Solution Brief

NetApp Solutions for Hadoop

Speed time to insights, scale without sacrifice, and enable hybrid cloud deployments

KEY BENEFITS

Increase Performance

- Reduce runtime from weeks to hours
- Experience over 500% better performance during data rebuilding with DDP technology than with RAID 5

Prevent Cluster Downtime

- Deliver 99.9999% availability
- Leverage enterprise-grade backup, restore, and compliance capabilities

Enable Linear Scalability

- Scale workloads without sacrificing performance

Reduce Operational Cost

- Lower initial hardware and software investments by up to 61% and five-year cost of operations by up to 52%
- Quickly deploy a presized, prevalidated storage solution
- Improve storage efficiency by 33% and lower operational expenses

Maintain Flexibility and Choice

- Leverage cloud compute for analytics while keeping your data safely on the premises
- Support Apache-compatible distributions, such as Cloudera, Hortonworks, and MapR
- Run analytics natively on NFSv3 data without adding infrastructure

The Challenge

Harness the power of big data

Apache Hadoop and its growing ecosystem of products enable organizations to extract valuable insights from large volumes of diverse data that cannot be analyzed with relational databases. With these insights, people across the organization can ask the right questions and get better answers, supporting more informed decisions that help promote business transformation.

However, because initial Hadoop deployments often rely on commodity servers with internal drives, infrastructure resilience and agility issues prevent organizations from realizing the full benefits of their Hadoop deployment. For example, a single disk failure degrades performance of the entire cluster. Managing disk replacements is continual and error-prone. Triple file replication and failure redistribution models increase network costs and complexity. Also, commodity internal drives inefficiently accommodate use cases that have a different mix of processing power and storage requirements on the same infrastructure.

The Solution

NetApp enterprise storage for Hadoop Distributed File System (HDFS) based deployments

NetApp solutions for Hadoop feature enterprise storage that is independent of the compute servers to offer an enterprise-class deployment with lower cluster downtime, higher data availability, and linear scalability. We provide two powerful solutions to choose from to deploy HDFS: the NetApp E-Series solution for Hadoop or the NetApp ONTAP SAN solution for Hadoop.

The NetApp E-Series running DDP technology maintains performance, with only a negligible impact, even if a disk fails. And recovery is 10 times faster than with typical RAID schemes on commodity servers with internal storage. With these solutions, new data nodes can be added nondisruptively, and external data protection reduces both the storage footprint and the data replication overhead.

The NetApp ONTAP SAN solution leverages the robust and proven features of NetApp ONTAP 9 software, the enterprise data management software that powers the NetApp engineered systems of All Flash FAS (AFF) and FAS, as well as software-only ONTAP Cloud.

Increase Performance

By sustaining up to 21GBps throughput and up to 1,000,000 IOPS, NetApp E-Series storage systems support performance-intensive operations. An ESG Lab Validation tested a 10-node Hadoop cluster on E-Series. In this configuration, the run time for a query of 24 billion unstructured records dropped from 4 weeks to 10.5 hours, more than

NetApp®

cloudera®
NetApp Alliance Partner

Hortonworks
NetApp Alliance Partner

MAPR
NetApp Alliance Partner

94% faster. A second query of 240 billion unstructured records that had previously timed out was completed in 18 hours.

Maximize Availability for Greater Insight and ROI

Storage media is the most fault-prone component in the Hadoop architecture. It is simply a matter of time before a disk fails. The larger the cluster, the more likely it is that a disk failure will occur. In an HDFS implementation, a disk failure initiates a restart of a job task, which leads to more downtime. With NetApp solutions for Hadoop, if a disk fails, the the DDP protection in the E-Series systems or RAID DP® protection in AFF eliminates the need for job restarts, thereby maintaining the predictability of job run times.

Enable Linear Scalability

In Hadoop deployments that use commodity servers with internal storage, performance is negatively affected as more data nodes are added. Because the NetApp enterprise ONTAP shared storage and E-Series building-block designs separate compute from storage, capacity and performance can scale independently. Servers are only added when more compute resources are required, and performance scaling is linear.

