CIO
Custom Solutions Group

eBOOK

Why BIG DATA is a BIG DEAL

Effectively mining big data with tools such as Hadoop can deliver real business benefits.

ENTERPRISES are generating, collecting and storing data at an astounding rate, in large part because they're collecting data from more sources than ever. Newer sources include social media streams; sensors of various types; and even call centers, which generate a seemingly never-ending stream of audio files. And, of course, enterprises still have all the traditional sorts of data they have long produced, of both a historical and transactional nature. It all adds up to big data.

The challenge is for enterprises to turn that big data into valuable information by mining it to find useful nuggets or analyzing it in new ways to answer questions and make predictions that previously were simply not possible. More and more, enterprises are finding that they can indeed extract value from big data by using a tool that makes the chore feasible: Hadoop, a platform for processing, storing and analyzing large volumes of unstructured data.

Volume and velocity

Large volumes of data are exactly what organizations are dealing with. Oracle last year estimated that data was growing at a 40 percent compound annual rate and would reach 45 zettabytes (ZB) by 2020[1]. One ZB is about one thousand exabytes, or a billion terabytes.

When you consider the sources of all that data, it's not hard to see how it can quickly add up. The sources include feeds from social media sites such Facebook, which can be heavy on photographs; Twitter; and even YouTube videos. Increasingly, companies are also storing more audio from their call centers, in hopes of mining it for tidbits that can help them improve customer service, sales and operational efficiency. There's also video from surveillance cameras.

All of these sources produce unstructured data—data that doesn't come neatly wrapped in rows and columns like that in a relational database. That makes the data far more difficult to analyze in meaningful ways. In fact, it was previously not economically feasible to collect and store all this unstructured data, never mind analyze it.

Hadoop is changing that equation. It is an open source framework that makes it possible to store huge volumes of data on many commodity computers, so there's no need for expensive, massive data stores. What's more, by dividing up processing chores into smaller chunks that can run simultaneously, Hadoop supports dramatically increased data analytics speeds (see "Hadoop FAQ").

Big Data + Hadoop

With Hadoop it's now possible to effectively query big data sources to find trends and other valuable business information.

VIDEO: **Big Data and other megatrends**

Oracle last year estimated that data was growing at a 40 percent compound annual rate and would reach **45 zettabytes** (ZB) **by 2020**[1].

Consider, for example, how an insurance company is using Hadoop to improve service in its telephone call center. The company takes all the audio files from its call center and uses Hadoop to analyze them in search of ways to route calls more efficiently so callers will more quickly get to an agent who can address their issues.

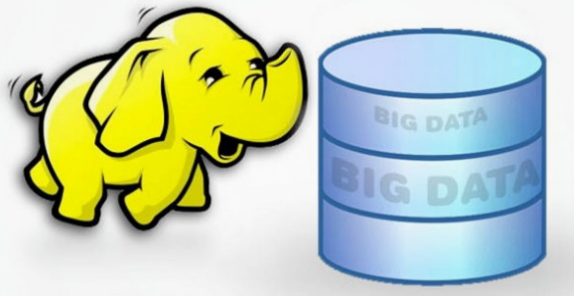
The company is also analyzing social media sites in an effort to improve service. If there's a storm in the Northeast, for example, it will search to find out if any customers have posted about damage to their home or automobile. If the company finds a hit, it will proactively reach out to help those customers. Finding such proverbial needles in haystacks was simply not feasible or cost-effective prior to Hadoop.

Companies such as credit card providers are also finding Hadoop valuable in offering promotions to customers going about their daily routine. For example, a credit card company might see transactions coming in from stores where a customer is shopping at midday on a Saturday. The company could push out a promotion to the customer's mobile device from nearby restaurants that accept the credit card. Such highly targeted, location-based, time-sensitive offers are relatively likely to meet with success—and may not even be feasible to deliver without Hadoop.

[1] SOURCE: A.T. Kearney, "Big Data and the Creative Destruction of Today's Business Models"

HADOOP FAQ

WHAT IS HADOOP?



Hadoop is a project under the auspices of Apache, the open source software development group. As defined on the Apache Hadoop Website[2]:

"The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

"The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures."

