CLOUDERA

HITACHI
Inspire the Next

Hitachi Analytics Infrastructure using Hitachi Unified Compute Platform RS with Cloudera Enterprise Data Hub

Reference Architecture Guide

By Hitachi Vantara

July 2020

Feedback

Hitachi Vantara welcomes your feedback. Please share your thoughts by sending an email message to SolutionLab@HitachiVantara.com. To assist the routing of this message, use the paper number in the subject and the title of this white paper in the text.

Revision History

Revision	Changes	Date
SL-020-00	Initial release	January 29, 2018
SL-020-01	Update title page to indicate authors' job title.	February 5, 2018
MK-SL-020-02	Add validation information. Updated component deployment options and content list of the first rack in a multiple rack configuration.	February 28, 2018
MK-SL-020-03	Update author information.	May 16, 2018
MK-SL-020-04	Add support for Hitachi Advanced Server DS220	August 22, 2018
MK-SL-020-05	Add Cascade Lake CPUs and new Advanced Server DS220 models.	February 21, 2020
MK-SL-020-06	Add additional, related products	July 28, 2020

Table of Contents

Hitachi Analytics Infrastructure using Hitachi Unified Compute Platform RS with Cloudera Enterprise Data Hub

Reference Architecture Guide

Accelerate the deployment of your analytics infrastructure. Leverage this reference architecture guide to implement Hitachi Analytics Infrastructure using Hitachi Unified Compute Platform RS with Cloudera Enterprise Data Hub. Use this guide to implement an architecture that maximizes the return on your investment.

This integrated big data infrastructure for advanced analytics uses the following:

- **Hitachi Advanced Server DS120** – This is a flexible 1U server designed for optimal performance across multiple applications.
- **Hitachi Advanced Server DS220** – This is a flexible 2U server designed for optimal performance across multiple applications.
- **Cloudera Enterprise Data Hub (CDH)** – Cloudera Enterprise Data Hub, powered by Apache Hadoop, enables an enterprise data hub together with the security, governance, management, support, and commercial ecosystem.
- **Cisco Nexus 3048** – This 48-port 1 GbE switch provides a management network. It is used both as a leaf switch and a spine switch.
- **Cisco Nexus 93180YC-E/FX** – This 48-port switch provides 10 GbE connectivity for intra-rack networks. It is used as the leaf switch for the data network. Designed with Cisco Cloud Scale technology, it supports highly scalable cloud architectures.
- **Cisco Nexus 9336** – This 100 GbE switch provides connectivity for inter-rack networks. It is used as the spine switch for the data network, supporting flexible migration options. It is ideal for highly scalable cloud architectures and enterprise datacenters.

Note – Testing of this configuration was in a lab environment. Many things affect production environments beyond prediction or duplication in a lab environment. Follow the recommended practice of conducting proof-of-concept testing for acceptable results in a non-production, isolated test environment that otherwise matches your production environment before your production implementation of this solution.

Key Solution Elements

These are the key hardware and software components to power this big data solution. You can create a scale-out configuration to power your Cloudera environment.

This solution supports using the Hitachi Advanced Server DS120 and the Hitachi Advanced Server DS220.

Hitachi Advanced Server DS120

Optimized for performance, high density, and power efficiency in a dual-processor server, Hitachi Advanced Server DS120 delivers a balance of compute and storage capacity. This 1U rack mounted server has the flexibility to power a wide range of solutions and applications.

The highly-scalable memory supports up to 3 TB using 24 slots of 2666 MHz DDR4 RDIMM. Advanced Server DS120 is powered by the Intel Xeon Scalable processor family for complex and demanding workloads. There are flexible OCP and PCIe I/O expansion card options available. This server supports up to 12 small form factor storage devices with up to 4 NVMe drives.

This solution allows you to have a high CPU to storage ratio. This is ideal for balanced and compute-heavy workloads.

Multiple CPU and storage devices are available. Contact your Hitachi Vantara sales representative to get the latest list of options.

Hitachi Advanced Server DS220

With a combination of two Intel Xeon Scalable processors and high storage capacity in a 2U rack-space package, Hitachi Advanced Server DS220 delivers the storage and I/O to meet the needs of converged solutions and high-performance applications in the data center.

The Intel Xeon Scalable processor family is optimized to address the growing demands on today's IT infrastructure. The server provides 24 slots for high-speed DDR4 memory, allowing up to 3 TB of memory per node when 128 GB DIMMs are used. This server supports up to 12 large form factor storage devices and an additional 2 small form factor storage devices.

This server has three storage configuration options:

- 12 large form factor storage devices and an additional 2 small form factor storage devices in the back of the chassis
- 16 SAS or SATA drives, 8 NVMe drives, and an additional 2 small form factor storage devices in the back of the chassis
- 24 SFF devices and an additional 2 SFF storage devices in the back of the chassis

This server allows you to have a dense storage solution with lower power consumption. The small form factor options allow you to have more drives per chassis. The larger form factor also provides you with more expansion options.

Multiple CPU and storage devices are available. Contact your Hitachi Vantara sales representative to get the latest list of options.

Hitachi Unified Compute Platform Advisor

Hitachi Unified Compute Platform Advisor (UCP Advisor) brings simplified IT administration to virtualized, converged, and hyperconverged systems from Hitachi. Unified Compute Platform Advisor supports guided life-cycle management to the server, network, and storage elements within Unified Compute Platform systems.

Unified Compute Platform Advisor is used to discover and provision servers initially, and later to manage the compute nodes:

- Identify Unified Compute Platform servers for remote management.
- Provision servers.
- Image the custom BIOS settings on the server
- Install the operating system.
- Upgrade the installed firmware,
- Power cycle a compute node remotely.
- Launch a remote console for a server.
- Provides remote access to general system information.

After deploying a bare metal operating system on the physical server with the network configured, the environment is ready for configuration management. Any tools, such as Puppet, Chef, or Ansible can be used to add additional packages, services, and patches to the operating system.

Cisco Switches

Cisco Nexus data center switches are built for scale, industry-leading automation, programmability, and real-time visibility.

