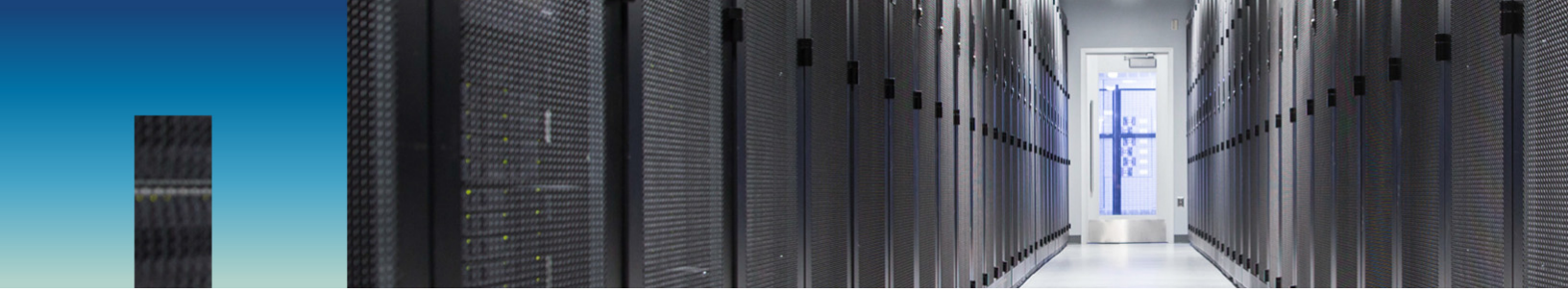


Technical Report

NetApp Solutions for Hadoop

Reference Architecture: Hortonworks

Faiz Abidi (NetApp), Ali Bajwa (Hortonworks), and Harsh Shah (Hortonworks)
September 2018 | TR-4716

In partnership with

HORTONWORKS®

Abstract

There has been exponential growth in data over the past decade, and enterprises are challenged with analyzing huge amounts of data within a reasonable time. Apache Hadoop is an open-source tool that can help your organization efficiently mine big data and extract meaningful patterns from it. However, enterprises face several technical challenges when deploying Hadoop, specifically in the areas of cluster availability, operations, and scaling. NetApp has developed a reference architecture with Hortonworks to deliver a solution that overcomes some of these challenges so that businesses can ingest, store, and manage big data with greater reliability and scalability, and with minimal time spent on operations and maintenance. This technical report discusses a flexible, validated, enterprise-class Hadoop architecture that is based on NetApp® E-Series storage using Hortonworks' Hadoop distribution.

NetApp®

TABLE OF CONTENTS

1 Introduction 4

1.1 Big Data 4

1.2 Hadoop Overview 4

2 NetApp E-Series Overview 5

2.1 E-Series Hardware Overview 5

2.2 SANtricity Software 6

2.3 Performance and Capacity 12

3 Hortonworks Overview 14

3.1 Hortonworks Data Platform Overview 14

3.2 Hortonworks Data Platform Products 14

4 Hadoop Enterprise Solution with NetApp 15

4.1 Data Locality and Its Insignificance for Hadoop 15

4.2 NetApp E-Series 16

4.3 Hadoop Replication Factor and TCO 16

4.4 Enterprise-Class Data Protection 16

4.5 Enterprise-Level Scalability and Flexibility 17

4.6 Easy to Deploy and to Use 17

4.7 Hortonworks Certified 17

5 Solution Architecture and Setup 17

5.1 Architectural Pipeline and Hardware Details 17

5.2 Effect of HDFS Block Size, Network Bandwidth, and Replication Factor 18

5.3 Hadoop Tuning 19

5.4 NameNode High Availability 19

5.5 Rack Awareness 19

5.6 Operating System Tuning 19

6 Certification Tests 21

7 Summary 21

Where to Find Additional Information 21

Acknowledgments 22

Version History 22

LIST OF TABLES

Table 1) Hadoop components. 5
Table 2) E5700 controller shelf and drive shelf models. 6
Table 3) Available drive capacities for the E5700 storage system. 14
Table 4) Architectural requirements. 17
Table 5) Alternative products supported by NetApp and its partners. 18
Table 6) Operating system tuning. 20

LIST OF FIGURES

Figure 1) MapReduce architecture. 4
Figure 2) E5700 controller drive shelf options. 6
Figure 3) Components of a pool created by the DDP feature. 8
Figure 4) DDP pool drive failure. 8
Figure 5) Technical components of NetApp E-Series FDE with an internally managed security key. 10
Figure 6) Technical components of NetApp E-Series FDE with an externally managed security key. 10
Figure 7) Write-heavy workload expected system performance on the E5700 storage system. 13
Figure 8) Read-heavy workload expected system performance on the E5700 storage system. 13
Figure 9) Overview of the setup. 18

1 Introduction

This report briefly discusses the various components of the Hadoop ecosystem. It also presents an overview of NetApp® E-Series solutions and tells why you should choose NetApp for Hadoop. It also includes best practices for configuring a Hadoop cluster and the recommendations by Hortonworks for extracting optimum performance.

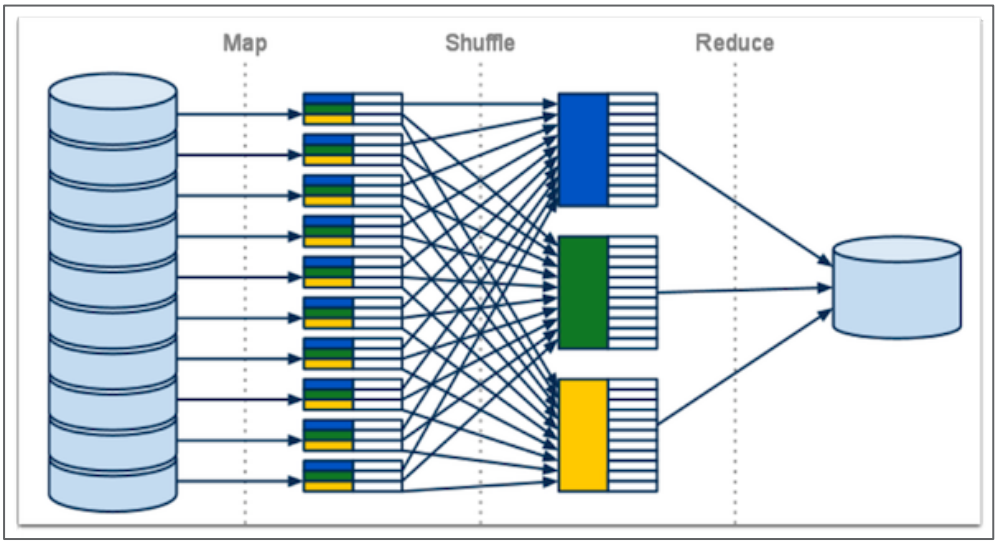
1.1 Big Data

Data is growing at a speed that no one could have predicted 10 years ago. The constant influx of data that is produced by technologies such as CCTV cameras, driverless cars, online banking, credit card transactions, online shopping, machine learning, and social networking must be stored somewhere. In 2013, it was estimated that 90% of the world's data had been generated over the past two years[1]. With such large amounts of data being generated, it is imperative for organizations to analyze this data to discover hidden patterns and behavior. The mining of big data for meaningful insights has several use cases in different industries. Just one example is e-commerce companies, such as Amazon, who can use this information to tailor advertisements to a specific audience.

