Solution Brief

NetApp Open Solution for Hadoop

Expandable storage for Enterprise Hadoop

KEY BENEFITS

- Fully tested, installed, and supported for rapid deployment
- Balanced and integrated network, compute, and storage on an open application stack
- Advanced analytics for any type of data: structured, unstructured, polystructured
- High reliability for Hadoop implementations with improved failure handling, excellent Name Node metadata protection and recovery, and double-parity RAID 6 on a Data ONTAP® storage platform operating system
- Lower total cost of ownership (TCO) through flexibility, choice of lower-cost components, and independently scalable compute and storage
- Extendable data racks for increased Hadoop workloads

The Challenge

Bringing clarity to big data

Big data is defined as data that is so large and arrives so fast that it cannot be affordably processed and analyzed using traditional relational database tools. Typically, machine-generated data combined with other data sources creates challenges for both businesses and their IT organizations. With data in organizations growing explosively and most of the new data unstructured, companies and their IT groups are facing an extraordinary number of scalability and complexity issues.

Lines of business are motivated by top-line business benefits to solve unaffordable or unsolvable problems involving machine-generated data, which is often cross-referenced with traditional relational data. They exploit big data to derive competitive advantage, provide better customer experiences, and make the right decision faster. Big data can be used to prevent fraud, improve logistics by correlating buyer behavior with inventory, correlate patient treatments to their cures, and improve homeland security and government intelligence, typically by cross-correlating very large datasets from credit card transactions, RFID scans, video surveillance, and many other sources. According to Philip Carter, associate vice president at IDC Asia/Pacific, "What is becoming clearer is that the real value from big data will be derived from high-end analytics, predominantly using data mining, statistics, optimization, and forecasting-type capabilities to proactively turn this data into intelligence to drive business benefits and better decision making capabilities."[1] More specifically, what's needed is a Hadoop workload or cluster.

There is, of course, a risk in not taking advantage of the opportunities big data presents, since big data is more about business opportunities than reducing costs. To address both the challenges and the risks of big data, companies need big data analytic solutions that:

- Provide resilient and reliable storage for Hadoop.
- Implement high-performance Hadoop clusters.
- Are compatible with other analytical tools.

[1] IDC press release: "IDC Expects 'Big Data' to Drive the Next Wave of Investments in Business Analytics for CIOs in Asia/Pacific," October 3, 2011.

CHALLENGE	NETAPP OPEN SOLUTION FOR HADOOP ANSWER
Resilient storage for Hadoop, highly reliable HDFS	• FAS serves HDFS, providing RAID protection • High-availability Name Node
Management of Hadoop clusters can be difficult	• "Set it and forget it" nonstop transparent RAID operation • Rapid deployment of NetApp Open Solution for Hadoop Master Rack and easy-to-configure expansion racks
Scalability as big data increases	• Expansion rack provides 16 compute servers, 16 storage modules, and 2 network switches
Limited choice and flexibility	• Can use one or many clusters, not tied to a configuration • Use of open-source Hadoop provides interoperability
Cost-effective Hadoop deployments	• NetApp storage separates compute from storage so you can buy as many or as few components of either • Lower base price means lower initial cost and easier administration means lower ongoing costs
Scalability is nondisruptive	• Hot-pluggable disk shelves let you add storage without disrupting service
Scarce training, support resources in industry	• Installation and deployment support is included • Expertise from NetApp Professional Services also available from day one

Table 1) NetApp Open Solution for Hadoop addresses the challenges of big data analytics..

- Allow efficient Hadoop clustering.
- Scale compute and storage independently and quickly and expand as data grows in volume and ages.
- Are cost effective.

The Solution

NetApp Open Solution for Hadoop Rack

The NetApp® Open Solution for Hadoop Rack combines leading-edge technologies to deliver a solution that exceeds the requirements of emergent big data analytics so that businesses can manage process and unlock the value of the new and very large types of data they generate. See Table 1 for the way that the NetApp Open Solution for Hadoop Rack answers the needs of big data analytics.

NetApp Open Solution for Hadoop Rack allows you to spend time extracting value from your data, rather than fixing or maintaining cluster failure, and provides enterprise-grade storage systems, offering data protection and addressing the single points of failure that plague current Hadoop deployments. Enterprise-grade firmware RAID provides near-nonstop transparent RAID operations without the burden of server-based replication, typically done via software triple mirroring.

NetApp Open Solution for Hadoop Rack is based on the NetApp Open Solution for Hadoop, adding compute and networking resources to that solution. Figure 2 provides an overview of the NetApp Open Solution for Hadoop architecture. The key components of the solution are the E-Series E2660, the FAS2040, and two file systems: NFS in the FAS2040 Name Node and HDFS with support for Apache Hadoop in the E2660.

