White Paper

Cisco, Intel, and Apache Hadoop: Industry's First Reference Design with Advanced Access Control and Encryption

November 2013

What You Will Learn

This white paper describes how the Cisco UCS® Common Platform Architecture (Cisco® CPA) for Big Data can be used to deploy an infrastructure that is optimized for the Apache Hadoop framework. Topics include the business and technical value of the solution and how it can support big data workloads.

In Collaboration With

Cisco, Intel, and Apache Hadoop: Industry's First Reference Design with Advanced Access Control and Encryption

White Paper
November 2013

Highlights

Comprehensive Data Protection

- Cisco UCS® with the Intel® Distribution for Apache Hadoop software offers built-in support for enterprise-class access controls, Kerberos authentication, and data encryption.

Hardware-Accelerated Encryption

- Cisco UCS servers with versatile Intel Xeon® processors accelerate encryption for the most challenging MapReduce jobs.

Fine-Grained Access to Services

- The Intel Distribution for Apache Hadoop software provides a flexible and efficient framework for managing and controlling user access to data and services using existing Kerberos authentication solutions and role-based access control (RBAC) lists to authorize individual users for specific data tables and services.

Scalability for Big Data Workloads

- Cisco UCS Common Platform Architecture (CPA) for Big Data offers linear scalability along with essential operation simplification for single-rack and multirack deployments.

Ease of Deployment

- Cisco UCS Manager automates configuration, reducing the risk of errors that can cause downtime.

Enterprise-Class Support

- Controlled software releases and Intel design, deployment, and ongoing support services increase reliability.

Cisco Unified Computing System™ (Cisco UCS®) with the Intel® Distribution for Apache Hadoop software encrypts data at rest to deliver excellent protection and performance.

The Rise of Big Data Technology

Big data technology, and Apache Hadoop software in particular, is used in an enormous number of applications and is being evaluated and adopted by enterprises of all sizes. As this important technology helps transform large volumes of data into actionable information, organizations struggle to deploy an effective and reliable Hadoop infrastructure. Many of the challenges arise from the friction between the rapid pace of change inherent in open-source software and the need for enterprise-class performance, reliability, and support.

A Unique Solution from Industry Leaders

Cisco and Intel have a long history of collaboration and innovation. The two companies have now worked together to design and deliver a big data solution that combines the Cisco CPA for Big Data and a feature-enhanced, hardware-accelerated, and supported distribution of the Apache Hadoop from Intel. The resulting enterprise-class solution improves performance, capacity, access control, and data protection while accelerating deployment and reducing business and operational risk. Cisco® Common Platform Architecture v2 (CPA v2) for Big Data uses the versatile Intel Xeon E5-2600 v2 family to further improve both performance and capacity.

Cisco and Intel have taken the time and risk out of Apache Hadoop deployment. Enhanced features and a controlled release cycle along with optimized software

deliver excellent performance. With enterprise-class support, the customer-centered solution can be rapidly deployed, scaled on demand, and secured.

Cisco UCS with the Intel Distribution for Apache Hadoop software provides:

- **Powerful computing infrastructure:** Cisco UCS servers with the Intel Xeon Processor E5 family form the core of a flexible and efficient data center that meets diverse business needs. This family of processors is designed to deliver versatility, offering an optimal combination of performance, built-in capabilities, and cost-effectiveness. Integrated I/O capabilities dramatically reduce I/O latency, eliminating data bottlenecks, streamlining operations, and increasing agility. Complementing the processing power of these servers is the massive storage capacity of Cisco UCS C240 M3 Rack Servers.
- **High-performance unified fabric:** The solution's low-latency, lossless 10-Gbps unified fabric is fully redundant and, through its active-active configuration, delivers higher performance than other vendor solutions.
- **Hardware-assisted data encryption:** Encryption and decryption are computing-intensive processes that traditionally add considerable latency and consume substantial processing resources. The Intel Distribution for Apache Hadoop software is optimized for Intel Advanced Encryption Standard New Instructions (AES-NI), a technology that is built into Intel Xeon processors. This approach helps eliminate much of the latency and significantly reduces the load on processors.
- **Ease of deployment:** Cisco UCS is the first unified system built from the beginning so that every aspect of server personality, configuration, and connectivity is set on demand through Cisco UCS Manager. Through the powerful framework of Cisco service profiles, Hadoop cluster servers can be configured rapidly and automatically without the risk of configuration drift that can lead to errors causing downtime. Unified management in Cisco UCS enables great agility and rapid deployment.
- **Robust manageability:** Big data environments can consist of hundreds of servers, resulting in immense management complexity. Whether an implementation uses blade servers to support enterprise applications or rack servers for big data applications, Cisco UCS provides a single point of management for the entire system. With the system's self-aware, self-integrating infrastructure, IT departments can proactively monitor the system and reduce operational costs.
- **Integration with enterprise applications:** Big data environments need high-speed connectivity to transfer results to enterprise applications. The Cisco solution can host the Intel Distribution for Apache Hadoop software and enterprise applications from vendors including Microsoft, Oracle, and SAP in the same management and connectivity domains, further simplifying data center management (Figure 1).
- **Architectural scalability:** The system is designed with logically centralized connectivity management that is physically distributed across the racks and blade chassis that house big data and enterprise applications. After the initial system is established, it can grow and scale without the need to add switching components or to redesign the system's connectivity in any way. The solution can be deployed a rack at a time, with the initial rack hosting the system's fabric interconnects. (These interconnects are described in the section titled "Cisco CPA v2 for Big

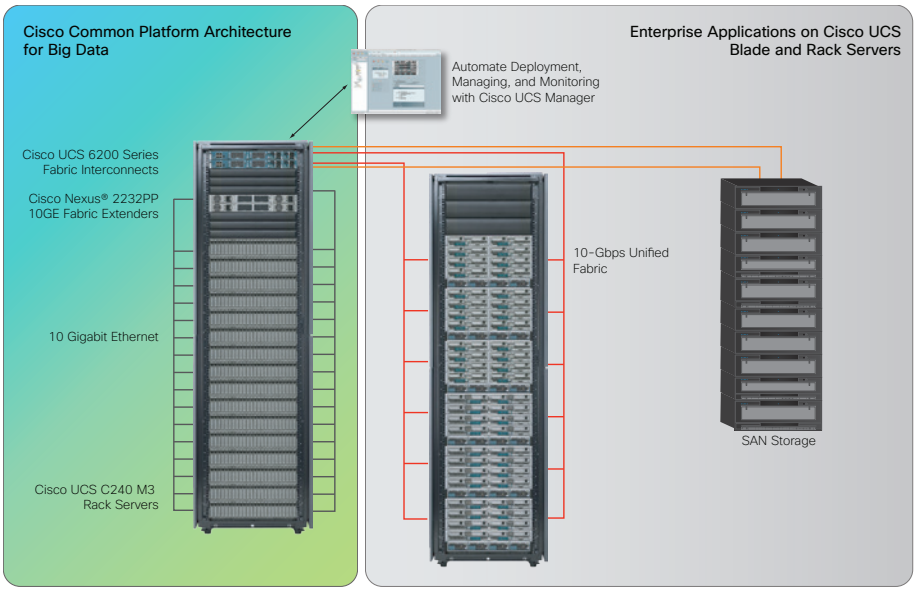

Figure 1. Cisco CPA for Big Data Integrates with Enterprise Applications in a Single Management Domain

Data" on page 6.) Subsequent racks use Cisco fabric extenders, low-cost, low-power-consuming devices that bring the unified fabric to each server in the rack with no additional points of management.

- **Enterprise service and support:** Enterprises using Apache Hadoop to help with business-critical decisions want to know that the vendors providing the solution have the expertise to help them quickly proceed through the initial design, deployment, and testing phases. They also need to have confidence that if a critical component fails, they will receive timely and professional support. One of the factors making this solution unique is the collaboration between Cisco and Intel support teams to make Cisco UCS with the Intel Distribution for Apache Hadoop software a fully supported, enterprise-class solution.