What can we do with Hadoop?

Hadoop is designed to enable businesses to query extremely large data sets to find trends or answer specific questions. It essentially takes batch processes and enables them to be performed in a parallel processing environment, thus greatly increasing performance.

What this means in practice is open to the imagination of implementers. Examples include healthcare firms collecting detailed data on patient diet, exercise patterns and more to develop individualized programs and promote healthier lifestyles. A retail company used Hadoop to reduce the cost and improve the performance of a price promotion analysis job that required some 10 steps. The company identified the two or three longest, most expensive steps from a processing perspective and moved them to Hadoop, resulting in a cost reduction of pennies on the dollar in terms of processing while significantly speeding up the process.

Companies are also using Hadoop to provide insight into product development, by synthesizing customer opinions from various sources and collecting massive amounts of data from sensors to improve operations in everything from manufacturing to delivery services.[3]

Who uses Hadoop?

Many organizations— including Amazon, eBay, Facebook, Hulu, *The New York Times* and Twitter—use Hadoop. You can see a partial list **here**.

How do you implement Hadoop?

Hadoop is an open source software framework, not an off-the-shelf product. As such, it is relatively bare-bones. Implementing the open source version requires becoming familiar with some related tools—such as Pig, a programming language; YARN, a framework for job scheduling and cluster resource management; and MapReduce, a programming paradigm for parallel processing of large data sets—that perform various functions.

As with any other open source software, many companies will likely opt for commercial implementations that are easier to accomplish and commercial tools that aid in operations, including integration and automated processing, such as BMC Control-M for Hadoop (see "Connecting Hadoop to the Enterprise").

[2] SOURCE: http://hadoop.apache.org

[3] SOURCE: *The Wall Street Journal*, "How Big Data Is Changing the Whole Equation for Business," March 10, 2013

VIDEO: **Keeping it simple with Hadoop**

Hadoop is not an island

To be most effective, Hadoop has to be tightly integrated with other enterprise data sources and applications, whether social media or conventional corporate information. Similarly, Hadoop often has to take the results it produces and feed them to other databases or applications, so companies need to be able to manage Hadoop in a way that makes it as easy as possible to control all these interactions.

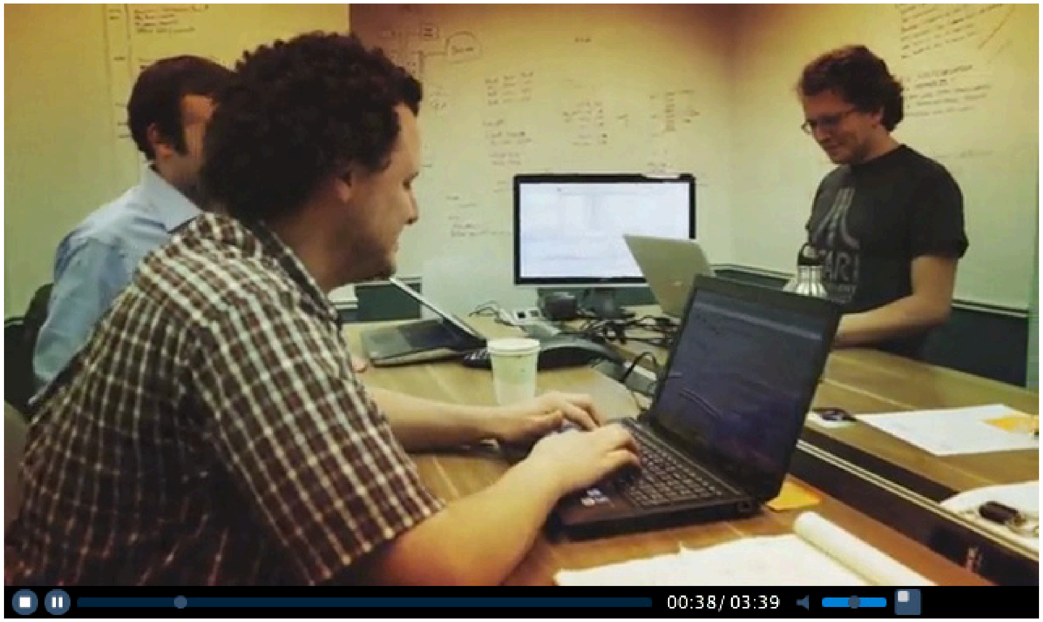
MetaScale is a company that understands this challenge well. The company was born out of Sears Holdings, where the IT group used Hadoop extensively to augment its own data warehouse initiatives and deal with big data challenges. Sears Holdings gained enough expertise with Hadoop that it eventually formed MetaScale to help other companies implement the platform.