This solution uses the following Cisco switches to provide Ethernet connectivity:

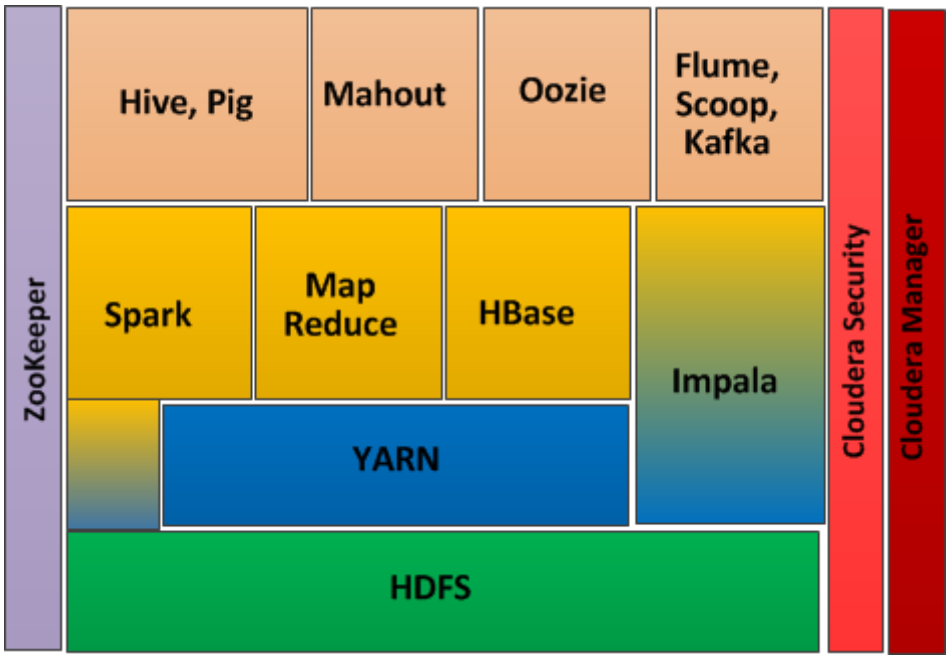
- Cisco Nexus 3048
- Cisco Nexus 93180YC-E/FX
- Cisco Nexus 9336

This solution uses a leaf-spine network architecture. This architecture can be replaced to match the rest of the network configuration.

Cloudera Enterprise Data Hub

Cloudera is the leading provider of enterprise-ready, big data software and services. Cloudera Enterprise Data Hub is the market-leading Hadoop distribution. It includes Apache Hadoop, Cloudera Manager, related open source projects, and technical support. Figure 1 provides an overview of the Cloudera software.

Figure 1



Big Data is a generic term to cover a set of components that are used with very large data sets to provide advanced data analytics. Big data usually refers to large volumes of unstructured or semi-structured data.

Usually, a big data solution is part of the Apache Hadoop project. However, big data can include components from many different software companies.

This reference architecture uses Red Hat Enterprise Linux 7.6 and Cloudera Enterprise Data Hub (CDH) 6.3.

This reference architecture also supports Cloudera Data Platform (CDP) 7.x. This is the merged release of Cloudera Data Hub and Hortonworks Data Platform Enterprise. While node configurations do not change, some of the software and software versions do change.

When Cloudera Private Cloud is released, this architecture will support the core storage portion (CDP-DC). Some of the processing will move from CDP-DC to a Kubernetes cluster called CDP-PVC, a required and separate standalone cluster.

The following is a partial list of the software components and modules that can be used in a Cloudera Enterprise Data Hub deployment:

- **Apache Flume**

  Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

- **Apache Hadoop Distributed File System**

  Hadoop Distributed File System (HDFS) is a distributed high-performance file system designed to run on commodity hardware.

- **Apache Hadoop Common**

  These common utilities support the other Hadoop modules. This programming framework supports the distributive processing of large data sets.

- **Apache Hadoop YARN**

  Apache Hadoop YARN is a framework for job scheduling and cluster resource management. This splits the functionalities of the following into separate daemons:

  - **ResourceManager** interfaces with the client to track tasks and assign tasks to **NodeMangers** management
  - **NodeManager** launches and tracks execution on the worker nodes

- **Apache HBase**

  Apache HBase is a datastore built on top of HDFS.

- **Apache Hive**

  Apache Hive is data warehouse software that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

- **Apache Impala**

  Apache Impala is a distributed SQL Query engine for Apache Hadoop.

- **Apache Kafka**

  Apache Kafka is a distributed streaming platform.

- **Apache Kudu**

  Apache Kudu provides fast inserts, updates, and analytics on top of Apache Impala or Apache Spark.

- **Apache Mahout**

  Apache Mahout is a framework for building scalable machine learning applications.

- **Apache Oozie**

  Apache Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

- **Apache Pig**

  Apache Pig is a platform for analyzing large data sets that consists of coupling the following:

  - A high-level language for expressing data analysis programs
  - An infrastructure for evaluating these data analysis programs

- **Apache Spark**

  Apache Spark is a fast, general-purpose engine for large-scale data processing.

  - **Spark Master Node**

    In a Spark cluster, the master node oversees assigning tasks for the worker nodes to execute. It checks the status of those tasks and retrieves the results.

    Also, the master node can be used as a worker node, if necessary. In this solution, the master node will assign tasks to itself, as well as to worker nodes. This can be useful if there are few overall nodes in the cluster.

  - **Spark Worker Node or Nodes**

    The worker node or nodes in a Spark cluster do the work assigned to them by the master node. They connect to the master node, are assigned tasks, and execute those tasks. These nodes use the CPU and available storage.

- **Apache Sqoop**

  Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores, such as relational databases.

- **Apache ZooKeeper**

  Apache ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

  - **ZooKeeper Master Node**

    ZooKeeper is a high-availability system, whereby two or more nodes can connect to a ZooKeeper master node. The ZooKeeper master node controls the nodes to provide high availability.

  - **ZooKeeper Standby Master Node**

    When ZooKeeper runs in a highly-available setup, there can be several nodes configured as ZooKeeper master nodes. Only one of these configured nodes is active as a master node at any time. The others are standby active nodes.

    If the currently-active master node fails, then the ZooKeeper cluster itself promotes one of the standby master nodes to the active master node.

- **Cloudera Manager**

  Cloudera Manager provides management and monitoring of the Cloudera Hadoop distribution cluster.

- **HUE**

  HUE (Hadoop User Experience) is a web interface for analyzing data with Apache Hadoop.

Solution Design

Use this detailed design to create an integrated infrastructure to implement your big data and business analytics solution using hardware and software from Hitachi Vantara and Cloudera.

- "Server Architecture," starting on page 6
  - Master node
  - Worker node
  - Utility node
  - Edge node
  - Hardware management server
- "Worker Node Storage Considerations," starting on page 11
  - Heterogeneous Storage
  - Erasure Encoding
- "Network Architecture," starting on page 16
  - Switches
  - Data network
  - Client network
  - Management network
- "Deployment Options," starting on page 18
  - Visit Cloudera to see the current recommended software deployment options.
- "Rack Configuration," starting on page 19
  - Single rack configuration
  - Multiple rack configuration

This design does not limit the maximum number of nodes. Your solution size depends on the resources you need to deploy.

For large deployments, the recommendation is to validate that your network meets your individual requirements.

Server Architecture

This solution uses multiple Hitachi Advanced Server DS120 systems or Hitachi Advanced Server DS220 systems. The architecture supports using servers in multiple configurations. This guide does not list all possible options.

There are these types of nodes:

- "Master Node" on page 7
- "Sample Hadoop Worker Nodes" on page 8
- "Edge Node" on page 10
- "Utility Node" on page 11
- "Hardware Management Server" on page 11

Master Node

Master nodes control other processes on the network. There are the following types of master nodes:

- Name node
- ZooKeeper node
- Spark master
- Hive mete data server

Table 1, "Hitachi Advanced Server DS120 Master Node Configuration," on page 7 and Table 2, "Hitachi Advanced Server DS220 Master Node Configuration," on page 8 list the default hardware configuration for each type of master node server. The number of master nodes depends on the specific implementation and software used. A general rule is to have at least three master nodes for each hundred of other types of nodes.