Intel Distribution for Apache Hadoop Software

The Intel Distribution for Apache Hadoop software is a quality-controlled distribution based on the Apache Hadoop source code, with feature enhancements and performance optimizations. Figure 2 shows the elements of the Intel Distribution for Apache Hadoop software, which includes:

- **Intel Manager:** The Intel Manager for Apache Hadoop software streamlines the configuration, management, and resource monitoring of Hadoop clusters. This powerful, easy-to-use, web-based tool allows IT departments to focus critical resources and expertise on generating business value from the environment rather than worrying about the details of cluster management. The software provides installation and configuration features, wizard-based cluster management, proactive cluster health checks, monitoring and logging, and highly secure authentication and authorization.

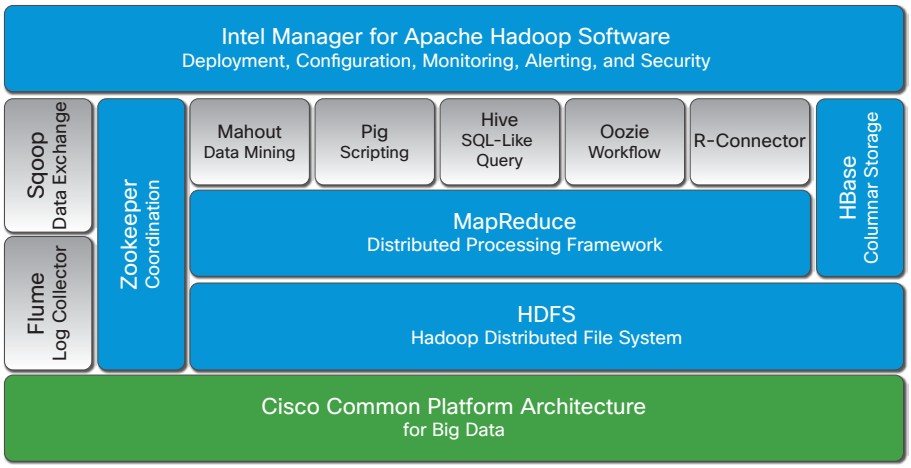

Figure 2. The Solution Combines the Intel Distribution for Apache Hadoop Software with the Cisco CPA

- **Hadoop Data Storage Framework (HDFS):** HDFS is a distributed, scalable, and portable file system that stores data on the cluster nodes. Compression and encryption help enhance security and performance.
- **Data processing framework:** Inspired by Google's MapReduce model, this massively parallel computing framework uses dynamic replication capabilities to intelligently increase and decrease the number of data replicas depending on workload characteristics.
- **Real-time query processing framework:** This component includes the scalable, distributed HBase columnar data storage system for large tables and the Hive data warehouse infrastructure for ad hoc query processing. Extensions support big tables across geographically distributed data centers and feature additions help accelerate HBase and Hive performance.
- **Data protection:** Three independent features provide comprehensive data protection to support regulatory compliance requirements, such as those of the Payment Card Industry security standard and the Health Insurance Portability and Accountability Act. These features include hardware-assisted AES encryption of data, Kerberos authentication of services and users, and precise role-based access control.

Cisco CPA v2 for Big Data

Cisco UCS with the Intel Distribution for Apache Hadoop software is optimized for high performance on the Cisco CPA for Big Data. The Cisco CPA is a highly scalable architecture designed to meet a variety of scale-out application demands with transparent data and management integration capabilities. The new Version 2 has added the latest Intel Xeon processor E5-2600 family, as well as new flash storage options.

The Cisco CPA is built on Cisco UCS, the first truly unified data center platform that combines industry-standard Intel x86-architecture servers with networking and storage access in a single system. Cisco UCS is smart infrastructure that is automatically configured through integrated, model-based management to simplify

and speed the deployment of enterprise-class applications and services running in bare-metal, virtualized, and cloud-computing environments. This approach allows Cisco UCS to unify both big data and enterprise applications in the same centralized management domain.