NetApp Open Solution for Hadoop Rack differs from other Hadoop storage offerings by delivering:

- A complete solution for Hadoop-style analytics in open storage ecosystems that works with third-party analytical tools
- Hadoop Name Node metadata enterprise-grade FAS protection
- Disk failure detection, global spares, and online repair with the E2660
- Nonstop storage operation with online repair and global hot spares
- Minimized production impact related to RAID group rebuild
- High-performance ingest and analytics with reduced server replication

Big Data Analytics with Enterprise TCO

The NetApp Open Solution for Hadoop Rack integrates big data analytics with out-of-the-box support for Hadoop with the performance, capacity, and reliability of NetApp E-Series storage. It is optimized for node-storage balance, cost, reliability/serviceability, performance, and storage capacity and density. Organizations that use Hadoop often use traditional server-based storage with inefficient, hard-to-scale, internal direct-access storage (DAS). The NetApp Open Solution for Hadoop employs the external DAS model, with higher scalability and lower total cost of ownership (TCO). The lower cost comes from choosing the right number of clusters and features and not spending money for storage and features you don't need. It also comes from nonproprietary options and decoupling compute nodes from storage, preventing unnecessary purchases of compute nodes for storage-intensive workloads.

Superior Flexibility Provides Choice

NetApp focuses on providing customers with flexible choices driven by important factors such as number of clusters, servers, and support options. With the NetApp Open Solution for Hadoop Rack, you're not locked in to a specific analytics tool to use in processing big data and working with Hadoop. You can choose from a number of leading-edge Hadoop and other big data analytical tools and not be restricted in the way you architect your big data analytic strategy. You have a choice, whether you're looking to work with the analytical tools you have, you don't have any Hadoop deployment solution, or you are just experimenting with it.

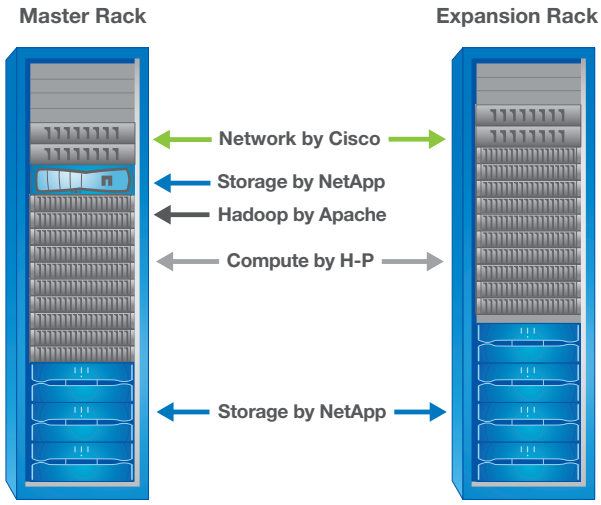

Figure 1) NetApp Open Solution for Hadoop Rack adds compute and networking to provide an integrated stack for enterprise Hadoop deployments.

NetApp also has a growing analytic partner ecosystem that businesses can choose to work with for their data analytic solution. This enables them to 1) work with existing analytics software, 2) work with analytic software from a NetApp partner, or 3) choose a combination of option 1 and option 2.

Fast Deployment

The NetApp Open Solution for Hadoop Rack comes as a complete and integrated set of components (see Figure 1) so you can start deploying it right away. The base (or master) configuration rack is a fully serviceable storage system that consists of:

- 4 Hadoop servers
- 2 FAS2040 storage modules
- 12 compute servers
- 3 E2660 storage modules that provide 360TB of storage
- 2 Ethernet switches

To help you deploy the system rapidly, deployment and installation services are bundled with the NetApp Open Solution for Hadoop Rack.

Expandable and Highly Scalable

The NetApp Open Solution for Hadoop Rack configuration can be expanded with data (also called expansion) racks. An expansion rack comes with:

- 4 NetApp E2660 storage modules
- 16 compute servers
- 2 Cisco® switches

Customers can increase their capacity by adding more data racks, which are simple to configure and set up. That enables them to handle different data types and increasing data volume.

Increased Performance to Handle Your Big Data Workloads

Increased performance is a result of NetApp's fast E-Series storage, fast connections and routes, and reduced software mirroring, which reduces cluster network congestion. One example of this is NetApp's own AutoSupport™ program, which monitors customer systems and proactively alerts customers to potential issues. Querying customer data took 4 weeks to run 24 billion unstructured records, and it was impossible to run the query on 240 billion unstructured records. With a small Hadoop cluster, NetApp was able to reduce the 4 weeks to 10.5 hours for 24 billion records and not only ran a query on 240 billion records but accomplished it in 18 hours.