ESG tested the scalability of NetApp solutions for Hadoop by using E-Series systems. When scaling from 4 data nodes and 60 drives to 8 data nodes and 120 drives, ESG found that the data-loading completion time was not affected, and the data-sorting completion time decreased by 11%. Not only did capacity scale to meet workload demands, but performance also increased.

Reduce Operational Costs

With NetApp E-Series solutions for Hadoop that are presized, preconfigured, and prevalidated, you can quickly deploy Hadoop and start gaining insights from your data. FlexPod® Select for Hadoop simplifies deployment and facilitates future scalability with a preconfigured solution that combines storage, networking, and servers, all validated for enterprise-class Hadoop environments. FlexPod Select leverages NetApp E-Series storage that is connected to Cisco UCS C-Series servers for high availability, seamless scalability, and improved storage efficiency.

A big data platform based on ONTAP software enjoys significant cost advantages compared to the traditional server-based storage. ONTAP solutions require far fewer and less costly servers (with fewer storage bays) and consequently fewer OS and application licenses. According to the IDC, these efficiencies can lower initial hardware and software investments by up to 61% and five-year cost of operations by up to 52%.

Maintain Availability with Less Storage Hardware

With the use of either the NetApp E-Series running DDP technology or NetApp ONTAP SAN solution for Hadoop , the replication count for HDFS can be reduced from three to two. This reduction in the number of data copies enables E-Series and ONTAP to store the same data with a much smaller footprint.

Maintain Flexibility and Choice

Each NetApp E-Series solution is certified to work with the primary Apache-compatible distributions: Cloudera, Hortonworks, and MapR technologies. This flexibility allows greater interoperability within the Hadoop framework. It also enables you to use Hadoop with the big data tools that you already have or with tools from any number of analytics platforms.

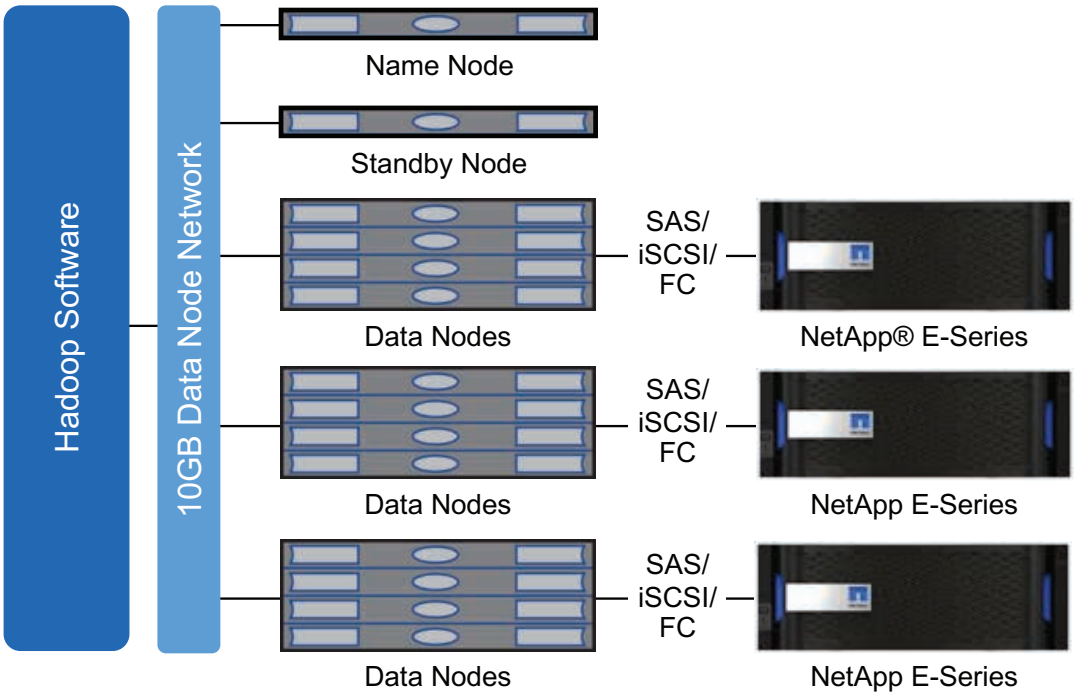

Figure 1) NetApp E-Series solution for Hadoop.

