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

The Hitchhiker's Guide to Hadoop 2

The new version of Hadoop opens up the distributed processing framework to support more than just MapReduce applications. That gives users more options—if they're ready to take advantage.

TechTarget

Less Talk, More Action With New Hadoop Release?

Hadoop doesn't lack for attention from prospective users and industry analysts. But that has yet to translate into a high adoption rate. For example, in a 2013 survey conducted by Enterprise Management Associates and 9sight Consulting, only 16% of the 259 respondents said their organizations were using Hadoop. Other surveys have shown similar results. Hadoop has some marquee users, primarily among large Internet companies—but it hasn't reached very far beyond the marquee.

That's partly because a lot of companies are still trying to figure out exactly what to do with it. In a January 2014 blog post, Gartner analyst Merv Adrian cited the results of a poll question that asked attendees of a webinar about the top barriers to Hadoop adoption. First on the list, chosen by just over 50% of the 213 respondents, was an "undefined value proposition."

Helping to feed the lack of business cases is the fact that Hadoop's first incarnation was limited, with only MapReduce applications supported and limited scalability. In another blog post, Adrian described talk about Hadoop taking on the role of an enterprise data hub as "aspirational marketing" by advocates.

A first step down that road was the late 2013 release of Hadoop 2, which undoes the MapReduce dependency and makes Hadoop clusters more scalable. This three-part guide explores the capabilities of the new version, to help readers decide if it's something they can work with. First, we answer some frequently asked questions about Hadoop 2. Next we examine the fine details of YARN, the redesigned resource manager that opens up Hadoop to non-MapReduce applications. We close with an assessment of Hadoop 2's ability to support real-time analytics.

Craig Stedman
Executive Editor, SearchDataManagement

Hadoop 2 Spins New YARN, Breaks MapReduce Bonds

As with most 2.0 releases, Apache Hadoop 2 is a step forward for the open source distributed processing framework. The first version of Hadoop has found growing uses, particularly for processing large amounts of unstructured data and acting as a staging area for incoming information. That being said, it also presented significant limitations to many users.

Hadoop 2, originally referred to as Hadoop 2.0, makes several major architectural advances, most notably additional support for running non-batch applications created with programming models other than MapReduce. It also supports federation of Hadoop Distributed File System operations and configuration of redundant HDFS NameNodes. Doing so increases scalability and eliminates a nasty single point of failure that was part of the original design. In great part, Hadoop 2 is meant to widen the technology's utility for enterprise applications.

Prospective users looking to kick the proverbial tires on the new car that is Hadoop 2 likely have a lot of questions about the upgrade. The following are some answers for the IT managers, data architects, developers and business executives involved in evaluating potential deployments of Hadoop clusters.

When can I get my hands on Hadoop 2?

The Apache Software Foundation released Hadoop 2 for general availability in October 2013, after a series of alpha releases that began in May 2012. In addition to the downloadable community version, commercial Hadoop distribution vendors have started making the new software available to their customers.

As with any open source software, though, bug reports and fixes are still part of the daily fare for Hadoop. So it's best to keep an eye open for issues.

What's the story with YARN?

It's worthwhile to keep in mind that "Hadoop, as it first came out, was a learning experience," said Dave Wells, an independent consultant at Infocentric and an instructor for The Data Warehousing Institute.

"It was more about things patched together than it was about design and structure." With Hadoop 2, some of that patchiness begins to subside—and a key contributor to that is a software layer known as YARN.

The most common knock on Hadoop 1.x, which coupled HDFS with the MapReduce parallel programming model, was that its batch-oriented format limited its use in interactive and iterative analytics—and it pretty much eliminated the possibility of using the technology in real-time operations. Hadoop 2 changes that, principally by the insertion of YARN.

Although its name is modest, YARN—short for Yet Another Resource Negotiator—casts a long shadow. It's a rebuilt cluster resource manager that ends Hadoop's total reliance on MapReduce and its batch processing format. YARN does that by separating the resource management and job scheduling capabilities handled by MapReduce from Hadoop's data processing layer. As a result, MapReduce becomes just one of many processing engines that can sit on top of YARN in Hadoop clusters.

