White Paper

NetApp Solutions for Hadoop

Reference Architecture: Cloudera

Faiz Abidi (NetApp) and Udai Potluri (Cloudera)

June 2018 | WP-7217

In partnership with

cloudera®

Abstract

There has been an exponential growth in data over the past decade and analyzing huge amounts of data in a reasonable time can be a challenge. Apache Hadoop is an open-source tool that can help your organization quickly 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 Cloudera 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 less time spent on operations and maintenance. This white paper discusses a flexible, validated, enterprise-class Hadoop architecture that is based on NetApp E-Series storage using Cloudera's Hadoop distribution.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

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

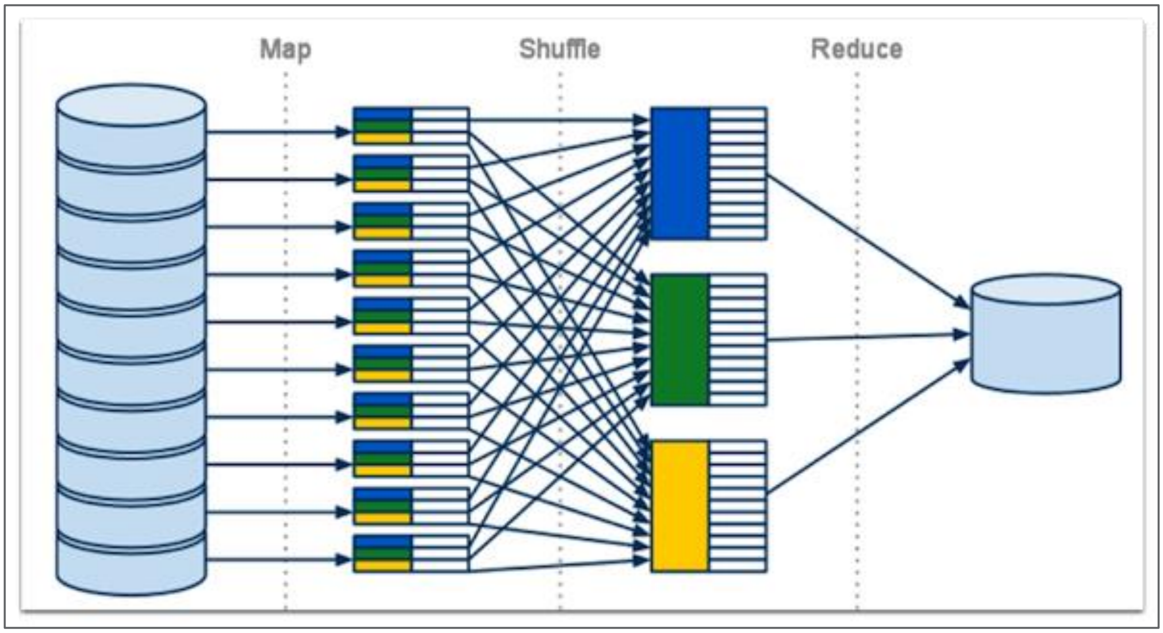
1.1 Big Data

Data has been growing at a speed that no one could have predicted 10 years ago. The constant influx of data 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 becomes imperative to analyze this data and 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, which can use this information to tailor advertisements for 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 (see Figure 1 for an overview). 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 on which the failed node was working gets passed on to another running node.

Figure 1) MapReduce architecture.[6]



The origin of Hadoop can be traced back to a 2003 paper released by Google that talks about the Google File System.[2] Since then, a lot of effort has gone into developing Hadoop into a robust, scalable, and highly reliable project. Companies such as Yahoo! IBM, Cloudera, 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) Components of Hadoop.[7]

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 bandwidth with low latency to support the demanding performance and capacity needs of science, 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 great 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
- An industry-leading level of performance, density, and economics
- Interface protocol flexibility to simultaneously support FC host and iSCSI host workloads
- 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

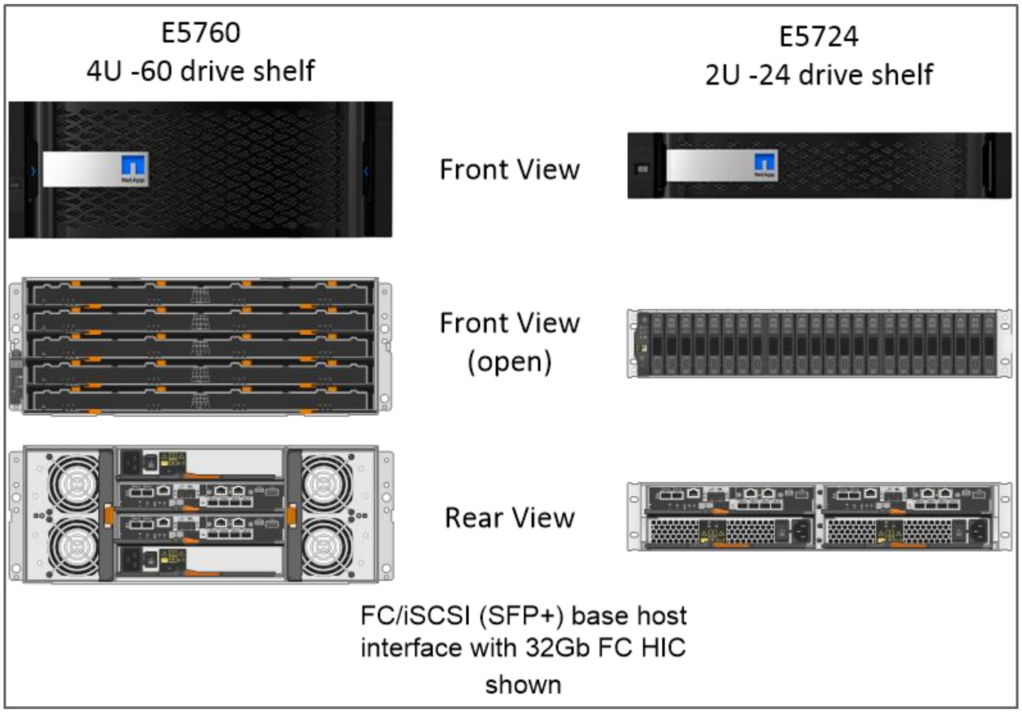
As shown in Table 2, the E5700 is available in two shelf options that support both HDDs and solid-state drives (SSDs) to meet a wide range of performance and application requirements.