1.2 Hadoop Overview

Apache Hadoop is open-source software that is used for processing big datasets by using a MapReduce architecture as shown in Figure 1. It enables parallel processing of data spread across nodes and can easily scale up to thousands of nodes. Hadoop is also fault tolerant in the sense that when a node faces downtime, the corresponding task that the failed node was working on gets passed on to another running node.

Figure 1) MapReduce architecture.[2]



[1] ScienceDaily. "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years." May 22, 2013. https://www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed December 22, 2017).
[2] AIMOTION. "Introduction to Recommendations with Map-Reduce and mrjob." http://aimotion.blogspot.com/2012/08/introduction-to-recommendations-with.html (accessed Aug 3, 2018).

The origin of Hadoop can be traced back to a 2003 paper released by Google that discusses the Google File System.[3] Since then, a lot of effort has gone into developing Hadoop into a robust, scalable, and highly reliable project. Companies such as Yahoo, IBM, Hortonworks, Facebook, Google, and others have been constantly contributing to the project. Table 1 discusses the four main projects (components) of Apache Hadoop. There are also other related projects, such as Spark, HBase, Impala, and Kudu, and each project has its own use case.

Table 1) Hadoop components.[4]

Component	Description
Hadoop Common	The common utilities that support the other Hadoop modules
Hadoop Distributed File System (HDFS)	A distributed file system that provides high-throughput access to application data
Hadoop YARN	A framework for job scheduling and cluster resource management
Hadoop MapReduce	A YARN-based system for parallel processing of large datasets

2 NetApp E-Series Overview

The industry-leading E-Series E5700 storage system delivers high IOPS and high bandwidth with consistently low latency. These features support the demanding performance and capacity needs of science and technology, simulation modeling, and decision support environments. The E5700 is equally capable of supporting primary transactional databases, general mixed workloads, and dedicated workloads such as video analytics in a highly efficient footprint, with extreme simplicity, reliability, and scalability.

E5700 systems provide the following benefits:

- Support for wide-ranging workloads and performance requirements
- Fully redundant I/O paths, advanced protection features, and proactive support monitoring and services for high levels of availability, integrity, and security
- Increased IOPS performance by up to 20% over the previous high-performance generation of E-Series products
- A level of performance, density, and economics that leads the industry
- Interface protocol flexibility to support FC host and iSCSI host workloads simultaneously
- Support for private and public cloud workloads behind virtualizers such as NetApp FlexArray® software, Veeam Cloud Connect, and NetApp StorageGRID® technology

2.1 E-Series Hardware Overview

As shown in Table 2, the E5700 storage array is available in two shelf options, which support both HDDs and solid-state drives (SSDs) to meet a wide range of performance and application requirements.

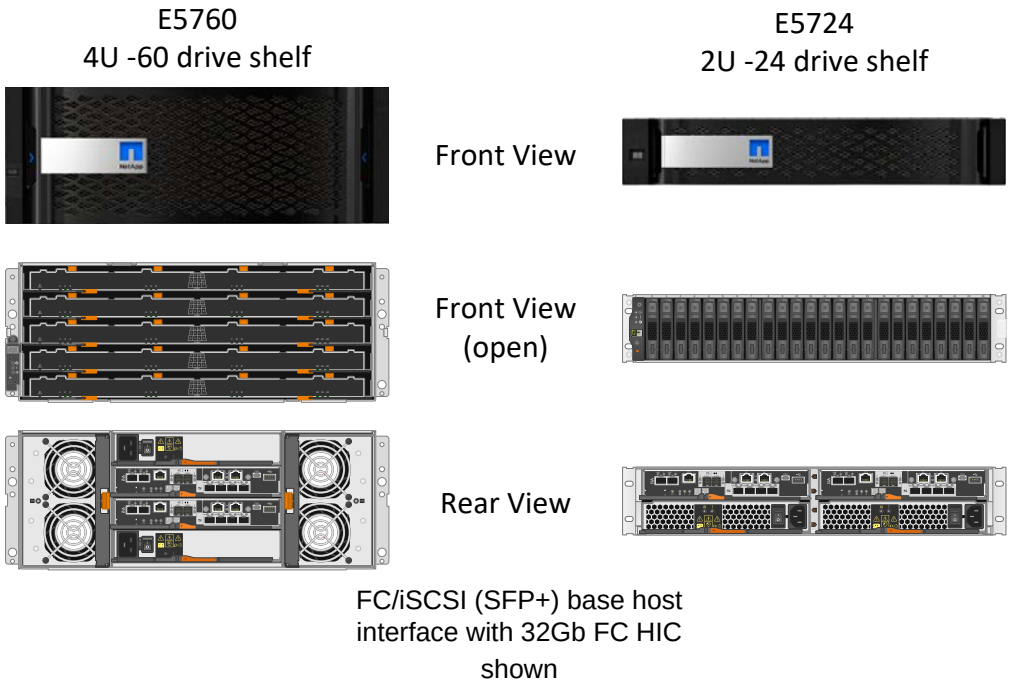
[3] Ghemawat, S., Gobioff, H., and Leung, S.-T., "The Google File System," ACM SIGOPS Operating Systems Review, vol. 37, no. 5., December 2003.
[4] Apache. "Welcome to Apache™ Hadoop®!" http://hadoop.apache.org/ (accessed Aug 3, 2018).

Table 2) E5700 controller shelf and drive shelf models.

Controller Shelf Model	Drive Shelf Model	Number of Drives	Type of Drives
E5760	DE460C	60	2.5" and 3.5" SAS drives (HDDs and SSDs)
E5724	DE224C	24	2.5" SAS drives (HDDs and SSDs)

Both shelf options include dual-controller modules, dual power supplies, and dual fan units for redundancy. The 24-drive shelf has integrated power and fan modules. The shelves are sized to hold 60 drives or 24 drives, as shown in Figure 2.

Figure 2) E5700 controller drive shelf options.



Each E5700 controller shelf includes two controllers, with each controller providing two Ethernet management ports for out-of-band management. The system has two 12Gbps (4-lane wide) SAS drive expansion ports for redundant drive expansion paths. The E5700 controllers also include two built-in host ports, which can be configured as either 16Gb FC or 10Gb iSCSI. The following host interface cards (HICs) can be installed in each controller:

- 4-port 12Gb SAS wide port (SAS-3 connector)
- 4-port 32Gb FC
- 4-port 25Gb iSCSI
- 2-port 100Gb InfiniBand

Note: Both controllers in an E5700 array must be identically configured.

2.2 SANtricity Software

E5700 systems are managed by NetApp SANtricity® System Manager 11.40, which is embedded on the controller.

To create volume groups on the array, the first step when you configure SANtricity is to assign a protection level. This assignment is then applied to the disks that are selected to form the volume group. E5700 storage systems support Dynamic Disk Pools (DDP) technology and RAID levels 0, 1, 5, 6, and 10. We used DDP technology for all the configurations that are described in this document.

To simplify the storage provisioning, NetApp SANtricity provides an automatic configuration feature. The configuration wizard analyzes the available disk capacity on the array. It then selects disks that maximize array performance and fault tolerance while meeting capacity requirements, hot spares, and other criteria that are specified in the wizard.

For more information about SANtricity Storage Manager and SANtricity System Manager, see the E-Series Systems Documentation Center.

