyze the explosion of data that we're generating," he said. "The cloud is winning out." —*Alex Barrett*

Big Data, Big Changes

IT PROS CALLED in on big data projects are finding that the typical approach doesn't play nice on enterprise-grade virtualized infrastructure.

Brace yourself for big data. If it hasn't already hit your data center, it will soon, putting new demands on IT infrastructure and operations.

Big data analytics are used by sites like eHarmony to bring couples together, by retailers to predict customers' buying behavior, and even by healthcare organizations to predict a person's lifespan and future ailments. It's a little bit Big Brother, but it's also revolutionizing the way computing is used to interpret and influence human behavior.

Big data isn't just data growth, nor is it a single technology; rather, it's a set of processes and technologies that can crunch through substantial data sets quickly to make complex, often real-time decisions.

"It really is the future of what health care should be, using predictive analytics to improve treatment," said Michael Passe, storage architect for Beth Israel Deaconess Medical Center (BIDMC) in Boston. "It could be really big anywhere you might want a kind of crystal ball."

Sounds good, but IT professionals involved with big data initiatives may find that the new plans contradict the last decade's worth of virtualization and consolidation efforts in the data center.

AN INFLUX OF COMMODITY SYSTEMS

Generally, big data analytics requires an infrastructure that spreads storage and compute power over many nodes to deliver near-instantaneous results to complex queries.

The most commonly used platform for big data analytics is the open source Apache Hadoop, which uses the Hadoop Distributed File System (HDFS) to manage storage. Distributed

databases, including NoSQL or Cassandra, are also commonly associated with big data projects.

These are relatively new technologies, and as such, come with some maturity problems.

For example, HDFS does not natively incorporate certain tenets of storage design that have become gospel to storage managers over the years: archive, backup, snapshot and high availability, said John Webster, senior partner for the Evaluator Group, based in Boulder, Colo.

"Experienced Hadoop users tend to work for social media companies, and they're coming at this with the idea that storage is dumb disk, where you throw in a node and pound I/O against it," Webster said. "All the storage intelligence developed over the last two decades, it's like it doesn't exist."

And regulatory compliance? "Forget it," Webster said. "There's no way to lock down a file."

Furthermore, Hadoop is most commonly deployed on a cluster of physical servers in which the storage network and compute network are one and the same, often leaving enterprise storage and infrastructure pros with another separate, physical infrastructure to manage.

At Mazda North America, headquartered in Irvine, Calif., the servers are 90% virtualized, and infrastructure architect Barry Blakeley is working to push that ratio higher.

In the meantime, however, at least one of Mazda's business units is considering big data projects using QlikView or SAP BusinessObjects, or some combination of the two, much of which requires physical servers with direct-attached storage (DAS).

"I'm trying to virtualize, and here we are putting in physical servers," Blakeley said.

That translates to management headaches. Separate environments and silos of data mean "a lot of dashboards, and there are so few of us it becomes unwieldy to manage it all on separate devices," he added.

INTO THE BIG DATA FOLD

Projects are afoot to counteract this trend, most notably VMware Inc.'s Project Serengeti. Like Hadoop itself, Serengeti is an Apache Software Foundation open source project.

The purpose of the project is to produce a freely downloadable offering that "enables rapid

deployment of standardized Apache Hadoop clusters on an existent virtual platform, using spare machine cycles, with no need to purchase additional hardware or software," according to a VMware blog post.

A virtualized Hadoop cluster can take advantage of VMware's native high availability and fault tolerance capabilities for availability as well, protecting critical components such as the HDFS NameNode, which keeps track of all the files in the file system and is a single point of failure. Hadoop does not yet natively offer high availability for the NameNode, which is another fingernails-on-a-chalkboard feeling for enterprise infrastructure admins, particularly failure-conscious storage pros.

Other vendors, including Symantec Corp. and Red Hat Inc., propose replacing HDFS with their own scale-out file systems: Clustered File System and the Gluster File System, respectively. These more mature file systems offer capabilities like snapshots and high availability.

At least one centralized storage vendor claims native integration with HDFS to solve its high-availability challenges—EMC Corp.'s Isilon scale-out network-attached storage (NAS) system. Incorporating HDFS into Isilon means providing Hadoop users with built-in data protection, greater storage efficiency and better performance than physical clusters built on DAS, claims EMC, in addition to eliminating single points of failure.