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 lanes wide) SAS drive expansion ports for redundant drive expansion paths. The E5700 controllers also include two built-in host ports that 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 when you configure SANtricity, you must first 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 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 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.

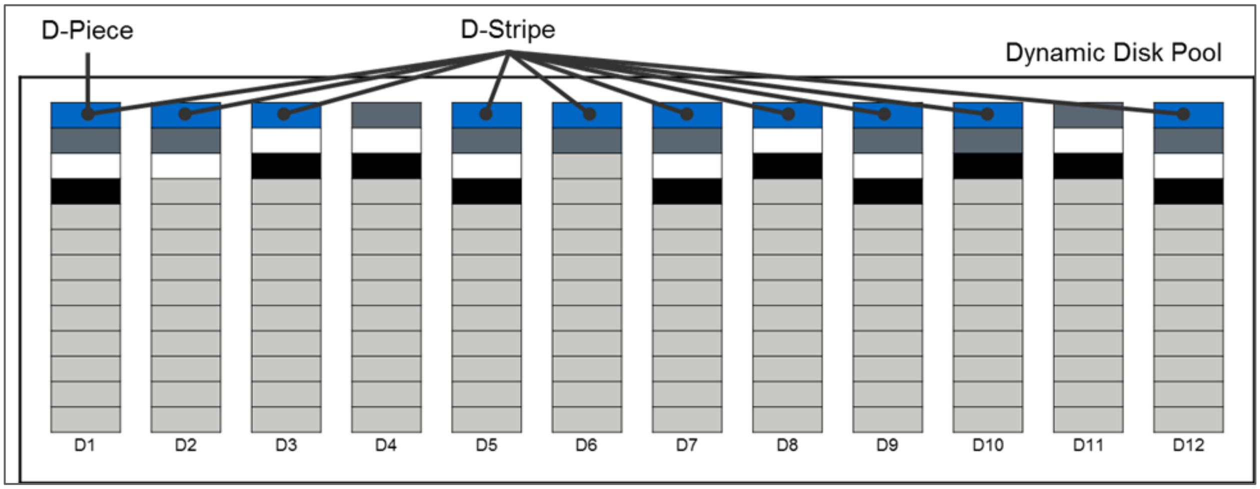
Dynamic Disk Pool Features

The 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. In a pool, 10 D-pieces are selected by using an intelligent optimization algorithm from selected drives in the pool. Together, the 10 associated D-pieces are considered to be a D-stripe, which is 4GB of usable capacity in size. In 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 in a pool. Figure 3 shows the relationship between these data structures.

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

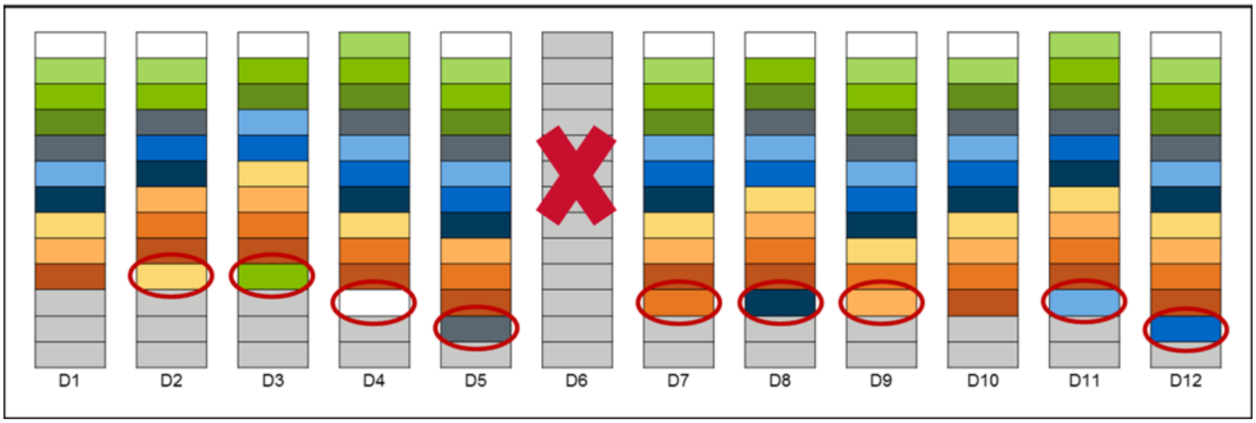


Another major benefit of a DDP pool is that rather than using dedicated stranded hot spares, the pool contains integrated preservation capacity to provide rebuild locations for potential drive failures. This approach simplifies management, because individual hot spares no longer need to be planned or 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 D-pieces from the failed drive are reconstructed 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 recreated 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 in a DDP pool, priority for reconstruction is given to any D-stripes that are missing two D-pieces, thus minimizing data availability risk. 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 in 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 must 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 for the base volume under a failure condition than with traditional volume groups.

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 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 FDE enabled, data is encrypted at rest on the media and is automatically decrypted on a read request.

When the FDE feature is enabled on the storage array, data at rest is protected 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 allowing 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 are invisible to users and do 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 Advanced Encryption Standard (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 at volume group creation, disk pool creation, or afterward. This selection does not affect existing data on the drives and can be used to secure the data after creation. 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 feature with an internally managed security key.