Dynamic Storage Functionality

From a management perspective, SANtricity offers several capabilities to ease the burden of storage management, including the following:

- New volumes can be created and are immediately available for use by connected servers.
- New RAID sets (volume groups) or disk pools can be created at any time from unused disk devices.
- Dynamic volume expansion allows capacity to be added to volumes online as needed.
- To meet any new requirements for capacity or performance, dynamic capacity expansion allows disks to be added to volume groups and to disk pools online.
- If new requirements dictate a change - for example, from RAID 10 to RAID 5 - dynamic RAID migration allows the RAID level of a volume group to be modified online.
- Flexible cache block and dynamic segment sizes enable optimized performance tuning that is based on a particular workload. Both items can also be modified online.
- Online controller firmware upgrades and drive firmware upgrades are possible.
- Path failover and load balancing (if applicable) between the host and the redundant storage controllers in the E5700 are provided. For more information, see the Multipath Drivers Guide.

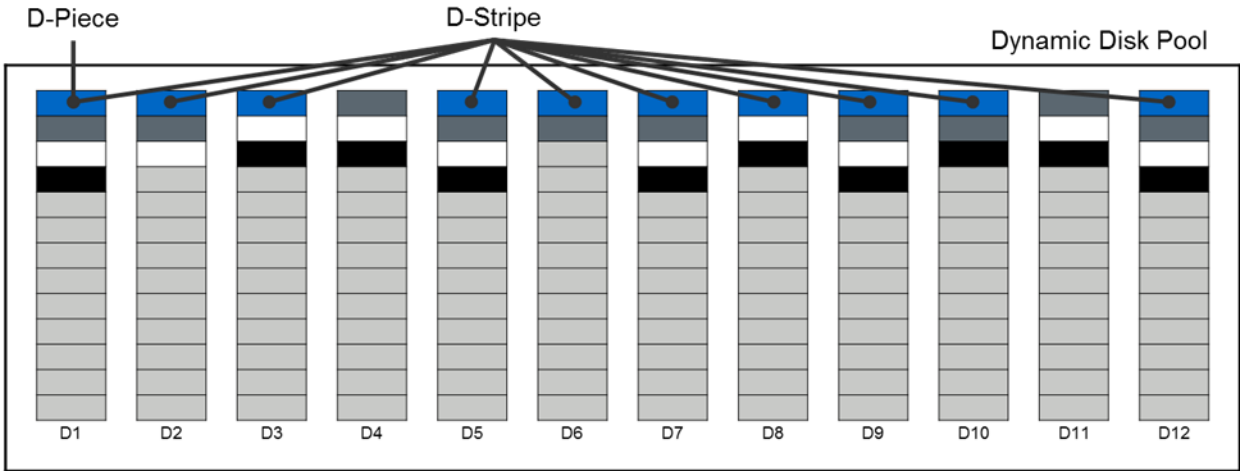
Dynamic Disk Pools Features

The Dynamic Disk Pools (DDP) feature dynamically distributes data, spare capacity, and protection information across a pool of disks. These pools can range in size from a minimum of 11 drives to all the supported drives in a system. In addition to creating a single pool, storage administrators can opt to mix traditional volume groups and a pool or even multiple pools, offering greater flexibility.

The pool that the DDP feature creates includes several lower-level elements. The first of these elements is a D-piece. A *D-piece* consists of a contiguous 512MB section from a physical disk that contains 4,096 segments of 128KB. Within a pool, the system selects 10 D-pieces by using an intelligent optimization algorithm from the selected drives. Together, the 10 associated D-pieces are considered to be a *D-stripe*, which is 4GB of usable capacity in size. Within the D-stripe, the contents are similar to a RAID 6 scenario of 8+2. Eight of the underlying segments potentially contain user data. One segment contains parity (P) information that is calculated from the user data segments, and one segment contains the Q value as defined by RAID 6.

Volumes are then created from an aggregation of multiple 4GB D-stripes as required to satisfy the defined volume size, up to the maximum allowable volume size within a pool. Figure 3 shows the relationship between these data structures.

Figure 3) Components of a pool created by the DDP feature.

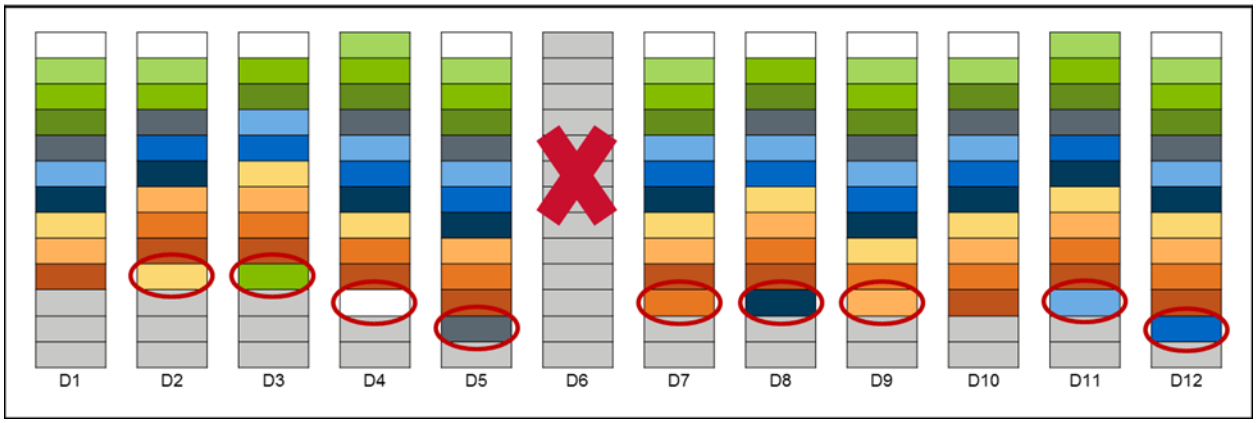


Another major benefit of a DDP pool is that rather than using dedicated stranded hot spares, the pool contains integrated preservation capacity to provide rebuild locations for potential drive failures. This approach simplifies management because individual hot spares no longer need to be planned or managed. The approach also greatly improves the time for rebuilds, if necessary, and enhances volume performance during a rebuild, as opposed to traditional hot spares.

When a drive in a DDP pool fails, the system reconstructs the D-pieces from the failed drive to potentially all other drives in the pool by using the same mechanism that is normally used by RAID 6. During this process, an algorithm that is internal to the controller framework verifies that no single drive contains two D-pieces from the same D-stripe. The individual D-pieces are reconstructed at the lowest available logical block address (LBA) range on the selected disk.

In Figure 4, disk 6 (D6) is shown to have failed. Subsequently, the D-pieces that previously resided on that disk are re-created simultaneously across several other drives in the pool. Because multiple disks participate in the effort, the overall performance impact of this situation is lessened, and the length of time that is needed to complete the operation is dramatically reduced.

Figure 4) DDP pool drive failure.



When multiple disk failures occur within a DDP pool, to minimize the data availability risk, priority for reconstruction is given to any D-stripes that are missing two D-pieces. After those critically affected D-stripes are reconstructed, the remainder of the necessary data is reconstructed.

From a controller resource allocation perspective, there are two user-modifiable reconstruction priorities within a DDP pool:

- **Degraded reconstruction priority** is assigned to instances in which only a single D-piece must be rebuilt for the affected D-stripes; the default for this value is **high**.
- **Critical reconstruction priority** is assigned to instances in which a D-stripe has two missing D-pieces that need to be rebuilt; the default for this value is **highest**.

