Cisco UCS Integrated Infrastructure for Big Data with IBM BigInsights for Apache Hadoop

Solution Overview

Highlights

Optimized for Enterprise Big Data Deployments

- Cisco UCS® Integrated Infrastructure for big data offers a balance of compute power, I/O bandwidth, and storage capacity for IBM BigInsights for Apache Hadoop, a modular framework that can scale from small to very large as needs change.

Built on Cisco UCS Advantages

- The solution offers unified fabric, unified management, and advanced monitoring capabilities.
- Consistent and rapid deployment using Cisco UCS service profiles delivers out-of-the-box performance.
- Simplified and Policy-Based Management
- Solution features include global inventory view, one-click system software management, and one-click configuration changes.

Reduced Risk

- Prevalidation, tighter integration, and performance optimization reduce integration and deployment risks.
- Extensive testing and validation of software distributions allow you to deploy the solution with confidence.

Speed and efficiency for data and analytics

- The Cisco® and IBM joint solution offers performance, capacity and scaling, plus extensive management and visualization capabilities.
- Comprehensive developer tools, and powerful analytics functions.
- Rapid time to analytics with Big SQL – a highly capable SQL engine for Hadoop
- Data preparation and views on Hadoop without coding, powered by IBM BigSheets.
- Native R functions run at big data scale with IBM BigInsights Big R, which automatically tunes machine learning performance over large-scale data.

Cisco and IBM deliver industry-leading solution to help businesses accelerate big data analytics

The Cisco Unified Computing System™ (Cisco UCS) solution for IBM BigInsights for Apache Hadoop is based on Cisco UCS Integrated Infrastructure for big data that integrates industry-leading computing, networking, and management capabilities into a unified, fabric-based architecture optimized for big data workloads. Optimized to deliver insights faster and reduce total cost of ownership (TCO), the joint solution enables you to unlock the intelligence in your data to help you create a sustainable, competitive business advantage.

Solution: Cisco UCS Integrated Infrastructure for Big Data with IBM Big Insights:

- Combines innovations from Cisco UCS such as programmable infrastructure with best of open source software with enterprise-grade capabilities in IBM BigInsights for Apache Hadoop
- Jointly designed, pre-tested, pre-validated and fully documented for predictable deployments that can scale as workload demands
- Cisco and IBM provide enterprises with transparent infrastructure management and integration of Hadoop with other information solutions to help enhance data manipulation and management tasks
- Deliver performance and scale with Enterprise SQL on Hadoop for businesses to accelerate data science and analytics to deliver the deepest possible insight into data
- Industry leading, world-wide support and services from Cisco, IBM and partners

In collaboration with:

Cisco UCS Solution for IBM BigInsights for Apache Hadoop

This solution is built on Cisco UCS infrastructure using Cisco UCS 6200 Series Fabric Interconnects and Cisco UCS C-Series Rack Servers optimized for IBM BigInsights for Apache Hadoop as shown in Figure 1.

Cisco UCS 6200 Series Fabric Interconnects

Cisco UCS 6200 Series Fabric Interconnects establish a single point of connectivity and management for the entire system. They provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management for all connected devices provided by Cisco UCS Manager. 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 in clusters that serve big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles, automating ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. It also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C-Series Rack Servers