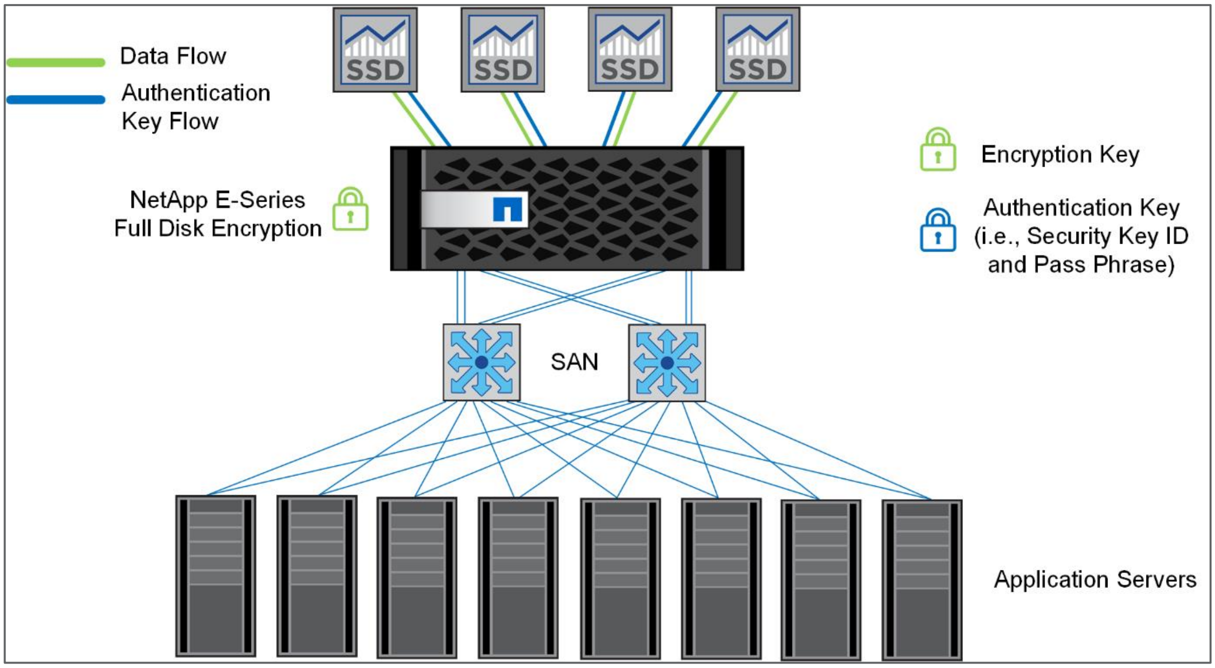
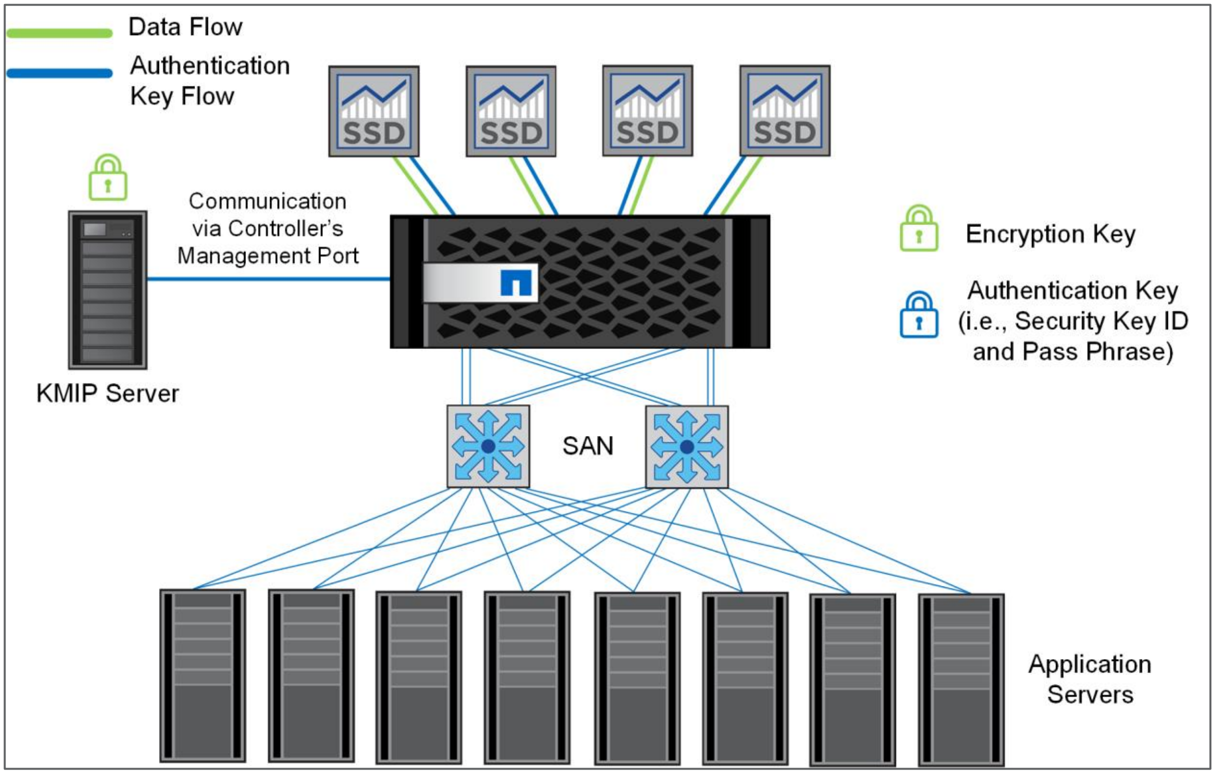