For large pools with two simultaneous disk failures, only a relatively small number of D-stripes are likely to encounter the critical situation in which two D-pieces must be reconstructed. As discussed previously, these critical D-pieces are identified and reconstructed initially at the highest priority. This approach returns the DDP pool to a degraded state quickly so that further drive failures can be tolerated.

In addition to improving rebuild times and providing superior data protection, DDP technology can also provide much better performance of the base volume under a failure condition than traditional volume groups can.

For more information about DDP technology, see TR-4115: SANtricity Dynamic Disk Pools Best Practices Guide.

E-Series Data Protection Features

E-Series systems have a reputation for reliability and availability. Many of the data protection features that are in E-Series systems can be beneficial in a Hadoop environment.

Encrypted Drive Support

E-Series storage systems provide at-rest data encryption through self-encrypting drives. These drives encrypt data on writes and decrypt data on reads regardless of whether the full disk encryption (FDE) feature is enabled. Without the FDE enabled, the data is encrypted at rest on the media, but it is automatically decrypted on a read request.

When the FDE feature is enabled on the storage array, the system protects data at rest by locking the drives from reads or writes unless the correct security key is provided. This process prevents another array from accessing the data without first importing the appropriate security key file to unlock the drives. It also prevents any utility or operating system from accessing the data.

SANtricity 11.40 has further enhanced the FDE feature by enabling you to manage the FDE security key by using a centralized key management platform. For example, you can use Gemalto SafeNet KeySecure Enterprise Encryption Key Management, which adheres to the Key Management Interoperability Protocol (KMIP) standard. This feature is in addition to the internal security key management solution from versions of SANtricity earlier than 11.40 and is available beginning with the E2800, E5700, and EF570 systems.

The encryption and decryption that the hardware in the drive performs is invisible to users and does not affect the performance or user workflow. Each drive has its own unique encryption key, which cannot be transferred, copied, or read from the drive. The encryption key is a 256-bit key as specified in the NIST AES. The entire drive, not just a portion, is encrypted.

You can enable security at any time by selecting the Secure Drives option in the Volume Group or Disk Pool menu. You can make this selection either during volume group or disk pool creation or afterward. It does not affect existing data on the drives and can be used to secure the data after it is created. However, you cannot disable the option without erasing all the data on the affected drive group or pool. Figure 5 and Figure 6 show the technical components of NetApp E-Series FDE.

Figure 5) Technical components of NetApp E-Series FDE with an internally managed security key.

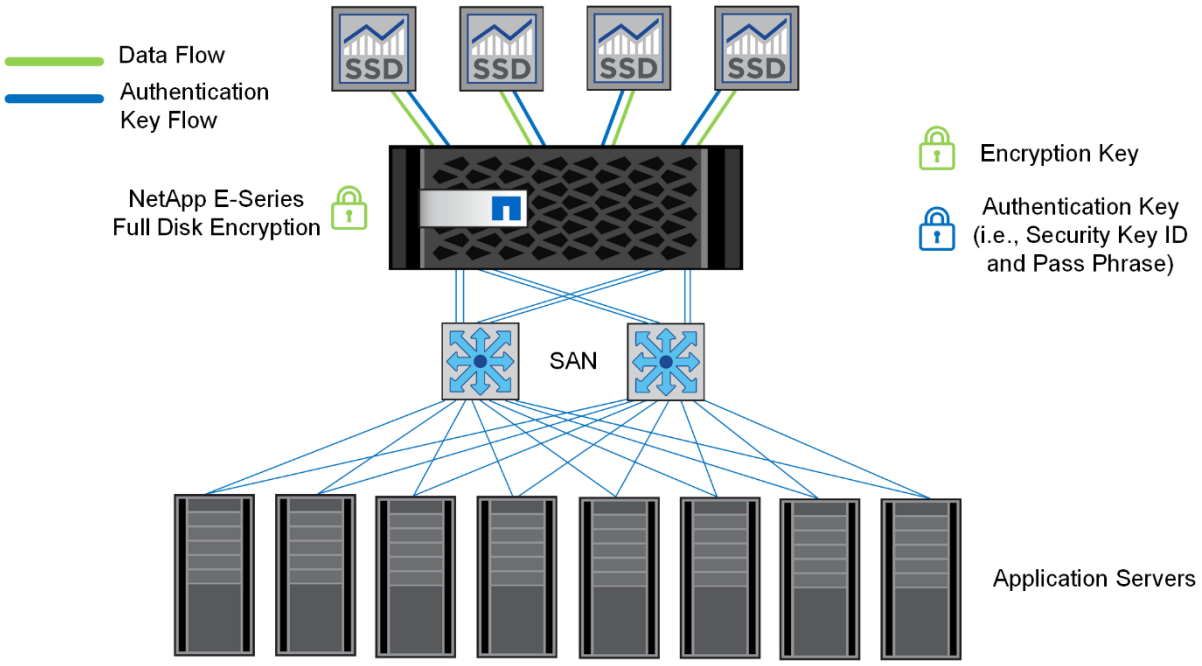
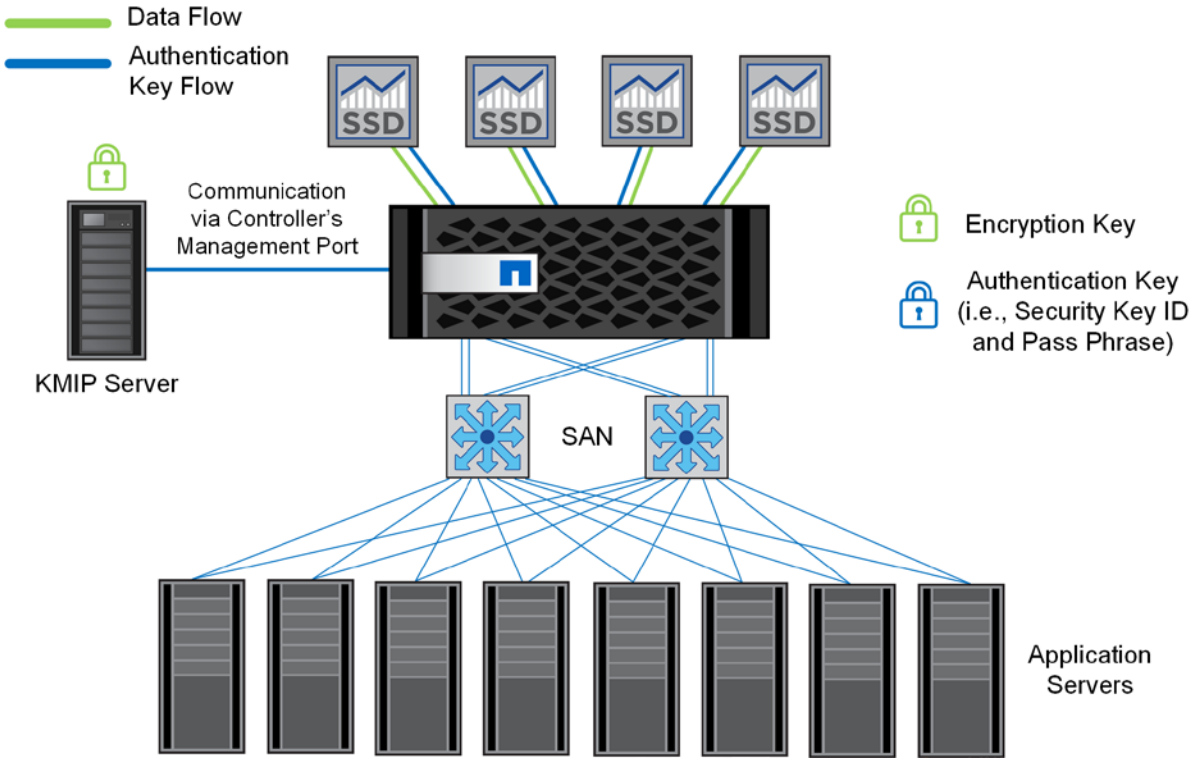


Figure 6) Technical components of NetApp E-Series FDE with an externally managed security key.