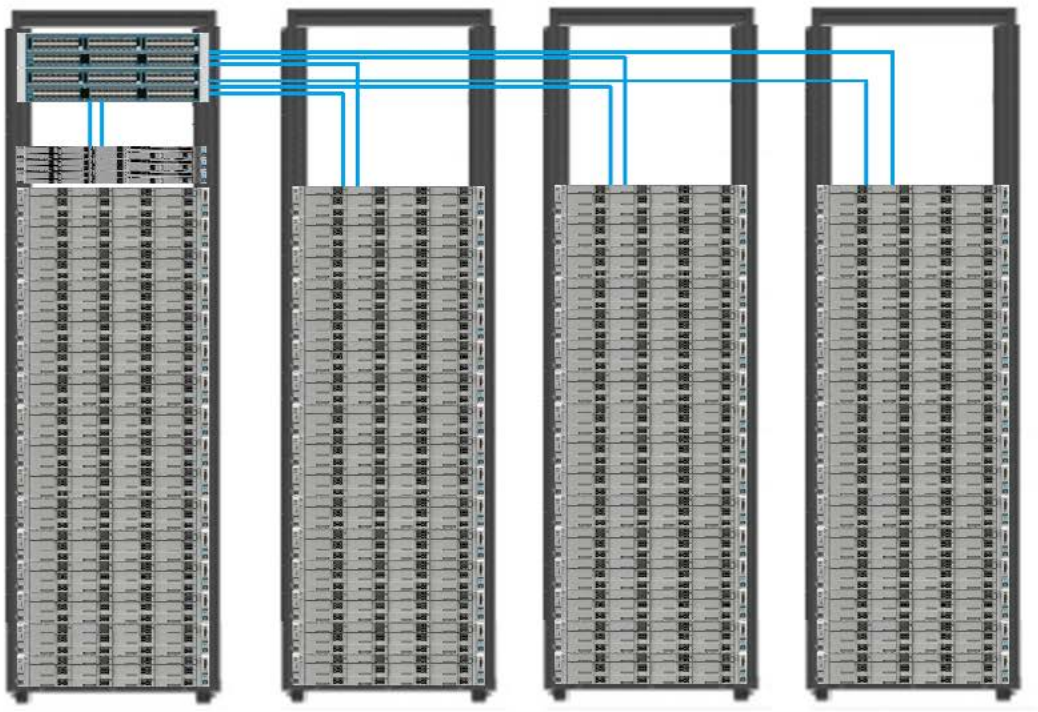
Cisco UCS C-Series Rack Servers deliver unified computing in an industry-standard form factor to reduce total cost of ownership and increase agility. Each product addresses varying workload challenges through a balance of processing, memory, I/O, and internal storage resources.

The Cisco UCS C240 M4 Rack Server is designed to support a wide range of computing, I/O, and storage-capacity demands in a compact design. The server is based on the Intel® Xeon® processor E5-2600 v3 series and offers 12-Gbps SAS throughput, delivering significant performance and efficiency gains over the previous generation of servers. The server uses dual Intel Xeon processor E5-2600 v3 series CPUs and supports up to 768 GB of main memory and a range of hard disk drive (HDD) and solid-state disk (SSD) drive options. Twenty-four small-form-factor (SFF) disk drives are supported in the performance-optimized option, and 12 large-form-factor (LFF) disk drives are supported in the capacity-optimized option, along with two 1 Gigabit Ethernet embedded LAN-on-motherboard (LOM) ports.

The Cisco UCS Virtual Interface Card (VIC) 1227 is designed for the M4 generation of Cisco UCS C-Series Rack Servers. The VIC is 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.

Figure 1: Cisco UCS Integrated Infrastructure for Big Data with IBM BigInsights: A 64 Data Node Cluster

- 2 x Cisco UCS 6296 Fabric Interconnect
- 3 x Cisco UCS C220 M4 Servers
- 64 x Cisco UCS C240 M4 Servers



IBM BigInsights for Apache Hadoop: A Complete Hadoop Platform

IBM BigInsights for Apache Hadoop introduces new analytics and enterprise capabilities for Hadoop, including machine learning using IBM InfoSphere BigInsights, Big R, IBM Big SQL enhancements, and current open-source Apache packages, to help data scientists, analysts, and administrators accelerate data science tasks.

Big SQL

Big SQL uses IBM's strength in SQL engines to provide ANSI SQL access to data transparently across any system from Hadoop, through Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC), whether that data exists in Hadoop or a relational database (Figure 2).

Developers familiar with the SQL programming language can access data in Hadoop without having to learn new languages or skills.

With Big SQL, all of your big data is accessible through SQL. It presents a structured view of your existing data, using an optimal processing strategy based on available resources. Big SQL can use MapReduce parallelism when needed for complex data sets and avoid it when it hinders performance, using direct access for less complicated, low-latency queries.