Data Flow
Authentication Key Flow
SSD
SSD
SSD
SSD
NetApp E-Series Full Disk Encryption
Encryption Key
Authentication Key (i.e., Security Key ID and Pass Phrase)
SAN
Application Servers

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


Data Flow
Authentication Key Flow
SSD
SSD
SSD
SSD
Communication via Controller's Management Port
KMIP Server
Encryption Key
Authentication Key (i.e., Security Key ID and Pass Phrase)
SAN
Application Servers

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. The primary purpose of a scan is to detect and repair media errors on disks that are infrequently read by user applications and where 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 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 normal I/O operation of the controller and during long-lived operations such as reconstructions. The 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 to see whether the fault condition can be cleared. 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 based on 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 approximately one-hour 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 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 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 performance requirements of a Hadoop cluster deployment.

An E5700 with 24 SSDs can provide up to one million 4K random read IOPS at less than 100µs 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, including 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 system under test used 48 SSDs, 4K and 16K block sizes, and 25% and 75% read workloads.

Figure 7) Expected system performance for write-heavy workloads on an E5700.

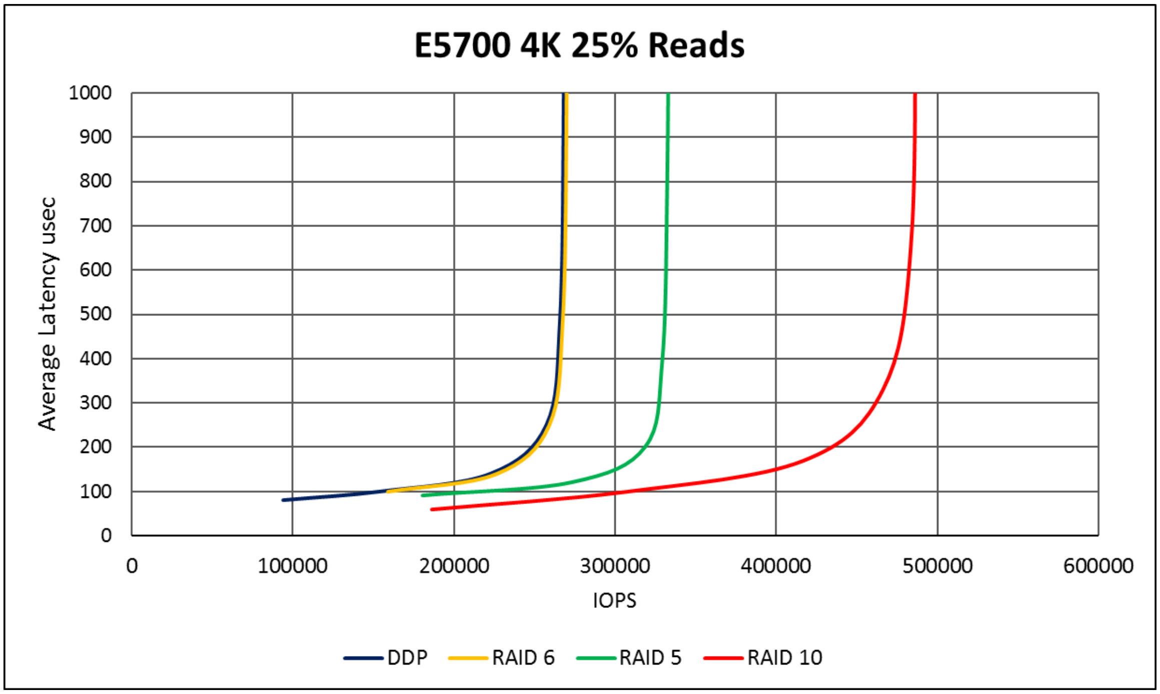
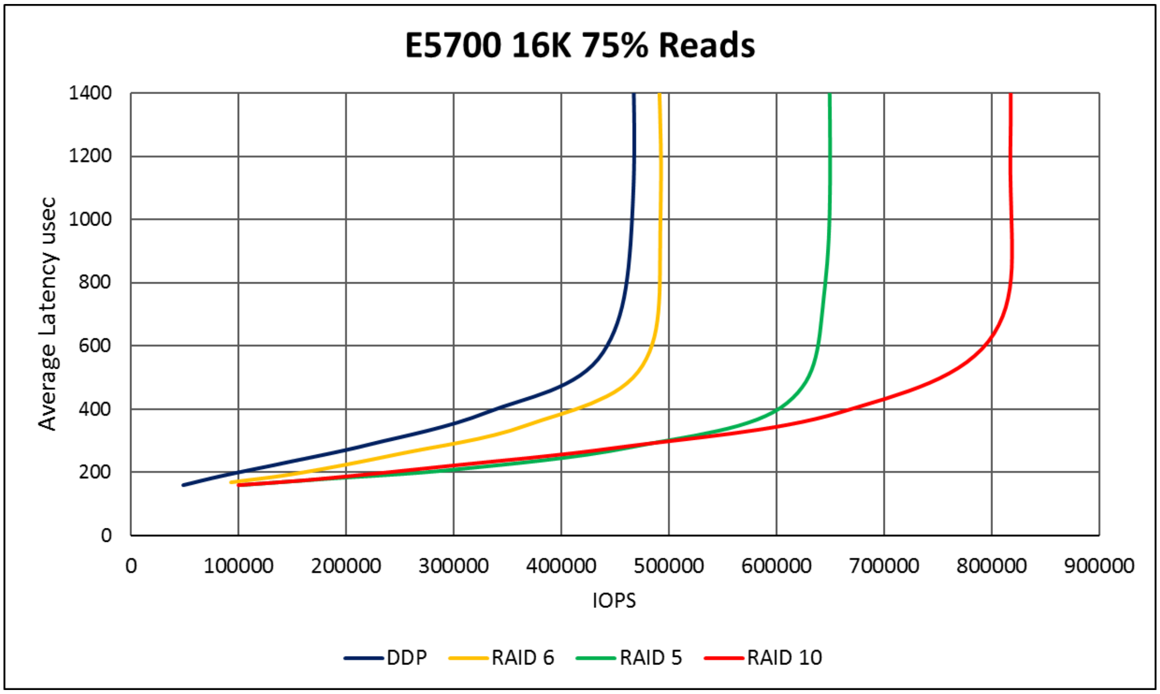


Figure 8) Expected system performance for a read-heavy workload on an E5700.