TABLE 1. HITACHI ADVANCED SERVER DS120 MASTER NODE CONFIGURATION

Component	Description
Server	▪ Hitachi Advanced Server DS120
CPU	▪ From the Intel Xeon Scalable processor family ▪ Default processors: 2 Intel 4210 processors, 10-core, 2.2 GHz
Memory Option	▪ Default: 128 GB: 4 × 32 GB DDR4 RDIMM at 2666 MHz ▪ Up to 3 TB
Network Connections	▪ 1 or 2 Intel XXV710, 10/25 GbE dual port SFP28 (LP-MD2) ▪ 1 GbE LOM management port
Disk Controller	▪ LSI 3516 RAID controller
Operating System Devices	▪ 2 × 128 GB MLC SATADOMS
Data Disks	▪ Either one of the following: ▪ Up to 12 SFF SAS drives D ▪ Up to 8 SFF SAS drives or SSD and up to 4 NVMe drives ▪ Default: 8 × 1.8 TB SFF SAS drives

TABLE 2. HITACHI ADVANCED SERVER DS220 MASTER NODE CONFIGURATION

Component	Description
Server	▪ Either one of the following: ▪▪ Hitachi Advanced Server DS220 using the LFF chassis ▪▪ Hitachi Advanced Server DS220 using the SFF chassis with 16 SAS or SATA drives, and 8 NVMe drives ▪▪ Hitachi Advanced Server DS220 using the SFF chassis with up to 24 SAS or SATA Drives ▪ Default: Hitachi Advanced Server DS220 using the SFF chassis with 24 SAS or SATA drives
CPU	▪ From the Intel Xeon Scalable processor family ▪ Default Processors: 2 Intel 4210 processors, 10-core, 2.2 GHz
Memory Option	▪ Default: 128 GB: 4 × 32 GB DDR4 RDIMM at 2666 MHz ▪ Up to 3 TB
Network Connections	▪ 1 or 2 Intel XXV710, 10/25 GbE dual port SFP28 (LP-MD2) ▪ 1 GbE LOM management port
Disk Controller	▪ SAS 3516-based RAID controller
Operating System Devices	▪ 2 × 480 GB SSDs in the rear cage
Disks	▪ Multiple sizes of SAS and SATA storage devices, both large and small form factor ▪ Default: 8 × 1.8 TB SFF SAS drives

Sample Hadoop Worker Nodes

Worker nodes are used to process data. These nodes have very diverse needs, having many different configurations and software packages running on the nodes.

Table 3, "Hitachi Advanced Server DS120 Worker Node Configuration Options," on page 8 and Table 4, "Hitachi Advanced Server DS220 Worker Node Configuration Options," on page 9 list sample worker node configuration options. To get the complete and up to date set of options, contact your Hitachi Vantara sales representative.

TABLE 3. HITACHI ADVANCED SERVER DS120 WORKER NODE CONFIGURATION OPTIONS

Component	Description
Server	Hitachi Advanced Server DS120
CPU	From the Intel Xeon Scalable processor family ▪ Default Processor: 2 Intel 4210 processors, 10-core, 2.2 GHz

TABLE 3. HITACHI ADVANCED SERVER DS120 WORKER NODE CONFIGURATION OPTIONS (CONTINUED)

Component	Description
Memory Options	Default 256 GB: 8 × 32 GB DDR4 RDIMM at 2666 MHz ▪ Up to 3 TB
Network Connections	1 or 2 Intel XXV710 10/25 GbE dual port SFP28 (LP-MD2) 1 GbE LOM management port
Disk Controllers	SAS 3516 RAID controller
Operating System Devices	2 × 128 GB MLC SATADOM for operating system
Data Disks	Either of the following: ▪ Up to 12 SFF SAS drives or SSD Or ▪ Up to 8 SFF SAS or SSD drives, and up to 4 NVMe drives Default: 12 × 1.8 TB SFF SAS drives

TABLE 4. HITACHI ADVANCED SERVER DS220 WORKER NODE CONFIGURATION OPTIONS

Component	Description
Server	Hitachi Advanced Server DS220 using one of the following options: ▪ 12 LFF chassis option ▪ 24 SFF SAS drives or SSD option ▪ 16 SFF SAS drives or SSD and 8 NVMe drives option
CPU	From the Intel Xeon Scalable processor family ▪ Default Processor: 2 Intel 4210 processor, 10-core, 2.2 GHz
Memory Options	Default: 256 GB: 8 × 32 GB DDR4 RDIMM at 2666 MHz ▪ Up to 3 TB
Network Connections	1 or 2 Intel XXV710 10/25 GbE dual port SFP28 (LP-MD2) 1 GbE LOM management port
Disk Controllers	SAS 3516-based RAID controller

TABLE 4. HITACHI ADVANCED SERVER DS220 WORKER NODE CONFIGURATION OPTIONS (CONTINUED)

Component	Description
Operating System Device	2 x 480 GB SATA SSDs in the rear cage
Disks	Storage devices for DS220 using LFF drives ▪ SFF SAS drives ▪ SATA SSD ▪ SATA LFF drives ▪ Default: 12 × 8 TB SATA LFF drives Storage device for DS220 with up to 16 SFF SAS or SATA drives, and up to 8 NVMe drives ▪ 16 SAS HDD or SSD ▪ 8 NVMe drives ▪ Default: 16 × 1.8 TB SAS drives Storage device for DS220 using 24 SFF drives ▪ 24 SAS HDD or SSD ▪ Default: 24 × 1.8 TB SAS drives

For a general configuration of nodes for different workloads and software, see Cloudera Enterprise 6.X Release Notes.

Edge Node

An edge node resides on the client network and the data network to initiate processing to Cloudera Enterprise Data Hub. This multi-homed node can do the following:

- Act as a gateway
- Run software than needs access to the Cloudera environment and corporate systems

Depending on the work being performed and what other software is running on the edge node, the configuration can vary significantly. In this solution, use an edge node either as a master node or in a worker node configuration.

These nodes are usually multi-homed, connecting to the client network and data network. In that case, each edge node should have a second Intel XXV710 NIC added to its configuration.

The number of edge nodes depends on the software and use. Some example edge nodes are the following:

- **Gateway node**
  - Allows for access to both client and data network
  - Runs Hadoop client processes

- **Pentaho node**
  - Transfer data from existing sources into Hadoop
  - Execute Hadoop data reports

Data transfer nodes, such as a node running Hadoop Sqoop, should be on edge nodes.

Utility Node

A utility node runs support and Cloudera software. Depending on the software hosted, a utility node could be an edge node and multi-homed, or only connect to the data network.

The number of utility nodes depends on the software and use. The following are examples:

- Cloudera Navigator
- Apache HUE
- Cloudera Manager
- Underlying database for management servers

Hardware Management Server

You can include an optional hardware management server in this solution. It allows access to the out-of-band management network. Table 5 lists the hardware used for this server.

