Lenovo Big Data Validated Design for Cloudera Enterprise and VMware with Bare-Metal and Virtualized ThinkSystem Servers

Last update: **15 December 2017**
Version 1.2
Configuration Reference Number BGDCL01XX74

Describes the reference architecture for Cloudera Enterprise, powered by Apache Hadoop and Apache Spark

Solution based on the powerful, versatile Lenovo ThinkSystem SR650 server, bare-metal and virtualized

Deployment considerations for high-performance, cost-effective and scalable solutions

Contains detailed bill of material for different servers and associated networking

Ajay Dholakia (Lenovo)
Dan Kangas (Lenovo)
Weixu Yang (Lenovo)
Dwai Lahiri (Cloudera)

Table of Contents

1 Introduction

This document describes the reference architecture for Cloudera Enterprise on bare-metal and on VMware vSphere with locally attached storage. It provides a predefined and optimized hardware infrastructure for the Cloudera Enterprise, a distribution of Apache Hadoop and Apache Spark with enterprise-ready capabilities from Cloudera. This reference architecture provides the planning, design considerations, and best practices for implementing Cloudera Enterprise with Lenovo products.

Lenovo and Cloudera worked together on this document, and the reference architecture that is described herein was validated by Lenovo and Cloudera.

With the ever-increasing volume, variety and velocity of data becoming available to an enterprise comes the challenge of deriving the most value from it. This task requires the use of suitable data processing and management software running on a tuned hardware platform. With Apache Hadoop and Apache Spark emerging as popular big data storage and processing frameworks, enterprises are building so-called Data Lakes by employing these components.

Cloudera brings the power of Hadoop to the enterprise. Hadoop is an open source software framework that is used to reliably manage large volumes of structured and unstructured data. Cloudera expands and enhances this technology to withstand the demands of your enterprise, adding management, security, governance, and analytics features. The result is that you get a more enterprise ready solution for complex, large-scale analytics.

VMware vSphere brings virtualization to Hadoop with many benefits that cannot be obtained on physical infrastructure or in the cloud. Virtualization simplifies the management of your big data infrastructure, enables faster time to results and makes it more cost effective. It is a proven software technology that makes it possible to run multiple operating systems and applications on the same server at the same time. Virtualization can increase IT agility, flexibility, and scalability while creating significant cost savings. Workloads get deployed faster, performance and availability increases and operations become automated, resulting in IT that is simpler to manage and less costly to own and operate.

The intended audience for this reference architecture is IT professionals, technical architects, sales engineers, and consultants to assist in planning, designing, and implementing the big data solution with Lenovo hardware. It is assumed that you are familiar with Hadoop components and capabilities. For more information about Hadoop, see “Resources” on page 50.

2 Business problem and business value

This section describes the business problem that is associated with big data environments and the value that is offered by the Cloudera solution that uses Lenovo hardware.

2.1 Business problem

The world is well on its way to generate more than 40 million TB of data by 2020. In all, 90% of the data in the world today was created in the last two years alone. This data comes from everywhere, including sensors that are used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone global positioning system (GPS) signals. This data is big data.

Big data spans the following dimensions:

- Volume: Big data comes in one size: large – in size, quantity and/or scale. Enterprises are awash with data, easily amassing terabytes and even petabytes of information.
- Velocity: Often time-sensitive, big data must be used as it is streaming into the enterprise to maximize its value to the business.
- Variety: Big data extends beyond structured data, including unstructured data of all varieties, such as text, audio, video, click streams, and log files.

Enterprises are incorporating large data lakes into their IT architecture to store all their data. The expectation is that ready access to all the available data can lead to higher quality of insights obtained through the use of analytics, which in turn drive better business decisions. A key challenge faced today by these enterprises is setting up an easy to deploy data storage and processing infrastructure that can start to deliver the promised value in a very short amount of time. Spending months of time and hiring dozens of skilled engineers to piece together a data management environment is very costly and often leads to frustration from unrealized goals. Furthermore, the data processing infrastructure needs to be easily scalable in addition to achieving desired performance and reliability objectives.

Big data is more than a challenge; it is an opportunity to find insight into new and emerging types of data to make your business more agile. Big data also is an opportunity to answer questions that, in the past, were beyond reach. Until now, there was no effective way to harvest this opportunity. Today, Cloudera uses the latest big data technologies such as the in-memory processing capabilities of Spark in addition to the standard MapReduce scale-out capabilities of Hadoop, to open the door to a world of possibilities.

2.2 Business value

Hadoop is an open source software framework that is used to reliably manage and analyze large volumes of structured and unstructured data. Cloudera enhances this technology to withstand the demands of your enterprise, adding management, security, governance, and analytics features. The result is that you get an enterprise-ready solution for complex, large-scale analytics.

How can businesses process tremendous amounts of raw data in an efficient and timely manner to gain actionable insights? Cloudera allows organizations to run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure can be used to tackle large data sets by breaking up the data into “chunks” and coordinating data processing across a massively parallel environment. After the raw data is stored across the nodes of a distributed cluster, queries and analysis of the data can be handled efficiently, with dynamic interpretation of the data formatted at read time. The bottom line: Businesses can

finally get their arms around massive amounts of untapped data and mine that data for valuable insights in a more efficient, optimized, and scalable way.

Cloudera that is deployed on Lenovo System x servers with Lenovo networking components provides superior performance, reliability, and scalability. The reference architecture supports entry through high-end configurations and the ability to easily scale as the use of big data grows. A choice of infrastructure components provides flexibility in meeting varying big data analytics requirements.

There is growing interest in deploying Hadoop on a virtualized infrastructure driven by the promise of ease of managing the cluster during initial deployment as well as adding more nodes when data storage and processing requirements grow. The ability to have virtualized Hadoop environment look and feel the same as it does on a bare-metal infrastructure allows flexibility in incorporating the solution within an enterprise's data management architecture.

3 Requirements

The functional and non-functional requirements for this reference architecture are desribed in this section.

3.1 Functional requirements

A big data solution supports the following key functional requirements:

- Ability to handle various workloads, including batch and real-time analytics
- Industry-standard interfaces so that applications can work with Cloudera
- Ability to handle large volumes of data of various data types
- Various client interfaces

3.2 Non-functional requirements

Customers require their big data solution to be easy, dependable, and fast. The following non-functional requirements are key:

- Easy:
  - Ease of development
  - Easy management at scale
  - Advanced job management
  - Multi-tenancy
  - Easy to access data by various user types
- Dependable:
  - Data protection with snapshot and mirroring
  - Automated self-healing
  - Insight into software/hardware health and issues
  - High availability (HA) and business continuity
- Fast:
  - Superior performance
  - Scalability
- Secure and governed:
  - Strong authentication and authorization
  - Kerberos support
  - Data confidentiality and integrity

4 Architectural overview

4.1 Cloudera Enterprise

Figure 1 shows the main features of the Cloudera reference architecture that uses Lenovo hardware. Users can log into the Cloudera client from outside the firewall by using Secure Shell (SSH) on port 22 to access the Cloudera solution from the corporate network. Cloudera provides several interfaces that allow administrators and users to perform administration and data functions, depending on their roles and access level. Hadoop application programming interfaces (APIs) can be used to access data. Cloudera APIs can be used for cluster management and monitoring. Cloudera data services, management services, and other services run on the nodes in cluster. Storage is a component of each data node in the cluster. Data can be incorporated into Cloudera Enterprise storage through the Hadoop APIs or network file system (NFS), depending on the needs of the customer.

A database is required to store the data for Cloudera manager, hive metastore, and other services. Cloudera provides an embedded database for test or proof of concept (POC) environments and an external database is required for a supportable production environment.

Figure 1. Cloudera architecture overview

4.2 Bare-metal Cluster

The solution described in this document is typically deployed on bare-metal infrastructure. This means that both the management nodes and the data nodes are implemented on physical host servers. The number of servers of each type is determined based on requirements for high-availability, total data capacity and desired performance objectives.

4.3 Virtualized Cluster with VMware vSphere

When Hadoop is virtualized, all of the components of Hadoop, including the NameNode, ResourceManager, DataNode, and NodeManager, are running within purpose-built Virtual Machines (VMs) rather than on the native OS of the physical machine. However, the Hadoop services or roles of the Cloudera software stack are installed with Cloudera Manager exactly the same way as with the physical machines. With a virtualization infrastructure, two or more VMs can be run on the same physical host server to improve cluster usage efficiency and flexibility.

The VMware-based infrastructure with direct attached storage for HDFS is used to maintain the storage-to-CPU locality on a physical node. VMs are configured for one-to-one mapping of a physical disk to a vSphere VMFS virtual disk - see Figure 2 below.

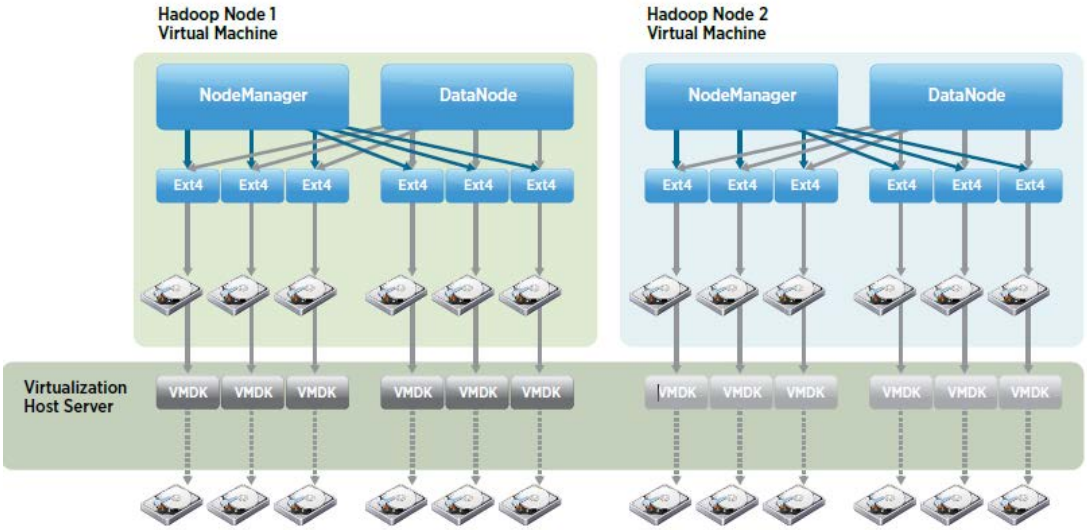


Figure 2. One-to-one mapping of local storage

5 Component model

Cloudera Enterprise provides features and capabilities that meet the functional and nonfunctional requirements of customers. It supports mission-critical and real-time big data analytics across different industries, such as financial services, retail, media, healthcare, manufacturing, telecommunications, government organizations, and leading Fortune 100 and Web 2.0 companies.

Cloudera Enterprise is the world's most complete, tested, and popular distribution of Apache Hadoop and related projects. All of the packaging and integration work is done for you, and the entire solution is thoroughly tested and fully documented. By taking the guesswork out of building out your Hadoop deployment, Cloudera Enterprise gives you a streamlined path to success in solving real business problems with big data.

The Cloudera platform for big data can be used for various use cases from batch applications that use MapReduce or Spark with data sources, such as click streams, to real-time applications that use sensor data.

Figure 3 shows the Cloudera Enterprise key capabilities that meet the functional requirements of customers.

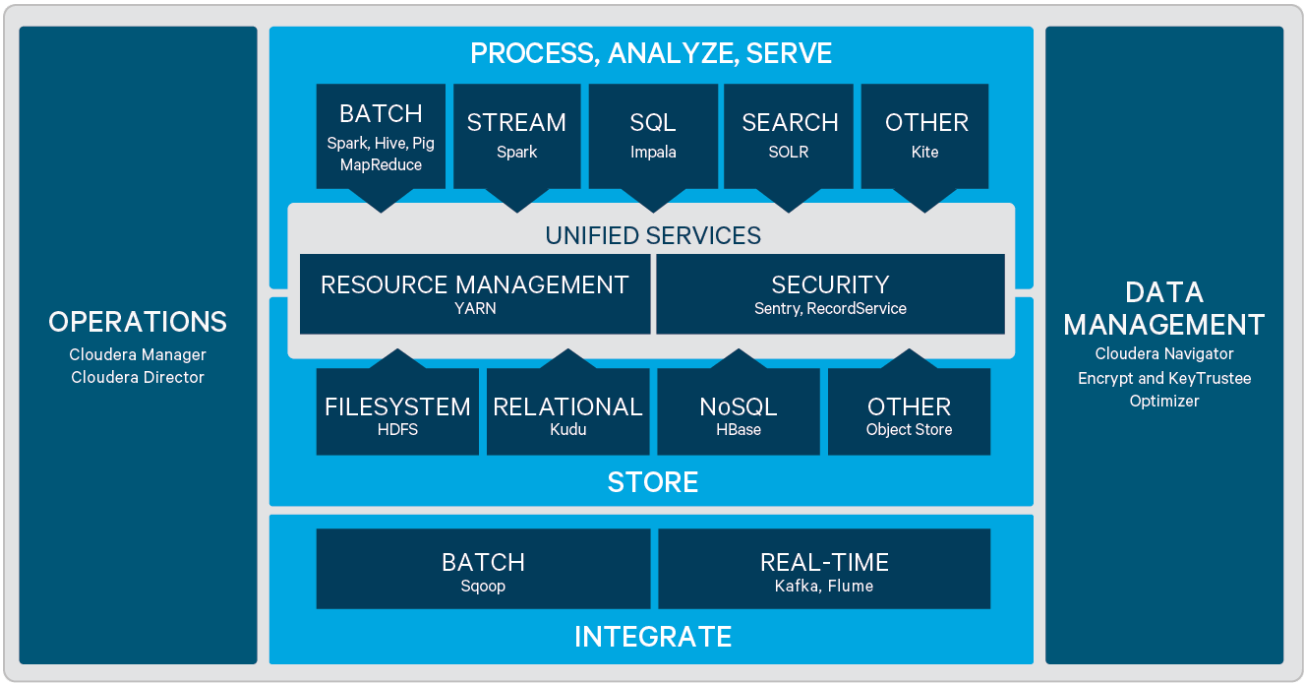


Figure 3. Cloudera Enterprise key capabilities

Cloudera Enterprise solution contains the following components:

- Analytic SQL: Apache Impala (incubating)

  Impala is the industry's leading massively parallel processing (MPP) SQL query engine that runs natively in Hadoop. Apache-licensed, open source Impala project combines modern, scalable parallel database technology with the power of Hadoop, enabling users to directly query data stored in HDFS and Apache HBase without requiring data movement or transformation. Impala is designed from the ground up as part of the Hadoop system and shares the same flexible file and data formats, metadata, security, and resource management frameworks that are used by MapReduce, Apache Hive, Apache Pig, and other components of the Hadoop stack.