The Cisco CPA is built using the following components:

- **Cisco UCS 6200 Series Fabric Interconnects** establish a single point of connectivity and management for the entire system. The fabric interconnects provide high-bandwidth, low-latency connectivity for servers. Cisco UCS Manager provides integrated, unified management for all connected devices. Deployed in redundant pairs, Cisco fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical of clusters serving big data applications. Using service profiles, Cisco UCS Manager enables rapid and consistent server configuration, automating system maintenance activities such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.
- **Cisco Nexus 2200 Series Fabric Extenders** bring the system's unified fabric to each rack, establishing a physically distributed but logically centralized network infrastructure. These low-cost, low-power-consuming devices act as remote line cards for the fabric interconnects, providing connectivity without adding the cost and management complexity required by top-of-rack switches. The result is highly scalable and cost-effective connectivity for many nodes.
- **Cisco UCS C240 M3 Rack Servers** support a wide range of computing, I/O, and storage-capacity demands in a compact two-rack-unit (2RU) design. The servers use dual Intel Xeon processor E5-2600 v2 CPUs and support up to 768 GB of main memory (128 GB or 256 GB is typical for big data applications) and a range of disk-drive options. Cisco UCS virtual interface cards (VICs) are optimized for high-bandwidth and low-latency cluster connectivity, with support for up to 256 virtual devices that are configured on demand through Cisco UCS Manager.

Choice of Configuration

The Cisco CPA for Big is offered as a reference architecture that can be deployed easily through Cisco UCS Solution Accelerator Paks. A single-rack configuration provides two Cisco UCS 6296UP 96-Port Fabric Interconnects (to connect up to 10 racks and 160 servers), along with two Cisco Nexus 2232PP 10GE Fabric Extenders and 16 Cisco UCS C240 M3 Rack Servers. The configurations are balanced either for performance and capacity, or capacity accelerated by flash memory. Multirack configurations include two Cisco Nexus 2232PP fabric extenders and 16 Cisco UCS C240 M3 servers for each additional rack.

Optimized Configurations

- The **performance and capacity balanced** configuration supports up to 320 cores, 384 TB of storage and 32 GBps I/O bandwidth.
- The **capacity optimized** configuration supports up to 256 cores, 768 TB of storage and 16 GBps I/O bandwidth.
- The **capacity optimized flash** configuration supports up to 320 cores, 768 TB of hard disk storage with 16 GBps I/O bandwidth and 3.125 TB of high-speed flash storage at 24 GBps I/O bandwidth.

Each server in the configuration connects to the Cisco Unified Fabric through two active-active 10 Gigabit Ethernet links using a Cisco UCS VIC. Each high-performance rack can support up to 256 cores and 32-GBps (SATA) or 48-GBps (SAS) I/O bandwidth. Each high-capacity rack can support up to 576 TB of raw storage.

Massive Scalability

The Cisco CPA for Big Data supports the massive scalability demanded by big data environments. Up to 160 servers are supported in a single management domain with a pair of Cisco fabric interconnects. Additional scaling can be accomplished by interconnecting multiple domains using Cisco Nexus 6000 or 7000 Series Switches. With Cisco UCS Central Software, thousands of servers and hundreds of petabytes of storage can be managed through a single interface with the same automation provided by Cisco UCS Manager (Figure 3).

Cisco UCS Solution Accelerator Paks

Cisco UCS Solution Accelerator Paks speed and simplify solution deployment (Table 1). With only a single part number to order, these packages make it easy to quickly deploy a powerful and highly secure big data environment without the cost or risk entailed in designing and building custom solutions.