Capacity

The E5760 has a maximum capacity of 4800TB (with expansion drive shelves), using 480 NL-SAS HDDs 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 E5700.

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 Cloudera Overview

Cloudera was founded in 2008 by some of the brightest minds at Silicon Valley's leading companies, including Google (Christophe Bisciglia), Yahoo! (Amr Awadallah), Oracle (Mike Olson), and Facebook (Jeff Hammerbacher). Cloudera's founders held at their core the belief that open source, open standards, and open markets are best. That belief remains central to their values. Doug Cutting, cocreator of Hadoop, joined the company in 2009 as chief architect and remains in that role. Today, Cloudera has more than 1,600 employees. The company has offices in 24 countries around the globe, with headquarters in Palo Alto, California.

Cloudera enables you to transform vast amounts of complex data into clear and actionable insights to enhance your business and exceed your expectations. The world's leading organizations choose Cloudera to grow their businesses, improve lives, and advance human achievement. Achieve the impossible with Cloudera.

3.1 CDH Overview

Cloudera Enterprise Data Hub is the fastest, most secure, and easiest big data software available. From data science and engineering, to powering an operational database, to running large-scale analytics, you get it all in this single, easy-to-use product. Use Cloudera Enterprise Data Hub to turn your data into real business value.

Fast for Business

Only Cloudera Enterprise enables more insights for more users, all in a single platform. With the most powerful open-source tools and the only active data optimization designed for Hadoop, you can move from big data to results faster.

Easy to Manage

Hadoop is a complex, evolving ecosystem of open-source projects. Only Cloudera Enterprise makes it simple so that you can run at scale across a variety of environments, all while meeting SLAs.

Secure Without Compromise

The potential of big data is huge, but not at the expense of security. Cloudera Enterprise is the only Hadoop platform to achieve compliance with its comprehensive security and governance.

3.2 Cloudera Products

Cloudera Analytic DB

Business intelligence and analytics continue to be the lifeblood of business. But many enterprises are stuck in a pattern of just trying to keep up with the status quo, instead of shifting for the future. Cloudera's modern analytics database, powered by Apache Impala, is the only solution that brings high-performance SQL analytics to big data. Power your business with the platform that can meet the needs of today and evolve to meet those of tomorrow.

Cloudera Operational DB

Relational or NoSQL, structured or unstructured, operational DB delivers insights at the speed of business. Data is driving modern business. Supplied with the right data at the right time, decision makers across industries can guide their organizations toward improved efficiency, new customer insights, better products and services, and decreased risk. Cloudera's operational database—powered by open-source technologies such as Apache HBase, Apache Kudu, and Apache Spark—delivers a secure, low-latency, high-concurrency experience that can extract the real-time insights you need from big data.

Cloudera Data Science and Engineering

Never leave your predictions to chance. Cloudera Data Science provides better access to Apache Hadoop data with familiar and performant tools that address all aspects of modern predictive analytics. Using Cloudera, your organization will be able to perform advanced data engineering, exploratory data science, and machine learning at scale. And that's regardless of where your data lives: on the premises, across public clouds, or both. Because the right insights today lead to better business decisions tomorrow.

Cloudera Data Science Workbench

Machine learning is all about the data, but it's often out of reach for analytics teams working at scale. Cloudera Data Science Workbench enables fast, easy, and secure self-service data science for the enterprise. CDSW helps accelerate data science from exploration to production using R, Python, Spark, and more for both data scientists and IT professionals.

4 Hadoop Enterprise Solution with NetApp

This section discusses the advantages of using NetApp to build Hadoop solutions and how you can decrease the total cost of ownership while also building a secure, stable, and scalable Hadoop cluster. We also briefly discuss how data locality for Hadoop is no longer important, and, in fact, segregating compute and storage makes your Hadoop cluster more flexible and stable.

4.1 Data Locality and Its Insignificance for Hadoop

Among the initial primary concepts in Hadoop to improve performance was to move compute to data or to colocate compute and storage. This approach meant 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 in 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 is only 8% faster than reading it from remote disks,[4] and this number only decreases with time. Networks are getting faster, but the disks are not. As an example, Ananthanarayanan et al.[4] analyzed logs from Facebook and concluded that "disk-locality results in little, if any, improvement of task lengths."

With all the advancements being made in improving the network infrastructure, data compression, and deduplication, under the right conditions, colocating 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 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 maximum disk I/O for low cost and delivers sustained high bandwidth and IOPS. E-Series systems run a separate operating environment, called NetApp SANtricity. 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 the Total Cost of Ownership

To be fault tolerant and reliable, HDFS allows users to set a replication factor for data blocks. A replication factor of three means that three copies of the data are 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 configuration, setting a replication factor of three is recommended to significantly decrease the probability of data loss. This also leverages the concept of data locality, thereby 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'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 two instead of three when using 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 allows you to spend less on 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 using RAID, the drive rebuild times can be several hours to several 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.

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 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 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. When compute and storage components are segregated, a cluster is more stable and easier to maintain than if compute and storage were grouped together, allowing you to concentrate on your business needs. This solution flattens the operational learning curve of Hadoop.