- Search Engine: Cloudera Search

  Cloudera Search is Apache Solr that is integrated with Cloudera Enterprise, including Apache Lucene, Apache SolrCloud, Apache Flume, Apache Tika, and Hadoop. Cloudera Search also includes valuable integrations that make searching more scalable, easy to use, and optimized for near-real-time and batch-oriented indexing. These integrations include Cloudera Morphlines, which is a customizable transformation chain that simplifies loading any type of data into Cloudera Search.

- NoSQL - HBase

  A scalable, distributed column-oriented datastore. HBase provides real-time read/write random access to very large datasets hosted on HDFS.

- Stream Processing: Apache Spark

  Apache Spark is an open source, parallel data processing framework that complements Hadoop to make it easy to develop fast, unified big data applications that combine batch, streaming, and interactive analytics on all your data. Cloudera offers commercial support for Spark with Cloudera Enterprise. Spark is 10 – 100 times faster than MapReduce which delivers faster time to insight, allows inclusion of more data, and results in better business decisions and user outcomes.

- Machine Learning: Spark MLlib

  MLlib is the API that implements common machine learning algorithms. MLlib is usable in Java, Scala, Python and R. Leveraging Spark's excellence in iterative computation, MLlib runs very fast, high-quality algorithms.

- Cloudera Manager

  Cloudera Manager is the industry's first and most sophisticated management application for Hadoop and the enterprise data hub. Cloudera Manager sets the standard for enterprise deployment by delivering granular visibility into and control over every part of the data hub, which empowers operators to improve performance, enhance quality of service, increase compliance, and reduce administrative costs. Cloudera Manager makes administration of your enterprise data hub simple and straightforward, at any scale. With Cloudera Manager, you can easily deploy and centrally operate the complete big data stack.

  Cloudera Manager automates the installation process, which reduces deployment time from weeks to minutes; gives you a cluster-wide, real-time view of nodes and services running; provides a single,

central console to enact configuration changes across your cluster; and incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization.

- Cloudera Manager Metrics

  Cloudera Manager monitors a number of performance metrics for services and role instances that are running on your clusters. These metrics are monitored against configurable thresholds and can be used to indicate whether a host is functioning as expected. You can view these metrics in the Cloudera Manager Admin Console, which displays metrics about your jobs (such as the number of currently running jobs and their CPU or memory usage), Hadoop services (such as the average HDFS I/O latency and number of concurrent jobs), your clusters (such as average CPU load across all your hosts) and so on.

- Cloudera Manager Backup And Disaster Recovery (BDR)

  Cloudera Manager provides an integrated, easy-to-use management solution for enabling data protection in the Hadoop platform. Cloudera Manager provides rich functionality that is aimed towards replicating data that is stored in HDFS and accessed through Hive across data centers for disaster recovery scenarios. When critical data is stored on HDFS, Cloudera Manager provides the necessary capabilities to ensure that the data is available at all times, even in the face of the complete shutdown of a data center. Cloudera Manager also provides the ability to schedule, save, and (if needed) restore snapshots of HDFS directories and HBase tables.

- Cloudera Manager API

  The Cloudera Manager API provides configuration and service lifecycle management, service health information and metrics, and allows you to configure Cloudera Manager. The API is served on the same host and port as the Cloudera Manager Admin Console, and does not require an extra process or extra configuration. The API supports HTTP Basic Authentication, accepting the same users and credentials as the Cloudera Manager Admin Console.

- Cloudera Navigator

  A fully integrated data management and security tool for the Hadoop platform. Cloudera Navigator provides three categories of functionality:

  - Auditing data access and verifying access privileges. Cloudera Navigator allows administrators to configure, collect, and view audit events, and generate reports that list the HDFS access permissions granted to groups. Cloudera Navigator tracks access permissions and actual accesses to all entities in HDFS, Hive, HBase, Hue, Impala, Sentry, and Solr.
  - Searching metadata and visualizing lineage. Metadata management features allow DBAs, data modelers, business analysts, and data scientists to search for, amend the properties of, and tag data entities. Cloudera Navigator supports tracking the lineage of HDFS files, datasets, and directories, Hive tables and columns, MapReduce and YARN jobs, Hive queries, Impala queries, Pig scripts, Oozie workflows, Spark jobs, and Sqoop jobs.
  - Securing data and simplifying storage and management of encryption keys. Data encryption and key management provide protection against potential threats by malicious actors on the network or

in the datacenter. It is also a requirement for meeting key compliance initiatives and ensuring the integrity of enterprise data.

- Cloudera Kafka

  Cloudera Distribution of Apache Kafka is a distributed commit log service. Kafka functions much like a publish/subscribe messaging system, but with better throughput, built-in partitioning, replication, and fault tolerance. Kafka is a good solution for large scale message processing applications. It is often used in tandem with Apache Hadoop, Apache Storm and Spark Streaming.

For more information, see this website:
cloudera.com/content/cloudera/en/products-and-services/product-comparison.html

The Cloudera solution is operating system independent. Cloudera supports many Linux® operating systems, including Red Hat Linux and SUSE Linux. For more information about the versions of supported operating systems, see this website:

http://www.cloudera.com/documentation/enterprise/latest/topics/cm_ig_cm_requirements.html.

5.1 Apache Spark in CDH 5.8

Spark has recently become very popular and is being adopted as a preferred framework for a variety of big data use-cases ranging from batch applications that use MapReduce or Spark with data sources such as click streams, to real-time applications that use sensor data.

The Spark stack is shown in Figure 4. As depicted, the foundational component is the Spark Core. Spark is written in the Scala programming language and offers simple APIs in Python, Java, Scala and SQL.

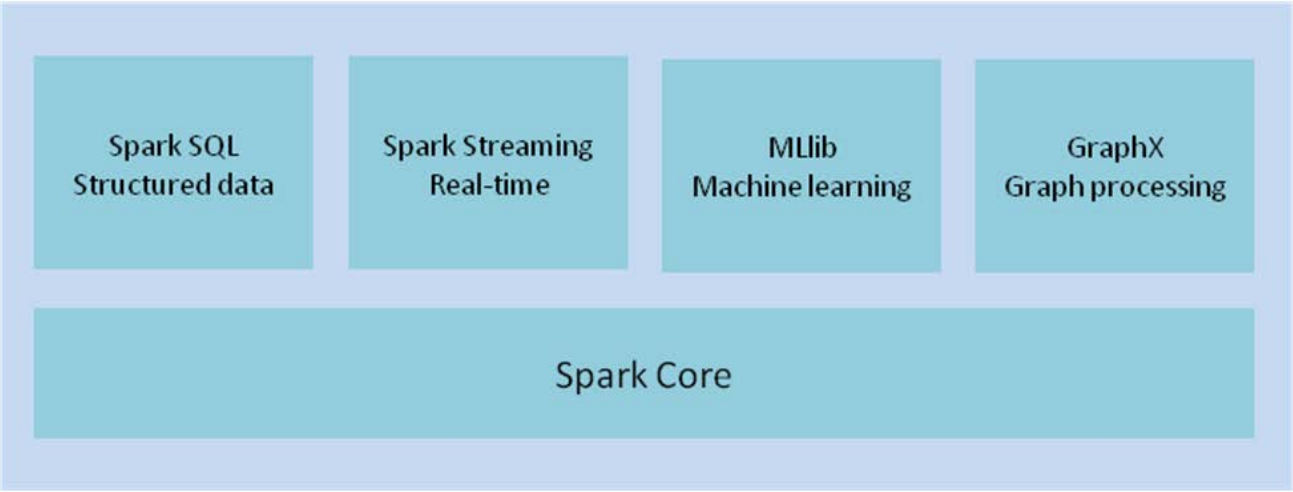


Figure 4. The Spark stack

In additional to the Spark Core, the framework allows extensions in the form of libraries. Most common extensions are Spark MLlib for machine learning, Spark SQL for queries on structured data, Spark Streaming for real-time stream-processing, and Spark GraphX for handling graph databases. Other extensions are also available. Cloudera does not currently support GraphX or SparkR. There are also caveats for Spark SQL support - please refer to Cloudera's Spark documentation.

The Spark architecture shown in Figure 4 enables a single framework to be used for multiple projects. Typical big data usage scenarios to date have deployed the Hadoop stack for batch processing separately from another framework for stream processing, and yet another one for advanced analytics such as machine learning. Apache Spark combines these frameworks in a common architecture, thereby allowing easier management of the big data code stack and also enabling reuse of a common data repository.

The Spark stack shown in Figure 4 can run in a variety of environments. It can run alongside the Hadoop stack, leveraging Hadoop YARN for cluster management. Spark applications can run in a distributed mode on a cluster using a master/slave architecture that uses a central coordinator called “driver” and potentially large number of “worker” processes that execute individual tasks in a Spark job. The Spark executor processes also provide reliable in-memory storage of data distributed across the various nodes in a cluster. The components of a distributed Spark application are shown in Figure 5.

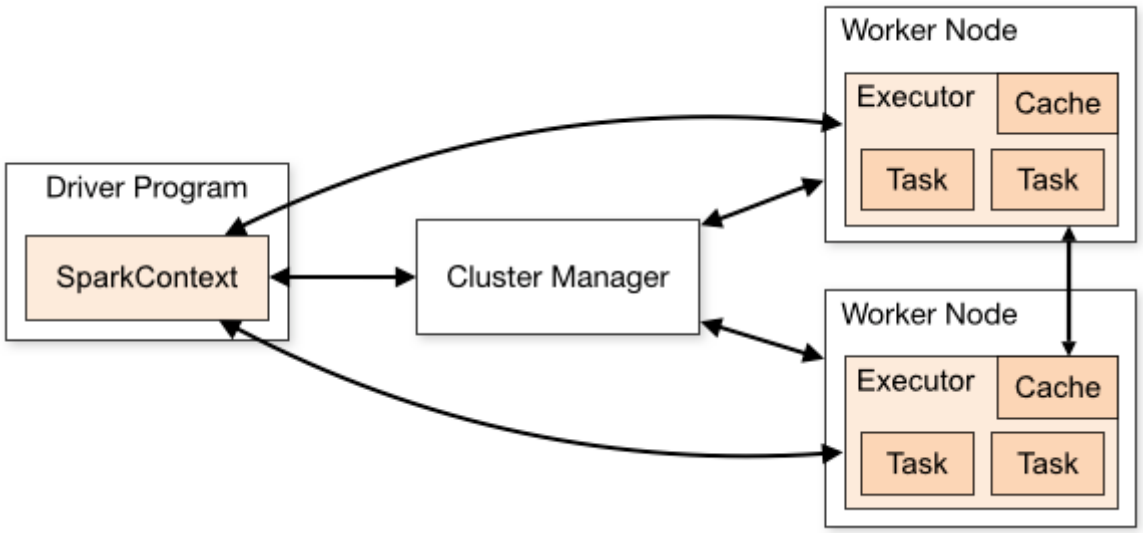


Figure 5. Distributed Spark application component model

A key distinguishing feature of Spark is the data model, based on RDDs (Resilient Distributed Datasets). This model enables a compact and reusable organization of data set that can reside in the main memory and can be accessed by multiple tasks. Iterative processing algorithms can benefit from this feature by not having to store and retrieve data sets from disks between iterations of computation. These capabilities are what deliver the significant performance gains compared to MapReduce.

RDDs support two types of operations: Transformations and Actions. Transformations are operations that return a new RDD, while Actions return a result to the driver program. Spark groups operations together to reduce the number of passes taken over the data. This so-called lazy evaluation technique enables faster data processing. Spark also allows caching data in memory for persistence to enable multiple uses of the same data. This is another technique contributing to faster data processing.

6 Operational model

This section describes the operational model for the Cloudera reference architecture. To show the operational model for different sized customer environments, four different models are provided for supporting different amounts of data. Throughout the document, these models are referred to as starter rack, half rack, full rack, and multi-rack configuration sizes. The multi-rack is three times larger than the full rack.

A Cloudera deployment consists of cluster nodes, networking equipment, power distribution units, and racks. The predefined configurations can be implemented as-is or modified based on specific customer requirements, such as lower cost, improved performance, and increased reliability. Key workload requirements, such as the data growth rate, sizes of datasets, and data ingest patterns help in determining the proper configuration for a specific deployment. A best practice when a Cloudera cluster infrastructure is designed is to conduct the proof of concept testing by using representative data and workloads to ensure that the proposed design works.

6.1 Hardware description

This reference architecture uses Lenovo servers SR630 (1U) and SR650 (2U) servers and Lenovo RackSwitch G8052 and G8272 top of rack switches.

6.1.1 Lenovo ThinkSystem SR650 Server

The Lenovo ThinkSystem SR650 is an ideal 2-socket 2U rack server for small businesses up to large enterprises that need industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. The SR650 server is particularly suited for big data applications due to its rich internal data storage, large internal memory and selection of high performance Intel processors. It is also designed to handle general workloads, such as databases, virtualization and cloud computing, virtual desktop infrastructure (VDI), enterprise applications, collaboration/email, and business analytics.