In effect, YARN opens the door for other programming frameworks and new types of applications, according to Douglas Moore, a principal consultant at Think Big Analytics. Until now, Moore said, "Hadoop has been like a freight train carrying freight." Hadoop 2, he added, will be able to support programming approaches that let it "go around a racetrack very quickly, like a Lamborghini."

What's with all the talk about HDFS high-availability and federation in Hadoop 2?

As it was originally built, Hadoop had some big drawbacks as a parallel processing platform. Clusters were dependent on a single namespace server, called a NameNode; it maintained a directory tree of files in HDFS and kept track of where in a cluster data was stored. That created a single point of control in a cluster, which caused real trouble if the NameNode went down. It also put fetters on the ability

of users to expand clusters and scale up their performance.

Those problems led to the development of the new high-availability and federation features for HDFS. Now developers can configure pairs of redundant NameNodes to provide a backup in case the active one crashes or requires maintenance work. And independent NameNodes that share a pool of data storage can be added at will to, in Moore's words, "spread the processing out."

Hadoop 2, in ending the reliance on MapReduce and introducing HDFS federation and high availability, is a big step toward maturity for the technology.

For William Bain, CEO of in-memory data grid vendor ScaleOut Software, the new capabilities were much-needed. "Single points of failure are not acceptable in any distributed environment," he explained.

The HDFS federation and high-availability features set the stage for processing of bigger and bigger data pools, said Sanjay Sharma, a principal architect at software development services provider Impetus Technologies. The federation scheme in particular is crucial for helping to grow Hadoop's data processing capacity "to the petabyte level."

Is Hadoop a mature, enterprise-ready technology now that Hadoop 2 is out?

Ending the reliance on MapReduce and introducing HDFS federation and high availability are big steps toward maturity for Hadoop. The technology also now supports Windows and point-in-time data snapshots for backup and disaster recovery purposes. But, it can still be a complicated platform to work with, in part because of its openness—and dependence upon tools to meet application needs.

Some assembly typically is still required in building out Hadoop-based environments. And, with Hadoop at the center of ongoing changes in data architecture, it seems a "Wild West" feel is guaranteed for some time to come.

Hadoop 2's release does show how thinking

about the framework has changed in recent years, according to Doug Cutting, one of Hadoop's original creators while working at Yahoo and now chief architect at Hadoop vendor Cloudera.

"In 2009, when the 0.20 release was created, most folks thought of Hadoop as a useful tool in and of itself," Cutting said in an email. "It primarily provided a MapReduce engine, making scalable, reliable batch computing available to enterprises." Now Hadoop can support a far wider range of workloads.

Still, even with Hadoop 2 now in the picture, the software remains new territory, holding both promise and pitfalls for prospective users.

—Jack Vaughan

YARN Adds More Application Threads for Hadoop Users

Even its most enthusiastic proponents might admit that Hadoop's marriage to MapReduce has limited what the open source technology can do. But with the advent of Hadoop 2 and its key component, the new YARN resource manager, the distributed processing framework has become a kind of launch pad for new applications incorporating a variety of related tools.

For example, Hadoop 2 is making real-time processing and analysis of streaming data possible for Synapse Wireless Inc., a Huntsville, Ala., maker of intelligent control and monitoring systems connected by a wireless mesh network. In present parlance, the company creates a "network of things" that uses the Internet to collect operational data from sensors and devices at customer sites. Some of the uses it supports include monitoring of health care operations and of large-scale commercial and residential lighting systems and solar panel fields.

Now, Synapse Wireless is looking to combine Hadoop 2 and Storm, an open source streaming data engine, to provide real-time business intelligence and analytics capabilities to its customers. "Our systems can capture high-velocity data streams coming off all these remote devices," said Bryan Stone, a cloud architect and lead platform developer at the company. With the pairing of Hadoop 2 and Storm, he added, "we don't just capture the data. We're also able to act on it. We can present it in a meaningful way so it can affect our customers' business decisions."

Using data integration tools from software vendor Pentaho, Stone and his colleagues at Synapse Wireless have created a pilot health care monitoring application that puts Storm on top of YARN in a Hadoop 2 cluster. The application is intended to ensure good handwashing hygiene in hospitals, as an example of what can happen when big data meets cloud

computing and the Internet of Things.