To help in that effort, MetaScale employs BMC Software's BMC Control-M for Hadoop, a workload automation solution that simplifies and automates Hadoop batch processing and connected enterprise workflows.

"Big data technology is not an island," says Scott LaCosse, head of technology operations for MetaScale. "It is a hub that has a lot of spokes, and Control-M provides us that control point to integrate all those spokes. It's a key component for success, because in large, complex environments, if you don't have that control point, you're going to have chaos."

> "Control-M…gives us clear visibility into all the processing that you have to do. So it's an efficiency and a productivity gain for us."
>
> —Scott LaCosse
> head of technology operations
> MetaScale

BMC Control-M for Hadoop brings order to the chaos by enabling companies to automate the various workflows involved in a typical Hadoop process, including processes that span Hadoop and other traditional IT environments, such as mainframes, servers and databases. BMC Control-M for Hadoop also greatly simplifies the task of writing the custom workflows that are at the heart of much of Hadoop processing. With a simple GUI, building workflows becomes a simple drag-and-drop process (see "Connecting Hadoop to the Enterprise") and obviates the need to write scripts.



VIDEO: **ChipRewards: Driving healthy outcomes with Big Data**

Extract maximum value from Big Data

Clearly there's great opportunity to gain business value from big data, but it takes appropriate tools to realize that opportunity. Hadoop is one such tool that is proving itself up to the task, but by itself, Hadoop is no panacea. Rather, Hadoop must be integrated with other enterprise data sources and applications that both feed data into the Hadoop platform and can accept results from it. This introduces complexities—in terms of managing job scheduling, workflows, custom application development and more—that companies must deal with.

BMC Control-M has been handling those complexities and functions for years for all sorts of enterprise data sources. Now BMC Control-M for Hadoop extends the functionality to this important big data platform.

In the following pages, you'll learn more about Hadoop and how to use it effectively to extract maximum value from your big data environment. You can also visit **www.bmc.com/hadoop** to learn more about BMC Control-M for Hadoop. ■

HADOOP PROJECT

SIX CONSIDERATIONS FOR A HADOOP PROOF-OF-CONCEPT PROJECT

1. On-premises or in the cloud?

One of the earliest Hadoop-related decisions you'll have to make is whether to implement it on your own premises or somebody else's. Companies such as Amazon, Rackspace and Savvis all have hosted Hadoop options, sometimes with flavors of Hadoop to which they've added value.

2. Which Hadoop distribution?

Hadoop is open source software so, much like Linux, it is offered in various flavors, or distributions. You can go to Apache and download the source code, but you will quickly realize that you'll also need tools from several other, related Apache projects with names such as MapReduce, Hive, YARN, ZooKeeper and Pig. All of these projects are at different revisions, so you'll have to figure out what runs with what.

If you'd rather not deal with all that, you can get a commercial distribution that comes with all the required tools packaged up, like an off-the-shelf software product. Major distributions come from companies such as Cloudera, Hortonworks, MapR, IBM, Intel, Microsoft, HP and Dell.

3. Consider your infrastructure

If you decide to host Hadoop on your own premises, you'll need to think about the supporting infrastructure. One consideration is whether to host on virtual machines or not. If you're already highly virtualized and can stand up resources quickly, perhaps that's a good choice. If you're anticipating high volume and need the best-possible performance, dedicated hardware may be the better option.

Also consider the load Hadoop will put on your network infrastructure and make sure it can accommodate the expected traffic. No matter how much bandwidth you've currently got, adding Hadoop will have an impact.