TABLE 5. HARDWARE MANAGEMENT SERVER

Component	Description
Server	▪ Hitachi Advanced Server DS120
CPU	▪ Intel 4210 10-core; 2.2 GHz;
Memory Option	▪ 64 GB: 2 × 32 GB DIMMs
Network Connections	▪ 2 Intel XXV710 10 GbE dual port SFP28 (LP-MD2) ▪ 1 GbE LOM management port
Disk Controllers	▪ Intel VROC on the motherboard
Disks	▪ 2 × 128 GB SATADOM configured as RAID-1

Worker Node Storage Considerations

For Hadoop Distributed File System, Cloudera's current recommendation is to use SAS HDD over SATA HDD. However, SSDs provided better performance.

With significant price reductions, SSDs are a viable option. The cost per terabyte for a 1DWDP SSD can be less than an SFF SAS drive. Large form factor SATA drives still provide the lowest cost per terabyte.

Hitachi Advanced Server DS220 has large format factor drives up to 14 TB. For better performance and to lower the impact of disk failure, the recommendation is to use disk sizes of 4 TB or less. With Cloudera 6.3, Cloudera only supports nodes with 100 TB or less and a maximum drive size of 8 TB.

There are many things that can impact your choice of drives and server types.

- **Storage per node or storage per rack**. With large drives, more data can be stored in a rack.
- **Cost per terabyte**. Large form factor drives are usually less expensive per terabyte than small form factor drives. Historically, hard disk drives are less expensive than solid state drives, but the price difference has been diminishing.
- **Performance**. For Hadoop Distributed File System and MapReduce, the bottleneck is usually storage performance. This is dealt with by having more drives or faster drives. If the CPU is an issue, Hitachi Advanced Storage DS120 allows for more computer power in a rack.
- **Software**. Software like Spark can see larger benefits from faster storage. Having solid state drives or NVMe drives for Spark temporary files and MapReduce temporaries can provide a significant improvement of performance
- **Recovery**. When a data node or a disk goes down, the data replicates to other nodes. The larger the device or more data on a node causes this replication time to increase. Recovery impacts the network performance and can impact everything that is being done on the cluster.

When sizing the cluster, there needs to be enough space to handle the following:

- Working files
- Extra storage when nodes or drives are down

A good rule of thumb is to plan on 80% free space for small clusters and 90% free space for large clusters.

Two features that can have a large impact on Hadoop Distributed File System sizing are the following:

- Heterogeneous Storage
- Erasure Encoding

Heterogeneous Storage

Heterogeneous storage introduces the concept of storage classes and data temperature. It allows you to assign a storage device to a class of storage or storage type. The classes are associated with data temperatures. Predefine temperatures and classes in Hadoop.

Software packages in Hadoop do not have to support all temperatures and classes. There could be temperatures and classes that are not used in Hadoop Distributed File System.

To have consistent performance, it is important that all devices assigned to a storage type have similar performance. There are four storage classes:

- **ARCHIVE** – This is your slower storage with data dense devices. It is useful for rarely accessed data. It is usually cheaper per terabyte. Some examples are an Amazon Simple Storage Service (S3) Bucket, Hitachi United Compute Platform CS, large SATA drives, and SAN storage. The items that can be archive devices have the widest performance variations.
- **DISK** – This is for your hard drives, being the default storage class. Even though hard drives variance in performance is small compared to the archive options, it is still recommended that you have similar devices.
- **SSD** – This is used for solid state drives or NVMe drives. It's for data that needs faster access.
- **RAM_DISK** – This is used for high speed single node writes that you can afford to lose. An example of this are an in-memory file system.

Use a storage policy to describe the storage types and fall back types to use. The fall back type is what to use if there is no space on the request storage types. For Hadoop Distributed File System, the fall back type is DISK.

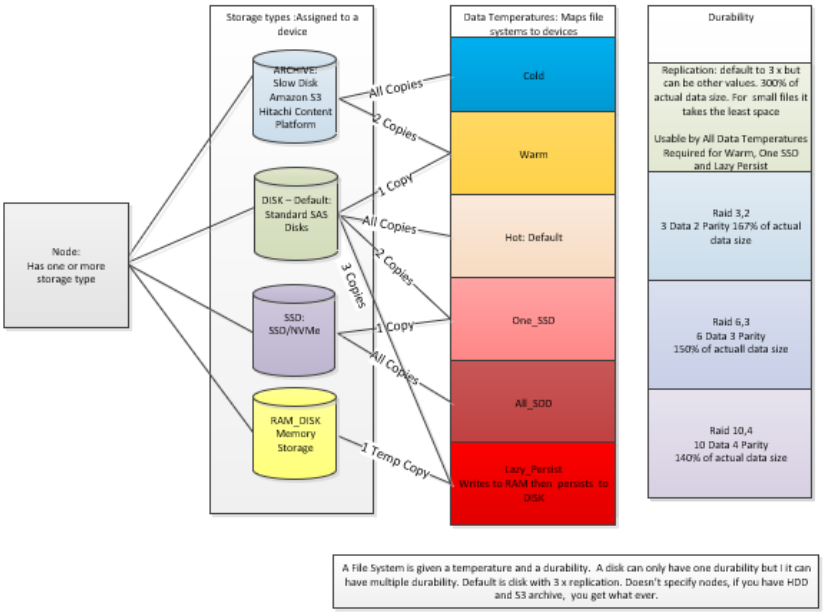
Hadoop Distributed File System supports six storage policies:

- **Hot** – All copies of the data are stored on DISK.
- **Cold** – All copies of the data are stored on ARCHIVE.
- **Warm** – One copy is stored on DISK and the rest of the copies are stored on ARCHIVE.
- **One_SSD** – One copy is stored on SSD and the rest of the copies are stored on DISK.
- **Lazy_Persist** – One copy is written to RAM_DISK. Some point later in time, all copies are persisted to DISK.
- **All_SSD** – All copies are stored on SSD.

Figure 2 shows the mapping between device types and temperatures. Two important things to note are the following:

- A directory is assigned a storage policy, not devices.
- Devices can be used in multiple directories and have different storage policies.

Figure 2



With different device types, storage policies, and multiple directories, storage needs must take all of these into consideration when sizing a system.

Erasure Encoding

"Erasure encoding" is Hadoop's way of saying "RAID." With Cloudera Data Hub, the default erasure coding is to use the standard "3-times" replication (mirroring). Including standard replication, Cloudera supports four methods for erasure coding:

- **Standard replication**
  - Defaults to three copies of teach storage block replicated across three nodes.
  - It has a 300% overhead verse storing one copy.
  - For high availability at the rack level, there should be at least three racks to spread the replica set across.
  - Data is local to processing, so – in most cases – it is faster.
- **Erasure Encoding**
  - For everything except small files, it takes less space.
  - It can lose more nodes without losing data.
  - Three levels are supported:
    - 6 data + 3 parity
    - 3 data + 2 parity
    - 10 data + 4 parity
  - For high availability at a rack level, each data device used in an erasure-encoded file system should be in a different rack. Processing is not co-located with data, so there is more network traffic which makes it slower.
  - It only supports the Reed-Solomon (RS) code algorithm.