4.7 Cloudera Certified

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

5 Solution Architecture and Setup

This section talks about the hardware used to build, test, and validate our solution with Cloudera's Hadoop distribution. We also briefly discuss some best practices to keep in mind when building a Hadoop cluster.

5.1 Architectural Pipeline and Hardware Details

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

Servers

- Eight Fujitsu servers with 48 vCPUs and 256GB of memory
- Three master servers with 24vCPUS and 64GB of memory

Storage

- One NetApp E5700 storage array
- Ninety-six 10TB HDDs (7200 RPM); 8 DDP pools with 12 drives each; each pool has one volume mapped to one host

Network

- All servers use a 10Gb dual-port network connection
- 12Gbps SAS connections to the E-Series storage array

Software

- RHEL 7.4 on all servers
- Cloudera distribution 5.13.1 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 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 4 shows all the alternative products supported by NetApp and its partners as of the writing of this paper.

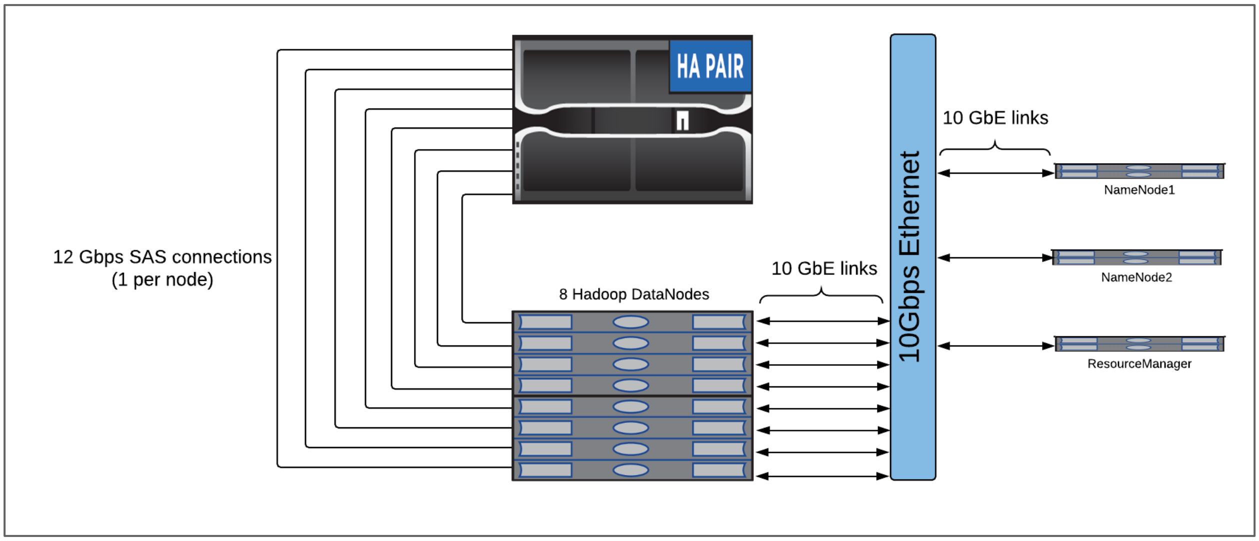
Table 4) Alternative products supported by NetApp and its partners.

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

Best Practice

For optimal performance from 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 eight compute nodes connected to it. We created eight pools by using the 96 drives, and each pool had 12 drives. We then 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 (4K, for example) means that 262,144 requests must be made to transfer a 1GB file. That many requests across a network would cause tremendous overhead. The block size that you select depends largely on the size of the input data, the access patterns, and the server configuration. It usually requires some trial and error to find out the block size that provides 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 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."[3]

5.3 Hadoop Tuning

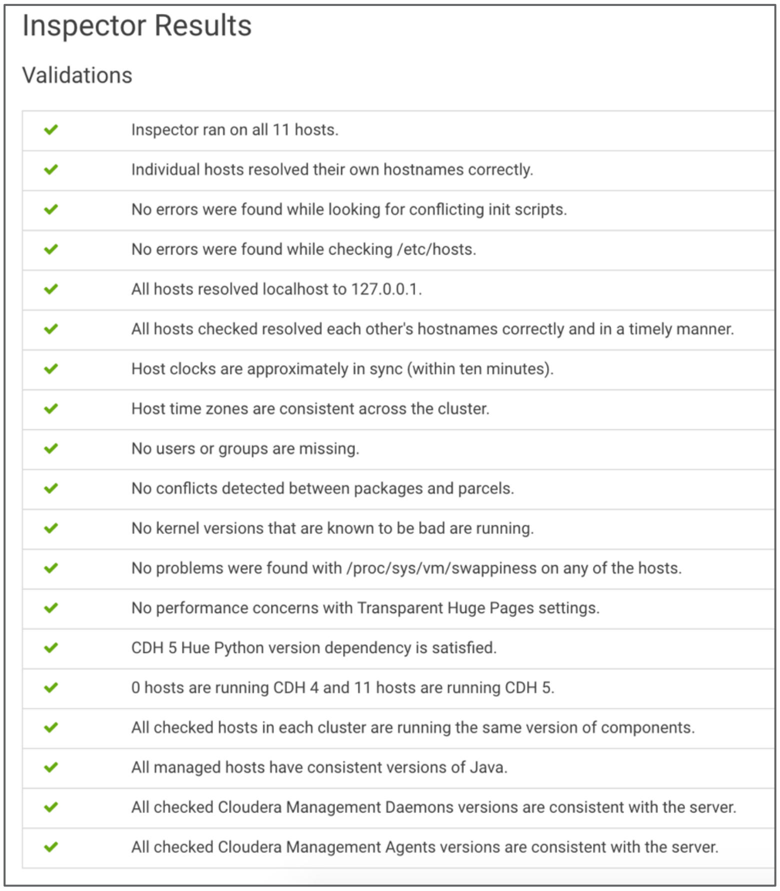
It can be challenging to optimally configure Hadoop for best performance. You must decide how many mappers to use, how much memory and how many cores to allocate to mappers and to reducers, how much buffer memory to use while sorting files, and so on. By properly configuring all these parameters, you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

Cloudera provides an Excel spreadsheet that can help you tune YARN clusters. For more information, see Tuning YARN by Cloudera.