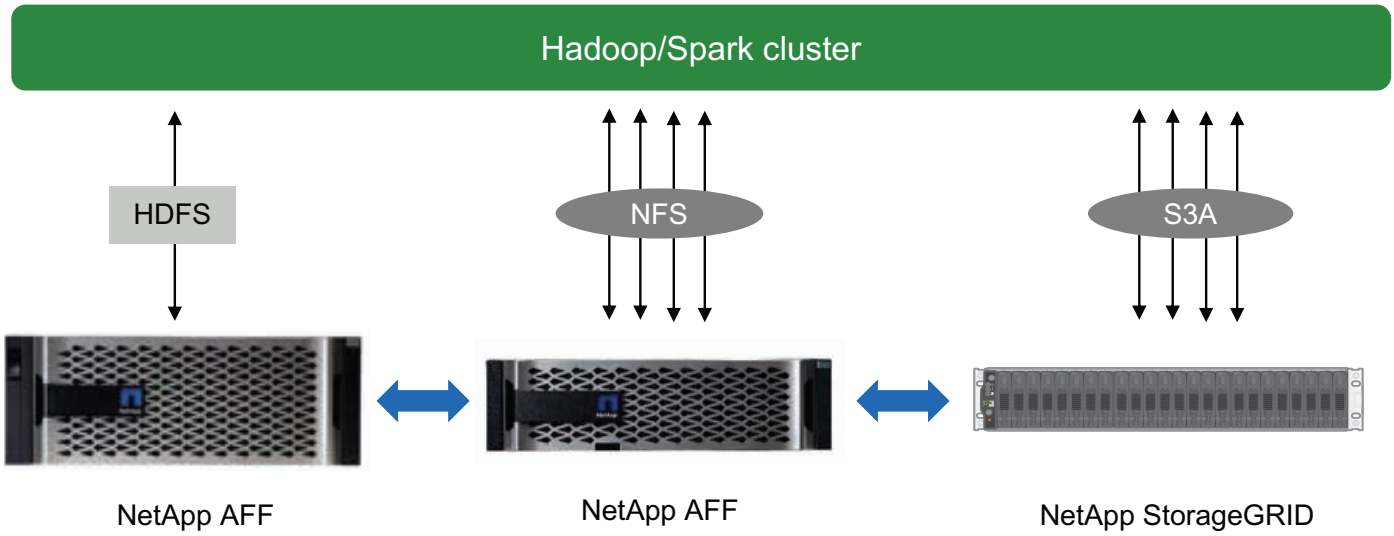

Figure 2) NetApp in-place analytics and data tiering for Hadoop.

NetApp and Zaloni have partnered to extend the NetApp Data Fabric to include data lifecycle management for the data lake. With this feature, you can define and execute data lifecycle policies that allow your organization to manage and move data across NetApp storage tiers.

In-place analytics with NetApp

You can use NetApp in-place analytics to run analytics on NFSv3 data without moving the data, creating a separate analytics silo, or setting up a separate HDFS cluster. You can switch from HDFS to NFS or run NFS alongside HDFS. Different Hadoop services are supported including YARN, MapReduce, Spark, HBase, Pig, and Hive, as well as being certified by Hortonworks for Hortonworks Data Platform.

These capabilities mean that you can support many types of workloads: batch, in memory, streaming, and more. An in-place analytics configured cluster can support S3 targets such as NetApp StorageGRID® Webscale or as Amazon Simple Storage Service (Amazon S3) by using the Hadoop S3A connector.

Increase Business Agility with Seamless Hybrid Cloud Deployments

The big data platform can take on many forms, depending on the needs of each enterprise. With ONTAP, choosing between an on-premises installation, one in the near cloud, or one in a private or public cloud is not necessary. ONTAP supports all options, whether a single option is chosen or more typically any combination of options.