BIDMC uses Isilon storage to explore big data analytics for use in its clinical practice, since the hospital has already purchased Isilon hardware for other purposes.

"I want to use the infrastructure because it's not a RadioShack science kit; it's purpose-built to do this kind of thing and it does it very well," Passe said. "Why would you want some generic thing with its own disks and a higher failure rate if you've already got Isilon in place?"

That's the plan, at least. At the moment, however, as the clinical practice experiments with Microsoft's SQL and Hadoop integration, called HDInsight, the software is still running on a separate physical cluster. Nor is centralizing storage the only issue with integrating big data into BIDMC's IT practice, Passe said—Microsoft hasn't fully integrated Active Directory with HDInsight yet, something the hospital is waiting for before proceeding.

"We're just starting to figure out how to use it and what makes sense for us, and then trying to figure out how we best posture ourselves from an infrastructure standpoint to support it," Passe said.

THE TROUBLE WITH VIRTUALIZING HADOOP

Still, some analysts say virtualization-centric solutions to the big data infrastructure problem pose their own challenges.

Virtualized Hadoop may work as advertised, but in terms of licensing and system costs, enterprises may find it's still cheaper to go with commodity, scale-out DAS for big data projects.

Virtualization management also isn't ideally suited to managing virtualized big data clusters yet, according to Jeff Boles, senior analyst for the Taneja Group, based in Hopkinton, Mass.

"We'll see some convergence, with virtualization vendors fighting their way back with solutions that allow you to virtualize all this stuff, but you still don't necessarily want to mix that into your main infrastructure pool," Boles said.

Meanwhile, according to Webster, "purists will say replacing the file system or using something like Isilon is too expensive. Using scale-out storage separate from Hadoop nodes can also add another network to the cluster, increasing complexity," he said.

As a result, some companies are considering external public clouds as an alternative to rolling out a separate infrastructure for big data within a data center, sidestepping the split-infrastructure problem altogether. That approach has the added bonus of enabling the sharing of data sets and analytical results with business or research partners if necessary. Cloud service providers such as Medio and Amazon Web Services have been offering such big data services for years.

But doing big data analytics in the cloud can also raise some of the same compliance and governance challenges enterprises are already dealing with when it comes to Infrastructure as a Service options, analysts say.

And sidestepping an internal infrastructure may also mean sidestepping IT altogether, resulting in shadow IT deployed on public cloud vendors' infrastructures that IT doesn't know

about, said Webster.

Even if the public cloud is used with the blessing of IT, "Whose data is it?" Webster asked. "And if data is covered under compliance of one sort or another, is the service provider going to cover you?"

TO BE CONTINUED

Eventually, companies like Intel Corp. predict that the scale-out infrastructures associated with big data and the centralized virtual infrastructures popular over the past decade will converge into what's becoming known as the software-defined data center.

"More and more companies are realizing there's a lot of value in the data they have that they're not taking advantage of," said Christie Rice, marketing director for Intel's storage division. "In time it will become a necessary thing if a lot of companies want to be able to stay in business and ... to expand the business."

Rice predicts that in the long run, the software-defined data center will commoditize hardware so that any friction between centralized storage systems and scale-out DAS becomes irrelevant—software, whether for compute, networking or storage, could allow server workloads to change on demand.

"We also see solid-state drives being used more and more, particularly as you're talking about real-time analytics—being able to get data in and out of the storage media faster becomes more important," Rice said.

For now, big data projects remain confined to a small niche of the enterprise—maybe 3% to 5% of companies, estimated Taneja Group's Boles. However, he expects that number to double in the next year and a half to two years, and for there to be an eventual "trickle-down effect" from the largest of Web and enterprise entities to small and medium-sized organizations.

"We're more serious about analytics than ever before and it's easier to deploy an analytics solution than ever before," he said. "That makes it practical for a whole new range of companies." —*Beth Pariseau*

Cloud Computing Brings Big Data to the Masses

As cloud computing evolves, one use for it is beginning to stand out: big data. What exactly big data means and how it fits into the cloud conversation, however, are questions not easily answered.