Table 6 shows the space used for the different erasure encoding options when calculating the space for an individual file system.

TABLE 6. STORAGE EFFICIENCY

Type	Minimum Cluster Size	Data Durability[a]	Storage Efficiency	Storage Needed for 100 TB
Single Copy	1	0	100%	100
3× replication	3	2	33%	300
RS (6,3)	9	3	67%	150
RS (3,2)	5	2	60%	167
RS (10,4)	14	4	71%	140

a. **Data durability** is the number of machines that can go down.

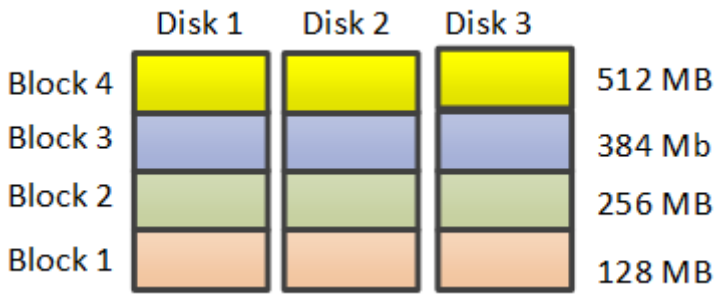
Note – Erasure encoding is at the directory level, not at the disk level. This means one disk can have data that has different replication strategies applied to it.

With small files, **replication** can take up less space than erasure encoding.

- Using replication files up to 128 MB requires one block on three disks: 384 MB.
- Using replication files from 128 MB to 256 MB requires two blocks on three disks: 768 MB.
- Using replication files from 384 MB to 512 MB requires three blocks on three disks: 1152 MB.
- Using replication files from 512 MB to 640 MB requires four blocks on three disks: 1536 MB.

The required number of blocks needed on how many disks for replication files is shown in Figure 3.

Figure 3

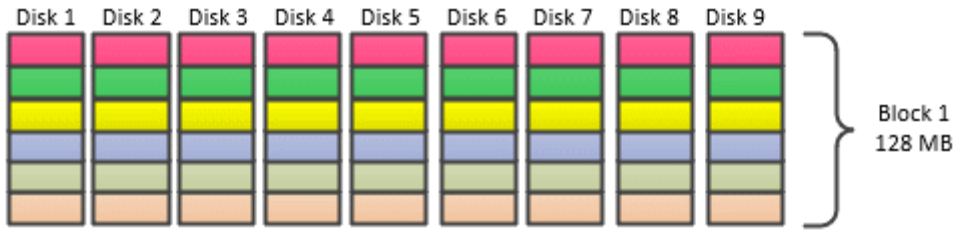


When using **erasure encoding** with six data and three parity, the block size is still 128 MB:

- Files up to 128 MB requires one block on all nine disks: 1152 MB.
- Files from 384 MB to 512 MB still requires one block on nine disks: 1152 MB.
- Files from 512 MB to 768 MB still requires one block on nine disks: 1152 MB.

The erasure encoding requirements are shown in Figure 4.

Figure 4



Both erasure encoding and the storage policy are applied to the directory. Not all combinations are supported. Warm, One_SSD, and Lazy_Persist require a replication encoding policy.

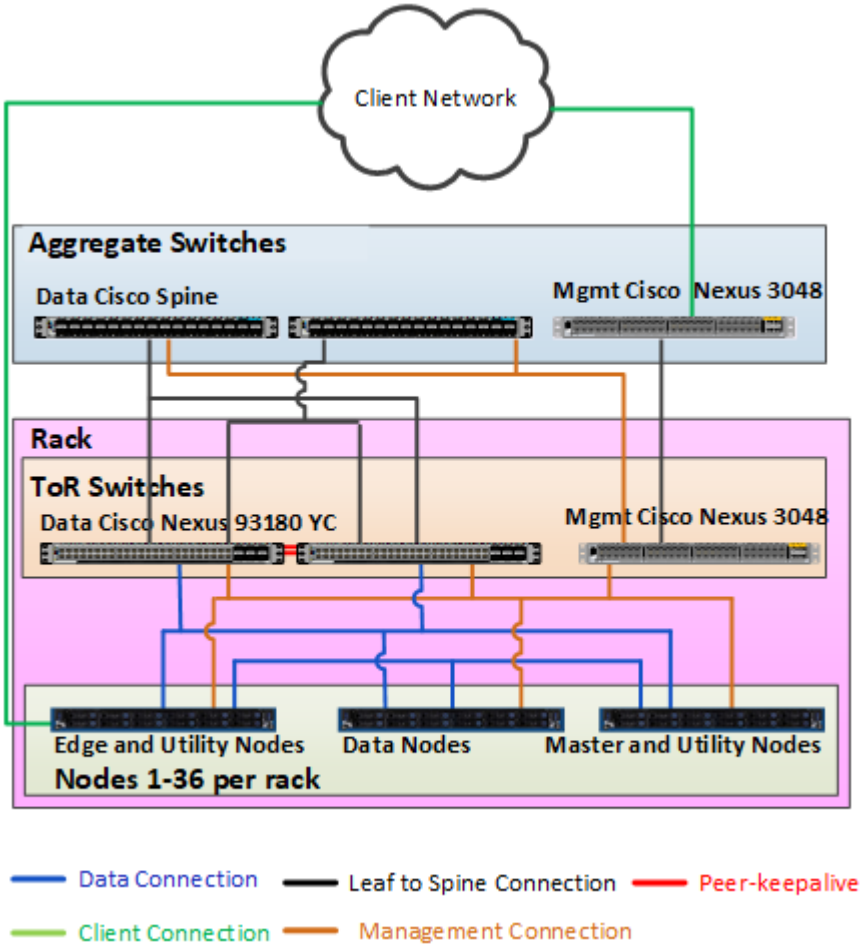
Network Architecture

This solution uses three logical networks. There can be multiple network configurations, depending on the Apache Hadoop deployment.

Figure 5 shows the standard configuration for the following networks:

- **Client Network** – Client access to edge nodes
- **Data Network** – Communication between nodes
- **Management Network** – Management of hardware

Figure 5



The network architecture uses these components:

- "Switches" on page 17
- "Data Network" on page 17
- "Client Network" on page 18
- "Management Network" on page 18

Switches

This solution requires the following types of switches:

- **Leaf Data Switches – Cisco Nexus 93180YC-E/FX**

  These leaf data switches connect all nodes in a rack together. Then, uplink the leaf switches to the spine data switches.

  Connect two switches together using stacking. This lets both switches act as one single logical switch. If one switch fails, there still is a path to the hosts.

- **Spine Data Switches – Cisco Nexus 93180LC-EX**

  A spine data switches interconnects leaf switches from different racks. The recommendation is that your place these switches in separate network rack.