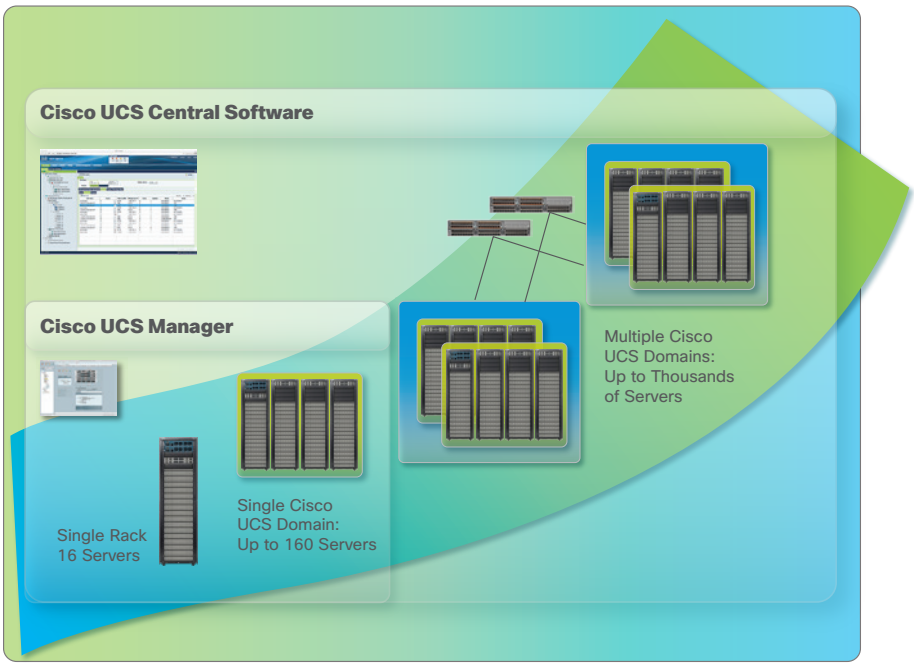

Figure 3. Cisco UCS with the Intel Distribution for Apache Hadoop Software Can Scale to Thousands of Servers

Table 1. Cisco UCS Solution Accelerator Paks for Cisco CPA v2 for Big Data

	Performance and Capacity Balanced (UCS-EZ-BD-PC)	Capacity Optimized (UCS-EZ-BD-C)	Capacity Optimized with Flash Memory (UCS-EZ-BD-CF)
Connectivity	• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects • 2 Cisco UCS 2232PP 10GE Fabric Extenders	• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects • 2 Cisco UCS 2232PP 10GE Fabric Extenders	• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects • 2 Cisco UCS 2232PP 10GE Fabric Extenders
Management	• Cisco UCS Manager	• Cisco UCS Manager	• Cisco UCS Manager
Servers	16 Cisco UCS C240 M3 Rack Servers, each with: • 2 Intel Xeon processors E5-2660 v2 • 256 GB of memory • LSI MegaRaid 9271CV 8i card • 24 1-TB 7.2K SFF SAS drives (384 TB total)	16 Cisco UCS C240 M3 Rack Servers, each with: • 2 Intel Xeon processors E5-2640 v2 • 128 GB of memory • LSI MegaRaid 9271CV 8i card • 12 4-TB 7.2K LFF SAS drives (768 TB total)	16 Cisco UCS C240 M3 Rack Servers, each with: • 2 Intel Xeon processors E5-2660 v2 • 128 GB of memory • Cisco UCS Nytro MegaRAID 200-GB Controller (3.125 TB total flash capacity) • 12 4-TB 7.2K LFF SAS drives (768 TB total)

Benchmark Results

Benchmark tests were run on a performance- and capacity-optimized configuration with 16 servers, each with dual Intel Xeon processor E5-2660 v2 CPUs and 256 GB of memory. The benchmark ran word counts on encrypted and unencrypted data for different data sizes. Encryption tasks included decrypting the input file, storing the intermediate encrypted file, and encrypting the final result. Figure 4 shows that hardware-assisted AES encryption and decryption imposed very little overhead, decreasing throughput by 5 percent and adding 1.5 percent CPU utilization over unencrypted jobs. Figure 5 illustrates the degree to which this overhead is consistently low across a long-running job.

