Lenovo ThinkServer and Cloudera Solution for Apache Hadoop

For next-generation Lenovo ThinkServer systems

Lenovo Enterprise Product Group

Version 1.0

December 2014

The Opportunity of Big Data Analytics

According to IDC, by 2020 it is projected that over 40ZB of data will be generated by the over 50 billion different devices ranging from smart phones and mobile devices to sensors used in automotive, smart city and industrial applications. This plethora of information flowing into today's data centers creates a massive opportunity for businesses to extract new insights about their customers.

For the last several years, the adoption of big data has steadily increased as large organizations see the value of investing in data analytics. The up-front capital investment of compute and networking clusters provides quick financial returns, as these organizations are able to leverage this IT investment to reduce escalating operational expenses and gain information-driven insights.

> The explosion of data presents an opportunity for businesses of all sizes to extract new insights about their customers.
>
> Cloudera helps enterprises see the value of their data in real time.
>
> Lenovo and Cloudera provide cost- and performance-optimized solutions that businesses can adopt with confidence.

Examples of implementing big data platforms to reduce costs and gain competitive advantage are everywhere. Financial firms use data analytics for fraud modeling to detect fraudulent activities in real time across millions of transactions and disparate systems, saving millions. Healthcare providers can enter patient history and symptomology and sift through millions of archived patient records for relevant outcomes — all while the patient is still in the office.

Many organizations even go one-step further by coupling big data with high-performance computing (HPC) clusters to create a foundation for high-performance data analytics (HPDA). In fact, IDC predicts that by 2015, 7.3 percent of all HPC installations will be focused on HPDA. Insurance companies use this approach to provide faster means of providing calling/online quotes down to 100ms, while oil and gas firms use similar clusters for analytics and visualizing of field data for internal and external use.

However, big data analytics is not just for the big guys. The truth is many small- to medium-sized businesses (SMBs) are looking at big data — as a necessity rather than an option — as their ability to stay competitive relies heavily on their use of the data they capture. The greater the data volume collected, the greater the need for an organization to manage and capitalize on its storage and compute capacity for accurate and timely decision-making. SMBs focused on logistics or shipping can use big data solutions to reroute trucks to create efficient routes, alert customers to deliveries and forecast and price services. Boutiques or restaurants can use data analytics to manage leads or inputs through social media for more efficient direct marketing, while smaller healthcare providers can leverage similar solutions to keep in accord with federal

guidelines, or to automate patient communications for proactive scheduling and improved relationship management.

A Solution Based on Cloudera's Enterprise Data Hub

Tapping into the potential from this wave of big data can be a complex task and has driven the need for new tools and frameworks for data processing. Hadoop has been at the forefront of these technologies helping to address this growth in data and emergence of new use cases. Driving Hadoop's adoption for these new workloads is its scalability and flexibility to handle data of various formats, generated by new devices, sensors, and systems.

Embarking on the journey of enabling Hadoop adoption for enterprise big data analysis can be challenging. It is a task that requires the right combination of technology partners to be successful. Cloudera, the market leader in enterprise analytics data management, has extensive experience enabling successful customer adoptions by providing best-in-class technology, support, and training to their customers. Cloudera Enterprise, Cloudera's enterprise data hub, provides a single location to store and process all data and all types of data with a variety of enterprise workloads. Built with Apache Hadoop™ at its core, it is a new, more powerful and scalable data platform. Furthermore, the inclusion of robust security, governance, and management capabilities allows businesses to adopt the solution with confidence, knowing that as their data management needs grow, their data platform will grow with them.

Big Data and Lenovo ThinkServer

One of Hadoop's key tenets is its ability to scale linearly as data storage and processing needs increase over time. With a mix of performance, flexibility and scalability, the new Lenovo® ThinkServer® RD550 and Lenovo ThinkServer RD650 servers provide affordable yet powerful solution building blocks to tackle today's big data needs. These servers are powered by the Intel® Xeon® E5 v3 Series family of processors and provide the computational horsepower your big data infrastructure needs. The RD550 and RD650 systems provide industry-leading storage that is ideal for big data analytics projects. The RD650 can support up to 74.4TB of internal storage capacity. The RD550 and RD650 also feature extreme flexibility in networking and I/O configurations. The RD650 has up to eight PCIe slots with plenty of bandwidth to scale I/O, while being cost- and performance- optimized. Designed for a wide variety of environments, the ThinkServer systems can run continuously at 45 degrees Celsius—with no impact on reliability.

Architecture Overview

The Lenovo and Cloudera Hadoop Solution provides a highly TCO optimized solution for large-scale big data projects, but is tailored for workgroups and departmental users. It consists of optimized Lenovo ThinkServer RD550 and RD650 servers using Cloudera Enterprise software as shown in Figure 1.