  Connect two switches together using an inter-switch link (ISL). This lets both switches act together as a single logical switch. If one switch fails, there still is a path to the hosts.

- **Cisco Nexus 9336c**

  Used in multi-rack configurations, this switch connects the leaf data switches to the spine data switches using redundant 100 GbE link from each top of rack switch.

- **Leaf and Spine Management Switches – Cisco Nexus 3048**

  A leaf and spine switch connects the management ports of the hardware to the management server. When there is more than one rack, use a spine switch to connect all management leaf switches together.

  Uplink the management network to the inhouse management network.

Note – Other switches can be used. If other switches are used, check the network over subscription rate, based upon the switches used

Data Network

Use the data network for communications between the nodes.

Provide redundancy with two network interfaces configured at the operating system level to use the **active-active** network-bonding mode. The default network speed of 10 GbE can be upgraded to 25 GbE by purchasing licenses for the ToR switches.

With Cloudera 6.3, the recommendation is to use 10 GbE or better NICs and an oversubscription ration of 1 to 1, or at close as you can.

Table 7, "Sample Oversubscriptions," on page 18 shows the correlation between network options over subscription rations and how many nodes can be in a rack. The more uplink ports you use, the more spine switches required. The default configuration is 4 uplink ports, 2 per top of rack switch, with a 10 GbE network.

More uplink ports used from the top of rack switches to the spine switches lowers the over subscription rate but means fewer racks for each spine switch. It is not recommended to have less than four uplinks from each rack to the pair of spine switches.

TABLE 7. SAMPLE OVERSUBSCRIPTIONS

Target Over Subscription	Uplink Speed	Total Uplink Ports	Max Uplink Bandwidth	Network	Max 1U Nodes	Max 2U Nodes
1:1	100	4	400	10 GbE	36	18
1:1	100	4	400	25 GbE	16	16
1:1	40	4	160	10 GbE	16	16
1:1	40	4	160	25 GbE	6	6
1:1	40	8	320	10 GbE	32	18
1:1	40	8	320	25 GbE	12	12
3:1	40	4	160	10 GbE	36	18
3:1	40	4	160	25 GbE	19	18

Client Network

The client network is an optional network used on edge nodes. Using a client network separates the Hadoop network traffic from the rest of the client network traffic. Cloudera recommends that only edge nodes are accessible from the client network.

Management Network

The management network allows access to the nodes using the 1 GbE LAN on motherboard (LOM) interface. This network provides out-of-band monitoring and management of the servers.

You can uplink this network to the client management network.

Deployment Options

Table 8 shows the nodes used to deploy different components. in a three-master node configuration. This is a subset of the components that can be used in a Cloudera Enterprise Data Hub solution. Software deployed on an edge node is specific to a deployment and is not shown in this table.

TABLE 8. COMPONENT DEPLOYMENT

Component	Master Node 1	Master Node 2	Master Node 3	Utility Nod	Data Nodes (Multiple Nodes)
ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper		
Hadoop Distributed File System	Name Node Quorum Journal Node	Name Node Quorum Journal Node	Quorum Journal Node		Data Nodes
YARN	Resource Manager	Resource Manager	History Server		Node Manager

TABLE 8. COMPONENT DEPLOYMENT (CONTINUED)

Component	Master Node 1	Master Node 2	Master Node 3	Utility Nod	Data Nodes (Multiple Nodes)
Hive			MetaStore WebHCat HiveServer		
Management	Cloudera Agent	Cloudera Agent	Ooze Cloudera Agent Management Services	Cloudera Manager Cloudera Agent	Cloudera Agent
Miscellaneous				Navigator Database instances KMS	
HUE				HUE	
Spark					Runs on YARN or Standalone

Rack Configuration

When determining how many racks are needed, there are many different deployment-specific considerations that need to be evaluated and balanced for an individual deployment.

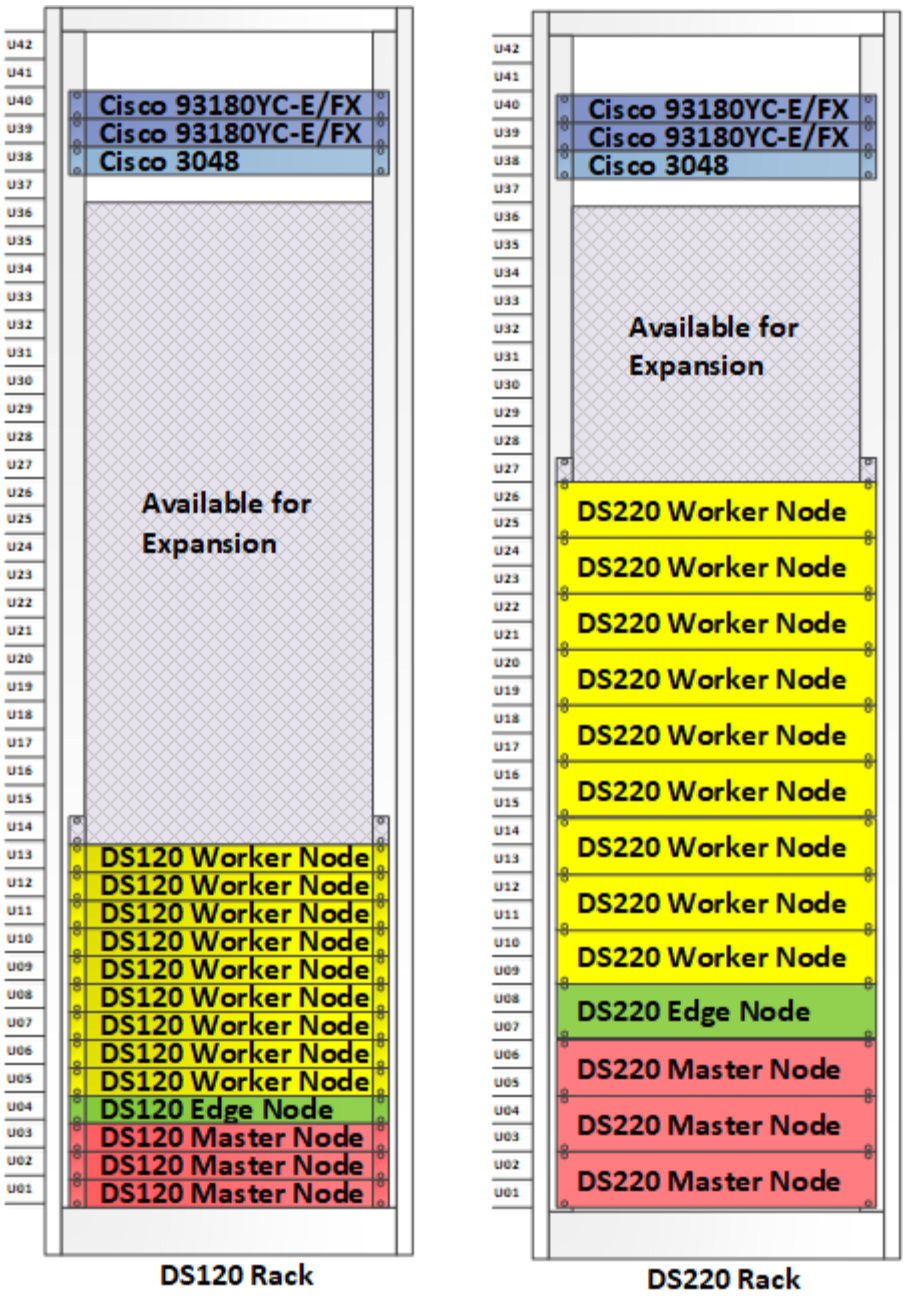
- **Storage per square foot.** In this case, racks are filled up as much as possible.
- **Rack high availability.** This Is recommended to reduce the likely hood of a complete system shut down.
  - Nodes with the same data or processes needs to be spread out across multiple racks. For standard Hadoop Distributed File System 3-times replication, this means you should have at least two racks. For an erasure encoding of 6+3, this would require at least three racks.
  - A spine switch pair should have each switch in a different rack.
- **Software.** When there are multiple racks, spread the different Hadoop components across the racks.
- **Performance.** Nodes that work together should be as close as possible. If not in the same rack, they should be under the same spine switch.
- **Growth plans.** When adding new nodes, they are placed in new racks, existing racks, or both.
- **Network over configuration.** The number of nodes in a rack can be limited on the network over subscription.
- **Power and heat design.** The data center's power and heat requirements can increase the number of racks needed.