For more information about disk encryption, see TR-4474: SANtricity Full Disk Encryption.

Background Media Scan

A media scan is a background process that the controllers perform to detect errors on the drive media. Its primary purpose is to detect and repair media errors on disks that are infrequently read by user applications, and on which data loss might occur if other drives in the volume group fail. A secondary purpose is to detect redundancy errors such as data-parity mismatches. A background media scan can find media errors before they disrupt normal drive reads and writes.

Data Assurance (T10 PI)

The data assurance feature provides controller-to-drive data integrity protection through the SCSI direct-access block device protection information model. This model protects user data by appending protection information to each block of user data. The protection model is sometimes referred to as *data integrity field protection*, or T10 PI. This model confirms that an I/O operation has completed without any bad blocks having been written to or read from disk. It protects against displacement errors, data corruption resulting from hardware or software errors, and bit flips. It also protects against silent drive errors, such as when the drive delivers the wrong data on a read request or writes to the wrong location.

To protect your data, you need both data assurance and a media scan. The two features complement each other for superior data protection.

Unreadable-Sector Management

This feature provides a controller-based mechanism for handling unreadable sectors that are detected both during the normal I/O operation of the controller and during long-lived operations such as reconstructions. This feature is transparent to the user and does not require special configuration.

Proactive Drive-Health Monitor

Proactive drive-health monitoring examines every completed drive I/O and tracks the rate of error and exception conditions that are returned by the drives. It also tracks drive performance degradation, which is often associated with unreported internal drive issues, by using predictive failure analysis technology. When any error rate or degraded performance threshold is exceeded (indicating that a drive is showing signs of impending failure), SANtricity software issues a critical alert message and takes corrective action to protect the data.

Data Evacuator

With data evacuator, unresponsive drives are automatically power-cycled in an attempt to clear the fault condition. If the condition cannot be cleared, the drive is flagged as failed. For predictive failure events, the evacuator feature removes data from the affected drive; this action moves the data before the drive actually fails. If the drive fails, rebuild picks up where the evacuator was disrupted, thus reducing the rebuild time.

Hot Spare Support

The system supports global hot spares that can be automatically used by the controller to reconstruct the data of the failed drive if enough redundancy information is available. The controller selects the best match for the hot spare according to several factors, including capacity and speed.

SSD Wear-Life Monitoring and Reporting

If an SSD supports wear-life reporting, the GUI gives you this information so that you can monitor how much of the useful life of an SSD remains. For SSDs that support wear-life monitoring, the percentage of spare blocks that remain in solid-state media is monitored by controller firmware at approximate hourly intervals. Think of this approach as a fuel gauge for SSDs.

SSD Read Cache

The SANtricity SSD read cache feature uses SSD storage to hold frequently accessed data from user volumes. It is intended to improve the performance of workloads that are performance-limited by HDD IOPS. Workloads with the following characteristics can benefit from using the SANtricity SSD read cache feature:

- Read performance is limited by HDD IOPS.
- There is a higher percentage of read operations relative to write operations. More than 80% of the operations constitute read.
- Numerous reads are repeat reads to the same or to adjacent areas of the disk.
- The size of the data that is repeatedly accessed is smaller than the SSD read cache capacity.

For more information about the SSD read cache, see TR-4099: NetApp SANtricity SSD Cache for E-Series.

2.3 Performance and Capacity

Performance

An E5700 system that is configured with all SSD, all HDD, or a mixture of both drives can provide high IOPS and throughput with low latency. Through its ease of management, high degree of reliability, and exceptional performance, you can use E-Series storage to meet the extreme performance requirements of a Hadoop cluster deployment.

An E5700 system with 24 SSDs can provide up to 1 million 4K random read IOPS at less than 100µsec average response time. This configuration can also deliver 21GBps of read throughput and 9GBps of write throughput.

Many factors can affect the performance of the E5700 storage system. These factors include different volume group types, the use of DDP technology, the average I/O size, and the read versus write percentage that the attached servers provide. Figure 7 and Figure 8 show additional performance statistics across various data protection strategies on the system under generic random I/O workloads.

Note: The tested system used 48 SSDs, 4K and 16K block sizes, and 25% and 75% read workloads.

Figure 7) Write-heavy workload expected system performance on the E5700 storage system.

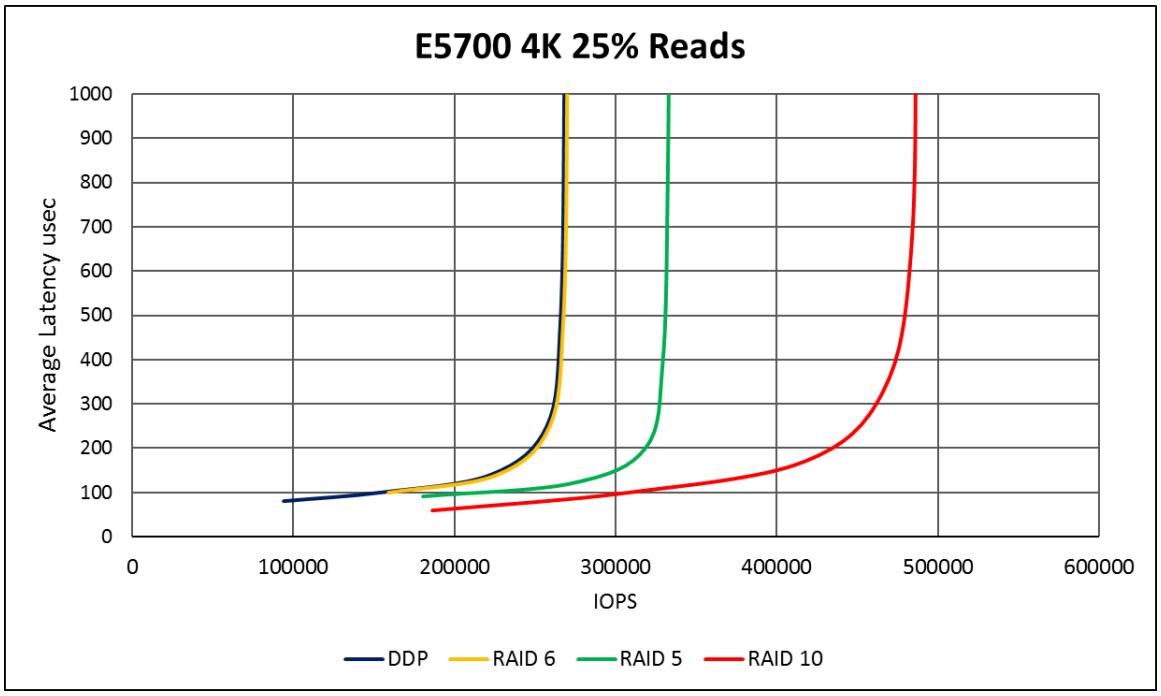


Figure 8) Read-heavy workload expected system performance on the E5700 storage system.