The SR650 server supports:

- Up to two Intel® Xeon® Scalable Processors
- Up to 1.5 TB 2666 MHz TruDDR4 memory (support for up to 3 TB is planned for future),
- Up to 24x 2.5-inch or 14x 3.5-inch drive bays with an extensive choice of NVMe PCIe SSDs, SAS/SATA SSDs, and SAS/SATA HDDs
- Flexible I/O Network expansion options with the LOM slot, the dedicated storage controller slot, and up to 6x PCIe slots

Figure 6. Lenovo ThinkSystem SR650

Combined with the Intel® Xeon® Scalable Processors (Bronze, Silver, Gold, and Platinum), the Lenovo SR650 server offers an even higher density of workloads and performance that lowers the total cost of ownership

(TCO). Its pay-as-you-grow flexible design and great expansion capabilities solidify dependability for any kind of workload with minimal downtime.

The SR650 server provides high internal storage density in a 2U form factor with its impressive array of workload-optimized storage configurations. It also offers easy management and saves floor space and power consumption for most demanding use cases by consolidating storage and server into one system.

This reference architecture recommends the storage-rich ThinkSystem SR650 for the following reasons:

- **Storage capacity**: The nodes are storage-rich. Each of the 14 configured 3.5-inch drives has raw capacity up to 10 TB and each, providing for 140 TB of raw storage per node and over 2000 TB per rack.
- **Performance**: This hardware supports the latest Intel® Xeon® Scalable processors and TruDDR4 Memory.
- **Flexibility**: Server hardware uses embedded storage, which results in simple scalability (by adding nodes).
- **PCIe slots**: Up to 7 PCIe slots are available if rear disks are not used, and up to 3 PCIe slots if the Rear HDD kit is used. They can be used for network adapter redundancy and increased network throughput.
- **Higher power efficiency**: Titanium and Platinum redundant power supplies that can deliver 96% (Titanium) or 94% (Platinum) efficiency at 50% load.
- **Reliability:** Outstanding reliability, availability, and serviceability (RAS) improve the business environment and helps save operational costs

For more information, see the Lenovo ThinkSystem SR650 Product Guide:

https://lenovopress.com/lp0644-lenovo-thinksystem-sr650-server

6.1.2 Lenovo ThinkSystem SR630 Server

The Lenovo ThinkSystem SR630 server (shown in Figure 7) is a cost and density-balanced 1U two-socket rack server. The SR650 features a new, innovative, energy-efficient design with up to two Intel® Xeon® Scalable processors (Bronze, Silver, Gold and Platinum), a large capacity of faster, energy-efficient TruDDR4 Memory, up to 14x 3.5" SAS drives or 24x 2.5" SAS drives, and up to three PCI Express (PCIe) 3.0 I/O expansion slots in an impressive selection of sizes and types. The server has improved feature set and exceptional performance is ideal for scalable cloud environments.

Figure 7: Lenovo ThinkSystem SR630

For more information, see the Lenovo ThinkSystem SR630 Product Guide:
https://lenovopress.com/lp0643-lenovo-thinksystem-sr630-server

6.1.3 Lenovo RackSwitch G8052

The Lenovo networking RackSwitch G8052 (as shown in Figure 8) is an Ethernet switch that is designed for the data center and provides a simple network solution. The Lenovo RackSwitch G8052 offers up to 48x 1 GbE ports and up to 4x 10 GbE ports in a 1U footprint. The G8052 switch is always available for business-critical traffic by using redundant power supplies, fans, and numerous high-availability features.

Figure 8. Lenovo RackSwitch G8052

Lenovo RackSwitch G8052 has the following characteristics:

- A total of **48x 1 GbE** RJ45 ports
- **Four 10 GbE** SFP+ ports
- Low 130W power rating and variable speed fans to reduce power consumption

For more information, see the Lenovo RackSwitch G8052 Product Guide:
https://lenovopress.com/tips1270-lenovo-rackswitch-g8052

6.1.4 Lenovo RackSwitch G8272

Designed with top performance in mind, Lenovo RackSwitch G8272 is ideal for today's big data, cloud and optimized workloads. The G8272 switch offers up to 72 10Gb SFP+ ports in a 1U form factor and is expandable with six 40Gb QSFP+ ports. It is an enterprise-class and full-featured data center switch that delivers line-rate, high-bandwidth switching, filtering and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Redundant power and fans and numerous HA features equip the switches for business-sensitive traffic.

The G8272 switch (as shown in Figure 9) is ideal for latency-sensitive applications. It supports Lenovo Virtual Fabric to help clients reduce the number of I/O adapters to a single dual-port 10Gb adapter, which helps reduce cost and complexity. The G8272 switch supports the newest protocols, including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) for support of FCoE and iSCSI and NAS.

Figure 9: Lenovo RackSwitch G8272

The enterprise-level Lenovo RackSwitch G8272 has the following characteristics:

- **48x SFP+ 10GbE ports** plus **6x QSFP+ 40GbE ports**
- Support up **to 72x 10Gb connections using break-out cables**
- 1.44 Tbps non-blocking throughput with low latency (~ 600 ns)
- Up to 72 1Gb/10Gb SFP+ ports
- OpenFlow enabled allows for easily created user-controlled virtual networks
- **Virtual LAG and LACP** for dual switch redundancy

For more information, see the Lenovo RackSwitch G8272 Product Guide:

https://lenovopress.com/tips1267-lenovo-rackswitch-g8272

6.1.5 Lenovo RackSwitch NE10032 - Cross-Rack Switch

The Lenovo ThinkSystem NE10032 RackSwitch that uses 100 Gb QSFP28 and 40 Gb QSFP+ Ethernet technology is specifically designed for the data center. It is ideal for today's big data workload solutions and is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth switching, filtering and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The NE10032 RackSwitch has 32x QSFP+/QSFP28 ports that support 40 GbE and 100 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. It is an ideal cross-rack aggregation switch for use in a multi rack big data Cloudera cluster.

Figure 10: Lenovo ThinkSystem NE10032 cross-rack switch

For further information on the NE10032 switch, visit this link:

https://lenovopress.com/lp0609-lenovo-thinksystem-ne10032-rackswitch

6.2 Cluster nodes

The Cloudera reference architecture is implemented on a set of nodes that make up a cluster. A Cloudera cluster consists of two types of nodes: Worker nodes and Master nodes. Worker nodes use ThinkSystem SR650 servers with locally attached storage and Master nodes use ThinkSystem SR630 servers.

Worker nodes run data (worker) services for storing and processing data.

Master nodes run the following types of services:

- Management control services for coordinating and managing the cluster
- Miscellaneous and optional services for file and web serving

1.1.1 Worker nodes

Table 1 lists the recommended system components for worker nodes demonstrated in this reference architecture.

Table 1. Worker node configuration

Component	Worker node configuration
Server	ThinkSystem SR650
Processor	2x Intel® Xeon® processors: 6130 Gold, 16-core 2.1Ghz
Memory - base	256 GB: 8x 32GB 2666MHz RDIMM
Disk (OS)	Dual M.2 128GB SSD with RAID1
Disk (data)	4 TB drives: 14x 4TB NL SAS 3.5 inch (56 TB total) Alternate HDD capacities available: 6 TB drives; 14x 6TB NL SAS 3.5 inch (84 TB total) 8 TB drives: 12x 8TB NL SAS 3.5 inch (96 TB total)* 10 TB drives: 10x 10TB NL SAS 3.5 inch (100TB total)*
HDD controller	OS: M.2 RAID1 mirror enablement kit HDFS: ThinkSystem 430-16i 12Gb HBA
Hardware storage protection	OS: RAID1 HDFS: None (JBOD). By default, Cloudera maintains a total of three copies of data stored within the cluster. The copies are distributed across data servers and racks for fault recovery.
Hardware management network adapter	Integrated 1G BaseT XCC management controller - dedicated or shared LAN port
Data network adapter	ThinkSystem 10Gb 4-port SFP+ LOM

* *Cloudera recommended maximum storage per worker node is 100 TB*

The Intel® Xeon® Scalable Processor recommended in Table 1 will provide a balance in performance vs. cost for Cloudera worker nodes. Higher core count and frequency processors are available for compute intensive workloads. A minimum of 256 GB of memory is recommended for most MapReduce workloads with 512 GB or more recommended for HBase, Spark and memory-intensive MapReduce workloads, and VMware virtualized environments.

The OS is loaded on a dual M.2 SSD memory module with RAID1 mirroring capability. Data disks are JBOD configured for maximum Hadoop and Spark performance with data fault tolerance coming from the HDFS file system 3x replication factor.

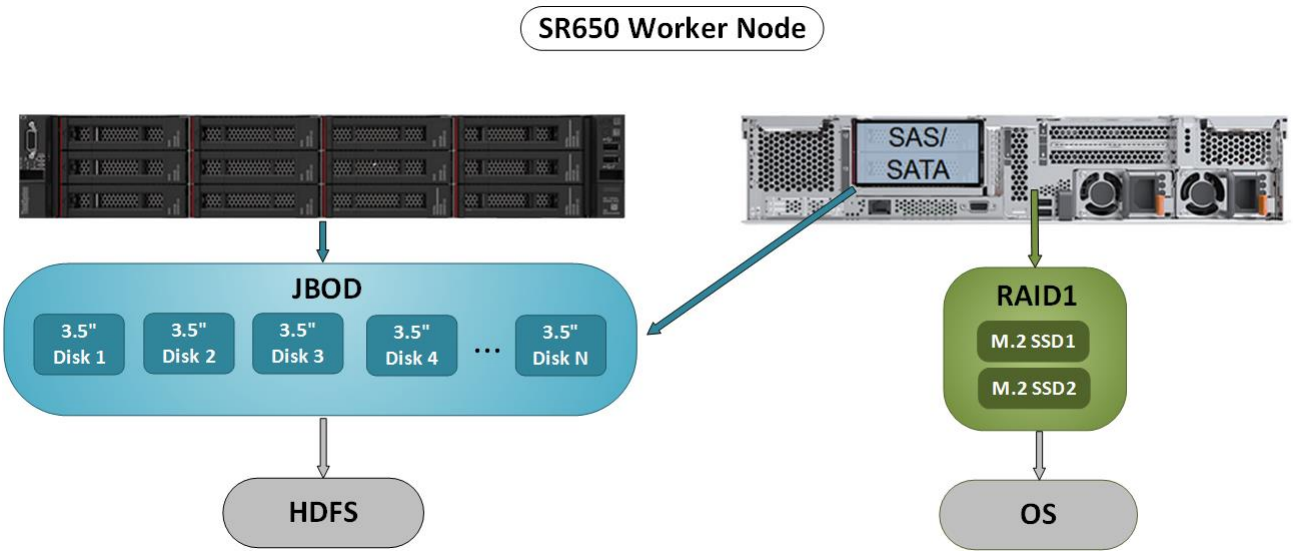


Figure 11: Worker node disk assignment

Each worker node in the reference architecture has internal directly attached storage. External storage is not used in this reference architecture. Available data space assumes the use of Hadoop replication with three copies of the data (reduces effective disk space by 3x) plus a 25% reserve capacity so the HDFS file system is not constrained near term usage growth.

A minimum of three worker nodes are required as Hadoop has three copies of data by default. Three should be used for test or Proof of Concept (POC) environments only. A minimum of five worker nodes are required for production environment to reduce risk from losing more than one node at a time

6.2.1 Master Nodes

The Master node is the nucleus of the Hadoop Distributed File System (HDFS) and supports several other key functions that are needed on a Cloudera cluster.

The Master node runs the following services:

YARN ResourceManager: Manages and arbitrates resources among all the applications in the system.

Hadoop NameNode: Controls the HDFS file system. The NameNode maintains the HDFS metadata, manages the directory tree of all files in the file system and tracks the location of the file data within the cluster. The NameNode does not store the data of these files.

ZooKeeper: Provides a distributed configuration service, a synchronization service and a name registry for distributed systems.

JournalNode: Collects, maintains and synchronize updates from NameNode.

HA ResourceManager: Standby ResourceManager that can be used to provide automated failover.

HA NameNode: Standby NameNode that can be used to provide automated failover.

Other non-master node services for Hadoop component management such as: Cloudera Manager, HBase master, HiveServer2, and Spark History Server.

Table 2 lists the recommended components for a Master node and they can be customized according to client needs.

Table 2. Master node configuration

Component	Master node configuration
Server	ThinkSystem SR630
Processor	2x Intel® Xeon® Scalable Processors: 4114 Silver, 12-core 2.1Ghz
Memory - base	128 GB – 8x 16 GB 2666MHz RDIMM
Disk (OS / local storage)	OS: Dual M.2 128GB SSD with RAID1 Data: 8x 2TB 2.5” SAS HDD
HDD controller	ThinkSystem RAID 930-16i 4GB Flash 12Gb controller
Hardware storage protection	OS: RAID1 NameNode/Metastore: RAID1 Database: RAID10 Zookeeper/QJN: No h/w protection; JBOD HDDs; multiple service instances across Master nodes provide redundancy
Hardware management controller	Integrated XCLARITY™ CONTROLLER (XCC) with 1GBaseT dedicated interface or shared LAN interface
Data network adapter	ThinkSystem 10Gb 4-port SFP+ LOM

The Intel® Xeon® Scalable Processors and minimum memory specified in Table 2 is recommended to provide sufficient performance as a Cloudera Master node. The M.2 SSD form factor is intended for Operating Storage in this reference architecture.

The Master node uses 10 drives for the following storage pools:

- Two drives (M.2 SSD modules) are configured with RAID 1 for operating system
- Two drives are configured with RAID 1 for NameNode metastore
- Four drives are configured with RAID 10 for database
- One drive is configured with RAID 0 for ZooKeeper
- One drive is configured with RAID 0 for Quorum Journal Node store

This design separates the data stores for different services and provides best performance. SSD drives in the 2.5" and 3.5" SAS/SATA form factor and PCIe card flash storage can be used to provide improved I/O performance for the database.