Software includes the Cloudera Distribution for Apache Hadoop (including Cloudera Manager, Impala, HBase, Cloudera Search, Spark, Cloudera Navigator), and Hadoop Ecosystem

Components (including Hadoop, Flume, HCatalog, Hive, Hue, Mahout, Oozie, Pig, Sentry, Sqoop, Whirr, ZooKeeper, Cloudera Back-up and Disaster Recovery).

The Cloudera software is hosted on Red Hat Enterprise Linux 6.5.

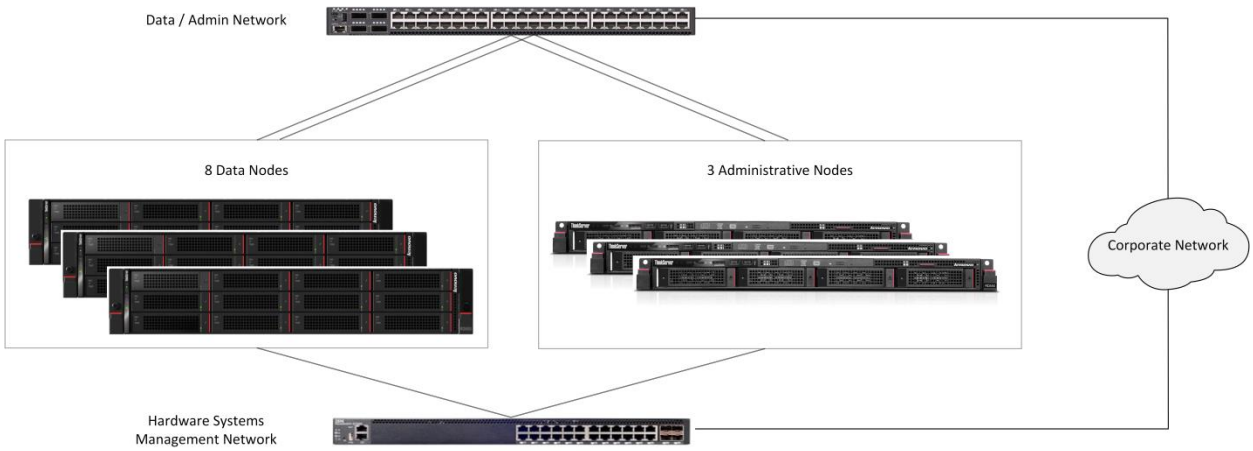

Figure 1 – Cloudera Solution Architecture

Cloudera Certified Solution

A configuration for the Cloudera solution includes server cluster nodes, networking, power, and a rack. This section describes the hardware included in the Cloudera certified solution.

ThinkServer Systems

The heart of the Cloudera solution is a set of server nodes in a clustered configuration. Two types of nodes are used – data nodes that provide data services for storing and processing data, and administrative nodes that manage and coordinate the workflow in the cluster. Management tasks are placed on dedicated administration nodes to optimize performance of the cluster.

For the data nodes, the certified configuration includes eight ThinkServer RD650 dual-CPU servers with locally attached storage. External storage is not used in the solution. The configuration for each of the ThinkServer RD650 system contains:

- Dual Intel Xeon E5-2640 v3 processors
- 128GB of DDR4 RAM
- Twelve 4 TB 3.5” SATA HDDs
- Dual 10 Gb Ethernet interfaces

The memory capacity is chosen to support HBase, Spark, and memory-intensive MapReduce workloads. HDDs are configured as just a bunch of disks (JBOD), rather than in a RAID configuration. This is the best choice for a Cloudera cluster. It provides excellent performance and, when combined with the Hadoop default of three times data replication, it 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 (it is acceptable to configure the operating system volume in a RAID mirror). In addition, SATA drives offer the same performance for less cost than SAS HDDs. Dual network connections are included for network bandwidth and load balancing.

For the administration nodes, the certified configuration includes three ThinkServer RD550 systems. The administrative nodes are the nucleus of the Hadoop Distributed File System (HDFS) and support several other key functions needed on a Cloudera cluster. Because the management node is responsible for many memory-intensive tasks, multiple management nodes are needed to split functions. The configuration for each of the RD550 servers contains:

- Dual Intel Xeon E5-2640 v3 processors
- 128GB of DDR4 RAM
- Four 2 TB 3.5” SATA HDDs
- Dual 10 Gb Ethernet interfaces

Ordering information for these servers is provided in Table 1, and Table 2.