Capacity

The E5760 has a maximum capacity of 4800TB (with expansion drive shelves), using 480 NL-SAS HDD drives of 10TB each. The E5724 has a maximum capacity of 1800TB (with expansion drive shelves), using 120 SSDs of 15.3TB each. See Table 3 for available drive capacities.

Table 3) Available drive capacities for the E5700 storage system.

Controller Shelf Model	Drive Shelf Model	Number of Drives	NL-SAS HDDs	SAS HDDs	SSDs
E5760	DE460C (4U60)	60	4TB 8TB 10TB	900GB 1.2TB 1.8TB	800GB 1.6TB 3.2TB
E5724	DE224C (2U24)	24	N/A	900GB 1.2TB 1.8TB	800GB 1.6TB 3.2TB 15.3TB

3 Hortonworks Overview

In 2011, Rob Bearden partnered with Yahoo to establish Hortonworks with 24 engineers from the original Hadoop team, including founders Alan Gates, Arun Murthy, Devaraj Das, Mahadev Konar, Owen O'Malley, Sanjay Radia, and Suresh Srinivas. Under the leadership of Arun Murthy, chief products officer, this core product team is enriched with enterprise software talent from the likes of Oracle, IBM, HP, VMware, and others to help make HDP and HDF meet the enterprise-grade requirements our customers expect. Hortonworks is headquartered in Santa Clara, California, and the business model is based on open-source software support subscriptions, services, solutions, training and consulting services. Hortonworks operates in 19 countries with approximately 1,110 employees. There are 29 out of 114 Apache Hadoop committers from Hortonworks and 208 committer seats across 20+ Apache projects, and they focus on the data access, security, operations, and the governance needs of the enterprise Hadoop market.

Hortonworks is a leading innovator in the industry, creating, distributing, and supporting enterprise-ready open data platforms and modern data applications. Its mission is to manage the world's data. It has a single-minded focus on driving innovation in open-source communities such as Apache Hadoop, NiFi, and Spark. Hortonworks, along with its 1,600+ partners, provide the expertise, training, and services that allow customers to unlock transformational value for their organizations across lines of business. Hortonworks's connected data platforms power modern data applications that deliver actionable intelligence from all data: data in motion and data at rest.

3.1 Hortonworks Data Platform Overview

HDP is the industry-secure, enterprise-ready open-source Apache Hadoop distribution based on centralized architecture (YARN). HDP addresses the complete needs of data at rest, powers real-time customer applications, and delivers robust big data analytics that accelerate decision making and innovation.

3.2 Hortonworks Data Platform Products

HDP products are divided into six categories: data management, data access, data governance and integration, security, operations, and cloud.

Data Management

YARN and HDFS are the cornerstone components of HDP for data at rest. Whereas HDFS provides the scalable, fault-tolerant, cost-efficient storage for your big data lake, YARN provides the centralized architecture that enables you to process multiple workloads simultaneously. YARN provides the resource management and pluggable architecture for enabling a wide variety of data access methods.

Data Access

HDP includes a versatile range of processing engines that empower you to interact with the same data in multiple ways at the same time. Therefore, applications for big data analytics can interact with the data in the best way, from batch to interactive SQL or low latency access with NoSQL. Emerging use cases for data science, search, and streaming are also supported with Apache Spark, Storm, and Kafka.

Data Governance and Integration

HDP extends data access and management with powerful tools for data governance and integration. These tools provide a reliable, repeatable, and simple framework for managing the flow of data in and out of Hadoop. This control structure, along with a toolset to ease and automate the application of schemas or metadata on sources, is critical for successful integration of Hadoop into your modern data architecture. Hadoop has engineering relationships with many leading data management providers to enable their tools to work and integrate with HDP. Data governance and integration are provided by Atlas, Falcon, Oozie, Sqoop, Flume, and Kafka components.

Security

Security is woven and integrated into HDP in multiple layers. Critical features for authentication, authorization, accountability, and data protection are in place to help secure HDP across these key requirements. Consistent with this approach throughout all of the enterprise Hadoop capabilities, HDP also enables you to integrate and extend your current security solutions to provide a single, consistent, secure umbrella over your modern data architecture. The authentication, authorization, and data protection are provided by Knox and Ranger components.

Operations

Operations teams deploy, monitor, and manage a Hadoop cluster within their broader enterprise data ecosystem. Apache Ambari simplifies this experience. Ambari is an open-source management platform for provisioning, managing, monitoring, and securing the HDP. It enables Hadoop to fit seamlessly into your enterprise environment. Ambari and Zookeeper manage a Hadoop cluster.

Cloud

Cloudbreak, as part of HDP and powered by Ambari allows simplified provisioning and Hadoop cluster management in any cloud environment including; Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and OpenStack. It optimizes your use of cloud resources as workloads change.

4 Hadoop Enterprise Solution with NetApp

4.1 Data Locality and Its Insignificance for Hadoop

One of the first Hadoop concepts to improve performance was to move compute to data or to co-locate compute and storage. This approach means moving the actual compute code to servers on which the data resides, and not the other way around. Because data is usually larger in size than the compute code, it might be a challenge to send big data across a network, especially with lower-bandwidth networks.

However, segregation of storage and compute is necessary to scale up and to maintain flexibility. In 2011, it was estimated that reading the data from local disks was only 8% faster than reading it from remote disks, and this number is only decreasing with time. Networks are getting faster, but the disks are

not. As an example, Ananthanarayanan et al. analyzed logs from Facebook and concluded that "disk-locality results in little, if any, improvement of task lengths."[5]

With all the advancements being made in improving the network infrastructure, data compression, and deduplication, under the right conditions, co-locating storage and compute does not add significant benefit. In fact, it is better to segregate storage and compute and to build a flexible and easy-to-scale solution. This area is where NetApp solutions can help.

4.2 NetApp E-Series

NetApp E-Series systems offer a block-storage solution that is designed for speed. E-Series is a better fit for applications that need dedicated storage such as SAN-based business apps, dedicated backup targets, and high-density storage repositories. The E-Series solution delivers performance efficiency with an excellent price/performance ratio, from entry-level to enterprise-level systems. It also provides inexpensive maximum disk I/O and delivers sustained high bandwidth and IOPS. E-Series systems run NetApp SANtricity, a separate operating environment. All E-Series models support both SSD and HDD and can be configured with dual controllers for high availability. You can find more details on the E-Series webpages.

4.3 Hadoop Replication Factor and TCO

To be fault tolerant and reliable, HDFS allows users to set a replication factor for data blocks. A replication factor of 3 means that three copies of the data would be stored on HDFS. Therefore, even if two drives that store the same data fail, the data is not lost. The trade-off here is an increase (three times or more) in space and network utilization. With a traditional JBOD (just a bunch of disks) configuration, setting a replication factor of 3 is recommended to significantly decrease the probability of data loss. This configuration also exploits the concept of data locality, achieving better performance in a traditional Hadoop architecture.

With DDP technology on a NetApp E-Series storage system, the data and parity information are distributed across a pool of drives. The DDP technology's intelligent algorithm defines which drives are used for segment placement, helping to fully protect the data. For more information, see the DDP datasheet.