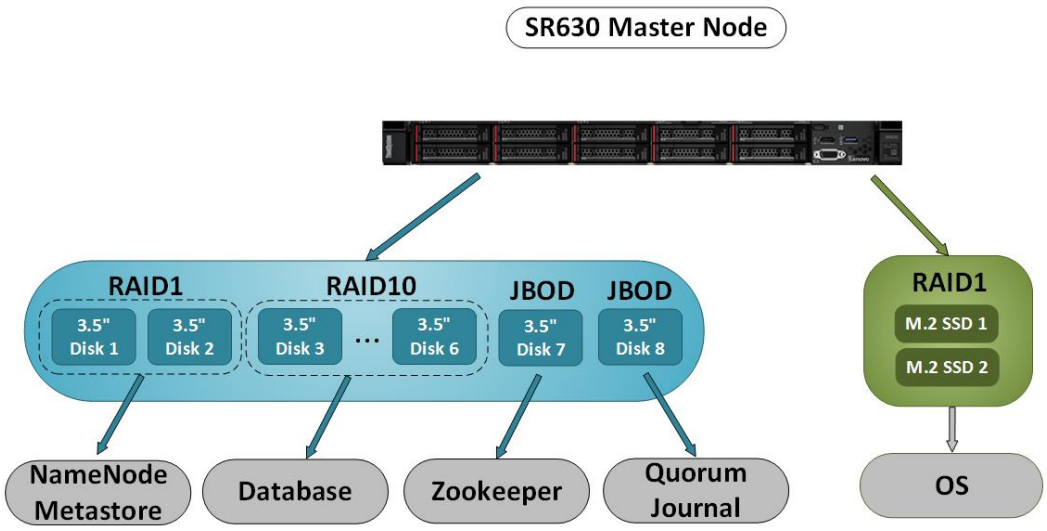


Figure 12: Cloudera Master node disk assignment

Because the Master node is responsible for many memory-intensive tasks, multiple Master nodes are needed to split out functions. For most implementations, the size of the Cloudera cluster is a good indicator of how many Master nodes are needed. Table 3 provides a high-level guideline for a cluster that provides HA NameNode and ResourceManager failover when configured with multiple Master nodes. For a medium size clusters approaching 200 worker nodes and beyond, Master nodes will need consideration for increased memory and CPU core size.

Table 3. Number of Master Nodes

Number of Data Nodes	Number of Master nodes	Breakout of function
< 100	3	Cloudera Manager, Journal Node, ZooKeeper ResourceManager, HA Hadoop NameNode, JournalNode, ZooKeeper HA ResourceManager, Hadoop NameNode, JournalNode, ZooKeeper
> 100	5	Cloudera Manager, Journal Node, ZooKeeper ResourceManager, HA Hadoop NameNode, JournalNode, ZooKeeper HA ResourceManager, Hadoop NameNode, JournalNode, ZooKeeper JournalNode, ZooKeeper, other roles JournalNode, ZooKeeper, other roles

Note: To ease scale-up the cluster with worker nodes, one can plan ahead by installing the next level of Master nodes to be ready for additional Worker nodes

Table 4. Service Layout Matrix

Node		Master Node	Master Node	Master Node	Data Nodes
Service/ Roles	**ZooKeeper**	ZooKeeper	ZooKeeper	ZooKeeper	
	HDFS	NN,QJN	NN,QJN	QJN	Data Node
	YARN	RM	RM	History Server	Node Manager
	Hive			MetaStore, WebHCat, HiveServer2	
	Management	Cloudera Agent	Cloudera Agent	Oozie, CM, Management Services	Cloudera Agent
	Navigator			Navigator, KMS	
	HUE			HUE	
	Spark				Runs on YARN
	Impala			Statestore	impalad
	HBASE	HMaster	HMaster	HMaster	Region Servers

Installing and managing the Cloudera Stack

The Hadoop ecosystem is complex and constantly changing. Cloudera makes it simple so enterprises can focus on results. Cloudera Manager is the easiest way to administer Hadoop in any environment, with advanced features like intelligent defaults and customizable automation. Combined with predictive maintenance included in Cloudera's Support Data Hub, Cloudera Enterprise keeps the business up and running.

Reference Cloudera's latest Installation documentation for detailed instructions on Installation

6.3 Systems management

Systems management of a cluster includes Operating System, Hadoop & Spark applications and hardware management. Systems management uses Cloudera Manager and is adapted from the standard Hadoop distribution, which places the management services on separate servers than the worker servers. The Master node runs important and high-memory use functions, so it is important to configure a powerful and fast server for systems management functions. The recommended Master node hardware configuration can be customized according to client needs.

Hardware management uses the Lenovo XClarity™ Administrator, which is a centralized resource management solution that reduces complexity, speeds up response and enhances the availability of Lenovo server systems and solutions. XClarity™ is used to install the OS onto new worker nodes; update firmware

across the cluster nodes, record hardware alerts and report when repair actions are needed.

Figure 13 shows the Lenovo XClarity™ Administrator interface in which servers, storage, switches and other rack components are managed and status is shown on the dashboard. Lenovo XClarity™ Administrator is a virtual appliance that is quickly imported into a server virtualized environment.

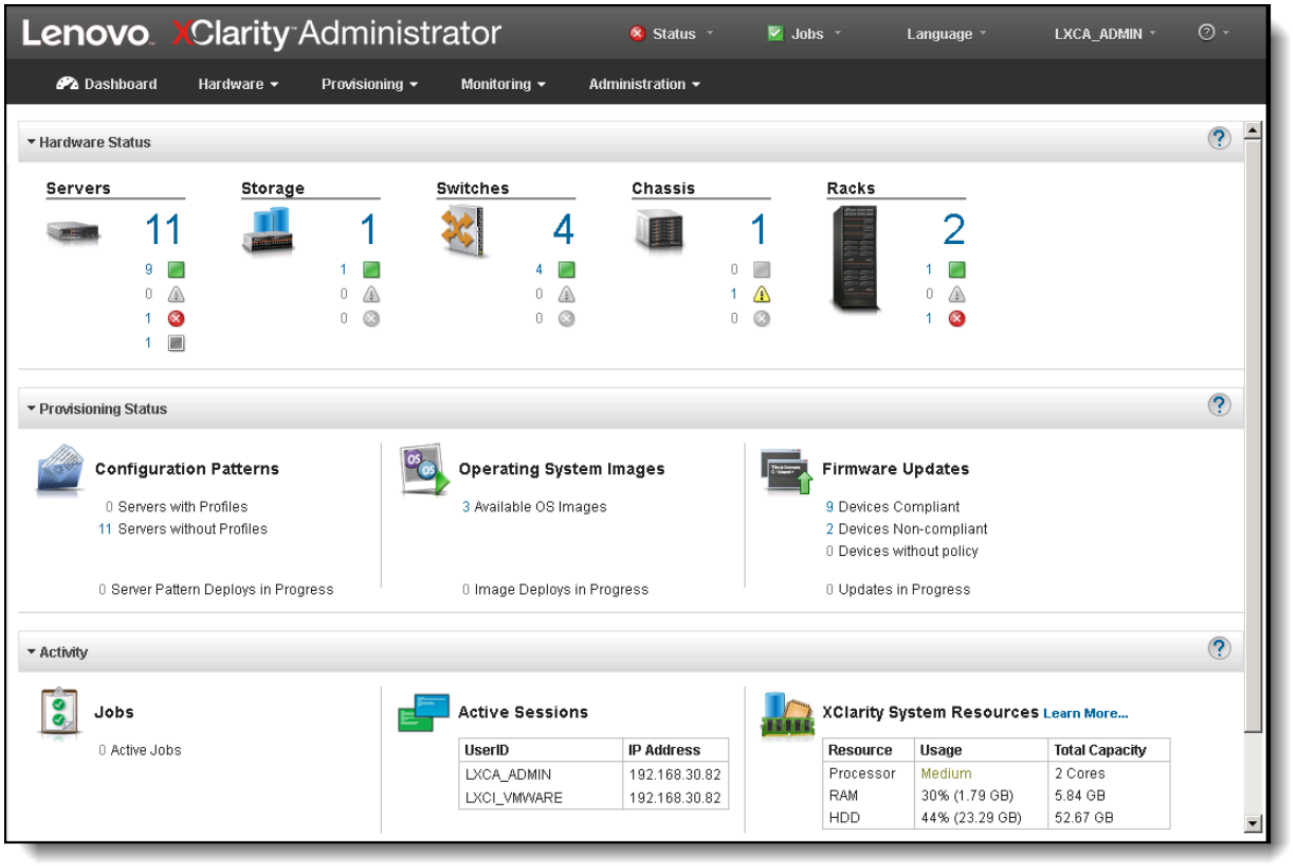


Figure 13: XClarity™ Administrator interface

In addition, xCAT provides a scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery and operating system deployment. It can be used to facilitate or automate the management of cluster nodes. For more information about xCAT, see “Resources” on page 50.

6.4 Networking

The reference architecture specifies two networks: a high-speed data network and a management network. Two types of top of rack switches are required; one 1Gb for out-of-band management and a pair of 10Gb for the data network with High Availability. See Figure 14 below.

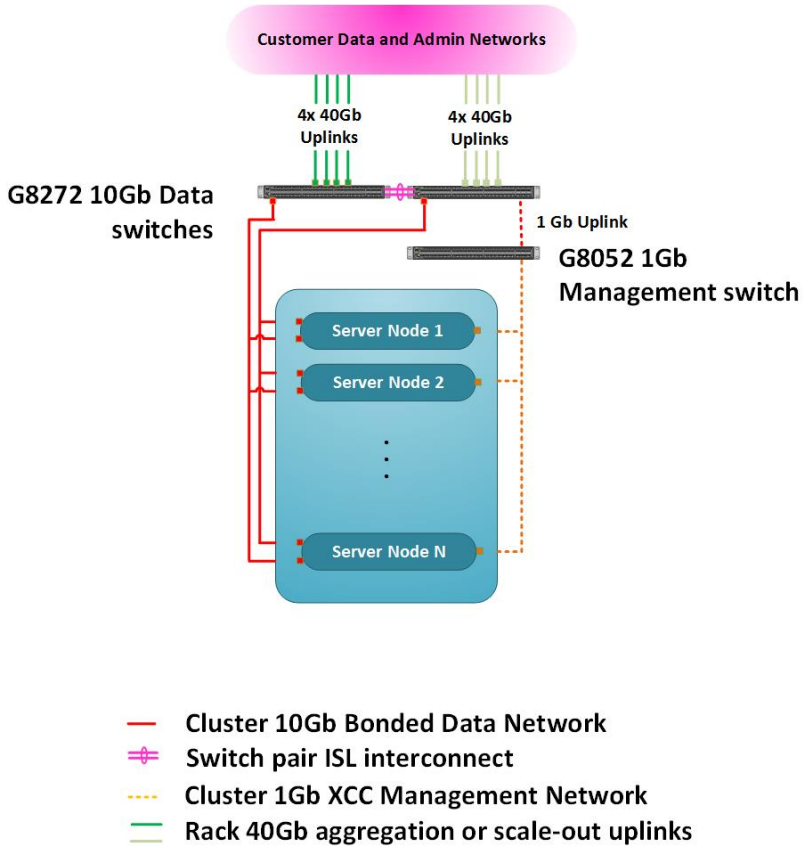


Figure 14 Cloudera network

6.4.1 Data network

The data network creates a private cluster among multiple nodes and is used for high-speed data transfer across worker and master nodes, and also for importing data into the Cloudera cluster. The Cloudera cluster typically connects to the customer's corporate data network. The recommended 10 GbE switch is the Lenovo System Networking RackSwitch™ G8272 that provides 48 10Gb Ethernet ports with 40Gb uplink ports.

The two 10GbE NIC ports of each node are link aggregated into a single bonded network connection. The two data switches are connected together as a Virtual Link Aggregation Group (vLAG) pair using LACP to provide the switch redundancy. Either G8272 switch can drop out of the network and the other G8272 continues transferring 10Gb traffic. The switch pairs are connected with dual 10Gb links called an ISL, which allows maintaining consistency between the two peer switches.

6.4.2 Hardware management network

The hardware management network is a 1GbE network for out-of-band hardware management. The recommended 1GbE switch is the Lenovo RackSwitch G8052 with 10Gb SFP+ uplink ports. Through the XClarity™ Controller management module (XCC) within the ThinkSystem SR650 and SR630 servers, the out-of-band network enables hardware-level management of cluster nodes, such as node deployment, UEFI firmware configuration, hardware failure status and remote power control of the nodes.

Hadoop has no dependency on the XCC management function. The Cloudera/OS management network can be shared with the XCC hardware management network, or can be separated via VLANs on the respective

switches. The Cloudera cluster and hardware management networks are then typically connected directly to the customer's existing administrative network to facilitate remote maintenance of the cluster.

6.4.3 Multi-rack network

The data network in the predefined reference architecture configuration consists of a single network topology. A rack consists of redundant G8272 access level 10Gb switches. Data and Master nodes are connected with bonded 10Gb links (NIC teaming) for further redundancy to each server node. Additional racks can be added as needed for scale out. Beginning with the third rack a core switch for rack aggregation is used and the Lenovo NE10032 core switch with 40Gb and 100Gb uplinks is the best choice for this purpose.

Figure 12 shows a 2-rack configuration. A single rack can be upgraded to this configuration by adding the second rack with the LAG network connection show.

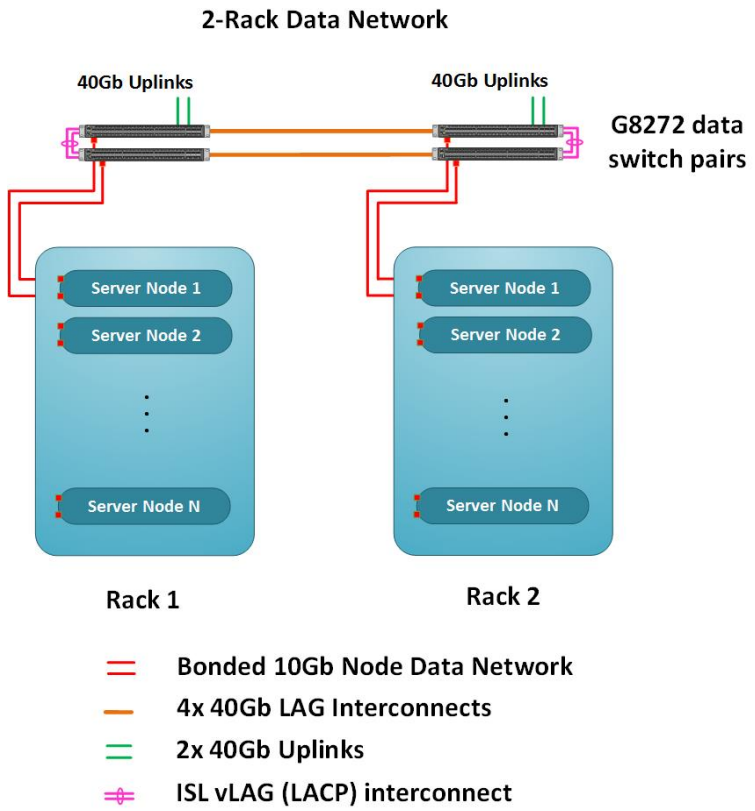


Figure 15. Cloudera 2-rack network configuration

Figure 13 shows how the network is configured when the Cloudera cluster contains 3 or more racks. The data network is connected across racks by four aggregated 40 GbE uplinks from each rack's G8272 switch to a core NE10032 switch. The 2-rack configuration can be upgraded to this 3-rack configuration as shown. Additional racks can be added with similar uplink connections to the NE10032 cross rack switch. Reference Figure 16 and Figure 18.

