ARCHITECTURE BRIEF

Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop

Easy to implement hardware, software and services for big data analytics architecture

HIGHLIGHTS

- Provides guidance for deploying Lenovo systems with a Cloudera Distribution for Hadoop to coordinate the processing of the data across a massively parallel environment.
- Includes the latest data center equipment available such as the Lenovo x3650 M5 and x3550 M5 and Lenovo RackSwitch Ethernet switches and Lenovo XClarity.
- Supports entry through high-end configurations and the ability to easily scale as the use of big data grows

The Big Data Challenge

Big data is more than a challenge. It is an opportunity to find new insights in data to make your business more agile and to answer questions that were previously beyond reach. To open the door to a world of possibilities, Cloudera uses the latest big data technologies with faster in-memory Apache Spark and the massive map-reduce scale-out capabilities of Hadoop MapReduce.

This Lenovo Big Data RA for Cloudera Distribution for Hadoop is certified by Cloudera and provides a thoroughly tested and integrated solution that combines the benefits of leading-edge technologies with mature, enterprise-ready features. Starting with a preconfigured hardware platform that is Cloudera-certified helps your team to be up and running analytics applications quickly.

Cloudera allows organizations to run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure can be leveraged to tackle very large data sets by breaking up the data into “chunks” and coordinating the processing of the data across a massively parallel environment.

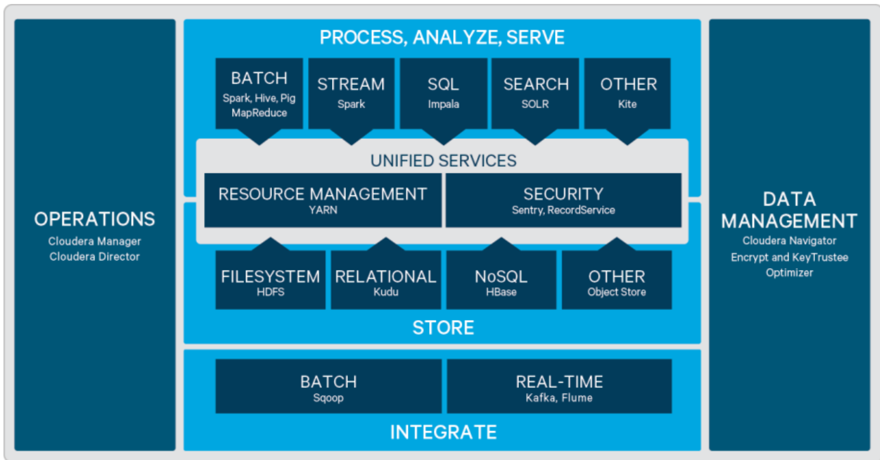
Cloudera deployed on this Lenovo reference architecture provides superior performance, reliability, and scalability. This architecture supports entry through high-end configurations and the ability to easily scale as the use of big data grows. A choice of infrastructure components provides the flexibility to meet a broad range of big data analytics requirements.

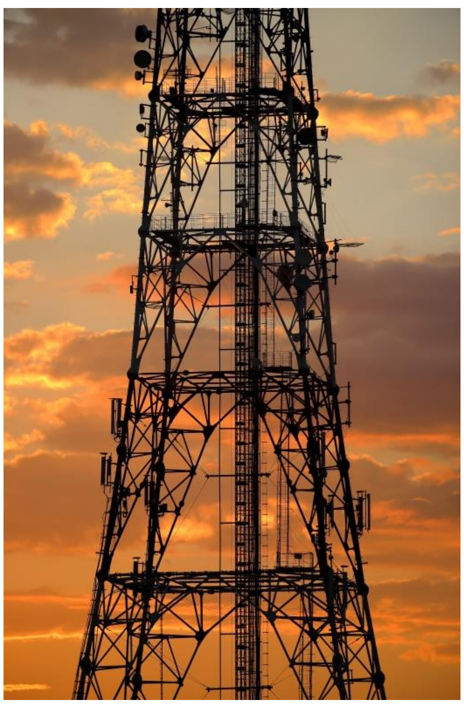
Why Lenovo and Cloudera

Lenovo and Cloudera have collaborated to deliver successful, integrated solutions to customers worldwide. These customers recognize Lenovo's leadership in the worldwide server market and Cloudera's leadership in Apache Hadoop as one of the largest contributors to this open source ecosystem.

Spark Support

Spark is now a fully integrated component of the Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop. An integrated part of CDH and supported with Cloudera Enterprise, Spark is the open standard for flexible in-memory data processing that enables batch, real-time, and advanced analytics on the Apache Hadoop platform. Take advantage of Spark's distributed in-memory storage for high performance processing across a variety of use cases, including batch processing, real-time streaming, and advanced modeling and analytics. With significant performance improvements over MapReduce, Spark is the tool of choice for data scientists and analysts to turn their data into real results.





Configuration Details

The Lenovo M5 servers, like the powerful two-socket x3650 M5 and x3550 M5, enhance performance and reduce power consumption of big data clusters. Purpose built for big data workloads, the 2U two-socket x3650 M5 server supports industry leading data storage capacity, the latest Intel Xeon E5-2600 v4 high performance compute processing power, flash storage options, and energy efficient features. The core Cloudera reference architecture leverages this server as a data node for scale-out clusters.

The versatile, two-socket 1U x3550 M5 rack server delivers outstanding performance as a big data master node in the Cloudera reference architecture. This compact, easy-to-use server features a pay-as-you-grow design to help lower costs and manage risks.