Because of these intelligent features, NetApp recommends using a replication factor of 2 instead of 3 when you use an E-Series storage system. The lower replication factor puts less load on the network, and jobs complete faster. Therefore, fewer DataNodes are required, which directly relates to spending less on the licensing fees if you use managed Hadoop software.

Also, because NetApp provides a decoupled Hadoop solution in which compute and storage are segregated, you no longer need to buy more servers to add more storage capacity.

4.4 Enterprise-Class Data Protection

Drive failures are common in data centers. Although Hadoop is built to be fault tolerant, a drive failure can significantly affect performance. Even when you use RAID, the drive rebuild times can be several hours to days, depending on the size of the disk. DDP technology enables consistent and optimal performance during a drive failure and can rebuild a failed drive up to eight times faster than RAID can. This superior performance results from the way that the DDP feature spreads the parity information and the spare capacity throughout the pool.

---

[5] Ananthanarayanan, G., Ghodsi, A., Shenker, S., and Stoica, I., "Disk-Locality in Datacenter Computing Considered Irrelevant." In HotOS '13, Proceedings of the 13th USENIX Conference on Hot Topics in Operating Systems, vol. 13, May 2011, pp. 12–12.

4.5 Enterprise-Level Scalability and Flexibility

By using E-Series storage solutions for HDFS, you can separate the two main components of Hadoop, compute and storage. This decoupled solution provides the flexibility of managing both components separately. For example, the SANtricity software that comes with the E-Series products provides an intuitive and user-friendly interface from which extra storage can be added seamlessly. This flexibility makes it convenient to scale the storage capacity up or down as needed without affecting any running jobs.

4.6 Easy to Deploy and to Use

There is a steep learning curve for customers who are new to Hadoop. Few enterprise applications are built to run on massively parallel clusters. However, the NetApp E-Series solution for Hadoop provides an operational model for a Hadoop cluster that does not require additional attention after its initial setup. The cluster is more stable and easier to maintain, allowing you to concentrate on meeting your business needs. This solution flattens the operational learning curve of Hadoop.

4.7 Hortonworks Certified

NetApp has partnered with Hortonworks to certify the NetApp Hadoop solutions. For more information, see the Hortonworks' website.

5 Solution Architecture and Setup

5.1 Architectural Pipeline and Hardware Details

Figure 9 shows a high-level overview of the architecture, which includes the components in Table 4:

Table 4) Architectural requirements.

Components	Requirement
Servers	• 8 x Fujitsu servers with 48 vCPUs and 256GB memory • 3 x Master servers with 24 vCPUs and 64GB memory
Storage	• 1 x NetApp E5700 storage array • 96 x 10TB HDD (7200 RPM); 8 DDP pools, each having 12 drives, and each pool having one volume mapped to one host
Network	• All servers using a 10Gb dual-port network connection • 12Gbps SAS connections to the E-Series storage array
Software	• Red Hat 7.4 on all the servers • Hortonworks Data Platform 2.6.5 or later • SANtricity 11.40 or later storage management software

Although we used a SAS connection between the DataNodes and the E-Series storage, you can also use other protocols, such as FC, iSCSI, and InfiniBand. For more information about the iSCSI validation, see the NetApp Solutions for Hadoop white paper. Similarly, depending on your storage requirements, you can also select other E-Series models such as the E2800 (lower end) or E5600 (medium end). Table 5 shows the alternative products currently supported by NetApp and its partners.

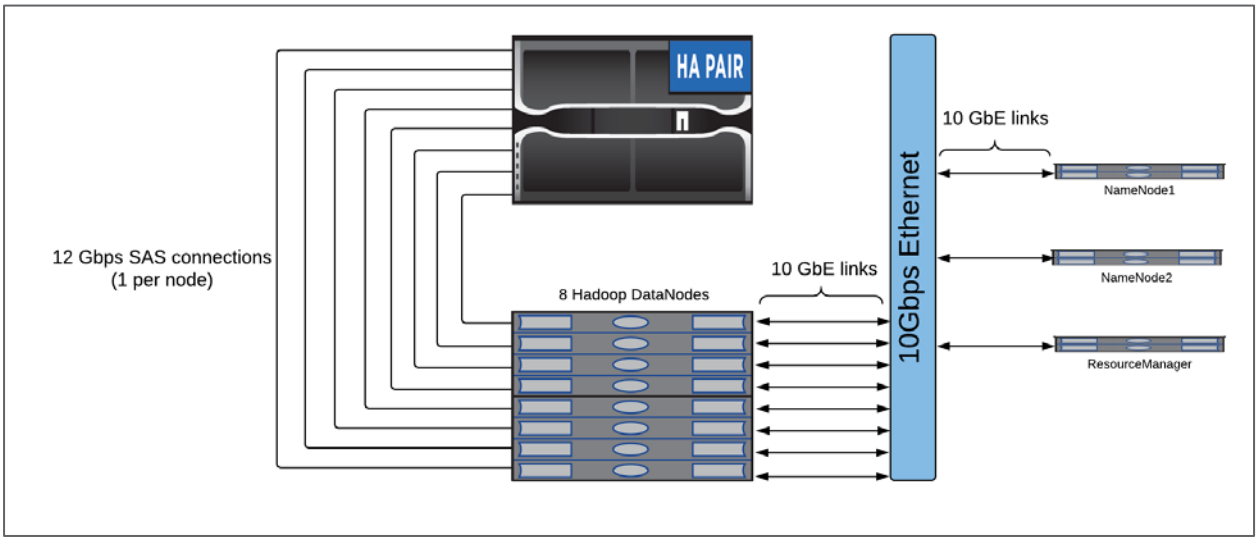
Table 5) Alternative products supported by NetApp and its partners.

Component	Supported Options	Details
Storage arrays	E5xxx E2xxx EFxxx	We tested with E5700 in this report, but all the other E-Series products are supported as well.
Disks and types	12, 24, 60, and up to 480 (depending on the type of the E-Series storage array)	Disks are supported with a capacity of 1TB, 2TB, 4TB, 6TB, or 10TB HDDs. SSDs with 800GB, 1.6TB, and 3.2TB are also available.
Protocols and connectivity	SAS Fibre Channel iSCSI InfiniBand	We tested with SAS connections in this report, but all the other network protocols are supported as well.

Best Practice

For optimal performance of an E-Series storage array, it is important to create one pool per host and one volume per pool, using all the disks. For example, in our test setup, we used 96 drives on the E5700 storage array and we had 8 compute nodes connected to it. We created 8 pools by using the 96 drives, and each pool had 12 drives. Then we created one volume per pool by using all 12 drives and mapped each to the host server.

Figure 9) Overview of the setup.



5.2 Effect of HDFS Block Size, Network Bandwidth, and Replication Factor

The HDFS block size differs from the underlying operating system block size. The HDFS block size is an abstraction on top of file systems such as XFS and ext4. HDFS block sizes are usually large to minimize the seek times. A small block size, for example 4K, means that to transfer a 1GB file, 262,144 requests must be made. That many requests across a network would cause tremendous overhead. The block size that you select largely depends on the size of the input data, the access patterns, and the server

configuration. It usually takes a little trial and error to determine the block size that gives the best performance.

Because the master nodes and DataNodes talk to each other when copying and transferring data, network speeds can often be the bottleneck in Hadoop performance, especially during the shuffle phase. For example, a 1Gbps connection might not be fast enough if the input data size is 100TB. In that case, the network might get saturated, and the Hadoop cluster performance would be adversely affected.