More Resiliency and Less Downtime

The NetApp Open Solution for Hadoop Rack consists of NetApp E-Series storage that serves as HDFS Data Node storage. In traditional HDFS implementations without the E-Series, a disk failure initiates the restart of a job, leading to downtime. E-Series RAID protection prevents such job restarts and prevents some of the downtime. A FAS2040 device also serves NFS for the Name Node storage. In order to provide better metadata protection, the NetApp Open Solution for Hadoop Rack keeps a copy of critical Name Node data on the FAS2040 system, so if the Name Node server fails, all the data is recovered instantaneously with minimum downtime.

Team With Experts

Accelerate your implementation and minimize your risk by taking advantage of NetApp Professional Services (PS). NetApp PS enables successful deployments of the NetApp Open Solution for Hadoop Rack, and gets it right the first time by using industry best practices to design, deploy, and integrate the solution to meet specific business needs. Businesses achieve the highest levels of efficiency, manageability, and agility whether the project is large or small, on one site or in multiple locations.

With NetApp's authorized Professional Services Delivery Partner Network, companies benefit from market-leading open-sourced solutions combined with the technical expertise and support our partners offer. NetApp provides exclusive access to the right resources at the right time to help maximize performance in a stable, resilient, and secure environment.

With hundreds of professionals worldwide and years of field experience, NetApp's expert teams work diligently

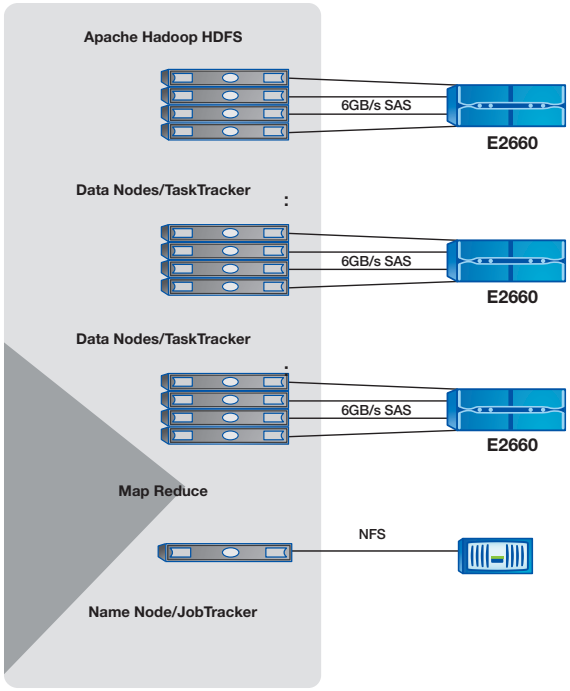

Figure 2) NetApp Open Solution for Hadoop is the building block.

to bring customers online with minimal time to production in a cost-effective manner, from the discovery and planning phase to solution deployment and acceptance. The portfolio of professional services available for the NetApp Open Solution for Hadoop includes Use-Case Discovery, Proof of Concept, Production Pilot, Solution Design and Implementation, and Optimization.

Easier Manageability and Administration

The NetApp Open Solution for Hadoop Rack does not require a lot of management or maintenance. It provides nonstop transparent RAID operation so customers can "set it and forget it." Another feature that makes it easy to manage the solution is online repair with global hot spares. This reduces mean time to repair and lets you swap storage without disrupting service.

NetApp Open Solution for Hadoop Fosters an Open Ecosystem

The NetApp Open Solution for Hadoop Rack includes support for the open-source Apache Hadoop framework, not a proprietary Hadoop distribution. This allows greater interoperability with other Hadoop tools and enables businesses to use analytical tools they already have or tools from any number of advanced analytical vendors, whether they're established or on the leading edge of Hadoop development. With the NetApp Open Solution for Hadoop Rack, customers aren't forced into a specific Map Reduce implementation or using a proprietary analytical tool.

NetApp Open Solution for Hadoop Rack can Help You Start Taking Advantage of Big Data Right Away

With the NetApp Open Solution for Hadoop Rack, NetApp offers an open and flexible Hadoop foundation that provides enterprise-class storage for advanced analytics platforms. You can deploy the NetApp Open Solution for Hadoop Rack rapidly and benefit from the reliability of NetApp E-Series storage in Hadoop. Because of the minimal cost to deploy and administer the solution and the choice of low-cost components you can also deploy the solution cost effectively.

About NetApp

NetApp creates innovative storage and data management solutions that deliver outstanding cost efficiency and accelerate business breakthroughs. Discover our passion for helping companies around the world go further, faster at *www.netapp.com*.

Go further, faster®

NetApp®

www.netapp.com