Single Rack Configuration

Figure 6 on page 21 shows a Hitachi Advanced Server DS120 deployment using and a Hitachi Advanced Server DS220 deployment with the following components:

- Top-of-rack data and management switches
- 3 master nodes
- 1 edge node
- 2 utility nodes
- 1 hardware management node
- 9 worker nodes
- Advanced Server DS120 worker node configuration
  - 2 Intel 4210 processors
  - 368 GB RAM
  - 12 × 1.8 TB SAS drives
- Advanced Server DS220 worker node configuration
  - 2 Intel 4210
  - 368 GB RAM
  - 12 × 6 TB SATA Drives

Figure 6


Cisco 93180YC-E/FX
Cisco 93180YC-E/FX
Cisco 3048
Available for Expansion
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Worker Node
DS120 Edge Node
DS120 Master Node
DS120 Master Node
DS120 Master Node
DS120 Rack
Cisco 93180YC-E/FX
Cisco 93180YC-E/FX
Cisco 3048
Available for Expansion
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Worker Node
DS220 Edge Node
DS220 Master Node
DS220 Master Node
DS220 Master Node
DS220 Rack

Multiple Rack Configuration

This sample deployment depicts a Hitachi Advanced Server DS120 solution and a similar Advanced Server DS220 solution. The spine switches placement is not shown.

To provide high availability in a multi-rack system. spread the node types out across multiple racks. Figure 7 on page 23 shows a sample three-rack Advanced Server DS120 configuration.

- **First rack**
  - Top-of-rack data and management switches
  - 2 master nodes
  - 1 edge node
  - 1 utility node
  - 1 hardware management node
  - 32 worker nodes
- **Second rack**
  - Top-of-rack data and management switches
  - 1 edge node
  - 2 master nodes
  - 33 worker nodes
- **Third rack**
  - Top-of-rack data and management switches
  - 1 master node
  - 1 utility
  - 32 worker nodes
- **Advanced Server DS120 worker node configuration**
  - 2 Intel 4210 processors
  - 368 GB RAM
  - 12 × 1.8 TB SAS drives
- **Total resource**
  - 99 worker nodes
  - 198 CPUs
  - 36432 GB RAM
  - 2138.4 TB raw data storage

Figure 7

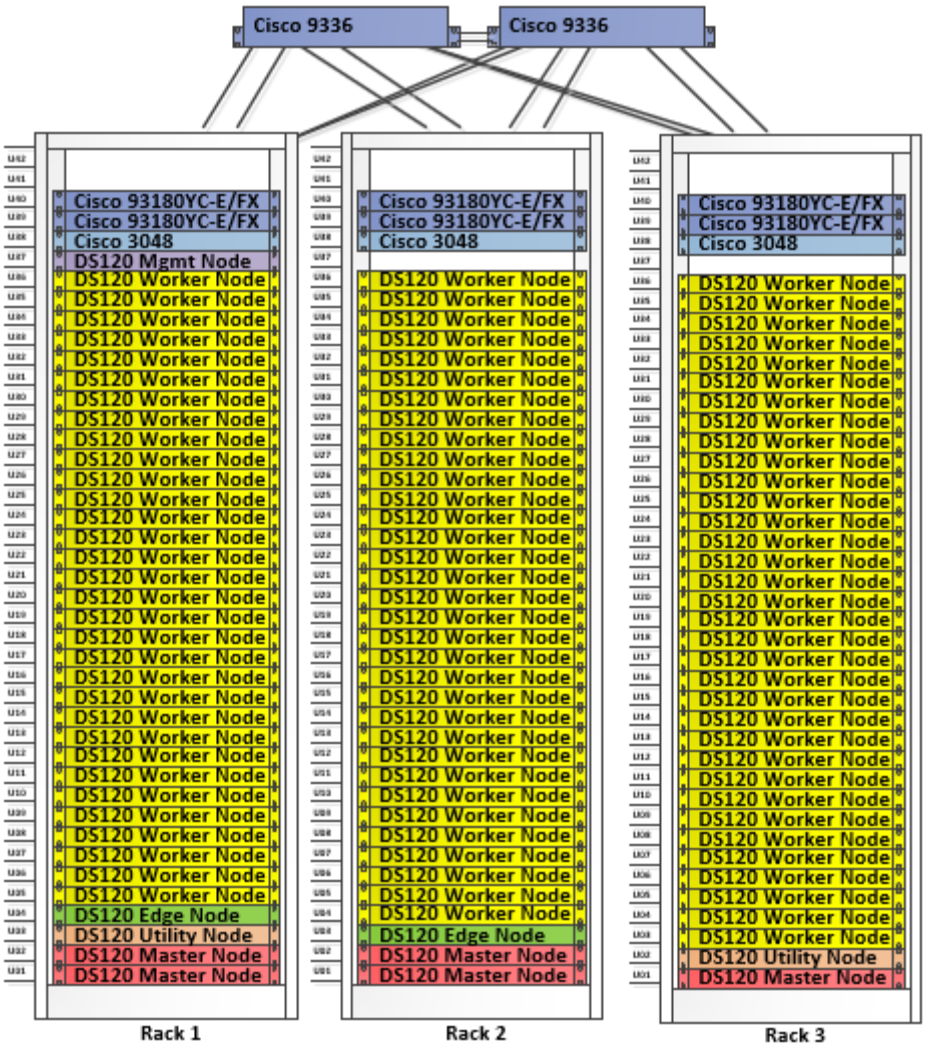

Cisco 9336
Cisco 9336
Cisco 93180YC-E/FX
Cisco 93180YC-E/FX
Cisco 3048
DS120 Mgmt Node
DS120 Worker Node
DS120 Edge Node
DS120 Utility Node
DS120 Master Node
Rack 1
Rack 2
Rack 3

Figure 8 on page 25 shows a similar three-rack deployment using a Hitachi Advanced Server DS220 configuration.