As part of the application, tags on the badges that nurses wear can track their movements around a hospital. Other tags collect data on the use of hand-cleanser dispensers. When a nurse enters a patient's room, a timer starts on the use of the dispenser there. If the application doesn't register that the dispenser has been used, Stone said, "we can send an alert down to the badge that the nurse is wearing as a reminder that she needs to wash her hands."

BATCH JOBS GET SOME COMPANY

While the original MapReduce-dependent version of Hadoop allowed Synapse Wireless to gather and analyze hand-washing data, the company couldn't act upon it immediately. Stone still sees value in MapReduce-based batch processing and analytics. But YARN "makes Hadoop more of [a platform] that you can build applications on top of," he said. "You can still use MapReduce in batch ways. But now you can roll out other applications, too."

Yahoo, the company where Hadoop first took flower, has been testing Hadoop 2 and YARN since September 2012. Yahoo built a Storm-on-YARN application to enable faster processing of website user activity data after a MapReduce batch program became unable to handle the information fast enough to meet the company's analytics and reporting needs. It released the application as an open source technology last year.

Speaking at the Hadoop Summit 2013, Bruno Fernandez-Ruiz, a senior fellow and vice president of platforms at Yahoo, described YARN as a flexible cog in the Hadoop framework—one that makes real-time processing in Hadoop clusters much more feasible than it was when they could only run MapReduce applications. "The problem with MapReduce computing is the batch window," he said, adding that users such as Yahoo can't afford to queue up data for processing while waiting for a three-hour batch job to finish running.

YARN's capabilities have even led the Apache Software Foundation, which manages the development of Hadoop, and vendors such as Yahoo spin-off Hortonworks to label it as an "operating system." That might be an overstatement, industry analysts said. But they agreed

that YARN provides an opportunity to broaden the use—and benefits—of Hadoop systems.

Calling YARN an operating system "is generous," said Gartner analyst Nick Heudecker. He compared it more to an application server, pointing to the Java middleware engines that began to gain ground in the late 1990s. And that's a good thing for users, according to Heudecker. "Developers can slide in different frameworks, some of which can be tightly integrated into the overall Hadoop stack," he said.

MAKING MORE WORK FOR HADOOP

Philip Russom, data management research director at The Data Warehousing Institute, said YARN's ability to concurrently execute and manage multiple processing jobs is something that "any decent operating system" should be expected to handle. "The concurrency feature alone makes Hadoop far more palatable to many [organizations]," he said, "because it enables multiple users with multiple application types to work simultaneously in the Hadoop environment."

Heudecker said YARN should also allow users to consolidate multiple Hadoop clusters, set up to process jobs simultaneously, into one large system. Instead of having the Hadoop equivalents of data marts, IT managers can combine systems and better rationalize technology, processing and management costs.

James Dixon, Pentaho's founder and chief technology officer, said YARN "will reduce the amount of MapReduce code people write," something he views as a step forward for users. Dixon minces few words in describing the limitations of MapReduce, claiming that it meets only a narrow set of processing needs.

"There are very few problems for which MapReduce is the right solution," he said. What YARN provides that MapReduce doesn't, he added, is the ability to "pick the right programming framework for individual problems." —*Jack Vaughan*

Analytics Finds Friendlier Turf in Hadoop 2 Systems

In a recent conversation with project team members from a client, one shared an internal slide deck used to promote the benefits of big data (in general) and Hadoop (in particular) among both key management decision makers and the IT development and implementation groups. One interesting aspect of the presentation was the comparison of Hadoop to earlier computing ecosystems and the casting of the open source distributed processing framework in the role of "operating system" for a big data environment.

At the time the slide deck was assembled, that characterization was perhaps somewhat of a stretch. The core components of the initial Hadoop release were the Hadoop Distributed File System (HDFS) for storing and managing data and an implementation of the MapReduce programming model. The latter included application programming interfaces, runtime support for processing MapReduce jobs, and an execution environment that provided the infrastructure for allocating resources in Hadoop clusters and then scheduling and monitoring jobs.

While those components acted as proxies for aspects of an operating system, the framework's processing capabilities were limited by its architecture, with the JobTracker resource manager and the application logic and data processing layers combined in MapReduce.