Founded in 2004 as an application service provider, Medio has recently reinvented itself as a cloud-hosted provider of big data analytics, with a focus on mobile platforms. The company is also marketing its inGenius Software as a Service, traditionally aimed at enterprise customers such as T-Mobile, Verizon, Disney and CBS, to small and medium-sized businesses. Brian Lent, co-founder and chief technology officer of Medio, and Ivan Sucharski, Medio's data strategist, discussed big data, the evolution of cloud computing and how the two trends affect each other.

How would you define big data?

Lent: I would define it as data on [such] a scale that you can't have a single department effectively manage it. Once you turn data into an operational process where you say, "The data is going to drive our commerce engine and our recommendations and our churn modeling and our financial forecast, now it becomes a different kind of asset, where you need to have the analytics passage sitting next to data, colocated because of the volume and the transfer. Once data is seen as a profit center, and hence there's a utility, then I think it moves more into the big data realm.

What distinguishes big data and cloud computing from its predecessors? Or is there a distinction?

Lent: As an application service provider, we would license software with a traditional approach, where you might do a license agreement for three years. Now what we're doing with the cloud-based approach is a license fee

per month. It's much more measured on volume, based on monthly active users. On the technical front, I think there are a lot of differences, but one would be us making available a lot of our services through REST-based APIs [application programming interfaces].

On the big data side, the fundamental difference is volume-based. So, when some of our customers get into petabytes, that puts you in that big data camp. The other aspect of big data is the velocity at which data is changing. One of our newer customers is Rovio, the makers of [the video game] *Angry Birds*. We're seeing more than 1.4 billion logging events in our cloud per day [in total]. They could launch a new app that could quadruple the traffic instantly.

How does big data fit into a cloud computing discussion, and vice versa?

LENT: The combination of cloud computing and big data is going to become more practical, simply because of the efficiencies of scale. The ability for any one company to keep up from an IT perspective, I think, is difficult. With big data typically come folks that know how to work on that data, and that may be the biggest gap—you can't just hire this talent. A friend that used to be at Google talked about data scientists as "pink unicorns" and said that there's only about 120 pink unicorns in the world: true data scientists that know how to manipulate and work with big data, delivering business value. So if you think about that as a commodity and a limited resource, the question becomes, "How do you centralize that into a cloud-based environment so everyone can get the value, but you don't have to have that person on-premises?"

SUCHARSKI: There's elasticity, too. When you're running analytical models, you want them to run as fast as possible, but you don't need to run them every 10 minutes. So you need as many machines as you can get for the next half hour. And then they're idle for 23.5 hours, until the next computing cycle. In a cloud situation, you've got the flexibility of that elasticity without the cost of being offline 95% of the time.

What is the most misunderstood thing about big data right now?

LENT: The notion that it's just about the size of data. The reality is, big data as a term is morphing to describe the complexities of data, and also how data is used in the enterprise. I think there's a connotation with big data that you're going to find operational uses for that data versus just storing it.

How do you see big data and the cloud evolving in the future?

LENT: You'll see the chief financial officer get engaged more into these types of decisions where they haven't to date, and get more engaged with the chief marketing officer (CMO) and the chief information officer (CIO). I think you'll see the cloud decision making moved to the CMO from the CIO. So rather than it being an IT artifact, big data and the cloud will be part of the core decision making when they go to roll out a new product.

SUCHARSKI: Many organizations are just at the baby steps of collecting the appropriate data. Turning data into a commodity will create a new generation of individuals who are interested in what I care about, which is quality, which means that the overall quality of information will jump. Today you're definitely mining. You're digging through a lot of garbage to glean small insights. —*Beth Pariseau*

ABOUT THE AUTHORS

ALEX BARRETT *is the editor in chief of* Modern Infrastructure *magazine.*

BETH PARISEAU *is a senior news writer for SearchCloud Computing.com and SearchServerVirtualization.com. Write to her at bpariseau@techtarget.com or follow her @PariseauTT on Twitter.*

In the Cloud, Big Data's a Big Deal
is a SearchCloudComputing.com e-publication.

Margie Semilof | *Editorial Director*

Lauren Horwitz | *Executive Editor*

Phil Sweeney | *Managing Editor*

Eugene Demaitre | *Associate Managing Editor*

Laura Aberle | *Associate Features Editor*

Linda Koury | *Director of Online Design*

Neva Maniscalco | *Graphic Designer*

Rebecca Kitchens | *Publisher*
rkitchens@techtarget.com

TechTarget
275 Grove Street, Newton, MA 02466
www.techtarget.com