The data replication factor (RF) can also play a significant role in Hadoop performance. A higher RF means that extra storage space is needed, and it means more stress on the network for data transfers. A higher RF, however, also means faster reads and slower writes. We get faster reads with a higher RF because of Hadoop's implementation: "To minimize global bandwidth consumption and read latency, HDFS tries to satisfy a read request from a replica that is closest to the reader. If there exists a replica on the same rack as the reader node, then that replica is preferred to satisfy the read request." [6]

5.3 Hadoop Tuning

It can be challenging to optimally configure Hadoop for best performance. You must decide how many mappers to use, how much memory and how many cores to allocate to mappers and to reducers, how much buffer memory to use while sorting files, and so on. By properly configuring all these parameters, you can expect to get better performance out of a Hadoop cluster.

Hortonworks provides some guidelines that can help you tune Hadoop clusters. For more details, refer to Determine YARN and MapReduce Memory Configuration Settings by Hortonworks.

5.4 NameNode High Availability

In Hadoop 1.x, the NameNode was a single point of failure. When a NameNode went down, the entire cluster would become unavailable and it would remain so until an administrator brought the NameNode back up. Hadoop 2.x introduced the concept of high availability (HA) for NameNodes, alleviating the problem of the single point of failure for HDFS. Therefore, with HA enabled, if a NameNode fails, the standby NameNode provides automatic failover. This approach ensures that HDFS does not go down and the Hadoop cluster remains online.

For more details on HA, refer to the NameNode High Availability document by Hortonworks.

5.5 Rack Awareness

Hadoop components are rack-aware in the sense that they know the cluster topology; in other words, they know how the data is distributed across different racks in a cluster. Rack awareness can be helpful when hundreds of nodes are spread across different racks and the worker nodes on the same rack are assumed to have lower latency and higher bandwidth. Rack awareness also increases data availability: If the master node knows that two nodes are in the same rack, it tries to put a copy of the data on a different rack. So, if one of the racks goes down, data can still be retrieved from the other rack.

5.6 Operating System Tuning

To maximize Hadoop performance, NetApp recommends that you configure some operating-system-level parameters, as discussed in Table 6. You can configure these settings in the `/etc/sysctl.conf` and `/etc/fstab` files on each worker node.

[6] Hadoop. "HDFS Architecture Guide." Last published August 4, 2013. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed December 22, 2017).

Table 6) Operating system tuning. [7] [8] [9]

Name	Description	Value to Be Set
`vm.swappiness`	Defines how aggressively the kernel swaps memory pages. Higher values increase aggressiveness, lower values decrease the amount of swap. A value of 0 instructs the kernel not to initiate a swap until the amount of free and file-backed pages is less than the high watermark in a zone.	`vm.swappiness=1`
`vm.dirty_background_ratio`	Contains, as a percentage of total available memory that contains free pages and reclaimable pages, the number of pages at which the background kernel flusher threads start writing out dirty data.	`vm.dirty_background_ratio=20`
`vm.dirty_ratio`	Contains, as a percentage of total available memory that contains free pages and reclaimable pages, the number of pages at which a process that is generating disk writes will itself start writing out dirty data.	`vm.dirty_ratio=50`
`vm.overcommit_ratio`	A percentage added to the amount of RAM when deciding how much the kernel can overcommit.	`vm.overcommit_ratio=100`

[7] Kernel.org. “Documentation for /proc/sys/vm/*.” https://www.kernel.org/doc/Documentation/sysctl/vm.txt (accessed August 3, 2018).
[8] Red Hat. “What Are Huge Pages and the Advantages of Using Them?” https://access.redhat.com/solutions/2592 (accessed August 3, 2018).
[9] ArchWiki. “fstab.” https://wiki.archlinux.org/index.php/fstab (accessed August 3, 2018).

Name	Description	Value to Be Set
`Transparent Huge Pages`	The Huge Pages feature allows the Linux kernel to use the multiple page size capabilities of modern hardware architectures. Linux creates multiple pages of virtual memory, mapped from both physical RAM and swap. A page is the basic unit of virtual memory, with the default page size being 4,096 bytes in the x86 architecture.	`Disabled`
`noatime`	Completely disables writing file access times to the drive each time that you read a file. `noatime` implies `nodiratime`. You do not need to specify both.	`Enabled`

In addition to the kernel-level tuning discussed in Table 6. NetApp recommends tuning the servers from the BIOS settings for `Performance`. For example, because we used Fujitsu servers, we tuned them following the recommendations from Fujitsu.

6 Certification Tests

To certify NetApp and Hortonworks Hadoop solutions, we used Hive 2.1 to run interactive queries by using a 10TB TPC-DS dataset. All the tests were run using Hortonworks Data Platform v2.6.5 and a NetApp E5700 storage array. The hardware details of our test cluster are mentioned in Section 5.1.

We tuned our Hive 2 server according to the recommendations by Hortonworks. Hive 2 testing included running 99 SQL interactive parallel queries using 1, 5, 10, and 20 users. Out of the 99 queries, 97 of the queries successfully completed.

7 Summary

This document discusses the best practices for creating a Hadoop cluster using Hortonworks's Hadoop distribution (Hortonworks Data Platform). It also describes how the NetApp E-Series storage system can help your organization attain maximum throughput when running Hadoop jobs. With the NetApp and Hortonworks certified Hadoop solution, you can decrease the overall cost of ownership and gain enhanced data security, increased flexibility, and simple scalability.

Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and websites:

- Multipath Drivers Guide
  https://library.netapp.com/ecmdocs/ECMP12404601/html/frameset.html
- NetApp Documentation Centers
  https://mysupport.netapp.com/info/web/ECMLP2557637.html

- E-Series Systems Documentation Center
  https://mysupport.netapp.com/info/web/ECMP1658252.html
- DDP technology
  TR-4115: SANtricity Dynamic Disk Pools Best Practices Guide
- Disk encryption
  TR-4474: SANtricity Full Disk Encryption
- NetApp Solutions for Hadoop
  www.netapp.com/us/media/wp-7196.pdf
- Determine Yarn and MapReduce Memory Configuration Settings by Hortonworks
  https://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.0.9.1/bk_installing_manually_book/content/rpm-chap1-11.html
- NameNode High Availability
  https://docs.hortonworks.com/HDPDocuments/Ambari-2.5.1.0/bk_ambari-operations/content/namenode_high_availability.html
- SSD read cache
  TR-4099: NetApp SANtricity SSD Cache for E-Series
- E-Series pages
  www.netapp.com/us/products/storage-systems/hybrid-flash-array/index.aspx#E-Series
- Hortonworks website
  https://hortonworks.com/partner/netapp/

Acknowledgments

I would like to thank the following NetApp and Hortonworks experts for their input and assistance:

- Mitch Blackburn, Technical Marketing Engineer, NetApp
- John Ryan, Big Data Global Alliances Director, NetApp
- Doug Reid, Director of Program Management, Hortonworks
- Gavin Welch, Channel Alliances Manager, Hortonworks
- Karthikeyan Nagalingam, Senior Architect, NetApp
- Nilesh Bagad, Senior Product Manager, NetApp
- Orla Specht, Editor, NetApp
- Anne Szabla, Editor, NetApp
- Lee Dorrier, Director, Data Fabric Group, NetApp
- Nadeem Asghar, Global CTO (Field), Hortonworks

Version History

Version	Date	Document Version History
Version 1.0	September 2018	Hortonworks certification for NetApp solutions for Hadoop.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.