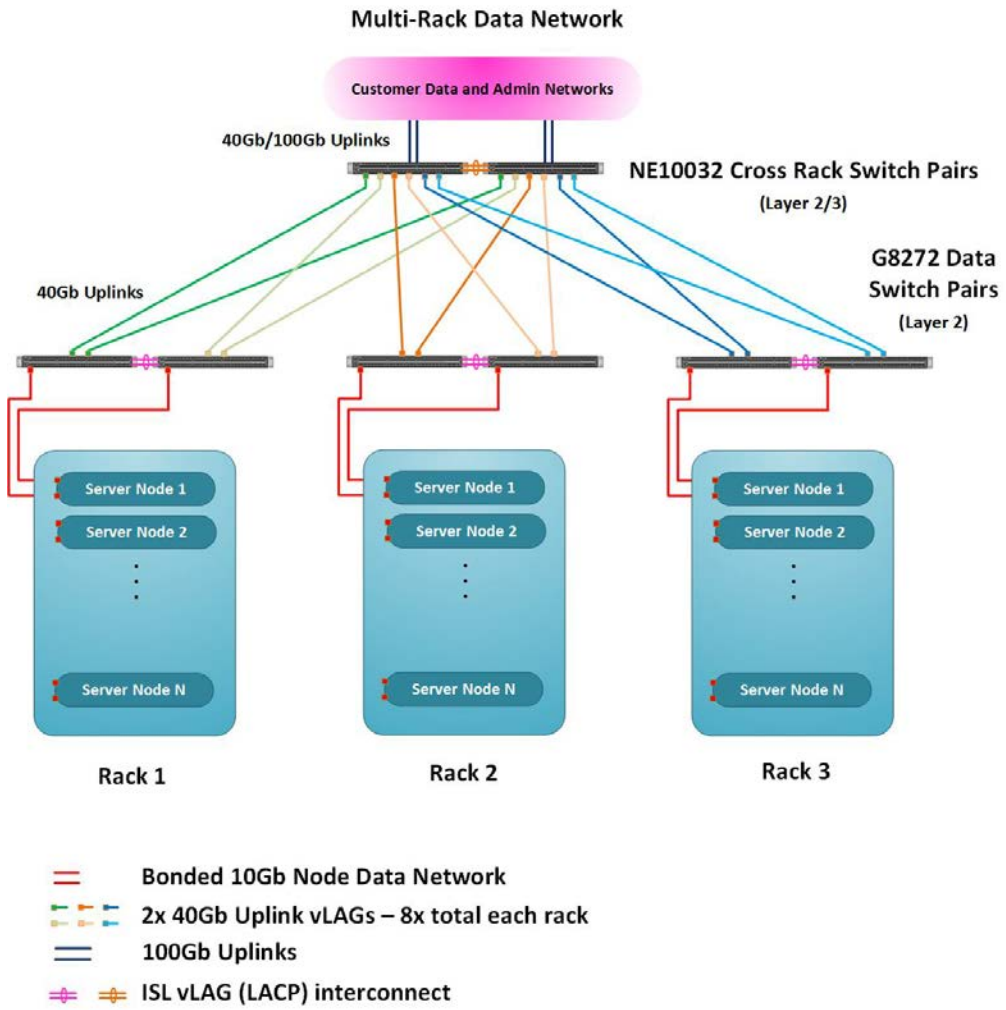


Figure 16 Cloudera multi-rack rack network configuration

Within each rack, the G8052 1Gb management switch can be configured to have two uplinks to the G8272 switch for propagating the management VLAN across cluster racks through the NE10032 cross-rack switch. Other cross rack network configurations are possible and may be required to meet the needs of specific deployments and to address clusters larger than three racks.

For multi-rack solutions, the Master nodes can be distributed across racks to maximize fault tolerance.

6.5 Predefined cluster configurations

The intent of the predefined configurations is to ease initial sizing for customers and to show example starting points for four different-sized workloads: the starter rack, half rack, full rack, and a 3 rack multi-rack configuration. These consist of Worker nodes, Master nodes, and network switches, and rack hardware. Table 5 below, Figure 17 and Figure 18 show the number of consists of three nodes and a both management and data rack switches. The half rack configuration consists of nine nodes and rack switches. The full rack configuration consists of 17 worker nodes, 3 Master nodes, and a Systems Management node. A three rack multi-rack contains a total of worker 55 nodes, 3 Master nodes, and a Systems Management node. Table 5 lists the four predefined configurations for the Cloudera reference architecture. The table also lists the amount

of space for data and the number of nodes that each predefined configuration provides. Storage space is described in two ways: the total amount of ***raw storage space*** when 4 TB or up to 10 TB drives are used and the amount of ***usable space*** available for customer data. Available data space assumes the use of Hadoop replication with three copies of the data and 25% reserve working capacity. The estimates that are listed in Table 5 are for uncompressed data. Compression rates can vary widely based on file contents and usable space must be calculated based on the specific compression rate used.

Table 5. Cluster Storage Capacity Examples, 3.5" HDDs

3.5" HDD Large Form Factor (LFF)	**Starter rack**	**Half rack**	**Full rack**	**Multi-rack (3x)**
Storage space using 4 TB drives				
Raw storage	168 TB	504 TB	952 TB	3080 TB
Usable w/ 25% reserve	42 TB	126 TB	238 TB	770 TB
Storage space using 6 TB drives				
Raw storage	252 TB	756 TB	1428 TB	4620 TB
Usable w/ 25% reserve	63 TB	189 TB	357 TB	1155 TB
Number of Nodes				
Number of data nodes	3	9	17	55
HDDs per data node	14	14	14	14
Storage space using 8 TB drives				
Raw storage	288 TB	480 TB	1632 TB	5280 TB
Usable w/ 25% reserve	72 TB	120 TB	408 TB	1320 TB
Number of Nodes				
Number of data nodes	3	5	17	55
HDDs per data node	12	12	12	12
Storage space using 10 TB drives				
Raw storage	300 TB	500 TB	1700 TB	5500 TB
Usable w/ 25% reserve	75 TB	125 TB	425 TB	1375 TB
Number of Nodes				
Number of data nodes	3	5	17	55
HDDs per data node	10	10	10	10

Note: Data compression techniques can reduce raw storage requirements.
Reference section Error! Reference source not found. Error! Reference source not found..

Figure 17 shows an overview of the reference architecture for the half rack and full rack configurations. Figure 18 shows a multi-rack-sized cluster.

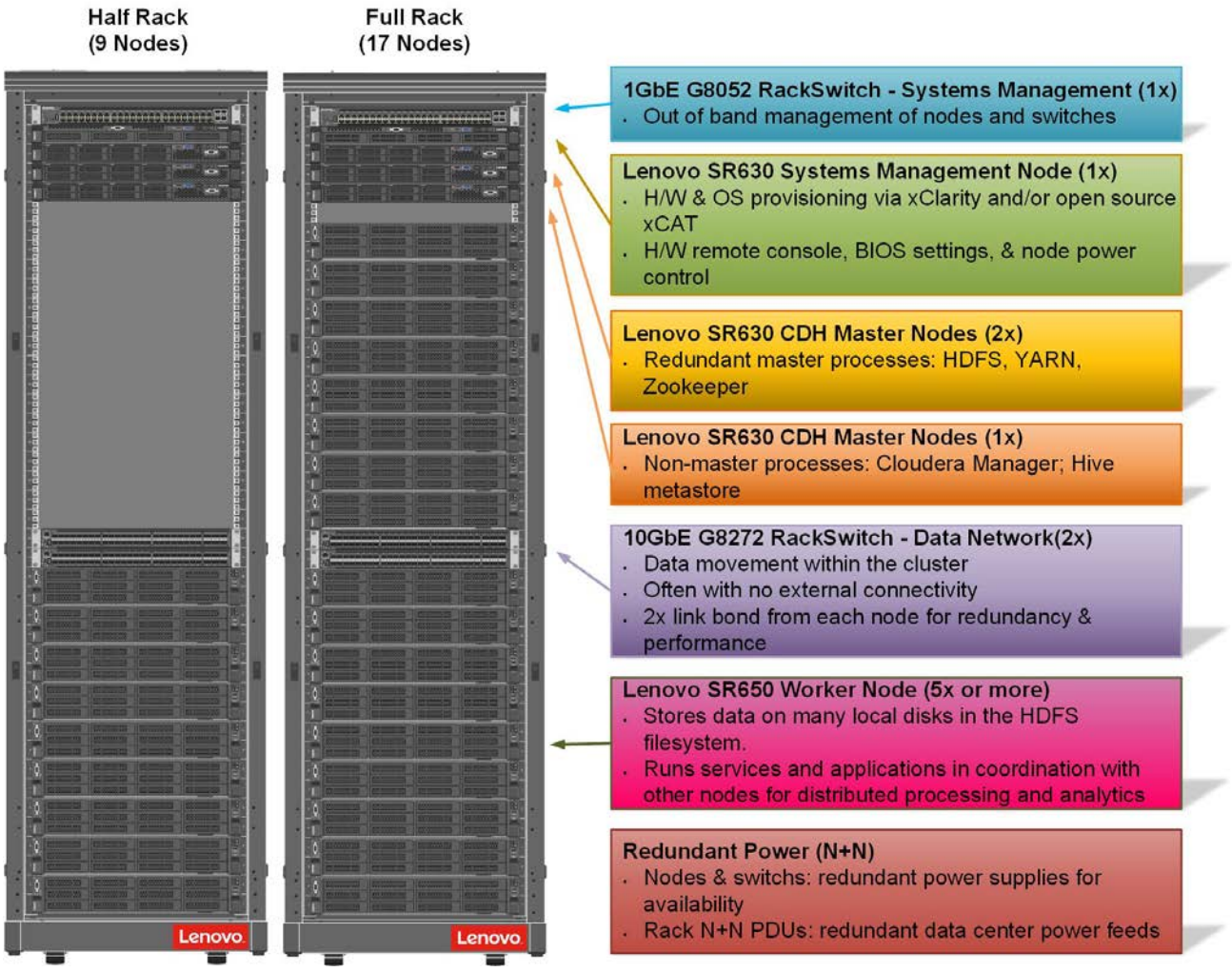


Figure 17: Half rack and full rack Cloudera predefined configurations

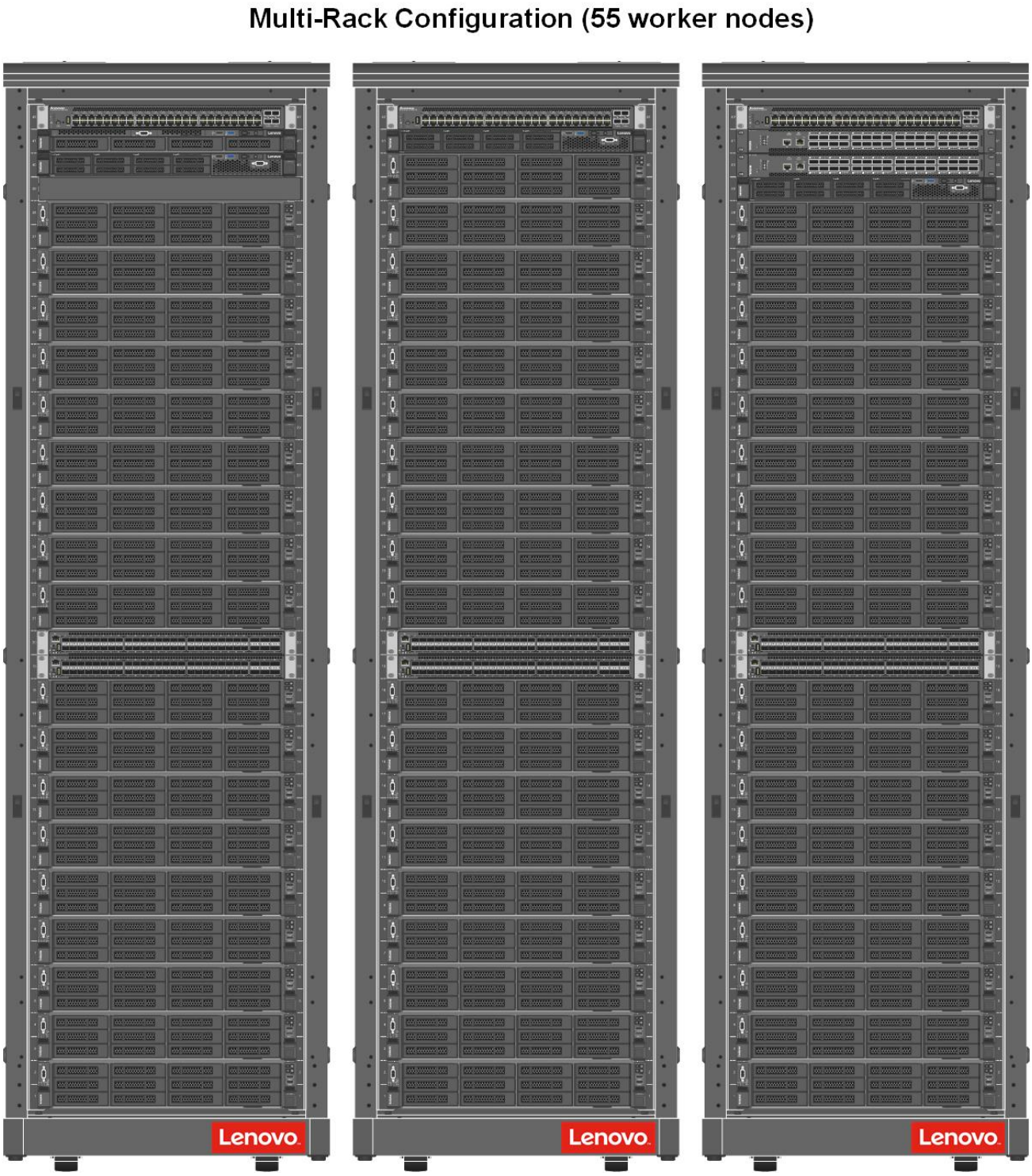


Figure 18: Multi-rack Cloudera configuration

7 Deployment considerations

This section describes other considerations for deploying the Cloudera solution.