4. Think about complementary tools

Like any other relatively sophisticated IT environment, your organization will need some tools to help make the Hadoop implementation manageable. They include some kind of monitoring tool and perhaps a configuration solution that will enable you to build clusters and dynamically add nodes. If you need to worry about compliance and auditing, a tool to help with patching, along with one to help restrict access to certain groups or individuals, may be important.

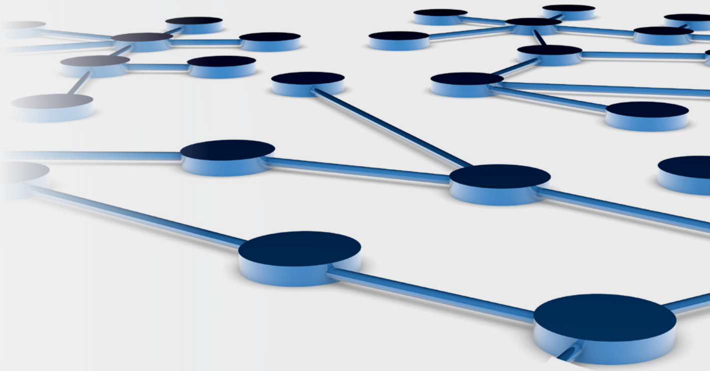
5. How will Hadoop fit in?

Few organizations run Hadoop in isolation from the rest of their IT environment, so think about what data you'll need to pull into Hadoop, how you're going to access it and whether it'll integrate natively. A tool that helps automate integration processes, such as BMC Control-M for Hadoop, may also prove valuable.

6. Consider required apps and skills

Finally, you need to think about the applications you'll be building to take advantage of Hadoop and whether you have the required skills in-house. Rarely do companies have developers on board who are already familiar with Hadoop tools such as MapReduce and the Pig programming language, so you'll likely need to devote time to learning those tools. Once again, a tool such as BMC Control-M for Hadoop adds value by simplifying the application development process by eliminating the need for scripting, freeing up time for developers to hone new Hadoop skills.

CONNECTING HADOOP TO THE ENTERPRISE



Hadoop, like any other new technology, does not live in a world all its own. To be truly valuable, it must integrate and interact with lots of other components in the IT environment, a process that can get tricky without a proper toolset.

First, it's helpful to understand how Hadoop works. For the most part, it's a platform meant to process large amounts of data that is not interactive in nature but, rather, mostly in batch form. In a batch environment, processing jobs typically happen sequentially, one after another.

Often one job has to complete before another begins. To automate this process, developers write scripts—often many of them—if an application entails numerous steps. This can be a tedious process that adds a significant amount of time, as much as 30 percent, to the application development process. Given that developers who are capable of writing for Hadoop are scarce to begin with, that's not a good use of their time.

So an important tool is one that can eliminate the need for all that scripting. BMC Control-M for Hadoop, for example, has a graphical user interface that reduces the scripting process to a simple point-and-click endeavor.

Another important consideration is managing the process of shifting data from one source to another, often using an extract, transform and load (ETL) process. Say, for example, you have a data warehouse and want to pull some data from it for processing in Hadoop. You may use a tool—say Informatica or DataStage—specific to the type of data warehouse. Then you may need to massage the data to get it into a format that is readable by the next application or database in the process. Each step of the way, you need to move the data, usually via a file transfer, until it finally reaches the Hadoop environment.

Once you've run your Hadoop application, you have to take the resulting data and push it somewhere else, likely to another application or a database.