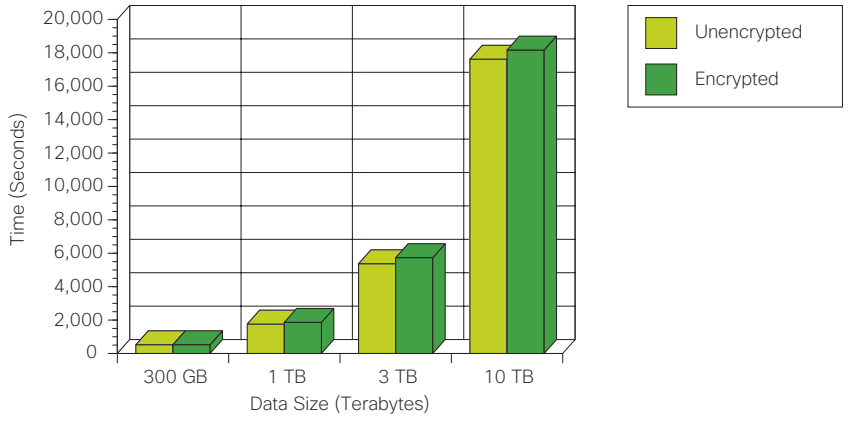

Figure 4. Word-Count Job Throughput on Different-Size Workloads

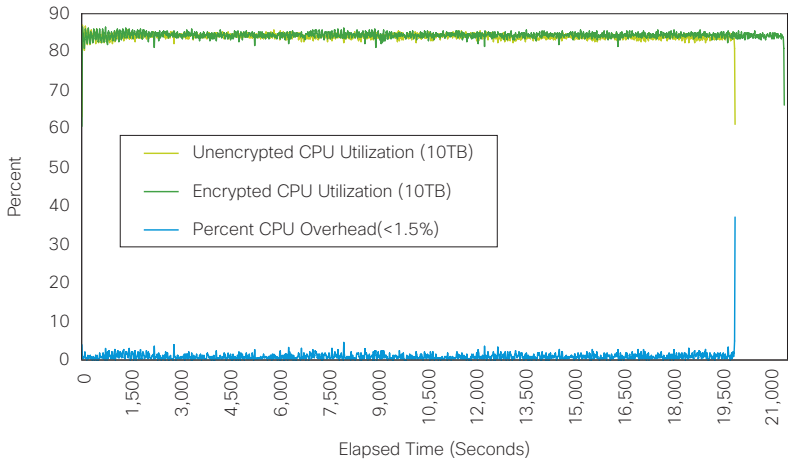

Figure 5. CPU Overhead for Encryption Is Consistently Low for Both Encrypted and Unencrypted Test Runs

Conclusion

Big data technology is becoming compelling for business organizations of all sizes. But there is understandable friction between the need for software that can stand up to mission-critical needs and the risk and stability concerns with using unsupported open-source software. Cisco UCS with the Intel Distribution for Apache Hadoop software provides critical technology enhancements that make it easy and safe to deploy big data applications in enterprise environments. Whether you are deploying a large data center or buying individual Cisco UCS Solution Accelerator Paks, the Cisco and Intel solution can be sized to help you meet your unique business challenges.

Acknowledgments

This white paper was made possible through the efforts of Karthink Kulkarni and Suyash Ramineni with assistance from Raghunath Nambiar and Girish Kulkarni.

For More Information

- For more information about the collaboration between Cisco and Intel, please visit http://www.cisco.com/go/intel.
- For more information about Cisco UCS, please visit http://www.cisco.com/go/ucs.
- For more information about the Cisco UCS Solution Accelerator Paks, please visit http://www.cisco.com/go/smartplay.
- For more information about Cisco UCS big data solutions, please visit http://www.cisco.com/go/bigdata.
- For more information about the Cisco UCS CPA v2 for Big Data, please visit http://blogs.cisco.com/datacenter/cpav2.

- The Cisco Validated Design for Intel Distribution for Apache Hadoop on Cisco UCS CPA v2 for Big Data is available at http://www.cisco.com/en/US/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_CPA_for_Big_Data_with_Intel.html
- Visit the Cisco big data design zone at http://www.cisco.com/go/bigdata_design.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices**.

LE-39604-00 11/13