- **First rack**
  - Top-of-rack data and management switches
  - 2 master nodes
  - 1 edge nodes
  - 1 utility nodes
  - 1 hardware management node
  - 14 worker nodes
- **Second rack**
  - Top-of-rack data and management switches
  - 2 master nodes
  - 1 edge node
  - 15 worker nodes
- **Third rack**
  - Top-of-rack data and management switches
  - 1 master node
  - 1 utility node
  - 16 worker nodes
- **Advanced Server DS220 worker node configuration**
  - 2 Intel 4210 processors
  - 368 GB RAM
  - 12 × 6 TB SATA drives
- **Total resources**
  - 45 worker nodes
  - 90 CPUs
  - 16560 GB RAM
  - 3240 TB raw data storage

Figure 8


Cisco 9336
Cisco 9336
Cisco 93180YC-E/FX
Cisco 93180YC-E/FX
Cisco 3048
DS120 Mgmt Node
DS220 Worker Node
DS220 Utility Node
DS220 Edge Node
DS220 Master Node
Rack 1
Rack 2
Rack 3

Engineering Validation

There are many different configurations of both hardware and software that can be used in this solution. A basic validation of this solution was done in the lab environment at Hitachi Vantara. It involved deploying a basic Apache Hadoop Distributed File System in a small cluster of mixed node configurations. Physical servers were used for storage and processing purposes.

The following servers were used:

- 4 Hitachi Advanced Server DS120
  - 4 drives for Hadoop Distributed File System
  - 2 Intel 4210 processors for the CPU
  - 386 GB of memory
- 1 Hitachi Advanced Server DS220 using large form factor drives
  - 4 drives for Hadoop Distributed File System
  - 2 Intel 4210 processors for the CPU
  - 386 GB of memory
- 1 Hitachi Advanced Server DS220 using 24 small form factor drives
  - 4 drives for Hadoop Distributed File System
  - 2 Intel 4210 processors for the CPU
  - 386 GB of memory
- Virtual machines for the master node

The follow software was used:

- Cloudera Enterprise Data Hub 6.3
- Red Hat Enterprise Linux 7.6

Operating System-Level Storage Testing

Two basic disk performance tests were performed. The tests are based upon Cloudera's Hardware verification tests. These were performed to get a basic understanding of the hardware.

The first set of tests used hdparm on SAS, SSD, and NVMe drives to validate disk read performance. Different model of drives will have different performance results.

- **SAS**
  - Timing cached reads: 14198 MB in 2.00 seconds for 7111.15 MB/s
  - Timing buffered disk reads: 670 MB in 3.00 seconds for 223.23 MB/s
- **SSD**
  - Timing cached reads: 15630 MB in 2.00 seconds for 7824.10 MB/s
  - Timing buffered disk reads: 1578 MB in 3.00 seconds for 525.85 MB/s

- **NVME**
  - Timing cached reads: 15716 MB in 2.00 seconds for 7867.91 MB/s
  - Timing buffered disk reads: 3924 MB in 3.00 seconds for 1307.98 MB/s

The second set of tests used dd on the same drives to verify disk read performance

- **SAS**
  - dd bs=16k count=1024000 if=/dev/zero of=/data2/img2.img conv=fdatasync
  - 1024000+0 records in
  - 1024000+0 records out
  - 16777216000 bytes (17 GB) copied, 88.4336 s, 190 MB/s
- **SSD**
  - dd bs=16k count=1024000 if=/dev/zero of=/ssd/img2.img conv=fdatasync
  - 1024000+0 records in
  - 1024000+0 records out
  - 16777216000 bytes (17 GB) copied, 41.3986 s, 405 MB/s
- **NVMe**
  - dd bs=16k count=1024000 if=/dev/zero of=/nvmetest/img2.img conv=fdatasync
  - 1024000+0 records in
  - 1024000+0 records out
  - 16777216000 bytes (17 GB) copied, 18.0212 s, 931 MB/s

Hadoop Distributed File System-Level Storage Testing

The next set of tests was to do a basic verification of the read and write performance of the Hadoop Distributed File System. These tests used TestDFSIO.

- **Write**
  - Number of files: 300
  - Total MBytes processed: 3072000
  - Throughput MB/s: 8.35
  - Average I/O rate MB/s: 8.58
  - I/O rate standard deviation: 1.34
- **Read**
  - Number of files: 300
  - Total megabytes processed: 3072000
  - Throughput MB/s: 15.56
  - Average I/O rate MB/s: 16.78
  - I/O rate standard deviation: 5.11

Processing Testing

There are many software packages to process Hadoop Distributed File System data. The two most popular are MapReduce and Spark. TeraGen and TeraSort was ran using both packages.

MapReduce

The following are sample commands for TeraSort and TeraGen. The commands were run for 1 TB data sets:

**yarn jar ${EXAMPLES_PATH}/hadoop-mapreduce-examples.jar teragen **

**-Dmapreduce.job.maps=600 -Dmapreduce.map.memory.mb=4096 10000000000 **

TS_input >>1TBgen.txt

**yarn jar ${EXAMPLES_PATH}/hadoop-mapreduce-examples.jar terasort **

**-Dmapreduce.job.maps=600 -Dmapreduce.map.memory.mb=4096 **

**-Dmapreduce.reduce.memory.mb=4096 -Dmapreduce.reduce.cpu.vcores=2 **

TS_input TS_output >1TBsort.txt

TeraGen 1tb CPU time spent (ms)=20331300

TeraSort 1 tb CPU time spent (ms)= 165554980

Spark

The following commands are sample of the commands for using SparkBench to generate and sort data:

spark-submit --executor-memory 150G --driver-cores 10 --executor-cores 40 sparkbench_2.11-1.0.14.jar generate 10000000000 /input1tb 40 > spark1tbgen.out

nohup spark-submit --executor-memory 150G --driver-cores 10 --executor-cores 40 sparkbench_2.11-1.0.14.jar sort /input1tb /output 1tb > spark1tbsort.out

The results are the following:

- 1 TB data generation: 9066 s
- 1 TB data sort: 8115 s

For More Information

Hitachi Vantara Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the Services website.

Demonstrations and other resources are available for many Hitachi products. To schedule a live demonstration, contact a sales representative or partner. To view on-line informational resources, see the Resources website.

Hitachi Academy is your education destination to acquire valuable knowledge and skills on Hitachi products and solutions. Our Hitachi Certified Professional program establishes your credibility and increases your value in the IT marketplace. For more information, see the Hitachi Vantana Training and Certification website.

For more information about Hitachi products and services, contact your sales representative, partner, or visit the Hitachi Vantara website.

Hitachi Vantara

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 95054 USA
www.HitachiVantara.com | community.HitachiVantara.com

Regional Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
HitachiVantara.com/contact

MK-SL-020-06, July 2020