5.4 Cloudera Hosts Inspection

Cloudera Manager provides an easy-to-use host inspection utility that performs basic sanity checking on all the hosts in the cluster. It makes sure that things such as swappiness, Java version, transparent huge pages, and other similar options are configured properly on all the hosts. Figure 10 shows the results of the host inspector running on our cluster.

Figure 10) Cloudera Manager running host inspection utility.



Inspector Results

Validations

✓	Inspector ran on all 11 hosts.
✓	Individual hosts resolved their own hostnames correctly.
✓	No errors were found while looking for conflicting init scripts.
✓	No errors were found while checking /etc/hosts.
✓	All hosts resolved localhost to 127.0.0.1.
✓	All hosts checked resolved each other's hostnames correctly and in a timely manner.
✓	Host clocks are approximately in sync (within ten minutes).
✓	Host time zones are consistent across the cluster.
✓	No users or groups are missing.
✓	No conflicts detected between packages and parcels.
✓	No kernel versions that are known to be bad are running.
✓	No problems were found with /proc/sys/vm/swappiness on any of the hosts.
✓	No performance concerns with Transparent Huge Pages settings.
✓	CDH 5 Hue Python version dependency is satisfied.
✓	0 hosts are running CDH 4 and 11 hosts are running CDH 5.
✓	All checked hosts in each cluster are running the same version of components.
✓	All managed hosts have consistent versions of Java.
✓	All checked Cloudera Management Daemons versions are consistent with the server.
✓	All checked Cloudera Management Agents versions are consistent with the server.

5.5 NameNode High Availability

In Hadoop 1.x, the NameNode was a single point of failure in that when a NameNode went down, the entire cluster would become unavailable for use and 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, in the case of a NameNode failure, the standby NameNode provides automatic failover, making sure that HDFS does not go down and the Hadoop cluster remains online.

The ZooKeeper's architecture supports HA through redundant services. For this reference architecture, we used three ZooKeeper servers, one acting as the leader and two as followers. We enabled HA from the path Cluster Name -> HDFS -> Actions -> Enable High Availability.

For more details about HA, refer to the HDFS High Availability document by Cloudera.

5.6 Rack Awareness

Hadoop components are rack aware in the sense that they have knowledge of the cluster topology. In other words, they know how data is distributed across different racks in a cluster. Rack awareness can be helpful in cases in which hundreds of nodes are spread across different racks, and the basic assumption is that the worker nodes on the same rack have lower latency and higher bandwidth. Rack awareness also increases data availability, because, 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. Therefore, if one of the racks goes down, data can still be retrieved from the other rack.

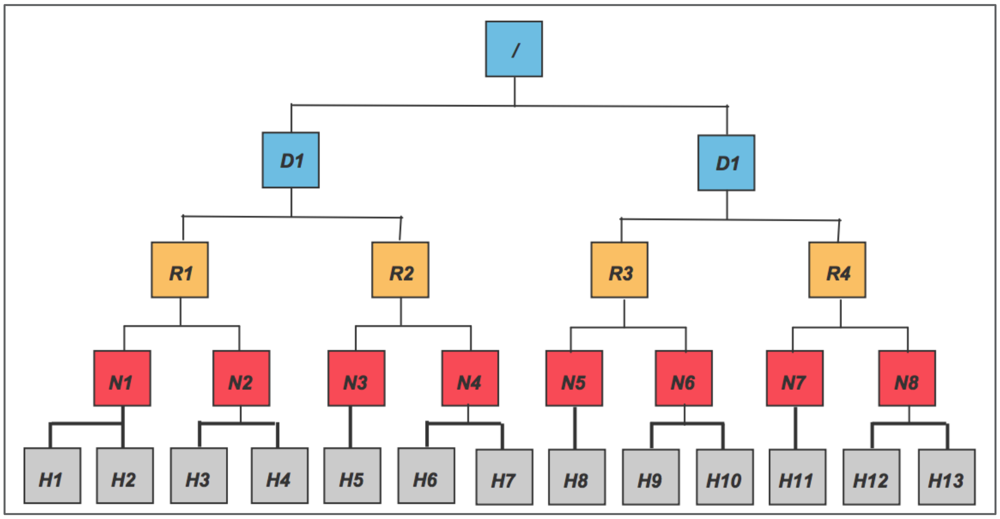
5.6.1 Hadoop Virtualization Extensions

The Hadoop Virtualization Extensions (HVE) feature extends rack awareness to leverage the storage arrays as another layer of abstraction for block placement. The storage arrays should be configured as the NodeGroup layer. The NodeGroups represent the physical hypervisor on which the nodes reside. Cloudera recommends using this additional level of rack awareness for storage arrays.

> "With the addition of the NodeGroup layer in the Hadoop topology to enable awareness of virtual nodes on the same physical host, HVE refines Hadoop's replica placement, task scheduling and balancer policies to achieve reliability and performance enhancements for Hadoop running on virtualized environments."[5]

HVE introduces a new layer in the topology hierarchy as shown in Figure 11, where D = data center, R = rack, N = NodeGroup, and H = host. For more details about HVE, refer to reference 5. Using HVE, other single points of failure such as power supplies and network switches can also be supported.