7.1 Increasing cluster performance

There are two approaches that can be used to increase cluster performance: increasing node memory and the use of a high-performance job scheduler and MapReduce framework. Often, improving performance comes at increased cost and you must consider the cost-to-benefit trade-offs of designing for higher performance.

In the Cloudera predefined configuration, node memory can be increased to 768 Gb with 24x 32GB RDIMMs, 1,536 GB using 24x 64GB LRDIMMs and up to 3,072 GB per node using 3DS RDIMMs and Intel processors that support 1.5TB each.

7.2 Designing for high ingest rates

Designing for high ingest rates is difficult. It is important to have a full characterization of the ingest patterns and volumes. The following questions provide guidance to key factors that affect the rates:

- On what days and at what times are the source systems available or not available for ingest?
- When a source system is available for ingest, what is the duration for which the system remains available?
- Do other factors affect the day, time and duration ingest constraints?
- When ingests occur, what is the average and maximum size of ingest that must be completed?
- What factors affect ingest size?
- What is the format of the source data (structured, semi-structured, or unstructured)? Are there any data transformation or cleansing requirements that must be achieved during ingest?

To increase the data ingest rates, consider the following points:

- Ingest data with MapReduce job, which helps to distribute the I/O load to different nodes across the cluster.
- Ingest when cluster load is not high, if possible.
- Compressing data is a good option in many cases, which reduces the I/O load to disk and network.
- Filter and reduce data in earlier stage saves more costs.

7.3 Designing for Storage Capacity and Performance

Selection of the HDD form factor, number of drives, and size of each drive can skew a worker node towards highest capacity or highest disk IO throughput.

7.3.1 Node Capacity

The 3.5" HDD form factor gives the maximum local storage ***capacity*** *for a node*. 10TB and larger HDDs are available and can be used to replace the 4TB HDDs used in this reference architecture to give a total of up to the 100 TBs per node (as the recommended maximum by Cloudera). The 4TB HDD size provides the best balance of HDD capacity and performance per node. When increasing data disk capacity, some workloads may experience a decrease in disk parallelism, creating a bottleneck at that node which negatively affects

performance. To increase capacity beyond the 4TB HDD size recommended in this reference architecture, the number of nodes in the cluster should be increased to maintain good I/O disk node performance

7.3.2 Node Throughput

The 2.5" HDD form factor gives the maximum local storage ***throughput*** for a node configuration. In cases where the maximum local storage throughout per node is required, the worker node can be configured with 24x 2.5-inch SAS drives. The 2.5-inch HDD has less total capacity per drive and gives less total capacity per node than the 3.5" form factor, but allows for higher parallel access to the drives - more data can be accessed simultaneously. The SR650 configuration using 2.5" and 3.5" HDDs is listed below as an example of maximum node capacity vs. parallel HDD connections for various drive sizes.

HDD Form Factor	HDD size	Max. node storage capacity	Parallel HDD Connections
3.5" HDDs, 14x HDDs	10 TB Drive	100 TB	10
	8 TB Drive	96 TB	12
2.5" HDDs, 24x HDDs	2.4 TB Drive	57.6 TB	24

Solid State Drives (SSDs) are also available in the 2.5" form factor for the SR650 with a higher capacity per drive than spinning HDDs, but at a significantly higher cost per drive.

In the 2.5" HDD configuration of the SR650, it is recommended to use 3 host bus adapters for maximum parallel throughput vs. a single host bus adapter.

7.3.3 HDD controller

For the type of HDD controller, a host bus adapter driving just-a-bunch-of-disks (JBOD) is the best choice for a worker node in the Cloudera cluster. It provides excellent performance and, when combined with the Hadoop default of 3x data replication, also provides significant protection against data loss. The use of RAID with data disks is discouraged because it reduces performance and the amount data that can be stored. The Hadoop file system, HDFS, provides data redundancy across the Cloudera cluster via the 3 replicas of each data block which makes RAID unnecessary.

Use of RAID0, as a secondary choice, is supported with a single HDD per RAID array for better fault tolerance.

RAID1 and RAID10 are used for certain disks in a Cloudera Master node; therefore, a RAID HDD controller is specified in this configuration.

7.4 Designing for in-memory processing with Apache Spark

Methods from the Lenovo Big Data Reference Architecture for Cloudera Enterprise apply for general Spark considerations as well; however, there are additional considerations. Conceptually, Spark is similar in nature to high performance computing.

It is important that memory capacity be carefully considered, as both the execution and storage of Spark should be able to reside fully in memory, to achieve maximum performance, however there continue to be performance benefits even when an application doesn't fully fit within memory Disk access, for storage or caching, is very costly to Spark processing. The memory capacity considerations are highly dependent on the application. To get an estimate, load an RDD of a desired dataset, into cache, and evaluate the consumption. Generally, for workloads with high execution and storage requirements, capacity is primary consideration.

Additional considerations for memory configuration include the bandwidth and latency requirements. Applications with high transactional memory usage should focus on DIMM configurations that are balanced across the CPU memory controllers and their memory channels. The following table provides ideal worker node memory configurations for bandwidth/latency sensitive workloads.

Table 6. Recommended memory configurations for 2-socket worker nodes

Capacity	DIMM Description	Quantity
128GB	16GB TruDDR4 Memory (1Rx4, 1.2V) 2666MHz RDIMM	8
256GB	32GB TruDDR4 Memory (2Rx4, 1.2V) 2666Mhz RDIMM	8
384GB	32GB TruDDR4 Memory (2Rx4, 1.2V) 2666Mhz RDIMM	12
512GB	32GB TruDDR4 Memory (2Rx4, 1.2V) 2666Mhz RDIMM	16
768GB	64GB TruDDR4 Memory (4Rx4, 1.2V) 2666MHz LRDIMM	12
1,536GB	64GB TruDDR4 Memory (4Rx4, 1.2V) 2666MHz LRDIMM	24
3,072GB *	128GB TruDDR4 Memory (8Rx4 1.2V) 2666Mhz 3DS RDIMM	24
	DIMM counts to be avoided: 2,6,10,14,18,20,22	

Best
Better
Avoid

Notes: DIMM quantity is of the same part number (speed, size, rank, etc.)
** Requires CPU part numbers that support 1.5TB of memory each.*

Some memory configurations are unbalanced and negatively affect memory interleaving ability of the memory controller. Although these DIMM configurations are supported by the hardware and will function, they should be avoided in favor of the higher performance configurations. For more information on balanced memory configurations for Intel Xeon Scalable Processors see the link in the Reference section to the Lenovo white paper, *Intel Xeon Scalable Family Balanced Memory Configurations.*

Similarly, processor selection may vary based on the level of desired level of parallelism for the workloads. For example, Apache recommends 2-3 tasks per CPU core. Large working sets of data can drive memory constraints, which can be alleviated through further increasing parallelism, resulting in smaller input sets per task. In this case, higher core counts can be beneficial. Naturally, the nature of the operations is considered, as they may be simple evaluations or complex algorithms.

7.5 Data Network Adapter Options

The cluster data network using 10Gb bonded NIC interfaces connected with dual 10Gb network switches

provides 20Gb of network connectivity between nodes in the cluster. The ThinkSystem 4 port 10Gb LAN on Motherboard (LOM) adapter is recommended in this reference architecture. Additional network adapters are available with higher data rates and are shown in the table below.

Table 7. Network adapters for cluster nodes

Code	Description
AT7S	Emulex VFA5.2 2x10 GbE SFP+ PCIe Adapter
AT7T	Emulex VFA5.2 2x10 GbE SFP+ PCIe Adapter and FCoE/iSCSI SW
ATPX	Intel X550-T2 Dual Port 10GBase-T Adapter
ATRN	Mellanox ConnectX-4 1x40GbE QSFP+ Adapter
AUAJ	Mellanox ConnectX-4 2x25GbE SFP28 Adapter
AUKN	ThinkSystem Emulex OCe1410B-NX PCIe 10Gb 4-port SFP+ Ethernet Adapter
AUKP	ThinkSystem Broadcom NX-E Pcie 10Gb 2-Port Base-T Ethernet Adapter
AUKS	ThinkSystem Broadcom NX-E PCIe 25GbE 1-Port SFP28 Ethernet Adapter
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter
B0WY	ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-Port SFP28 Ethernet Adapter

7.6 Designing for Hadoop in a Virtualized Environment

This section of the reference architecture shows the important configuration items for the vSphere VMs; the Hadoop Virtualized Extensions (HVE) which maintain the storage, memory and CPU locality on each physical node; and the Cloudera software stack.

7.6.1 VMware vSphere Design

Virtual Network Switch

Standard vswitches may be employed, which need to be configured for each ESXi host in the cluster. A key configuration parameter to verify is the MTU size to ensure that the same MTU size being set at the physical switches, guest OS, ESXi VMNIC and the vswitch layers and that it's set to Jumbo Frames for highest performance - the is recommended for Hadoop environments.

Storage Group

Each provisioned disk is either mapped to one vSphere datastore (which in turn contains one VMDK or virtual disk) or mapped to one raw device mapping (RDM). Configure virtual disks in “independent persistent” mode for optimal performance. Eager Zeroed Thick virtual disks provide the best performance. In addition, make sure to disable SIOC, and disable storage DRS.

vSphere Tuning Best Practices

Power Policy is an ESXi parameter. The balanced mode may be the best option. Evaluate your environment and choose accordingly. In some cases, performance might be more important than power optimization. Avoid memory and CPU over-commitment, and ensure Transparent Huge Pages setting is Disabled for the ESXi Hypervisor and Guest OS.

VMXNET3 Virtual Network Driver

This driver is supported in RHEL and CentOS with the installation of VMware tools. Verify MTU size for jumbo frames at the guest level as well as ESXi and switch level. Only VMXNET3 drivers at the Guest layer can leverage this. Similarly, other offload features can be leveraged only when using the VMXNET3 driver. Use regular platform tuning parameters, such as ring buffer size. However, RSS and RPS tuning must be specific to the VMXNET3 driver.

HBA Driver Type

Use the VMware PVSCSI storage adapter. This provides the best performance characteristics (reduced CPU utilization and increased throughput), and is optimal for I/O-intensive guests (as with Hadoop). Tune queue depth in the guest OS SCSI driver, as needed.

I/O Scheduler

The I/O scheduler used for the OS disks might need to be different if using VMDKS. Instead of using CFQ, use deadline or noop elevators (deadline should be the default with rhel OS). This varies and must be tested. Any performance gains must be quantified appropriately; for example, 1-2% improvement vs. 10-20% improvement.

Memory Tuning

Disable or minimize anonymous paging by setting vm.swappiness=0 or 1. When tuning memory from the default settings to a best performance condition, configure VMs so that one or more VMs fit within a NUMA node size associated with one CPU socket, as far as their collective memory goes. The goal is to not have a VM access memory across a NUMA boundary.

Figure 19 below shows the virtualized cluster layout with 2 VMs per physical node and locally attached disks for each VM.

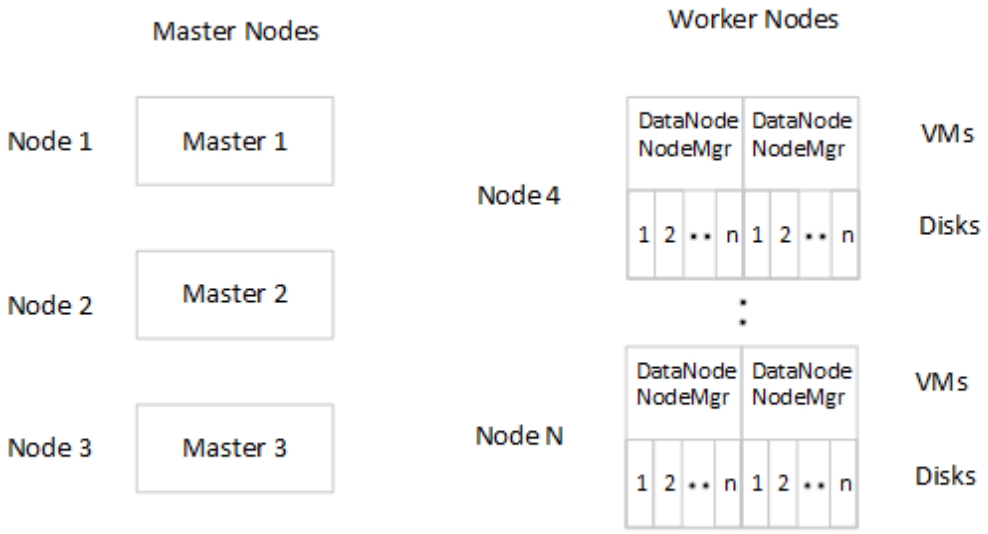


Figure 19. Example Virtualized Cluster Topology

7.6.2 Cloudera Software Stack Configuration

Guidelines for installing the Cloudera stack on a virtualized platform are nearly identical to those for bare-metal, except for enabling the Hadoop Virtualized Extensions (HVE) functionality.