So what did that mean for running business intelligence and analytics applications? It had a big hampering effect: Although the task scheduling capabilities allowed for parallel execution of MapReduce applications, typically only one batch job could execute at a single time. That basically prevented the interleaving of different types of analysis in Hadoop systems. Batch analytics applications would have to run on a set of cluster nodes separate from a front-end query engine accessing data in HDFS.

STATIC APPROACH SUPPRESSES PROCESSING

In addition, resource allocation was effectively static—nodes assigned as Reduce nodes might sit idle during an application's Map phase, with the reverse happening during the Reduce process. As a result, nodes that might have been used for real-time processing were unavailable.

Lastly, the serial scheduling of batch execution jobs in a cluster supported neither MapReduce multitasking nor running MapReduce applications simultaneously with ones developed using other programming models. Again, that affected the ability of Hadoop users to engage in any kind of ad hoc querying or real-time data analysis.

But when you review the details of the new Hadoop 2 release, you'll see that some of the JobTracker's responsibilities have been split off from MapReduce, a move that's intended to relieve some of the constraints inherent in Hadoop's initial development and execution architecture. That's good news for organizations looking to run analytical applications on Hadoop systems.

The primary idea behind YARN, one of Hadoop 2's key additions, is to divorce resource management from application management. Instead of relying on MapReduce for both scheduling and processing jobs, those tasks now are handled by separate components. YARN includes a ResourceManager that becomes the authority for scheduling jobs and allocating resources among applications across a cluster, plus a NodeManager agent that oversees operations on individual compute nodes. But the ResourceManager doesn't manage the application execution process. In Hadoop 2, each application is controlled by its own ApplicationMaster, which assesses resource requirements, requests the necessary level of resources and works with the node agents to launch jobs and track their progress.

Some have made the comparison of Hadoop to earlier computing ecosystems, casting the open source distributed processing framework in the role of "operating system" for a big data environment.

ONWARD AND UPWARD ON ANALYTICS

Those changes will have some positive effects on the Hadoop framework's ability to support real-time analytics and ad hoc querying. First, segregating resource management from application management and processing enables the ResourceManager to be more efficient and effective in allocating a cluster's inventory of CPU, disk and memory resources to applications.

But the segregation won't lead only to more balanced workloads across cluster nodes; it also makes it possible for users to simultaneously run MapReduce and non-MapReduce applications on top of YARN. In addition to MapReduce batch jobs, that could include the likes of event stream processing, NoSQL database, interactive querying and graph processing and analysis applications.

Also, allowing multiple types of applications to run at the same time in relative isolation begins to address an issue that is sometimes overlooked with open source technologies—data protection and system security. Embedding all oversight and monitoring of individual jobs in the ApplicationMaster prevents faulty or malicious code that gets into one application from affecting others, affording a greater degree of processing safety in a big data environment.

> *Hadoop 2 divorces resource management from application management, allowing users to simultaneously run MapReduce and non-MapReduce applications on top of YARN.*

The improvements provided by YARN are recognition of the need for "hardening" Hadoop and transitioning it toward a more general operating system model. They also greatly increase Hadoop's analytical flexibility: With Hadoop 2, real-time analytics, batch analysis and interactive data management can all find a place at the big data table. —*David Loshin*

JACK VAUGHAN *is news and site editor of SearchData Management. He covers topics such as big data management, data warehousing, databases and data integration. Vaughan previously was an editor for TechTarget's SearchSOA, SearchVB, TheServerSide and SearchDomino websites. Email him at jvaughan@techtarget.com or follow him on Twitter: @JackVaughanatTT.*

DAVID LOSHIN *is president of Knowledge Integrity Inc., a consulting, training and development services company that works with clients on big data, business intelligence and data management projects. He also is the author of numerous books, including* Big Data Analytics. *Email him at loshin@knowledge-integrity.com.*

TechTarget

The Hitchhiker's Guide to Hadoop 2 is a SearchDataManagement.com e-publication.

Scot Petersen | *Editorial Director*

Jason Sparapani | *Managing Editor, E-Publications*

Joe Hebert | *Associate Managing Editor, E-Publications*

Craig Stedman | *Executive Editor*

Melanie Luna | *Managing Editor*

Linda Koury | *Director of Online Design*

Neva Maniscalco | *Graphic Designer*

Doug Olender | *Publisher*
dolender@techtarget.com

Annie Matthews | *Director of Sales*
amatthews@techtarget.com

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