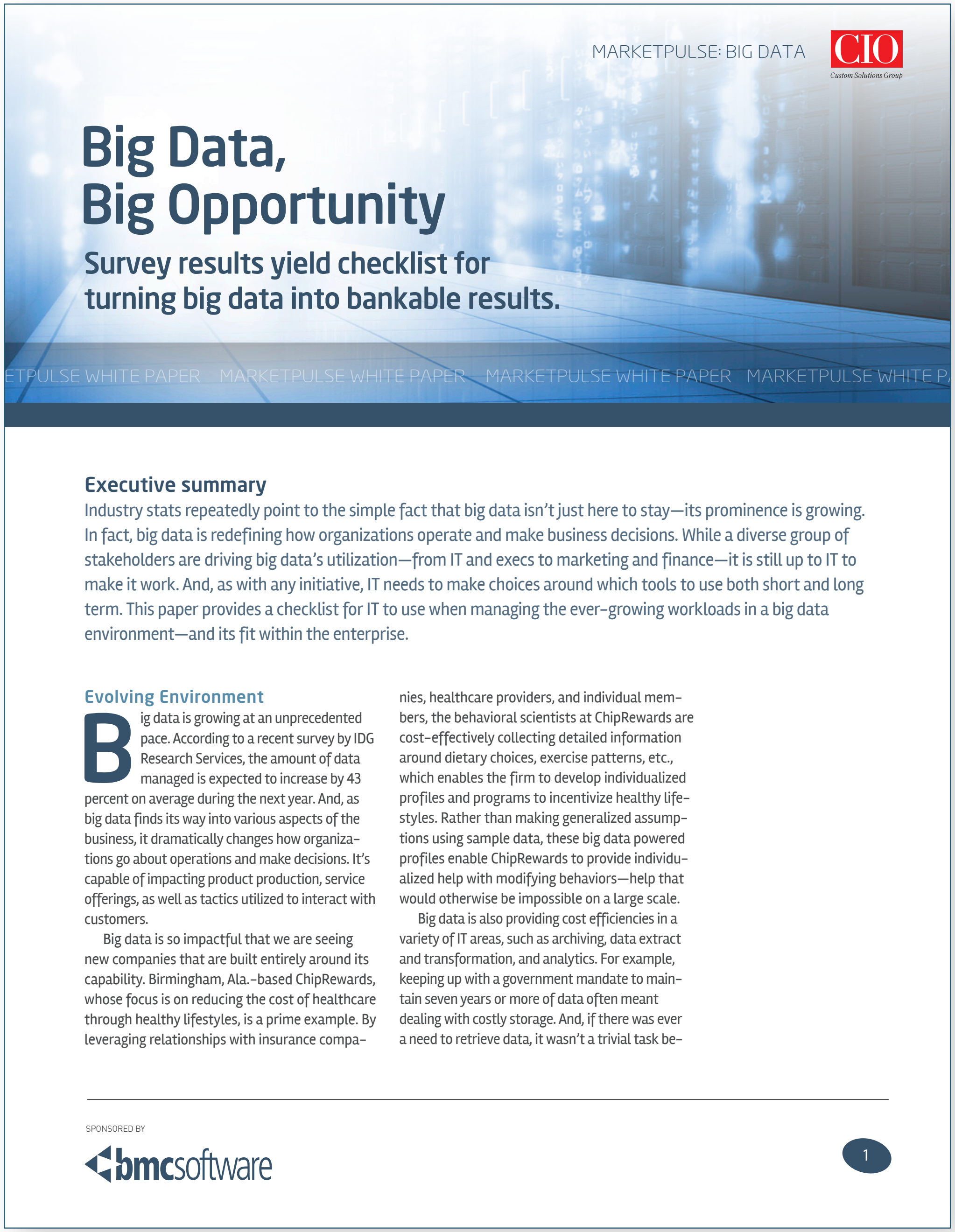
It may take four or five hops to complete this process. To manage and orchestrate each step, you need to be familiar with many tools, such as SQL or some other language; do file transfers; use Hadoop; and then do more file transfers and use whatever application will deal with the final data from Hadoop, such as a business intelligence application.

BMC Control-M has been managing these sorts of processes in traditional environments for years. It deals with all the job scheduling required to ensure that batch processes run smoothly while automating the scripting process. It also simplifies the multistep ETL and file transfer processes required to get data from one source to another. Rather than forcing users to learn multiple tools and languages to complete the process, they can use just one: BMC Control-M.

BMC Control-M for Hadoop extends BMC Control-M's benefits to the Hadoop environment. It gives architects and developers the ability to focus more on how to build applications that deliver business value from Hadoop than on all the behind-the-scenes tasks required to integrate Hadoop into the enterprise.

To learn more about BMC Control-M for Hadoop, visit www.bmc.com/hadoop

ADDITIONAL READING



MARKETPULSE: BIG DATA

CIO
Custom Solutions Group

Big Data, Big Opportunity

Survey results yield checklist for turning big data into bankable results.