Enabling Hadoop Virtualization Extensions (HVE)

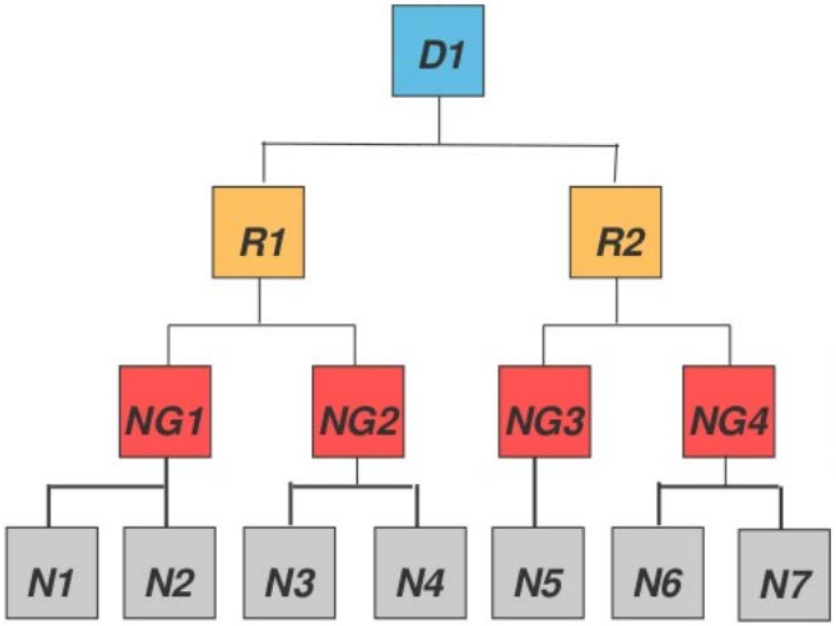
HVE is part of the Apache Project and adds physical node awareness to Hadoop and Spark for a virtualized environment. This enables HDFS to maintain all of the data block replicas across physical nodes as it does in the bare-metal environment. Refer to the Apache Project HVE descriptions and user guide at this link: (https://issues.apache.org/jira/browse/HADOOP-8468). Following are considerations for HVE:

1. Enable HVE when there is more than one Hadoop VM per physical node in virtualized environments.

2. Use the Node Group definition to group VMs that reside on the same physical node to enable HDFS to distribute block replication across physical nodes.

3. HVE extensions can be used to create node groups to further specify locality and awareness for spreading HDFS block replicas across physical servers of a common model type, with a certain level of power supply redundancy or nodes from certain hardware purchase cycles, for example.

The following diagram illustrates the addition of a new level of abstraction (in red) called Node Groups. The NodeGroups represent the physical hypervisor on which the nodes (VMs) reside.



Rx = Server Rack

NGx = Node Group

Nx = Physical Server Node

Figure 20. HVE Node groups

All VMs under the same node group run on the same physical host. With awareness of the node group layer, HVE refines the following policies for Hadoop on virtualization:

Replica Placement Policy

• No duplicated replicas are on the same node or nodes under the same node group.
• First replica is on the local node or local node group of the writer.
• Second replica is on a remote rack of the first replica.
• Third replica is on the same rack as the second replica.
• The remaining replicas are located randomly across rack and node group for minimum restriction.

Replica Choosing Policy

• The HDFS client obtains a list of replicas for a specific block sorted by distance, from nearest to farthest: local node, local node group, local rack, off rack.

Balancer Policy

• At the node level, the target and source for balancing follows this sequence: local node group, local rack, off rack.

• At the block level, a replica block is not a good candidate for balancing between source and target node if another replica is on the target node or on the same node group of the target node.

HVE typically supports failure and locality topologies defined from the perspective of virtualization. However, you can use the new extensions to support other failure and locality changes, such as those relating to power supplies, arbitrary sets of physical servers, or collections of servers from the same hardware purchase cycle.

7.6.3 Virtualized Configuration Summary

In this reference architecture, the following configuration was used to create the Cloudera software stack running on the ESXi hypervisor. As a starting point and to maintain best locality for HDD, CPU and memory, 2 VMs per physical node is demonstrated in this reference architecture. Additional configurations such as 4VMs or more per node are possible to meet specific customer requirements.

Cluster Hardware Design

Table 8. Cluster Design for Virtualized Cloudera on VMware ESXi

Component	Configuration
System	ThinkSystem SR650
Processor	2 x Intel® Xeon® Scalable processors 6130 Gold 2.1GHz 16-core
Memory - base	512GB: 16x 32GB 2666MHz RDIMM
Disk (OS)	Dual M.2 128GB SSD in a RAID1 mirrored array
Disk (data)	14x 4TB NL SAS 3.5 inch, 4TB each (56 TB Total)
HDD controller	OS: M.2 RAID1 mirroring kit HDFS: N2215 SAS/SATA HBA
Hardware storage protection	OS: RAID1 HDFS: None (JBOD). By default, Cloudera maintains a total of three copies of data stored within the cluster. The copies are distributed across data servers and racks for fault recovery.
Data Network	10Gb Ethernet, 2x bonded interfaces
Data Network adapter	ThinkSystem 10GbE integrated LOM
Hardware management network adapter	Integrated 1GBaseT XClarity™ Controller (XCC) interface

Cluster Software Stack

Table 9. Virtualized Cloudera Software Stack:

Component	Version
vSphere (ESXi + vCenter Server)	6.5.0
Guest Operating System	Red Hat rhel7.3
Cloudera Hadoop Distribution	CDH 5.12
Cloudera Manager	CDH 5.12
Java	Oracle 1.8.0

ESXi Hypervisor and Guest OS Configuration:

Below are the key configuration parameters used in this reference architecture. Many valid configurations exist and additional information can be obtained from the Cloudera and VMware whitepapers shown in the References section on page 50:

ESXi

- 4 VMs per physical node
- 8 physical nodes * 2 = 32 total VMs

Memory

- CPU to Memory locality: 2VMs per CPU
- 6% of node physical Memory allocated for ESXi; remainder Memory allocated to VMs
- Anonymous paging: vm.swappiness=0

Disks

- 3 disks each first 2 VMs; 4 disks each 3rd and 4ths VMs (14 disks total per physical node)
- DataNode: 3 disks each first 2 VMs; 4 disks each 3rd and 4th VMs (14 disks total per physical node)
- VMware PVSCSI storage adapter used (all 4 virtual SCSI controllers used) for best I/O performance
- Queue depth in guest OS SCSI driver: 4294967295 (default value)
- Eager-zeroed thick VMDKs (on EXT4 filesystem in guest OS)

Network

- VMXNET3 network driver used with MTU=90000 for jumbo frames on guest OS and virtual switch
- Enabled TCP segmentation offload (TSO) at the ESXi level (should be enabled by default). Only VMXNET3 drivers at the Guest layer can leverage this.

7.7 Estimating disk space

When you are estimating disk space within a Cloudera Enterprise cluster, consider the following points:

For improved fault tolerance and performance, Cloudera Enterprise replicates data blocks across multiple cluster worker nodes. By default, the file system maintains three replicas.
Compression ratio is an important consideration in estimating disk space and can vary greatly based on file contents. If the customer's data compression ratio is unavailable, assume a compression ratio of 2.5:1.
To ensure efficient file system operation and to allow time to add more storage capacity to the cluster if necessary, reserve 25% of the total capacity of the cluster.

Assuming the default three replicas maintained by Cloudera Enterprise, the raw data disk space and the required number of nodes can be estimated by using the following equations:

Total raw data disk space = (User data, uncompressed) * (4 / compression ratio)

Total required worker nodes = (Total raw data disk space) / (Raw data disk per node)

You should also consider future growth requirements when estimating disk space.

Based on these sizing principals, Table 10 shows an example for a cluster that must store 500 TB of uncompressed user data. The example shows that the Cloudera cluster needs 800 TB of raw disk to support 500 TB of uncompressed data. The 800 TB is for data storage and does not include operating system disk space. A total of 15 nodes are required to support a deployment of this size.

Total raw data disk space = 500TB * (4 / 2.5) = 500 * 1.6 = 800TB

Total required worker nodes = 800TB / (4TB * 14 drives) = 800TB / 56TB = 14.2 => 15 nodes

Table 10. Example of storage sizing with 4TB drives

Description	Value
Data storage size required (uncompressed)	500 TB
Compression ratio	2.5:1
Size of compressed data	200 TB
Storage multiplication factor	4
Raw data disk space needed for Cloudera cluster	800 TB
Storage needed for Cloudera Hadoop 3x replication	600 TB
Reserved storage for headroom (25% of 800TB)	200 TB
Raw data disk per node (with 4TB drives * 14 drives)	56 TB
Minimum number of nodes required (800/56)	15

7.8 Scaling considerations

The Hadoop architecture is linearly scalable but it is important to note that some workloads might not scale completely linearly, so planning ahead for these items will help ease the effort.

When the capacity of the infrastructure is reached, the cluster can be scaled out by adding nodes. Typically, identically configured nodes are best to maintain the same ratio of storage and compute capabilities. A Cloudera cluster is scalable by adding additional SR650 Worker nodes, Master nodes and network switches. As the capacity of a rack is reached, new racks can be added to the cluster.

When a Cloudera reference architecture implementation is designed, future scale out should be a key consideration in the initial design. There are two key aspects to consider: networking and management. These aspects are critical to cluster operation and become more complex as the cluster infrastructure grows.

The cross rack networking configuration that is shown in **Error! Reference source not found.** provides robust network interconnection of racks within the cluster. As racks are added, the predefined networking topology remains balanced and symmetrical. If there are plans to scale the cluster beyond one rack, a best practice is to initially design the cluster with multiple racks (even if the initial number of nodes fit within one rack). Starting with multiple racks can enforce proper network topology and prevent future re-configuration and hardware changes. As racks are added over time, multiple G8332 switches might be required for greater scalability and balanced performance.

Also, as the number of nodes within the cluster increases, so do many of the tasks of managing the cluster, such as updating node firmware or operating systems. Building a cluster management framework as part of the initial design and proactively considering the challenges of managing a large cluster pays off significantly in the long run.

Proactive planning for future scale out and the development of cluster management framework as a part of initial cluster design provides a foundation for future growth that can minimize hardware reconfigurations and cluster management issues as the cluster grows.

7.9 High availability considerations

When a Cloudera cluster on Lenovo servers, is implemented, consider availability requirements as part of the final hardware and software configuration. Typically, Hadoop is considered a *highly reliable* solution. Hadoop, Cloudera and Lenovo best practices provide significant protection against data loss. Generally, failures can be managed without causing an outage. There is redundancy that can be added to make a cluster even more reliable. Some consideration must be given to hardware and software redundancy.

7.9.1 Networking considerations

The second redundant management network switch can be added to ensure HA of the hardware management network. The hardware management network does not affect the availability of the Cloudera Hadoop file system functionality, but it might affect the management of the cluster; therefore, availability requirements must be considered.

To support HA in the network, link aggregation is used between the 10Gb ports of a server network adapter (Bonded interfaces) and the top-of-rack switch. Virtual Link Aggregation Groups (vLAG) is configured between the two switches. This way, a single NIC, network cable or switch can fail and that network connection will continue with the remaining half of the network connection.

7.9.2 Hardware availability considerations

The redundancy of each individual worker node is not necessary with Hadoop. HDFS default 3x replication provides built-in redundancy and makes loss of data unlikely. If Hadoop best practices are used, an outage from a worker node loss is extremely unlikely as the workload can be dynamically re-allocated. The loss of a worker node will not cause a job to fail; workload is automatically re-allocated to another data note.

Multiple Master nodes are recommended so that if there is a failure, function can be moved to an operational

Master node. Having multiple Master nodes does not automatically resolve the issue of the NameNode being a single point of failure. For more information, see “Software availability considerations.”

Within racks, switches and nodes must have redundant power feeds with each power feed connected from a separate PDU.

7.9.3 Storage availability

HDFS 3x replication provides more than sufficient protection. Higher levels of replication can be considered if needed.

Cloudera also provides manual or scheduled snapshots of volumes to protect against human error and programming defects. Snapshots are useful for rollback to a known data set.

7.9.4 Software availability considerations

Operating system availability is provided by using mirrored drives for the operating system.

NameNode HA is recommended and can be achieved by using three master nodes. Active and standby nodes communicate with a group of separate daemons called JournalNodes to keep their state synchronized. When any namespace modification is performed by the active NameNode, it durably logs a record of the modification to most of these JournalNodes. The standby NameNode can read the edits from the JournalNodes and is constantly watching them for changes to the edit log. As the standby Node sees the edits, it applies them to its own namespace.

An external database is required for Cloudera Manager, Hive metastore and so on, and HA configuration of external database is recommended to avoid single point of failure. Embedded databases should only be used for test or POC environment.

7.10 Migration considerations

If migrating data or applications to Cloudera is required, you must consider the type and amount of data to be migrated. Most data types can be migrated, but you must understand migration requirements to verify viability. Cloudera Enterprise provides tools to move data between external SQL databases and Hadoop.

Other considerations should be given to whether applications must be modified to use Hadoop functionality. Significant effort might be required in some cases.

8 Appendix: Bill of Materials

This appendix includes the Bill of Materials (BOMs) for different configurations of hardware for the Big Data Solution from Cloudera deployments. There are sections for Master nodes, worker nodes and networking.

The BOM includes the part numbers, component descriptions and quantities. Table 5 lists how many core components are required for each of the predefined configuration sizes.

The BOM lists in this appendix are not meant to be exhaustive and must always be verified with the configuration tools. Any discussion of pricing, support and maintenance options is outside the scope of this document.

This BOM information is for the United States; part numbers and descriptions can vary in other countries. Other sample configurations are available from your Lenovo sales team. Components are subject to change without notice.

8.1 Master node

Table 11 lists the BOM for the Master node.