The data network is a private cluster data interconnect among nodes used for data access, moving data across nodes within a cluster, and ingesting data into the Cloudera cluster. While a 1 Gb Ethernet switch is sufficient for some workloads, a 10 Gb Ethernet switch can provide extra I/O bandwidth for added performance. The recommended 10GbE switch is the Lenovo RackSwitch G8272.

Regarding Storage, each server node in the reference architecture has an internal, directly attached storage. External storage is not utilized in this reference architecture. In situations where higher storage capacity is required, the design approach followed in this reference architecture is to increase the amount of data disk space per node or increase the number of nodes in the cluster.

For system management, this reference architecture includes one platform for hardware management and another for cluster system management. The mechanism for cluster systems management is done via Cloudera Manager and adapted from the standard Hadoop distribution which places the management services on separate servers. Because the Master Node hosts important and high-memory functions, it is important that it is a powerful and fast server.

Hardware management is addressed with the Lenovo XClarity Administrator. Lenovo XClarity is an agentless centralized resource management solution that that is aimed at reducing complexity, speeding response, and enhancing the availability of Lenovo server systems. The solution seamlessly integrates into Lenovo M5 rack servers. Through an uncluttered, dashboard- driven GUI, XClarity provides automated discovery, monitoring, firmware updates, pattern-based configuration management, hypervisor operating system deployments. Lenovo XClarity also features extensive REST APIs that provide deep visibility and control via higher-level cloud orchestration and service management software tools.

The example below shows the advanced configuration for this reference architecture. This architecture also offers Enhanced and Advanced configurations. Configurations may also be customized to best suit the workloads running in your environment. To accelerate time to value, Lenovo has service offerings and expertise to implement this reference architecture and to accommodate customization.

Assemblies and Components

The following is an abbreviated Bill of Materials for the reference architecture sample configuration:

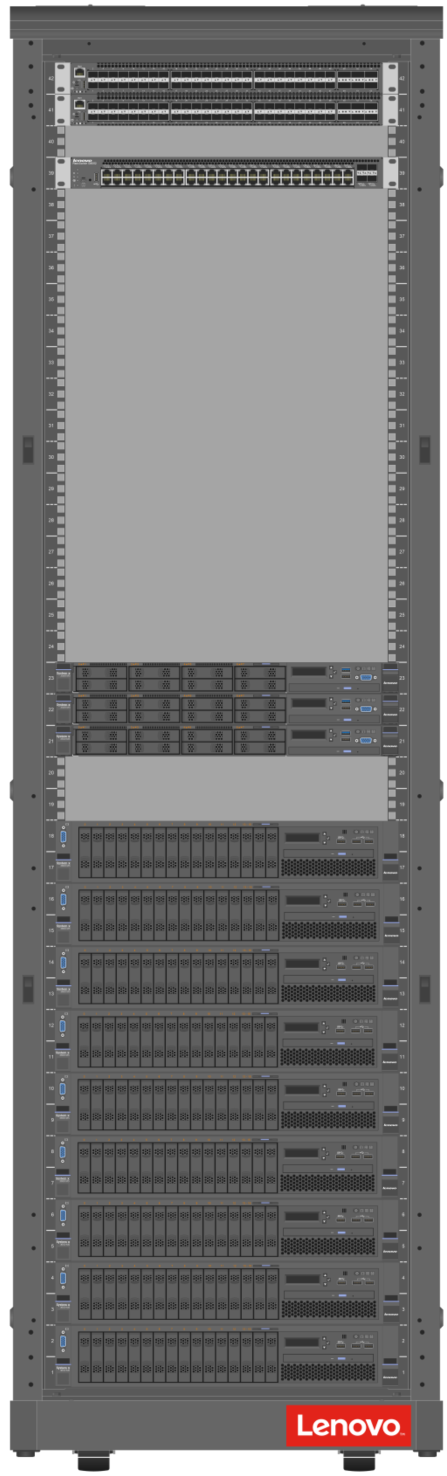
- 2 x Data Networking Switches
  - 1 x 10GbE Lenovo RackSwitch G8272 Network Switch

- 1 x Management Networking Switch
  - 1 x 1GbE Lenovo RackSwitch G8052 Networking Switch

- 3 x Master Node Servers
  - 1 x Lenovo System x3550 M5 Rack Server
  - 2 x E5-2650 v4 (12C, 2.20 GHz)
  - 8 x 16 GB of system memory
  - 2 x 2.5 inch HDD or SSD
  - 18x 2 TB SAS 2.5 inch

- 9 x Data Node Servers
  - 1 x Lenovo System x3650 M5 Rack Servers
  - 2 x E5-2680 v4 (14C, 2.40 GHz)
  - 16 x 32 GB of system memory
  - 2 x 2.5 inch HDD or SSD, OS Disks
  - 14 x 4 TB NL SATA 3.5 inch, Data Disks

The Lenovo x3650 M5 data nodes and x3550 M5 master nodes form the backbone of the Lenovo Big Data Reference Architecture for Cloudera CDH



Architecture Brief

Lenovo Big Data Reference Architecture for Cloudera Distribution for Hadoop

Big Data solutions with the Lenovo x3650 M5 server and Cloudera CDH

Why Lenovo

Lenovo is a leading provider of x86 servers for the data center. Featuring rack, tower, blade, dense and converged systems, the Lenovo server portfolio provides excellent performance, reliability and security. Lenovo also offers a full range of networking, storage, software, solutions, and comprehensive services supporting business needs throughout the IT lifecycle. With options for planning, deployment, and support, Lenovo offers expertise and services needed to deliver better service-level agreements and generate greater end-user satisfaction.

For More Information

To learn more about the Big Data Reference Architecture for Cloudera CDH, contact your Lenovo Business Partner or visit:

www.lenovo.com/systems/solutions