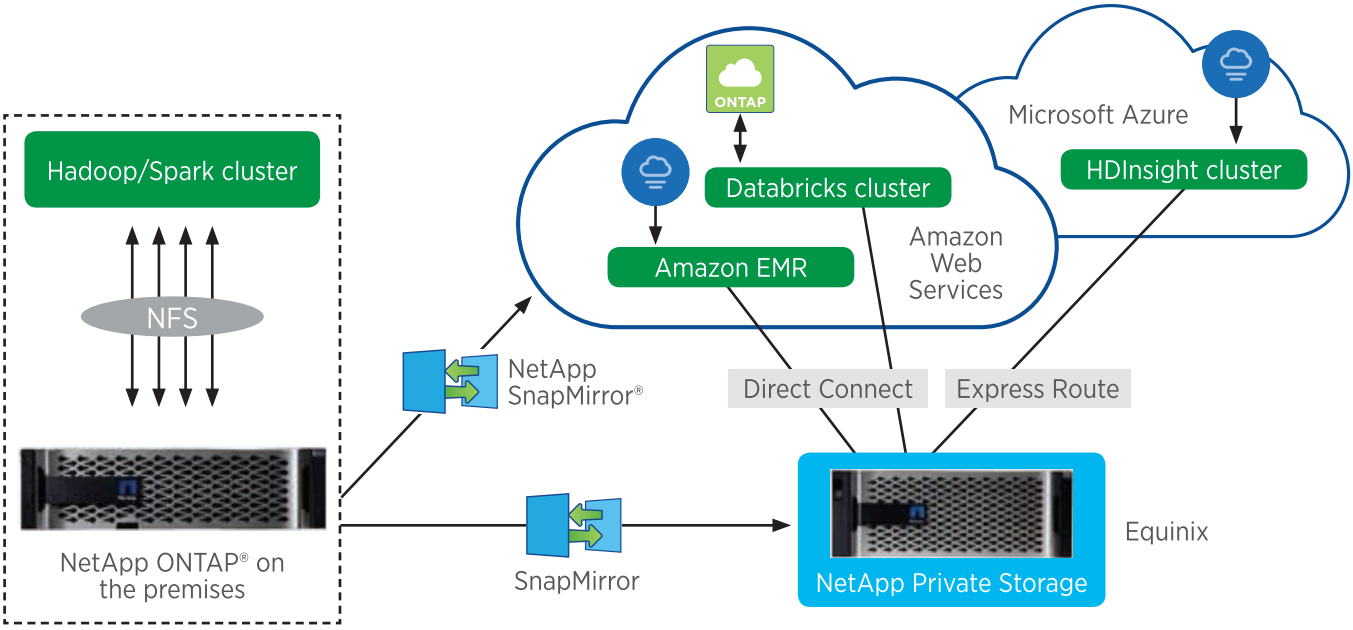

Figure 3) NetApp solution for analytics in a hybrid cloud.

Place your analytics compute tier on cloud architectures such as Amazon EC2 while keeping your data safely on storage that you control on the premises. The decoupled design allows independent scaling of compute and storage layers. This capability provides the flexibility to add storage capacity without adding compute nodes. With this design, just one copy of data is required—unlike HDFS, which requires three copies of data. In addition, the compute resources in a Hadoop cluster in a public cloud can be shut down when not used, because the data is stored in NetApp Private Storage (NPS). This configuration can result in a huge cost savings on cloud resources.

Enterprise Data Management

ONTAP provides integrated data protection to safeguard operations with near-instant backup and recovery using the highly efficient NetApp Snapshot™, SnapRestore®, and SnapCenter® technologies. These technologies help create efficient backup and replication of Hadoop data and FlexClone® technology helps to rapidly provision test and development environments.

Succeed with Big Data Analytics

From business owners and consumers of big data insights to data professionals, developers, and administrators, proven NetApp solutions for Hadoop can help everyone in your organization succeed with big data analytics.

If you need help in designing or deploying your NetApp solutions for Hadoop, NetApp Services experts and our certified partners can help.

Learn more about NetApp solutions for Hadoop at http://www.netapp.com/us/solutions/big-data/hadoop.aspx.

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organizations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation and optimize their operations. For more information, visit www.netapp.com. #DataDriven