ETPULSE WHITE PAPER MARKETPULSE WHITE PAPER MARKETPULSE WHITE PAPER MARKETPULSE WHITE P

Executive summary

Industry stats repeatedly point to the simple fact that big data isn't just here to stay—its prominence is growing. In fact, big data is redefining how organizations operate and make business decisions. While a diverse group of stakeholders are driving big data's utilization—from IT and execs to marketing and finance—it is still up to IT to make it work. And, as with any initiative, IT needs to make choices around which tools to use both short and long term. This paper provides a checklist for IT to use when managing the ever-growing workloads in a big data environment—and its fit within the enterprise.

Evolving Environment

Big data is growing at an unprecedented pace. According to a recent survey by IDG Research Services, the amount of data managed is expected to increase by 43 percent on average during the next year. And, as big data finds its way into various aspects of the business, it dramatically changes how organizations go about operations and make decisions. It's capable of impacting product production, service offerings, as well as tactics utilized to interact with customers.

Big data is so impactful that we are seeing new companies that are built entirely around its capability. Birmingham, Ala.-based ChipRewards, whose focus is on reducing the cost of healthcare through healthy lifestyles, is a prime example. By leveraging relationships with insurance companies, healthcare providers, and individual members, the behavioral scientists at ChipRewards are cost-effectively collecting detailed information around dietary choices, exercise patterns, etc., which enables the firm to develop individualized profiles and programs to incentivize healthy lifestyles. Rather than making generalized assumptions using sample data, these big data powered profiles enable ChipRewards to provide individualized help with modifying behaviors—help that would otherwise be impossible on a large scale.

Big data is also providing cost efficiencies in a variety of IT areas, such as archiving, data extract and transformation, and analytics. For example, keeping up with a government mandate to maintain seven years or more of data often meant dealing with costly storage. And, if there was ever a need to retrieve data, it wasn't a trivial task be-

SPONSORED BY
bmcsoftware

1

MARKET PULSE:
This white paper offers advice for choosing the most effective tools in a big data environment, as well as an IT checklist for managing increased big data workloads.

▸ **READ THE FULL PAPER**

ADDITIONAL READING

NEWS:

BMC puts Hadoop on the company timesheet

BMC's Control-M automated workflow software can incorporate big data-styled Hadoop jobs as well

By Joab Jackson
September 2013

IDG News Service — Now that many organizations see the utility in big data, BMC Software has provided a way to incorporate jobs from the Hadoop data processing platform into larger enterprise workflows.

"We provide a finer level of granularity for Hadoop workflows, and not just within Hadoop, but across the enterprise," said Shamoun Murtza, BMC director in the company's office of the chief technology officer.

BMC has released a new module for its Control-M job scheduling software that's designed to allow its users -- typically large organizations -- to manage Hadoop jobs as well, alongside the traditional IT jobs that they already manage using Control-M.

Murtza explained that even as enterprises start to use Hadoop, they don't have a lot of utilities to fit it into their existing IT operations.

Hadoop works mostly by batch processing, in that it ingests a chunk of data, analyzes it, and returns some output. This approach makes it well-suited for running in serial with other applications that can either feed Hadoop data, or use the results from Hadoop in some other computational operation.

Batch work "has to be coordinated not just within Hadoop, but across the enterprise. There are other workflows with results that can be pushed into Hadoop. Once you get the data out of Hadoop, you run some more [applications] to get value out of the data," Murtza said.

Originally designed for mainframe computers, BMC's Control-M workload automation tool provides a way for administrators to build workflows by linking different applications together into one task, without having to write a lot of scripts. BMC now offers the software for most modern enterprise platforms, including Linux, Unix and Microsoft Windows, and can work with most enterprise software, such as databases, enterprise resource planning (ERP) and ETL (extract, transform, load) software.

The new Control-M module now recognizes commonly used Hadoop components such as the HDFS (Hadoop File System), Pig, Scoop, Hive and MapReduce, which eliminates the need for administrators to write scripts to wire these applications into their workflow. Hadoop has its own set of job scheduling tools, although they work mostly for managing jobs only within the Hadoop environment, rather than for managing all the software being used in an organization.

▸ READ THE FULL ARTICLE