Table 11. Master node

Code	Description	Qty
7X01CTO1WW	-SB- ThinkSystem SR630 - 1yr Warranty	1
6570	2.0m, 13A/125-10A/250V, C13 to IEC 320-C14 Rack Power Cable	2
A2K7	Primary Array - RAID 1	1
AUM7	ThinkSystem 2.5" 2TB 7.2K SAS 12Gb Hot Swap 512n HDD	8
AVWA	ThinkSystem 750W(230/115V) Platinum Hot-Swap Power Supply	2
AUW9	ThinkSystem SR630/SR570 2.5" AnyBay 10-Bay Backplane	1
AUWC	ThinkSystem SR530/SR570/SR630 x8/x16 PCIe LP+LP Riser 1 Kit	1
5978	Select Storage devices - configured RAID	1
B0MK	Enable TPM 2.0	1
AXCA	ThinkSystem Toolless Slide Rail	1
AUKK	ThinkSystem 10Gb 4-port SFP+ LOM	1
AUPW	ThinkSystem XClarity™ Controller Standard to Enterprise Upgrade	1
AUNK	ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter	1
AWER	Intel® Xeon® Silver 4116 12C 85W 2.1GHz Processor	2
AUWQ	Lenovo ThinkSystem 1U LP+LP BF Riser BKT	1
AUNB	ThinkSystem 16GB TruDDR4 2666 MHz (1Rx4 1.2V) RDIMM	8
AUW1	ThinkSystem SR630 2.5" Chassis with 10 bays	1
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
AUUV	ThinkSystem M.2 CV3 128GB SATA 6Gbps Non-Hot Swap SSD	2
A2KL	Secondary Array - RAID 10	1
AUWW	-SB- Front VGA Cable for 1U 2.5"	1
2305	Integration 1U component	1
AURN	Lenovo ThinkSystem Super Cap Box	1
AULP	ThinkSystem 1U CPU Heatsink	2

AVWJ	ThinkSystem 750W Platinum RDN PSU Caution Label	1
AUWL	Lenovo ThinkSystem 1U LP Riser Dummy	1
AUW7	Lenovo ThinkSystem 1U Cable 4056 Fan module	2
AVWK	ThinkSystem EIA plate with Lenovo logo	1
AWF9	ThinkSystem Response time Service Label LI	1
AUX4	MS 1U Service label LI	1
AUX3	ThinkSystem SR630 Model Number Label	1
AUWV	10x2.5"Cable Kit (1U)	1
AVKG	ThinkSystem SR630 MB to 10x2.5" HDD BP NVME cable	1
AV00	Super Cap Cable-680mm	1
AWGE	ThinkSystem SR630 WW Lenovo LPK	1
AUW3	Lenovo ThinkSystem Mainstream MB - 1U	1
B0ML	Feature Enable TPM on MB	1
B173	XClarity™ Controller Standard to Enterprise Upgrade in factory	1

8.2 Worker node

Table 12 lists the BOM for the Worker node.

Table 12. Worker node

Code	Description	Qty
7X05CTO1WW	-SB- ThinkSystem SR650 - 1yr Warranty	1
6570	2.0m, 13A/125-10A/250V, C13 to IEC 320-C14 Rack Power Cable	2
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
AUR9	ThinkSystem SR650/SR550/SR590 3.5" SATA/SAS 12-Bay Backplane	1
AURC	ThinkSystem SR550/SR590/SR650 x16/x8(or x16) PCIe FH Riser 2 Kit	1
5977	Select Storage devices - no configured RAID required	1
B0MK	Enable TPM 2.0	1
AXCA	ThinkSystem Toolless Slide Rail	1
AUKK	ThinkSystem 10Gb 4-port SFP+ LOM	1
AUPW	ThinkSystem XClarity™ Controller Standard to Enterprise Upgrade	1
AUNM	ThinkSystem 430-16i SAS/SATA 12Gb HBA	1
AWEN	Intel® Xeon® Gold 6130 16C 125W 2.1GHz Processor	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	8
A484	Populate Rear Drives	1
AURZ	ThinkSystem SR590/SR650 Rear HDD Kit	1
AUVW	ThinkSystem SR650 3.5" Chassis with 8 or 12 bays	1
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
AUUV	ThinkSystem M.2 CV3 128GB SATA 6Gbps Non-Hot Swap SSD	2
AUU6	ThinkSystem 3.5" 4TB 7.2K SAS 12Gb Hot Swap 512n HDD	14
AUS8	ThinkSystem SR550/SR590/SR650 EIA Latch w/ VGA Upgrade Kit	1

2306	Integration >1U Component	1
AURS	Lenovo ThinkSystem Memory Dummy	16
AURP	Lenovo ThinkSystem 2U 2FH Riser BKT	1
AVWK	ThinkSystem EIA plate with Lenovo logo	1
AWF9	ThinkSystem Response time Service Label LI	1
AWFF	ThinkSystem SR650 WW Lenovo LPK	1
AURM	ThinkSystem SR550/SR650/SR590 Right EIA Latch with FIO	1
B0ML	Feature Enable TPM on MB	1
B173	XClarity™ Controller Standard to Enterprise Upgrade in factory	1

8.3 Systems Management Node

Table 13 lists the BOM for the Systems Management Node.

Table 13. Systems Management Node

Code	Description	Qty
7X01CTO1WW	-SB- ThinkSystem SR630 - 1yr Warranty	1
6570	2.0m, 13A/125-10A/250V, C13 to IEC 320-C14 Rack Power Cable	2
AVWA	ThinkSystem 750W(230/115V) Platinum Hot-Swap Power Supply	2
AUWB	ThinkSystem SR530/SR630/SR570 2.5" SATA/SAS 8-Bay Backplane	1
AUWC	ThinkSystem SR530/SR570/SR630 x8/x16 PCIe LP+LP Riser 1 Kit	1
5977	Select Storage devices - no configured RAID required	1
B0MK	Enable TPM 2.0	1
AXCA	ThinkSystem Toolless Slide Rail	1
AUKK	ThinkSystem 10Gb 4-port SFP+ LOM	1
AUPW	ThinkSystem XClarity™ Controller Standard to Enterprise Upgrade	1
AUNG	ThinkSystem RAID 530-8i PCIe 12Gb Adapter	1
AWEH	Intel® Xeon® Bronze 3106 8C 85W 1.7GHz Processor	1
AUWQ	Lenovo ThinkSystem 1U LP+LP BF Riser BKT	1
AUNB	ThinkSystem 16GB TruDDR4 2666 MHz (1Rx4 1.2V) RDIMM	1
AUW0	ThinkSystem SR630 2.5" Chassis with 8 bays	1
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
AUUV	ThinkSystem M.2 CV3 128GB SATA 6Gbps Non-Hot Swap SSD	2
AUWW	-SB- Front VGA Cable for 1U 2.5"	1
2305	Integration 1U component	1
AUS6	Lenovo ThinkSystem 1U height CPU HS Dummy	1
AULP	ThinkSystem 1U CPU Heatsink	1
AVWJ	ThinkSystem 750W Platinum RDN PSU Caution Label	1
AUWF	Lenovo ThinkSystem Super Cap Holder Dummy	1
AVKJ	ThinkSystem 2x2 Quad Bay Gen4 2.5" HDD Filler	1
AUWK	Lenovo ThinkSystem 4056 Fan Dummy	1
AUWL	Lenovo ThinkSystem 1U LP Riser Dummy	1

AVWK	ThinkSystem EIA plate with Lenovo logo	1
AWF9	ThinkSystem Response time Service Label LI	1
AUX4	MS 1U Service label LI	1
AUX3	ThinkSystem SR630 Model Number Label	1
AUWX	8x2.5" HDD BP Cable Kit	1
AWGE	ThinkSystem SR630 WW Lenovo LPK	1
AUW3	Lenovo ThinkSystem Mainstream MB - 1U	1
B0ML	Feature Enable TPM on MB	1
B173	XClarity™ Controller Standard to Enterprise Upgrade in factory	1

8.4 Management network switch

Table 14 lists the BOM for the Management/Administration network switch.

Table 14. Management/Administration network switch

Code	Description	Qty
7159HC1	Lenovo RackSwitch G8052 (Rear to Front)	1
ASY2	Lenovo RackSwitch G8052 (Rear to Front)	1
A3KR	Air Inlet Duct for 442 mm RackSwitch	1
A3KP	Adjustable 19" 4 Post Rail Kit	1
6201	1.5m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
2305	Integration 1U component	1

8.5 Data network switch

Table 15 lists the BOM for the data network switch.

Table 15. Data network switch

Code	Description	Qty
7159HCW	Lenovo RackSwitch G8272 (Rear to Front)	2
ASRD	Lenovo RackSwitch G8272 (Rear to Front)	2
ASTN	Air Inlet Duct for 487 mm RackSwitch	2
6201	1.5m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	4
A3KP	Adjustable 19" 4 Post Rail Kit	2
2305	Integration 1U component	2
3792	1.5m Yellow Cat5e Cable	2

8.6 Rack

Table 16 lists the BOM for the rack.

Table 16. Rack

Code	Description	Qty
9363RC	-SB- 42U 1100mm Enterprise V2 Dynamic Rack	1

4		
A1RC	-SB- 42U 1100mm Enterprise V2 Dynamic Rack	1
5895	1U 12 C13 Switched and Monitored 60A 3 Phase PDU	4
2304	Integration Prep	1
AU8J	Integrated Rack Miscellaneous Parts Kit	1
AU8K	LeROM Validation	1
91Y9793	Foundation Service - 5Yr Next Business Day Response	1
4271	1U black plastic filler panel	2
4275	5U black plastic filler panel	3

Different cluster sizing leaves different unused rack space; therefore, consider the use of blank plastic filter panels for the rack to better direct cool air flow.

The number of PDUs in the rack depends on the server numbers in the rack. Four PDU should be used for the half rack configuration and six PDUs for a full rack.

8.7 Cables

Table 17 lists the BOM for the cables, for each node.

Table 17. Cables

Code	Description	Qty
AT2S	-SB- Lenovo 3m Active DAC SFP+ Cables	2
A3RG	0.5m Passive DAC SFP+ Cable	2
A51N	1.5m Passive DAC SFP+ Cable	13
3792	1.5m Yellow Cat5e Cable	4
A51P	2m Passive DAC SFP+ Cable	11
3793	3m Yellow Cat5e Cable	9
AT2S	-SB- Lenovo 3m Active DAC SFP+ Cables	2
A3RG	0.5m Passive DAC SFP+ Cable	2
A51N	1.5m Passive DAC SFP+ Cable	13
3792	1.5m Yellow Cat5e Cable	4

9 Acknowledgements

This reference architecture document has benefited very much from the detailed and careful review comments provided by colleagues at Lenovo and Cloudera.

Lenovo business review

- Prasad Venkatachar – Sr. Solutions Product Manager

Cloudera technical review

- Alex Moundalexis – Software Engineer - Partner Engineering
- Calvin Goodrich – Engineering Manager - Partner Engineering

Cloudera business review

- Sean Gilbert – Director - Business Development/ Global Partner Sales

VMware technical review

- Technical staff members at VMware

Resources

For more information, see the following resources:

Lenovo ThinkSystem SR650 (Cloudera Worker node):

- Lenovo Press product guide: https://lenovopress.com/lp0644.pdf
- 3D Tour: https://lenovopress.com/lp0673-3d-tour-thinksystem-sr650

Lenovo ThinkSystem SR630 (Cloudera Master node):

- Lenovo Press product guide: https://lenovopress.com/lp0643-lenovo-thinksystem-sr630-server
- 3D Tour: https://lenovopress.com/lp0672-3d-tour-thinksystem-sr630

Lenovo RackSwitch G8052 (1GbE Switch):

- Product page: https://lenovopress.com/tips1270-lenovo-rackswitch-g8052
- Lenovo Press product guide: https://lenovopress.com/tips1270.pdf

Lenovo RackSwitch G8272 (10GbE Switch):

- Product page: https://lenovopress.com/tips1267-lenovo-rackswitch-g8272
- Lenovo Press product guide: https://lenovopress.com/tips1267.pdf

Lenovo ThinkSystem NE10032 (40GbE/100GbE Switch):

- Product page: https://lenovopress.com/lp0609-lenovo-thinksystem-ne10032-rackswitch
- Lenovo Press product guide: https://lenovopress.com/lp0609.pdf

Intel Xeon Scalable Family Balanced Memory

- https://lenovopress.com/lp0742-intel-xeon-scalable-family-balanced-memory-configurations

Lenovo XClarity Administrator:

- Product page: https://lenovopress.com/tips1200-lenovo-xclarity-administrator
- Lenovo Press product guide: https://lenovopress.com/tips1200.pdf

Cloudera:

- Cloudera Distribution for Hadoop (CDH): cloudera.com/content/cloudera/en/products-and-services/cdh.html
- Cloudera products and services: https://www.cloudera.com/products.html
- Cloudera solutions: http://www.cloudera.com/content/cloudera/en/solutions.html
- Cloudera resources: https://www.cloudera.com/resources.html
- Cloudera RA with VMware and local attached storage: cloudera.com/documentation/other/reference-architecture/PDF/cloudera_ref_arch_vmware_local_storage.pdf

VMware:

- VMware Hadoop Deployment Guide
- Big Data Performance on vSphere
- Virtaulized Hadoop Performance with VMware vSphere 6.0 on High-Performance Servers

Open source software:

- Hadoop: hadoop.apache.org
- Spark: spark.apache.org
- Flume: flume.apache.org
- HBase: hbase.apache.org
- Hive: hive.apache.org
- Hue: gethue.com
- Impala: rideimpala.com
- Oozie: oozie.apache.org
- Mahout: mahout.apache.org
- Pig: pig.apache.org
- Sentry: entry.incubator.apache.org
- Sqoop: sqoop.apache.org
- Whirr: whirr.apache.org
- ZooKeeper: zookeeper.apache.org
- Parquet: parquet.apache.org
  Hadoop Virtualization Extensions (HVE): https://issues.apache.org/jira/browse/HADOOP-8468
- xCat: xcat.org

Document history

Version 1.0	12 Oct 2017	First version
Version 1.1	28 Nov 2017	Updated storage charts to show Cloudera recommendation of 100 TB per node max.
Version 1.2	15 Dec 2017	Updates to networking sections

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkCenter, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.