Figure 11) Network topology extension.[5]



For testing purposes in this report, we connected eight servers to an E5700 storage array, which means that four servers were connected to each controller. In a scenario in which one of the storage controllers failed, we wanted to make sure that the entire cluster did not go down and the data could still be accessed from the other storage controller (we used a replication factor of two). Therefore, we created two NodeGroups in the same rack: NodeGroup1 and NodeGroup2, each belonging to one controller. This meant that even if one of the controllers went offline, data could still be accessed from the other controller, and Hadoop jobs would continue to run successfully. However, if you want to use multiple E-Series storage arrays in a Hadoop cluster, you would allocate a Rack R to each storage array and further group the controllers on each storage array into two NodeGroups: NodeGroup1 and NodeGroup2. Having multiple racks increases data availability.

Note that in the case where only four servers are connected to an E-Series storage array, the controllers are no longer a single point of failure because each server is connected to both the controllers, providing failover capabilities, and HVE implementation is not needed.

Best Practice

When using eight servers connected to an E-Series box, HVE must be enabled. This configuration makes sure that the data is not lost if a controller goes offline. If a server is connected to a controller, for example, A, the corresponding volumes mapped to the server must also be owned by the same controller. You can change ownership of volumes by using SANtricity.

Steps to Enable HVE Using the Cloudera Manager

1. Edit HDFS `core-site.xml` (ClusterName > HDFS > Configuration. Search for `core-site.xml`). In the property Cluster-wide Advanced Configuration Snippet (Safety Valve) for `core-site.xml`, paste the following information:

```
<property>
    <name>net.topology.impl</name>
    <value>org.apache.hadoop.net.NetworkTopologyWithNodeGroup</value>
</property>
<property>
    <name>net.topology.nodegroup.aware</name>
    <value>true</value>
</property>
<property>
    <name>dfs.block.replicator.classname</name>

<value>org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyWithNodeGroup</value>
</property>
```

2. Edit YARN `mapred-site.xml` (ClusterName > YARN (MR2 Included) > Configuration). Search for `mapred-site.xml`. In the property, HDFS Replication Advanced Configuration Snippet (Safety Valve) for `mapred-site.xml`, paste the following information:

```
<property>
    <name>mapred.jobtracker.nodegroup.aware</name>
    <value>true</value>
</property>
<property>
    <name>mapred.task.cache.levels </name>
    <value>3</value>
</property>
```

3. Set the rack location of the hosts (ClusterName > Hosts). See Specifying Racks for Hosts to learn how to perform this task. In our case, we created two NodeGroups, 1 and 2, each belonging to the same rack. Each NodeGroup had four servers. The master servers can be in any NodeGroup.

6 Certification Tests

To certify NetApp and Cloudera Hadoop solutions, we ran some basic validation tests along with disruptive tests using Cloudera's Hadoop distribution and a NetApp E5700 storage array.

The hardware details of our test cluster are described in section 5.1. Validation tests were run that included long-running TeraGen, TeraSort, TeraValidate, and DFSIO read/write jobs. Specific metrics were monitored to verify that good throughput was achieved with no failures.

In addition to the validation tests, we also performed disruptive testing that included the following:

- **Disk failure.** While running Hadoop jobs, two drives were failed from SANtricity. The jobs still completed successfully without any task failures. The two spare drives in the disk pool replaced the two failed drives, and no data was lost.
- **Controller failure.** While running Hadoop, we failed one of the controllers from SANtricity. As a result, four of the worker nodes managed by that controller went offline, and the tasks running on

those worker nodes were rescheduled to run on the other four worker nodes in the cluster. When four of the worker nodes went down, there were some task failures, but overall there was no data lost because we used a replication factor of two and had enabled HVE. Therefore, one block of the data was still available and was managed by the second controller. Therefore, the job completed successfully.

Conclusion

This document discusses best practices for creating a Hadoop cluster using Cloudera's Hadoop distribution and how the NetApp E-Series storage system can help your organization attain maximum throughout when running Hadoop jobs. With the NetApp and Cloudera certified Hadoop solution, you can decrease the overall cost of ownership and gain enhanced data security, flexibility, and scalability.

Acknowledgments

We would like to thank the following NetApp and Cloudera experts for their input and assistance:

- Karthikeyan Nagalingam, senior architect (big data analytics and databases), NetApp
- Mitch Blackburn, technical marketing engineer, NetApp
- Alex Moundalex, software engineer, Cloudera
- Dwai Lahiri, software engineer, Cloudera
- Calvin Goodrich, engineering manager, Cloudera
- Vinod Singh, program manager, Cloudera
- Dorian Henderson, technical editor, NetApp
- Lee Dorrier, director, Data Fabric group, NetApp
- Nilesh Bagad, senior product manager, NetApp

Where to Find Additional Information

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

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. Ghemawat, S., Gobioff, H., and Leung, S.-T., “The Google File System,” *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5., December 2003.
3. Hadoop. “HDFS Architecture Guide.” Last published August 4, 2013. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed December 22, 2017).
4. 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.
5. Apache Hadoop. “Umbrella of enhancements to support different failure and locality topologies.” https://issues.apache.org/jira/browse/HADOOP-8468 (accessed March 26, 2018).
6. Marcel Caraciolo. “Introduction to Recommendations with Map-Reduce and mrjob.” http://aimotion.blogspot.com/2012/08/introduction-to-recommendations-with.html (accessed May 14, 2018).
7. Apache Hadoop. “What Is Apache Hadoop?” http://hadoop.apache.org/ (accessed May 15, 2018).

Version History

Version	Date	Document Version History
Version 1.0	July 2015	Cloudera certification for NetApp solutions for Hadoop
Version 2.0	June 2018	Cloudera certification for NetApp solutions for Hadoop using E5700

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.

Copyright Information

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Trademark Information

NetApp®