Table 1 – Data Node Server Configuration

Part Number	Description	Quantity
70D00029UX	RD650 (2U rack server with 12x 3.5-inch hot-swap HDD bays) - 2x Intel Xeon processor E5-2640 v3 (8-cores, 20MB cache, 2.6GHz, 8.0GT) - 16x 8GB DDR4-2133MHz (1Rx4) RDIMM - ThinkServer RAID 720ix with expander and 1 GB cache - Lenovo ThinkServer X540-T2 AnyFabric 10Gb 2 Port Base-T Ethernet Adapter by Intel - 12x HS 4 TB SATA - Enterprise 7200 rpm, 6Gb/s 3.5-inch HDD - ThinkServer Management Module Premium - 2x 1100W hot-swap redundant power supply - ThinkServer tool-less rail kit - Next Business Day On-site Warranty, 3 Years Parts and Labor	8

Table 2 – Administration Node Server Configuration

Part Number	Description	Quantity
70CV001RUX	RD550 (1U Rack with 4 x 3.5-inch hot swap HDD bays) - 2x Intel Xeon processor E5-2640 v3 (8-cores, 20MB cache, 2.6GHz, 8.0GT) - 16x 8GB DDR4-2133MHz (1Rx4) RDIMM - ThinkServer RAID 510i AnyRAID adapter - Lenovo ThinkServer X540-T2 AnyFabric 10Gb 2 Port Base-T Ethernet Adapter by Intel - 4x HS 2 TB SATA - Enterprise 7200 rpm, 6Gb/s 3.5-inch HDD - ThinkServer Management Module Premium - 2x 750W hot-swap redundant power supply - ThinkServer tool-less rail kit - Next Business Day On-site Warranty, 3 Years Parts and Labor	3

Networking

Two networks are used in the Cloudera solution. The first network supports data and administrative nodes in the cluster. The second network is used to isolate hardware systems management tasks.

The data and administrative network connects all nodes in the cluster and is used for data access and moving data across nodes in the cluster. This network is typically connected to the customer's corporate network. A 10GbE switch is selected to insure adequate bandwidth is available for added performance. A single switch is required; however, a second switch can be added for network redundancy.

For the 10GbE switch, the certified configuration includes the Lenovo System Networking RackSwitch™ G8264T. The enterprise-level RackSwitch G8264T has the following characteristics:

- Low cost RJ45 cables/connections
- Forty-eight 10GBase-T ports
- Four QSFP+ 40GbE ports
- HS redundant fans and power

The hardware systems management network is a 1GbE network used out-of-band hardware management. Out-of-band management of the servers in the cluster is provided by the ThinkServer System Manager, which allows server configuration, monitoring of server health and status. A 1GbE switch is selected for the hardware system management network. Again, a single switch is required; however, a second switch can be added for network redundancy. On each of the server nodes, the systems management link is connected to the dedicated management port for the ThinkServer System Manager.

For the 1GbE switch, the certified configuration includes the Lenovo System Networking RackSwitch™ G7028. The enterprise-level RackSwitch G7028 has the following characteristics:

- 24 ports 1G, RJ-45
- 4 ports 10G, SFP+ uplinks standard

Ordering information for these switches is provided in Table 3.

Table 3 – Network Switch Configurations

Part Number	Description	Quantity
7309CR9	Lenovo System Networking RackSwitch G8264T	1
7309BAX	Lenovo System Networking RackSwitch G7028	1

Additional Recommended Options

To complete the Cloudera solution, additional components should be considered, including racks, power distribution and backup, and console managers. Table 4 provides ordering information for these items.

Table 4 – Recommended Optional Equipment

Part Number	Description	Quantity
9307-RC2 1042	25U S2 Dynamic Rack Cabinet	1
9307-RC2 6012	Enterprise C13 PDU with NA Line Cord	1
39Y8951	DPI Universal Rack PDU w/ US LV and HV line cords	2
55945KX	RT5kVA 3U Rack or Tower UPS (200V-240VAC)	1
5594-RU6 6500	North America Line Cord	1
1754-HC3 0725	Local 1X8 Console Manager (LCM8)	1
1754-HC3 3756	USB Four Pack of USB KVM Cables	3
1723-HC1 A3EK	1U 18.5-inch Standard Console	1
1723-HC1 A50G	Keyboard w/ Int. Pointing Device USB (US Eng)	1

ThinkServer and Cloudera

The Lenovo and Cloudera big data solution is a cost- and performance-optimized solution for big data analytics projects. The ThinkServer RD550 and RD650 systems are an excellent choice for the compute- and storage-intensive Hadoop application workloads. Lenovo enterprise switches provide world-class performance, power-efficient designs, and extensive standard features at an affordable price.

Cloudera Enterprise provides the speed, scale, and centralized management you need to build an enterprise data hub, while enabling you to operate Hadoop as the core of your big data infrastructure.

Together, Lenovo and Cloudera provide a solution that can scale to service large enterprise deployments, but can also be implemented by smaller businesses that are starting, or even investigating the benefits of a big data deployment.

For more information on Lenovo Enterprise servers, networking, storage products, and solutions, visit http://www.lenovo.com/servers.

For more information on Cloudera Enterprise Data Hub Edition and Hadoop distribution, visit www.cloudera.com/content/cloudera/en/home.html.