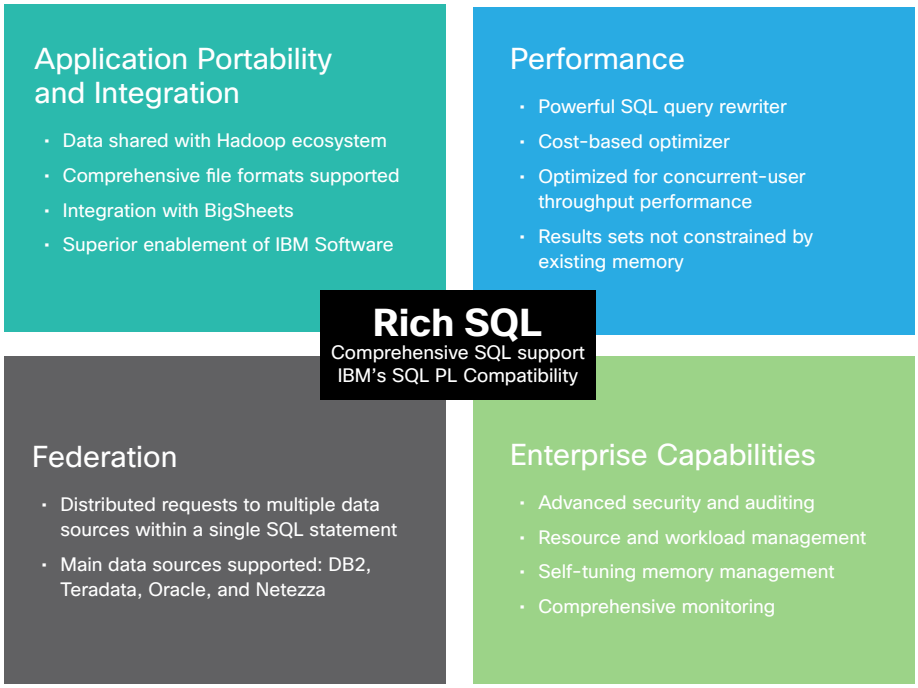
Big SQL offers the following capabilities:

- Low-latency queries enabled by massively parallel processing (MPP) technology.
- Query rewrite optimization and cost-based optimizer.
- Integration of both Hive and HBase data sources.
- Exceptional support for ANSI SQL Standard
- Federated query access to IBM DB2, Oracle, Teradata, and ODBC sources

Big SQL supports the most common use cases for modernizing and building next-generation logical data warehouses:

- Offload data and workloads from existing data warehouses.
- Move rarely used data out of high-cost data warehouses by creating queryable archives in Hadoop.
- Enable rapid prototyping of business intelligence reports.
- Securely support rapid adoption of Hadoop by using existing SQL skills.

Figure 2: IBM Big SQL Capabilities



IBM BigInsights for Apache Hadoop offers these features and benefits:

- Provides advanced analytics built on Hadoop technology to meet big data analysis requirements
- Is designed for performance and usability, with performance-optimized capabilities, visualization capabilities, comprehensive developer tools, and powerful analytics functions
- Delivers management, security, and reliability features to support large-scale deployments and help accelerate time to value
- Integrates with information solutions from IBM and other vendors to help enhance data manipulation and management tasks

BigSheets

BigSheets makes do-it-yourself-analytics a reality by going beyond structured database management to unstructured data management. The capability to see the whole picture can help people at all levels of the organization make better decisions.

BigSheets provides a web-based, spreadsheet-style view into collections of files in Hadoop. Users can perform data transformations, filter data, and create views at massive scale. No coding is required because BigSheets translates spreadsheet actions into MapReduce processes to use the computational resources of the Hadoop cluster. This feature helps analysts discover value in data quickly and easily.

BigSheets is an extension of the model that:

- Integrates gigabytes, terabytes, or petabytes of unstructured data from web-based repositories
- Collects a wide range of unstructured web data from user-defined seed URLs
- Extracts and enriches that data using the unstructured information management architecture you choose (LanguageWare, OpenCalais, etc.)
- Lets you explore and view this data in specific, user-defined contexts (such as ManyEyes)

BigSheets benefits include the following:

- It gives business users a new approach to keep pace with data escalation. By taking the structure to the data, BigSheets helps mine petabytes of data without the need for additional storage.
- It gives business users a new approach that allows them to divide data into consumable, situation-specific frames of reference. This feature helps enable organizations to translate untapped, unstructured, and often unknown web data into actionable intelligence.
- It uses all the computing resources of the Hadoop cluster to provide insights and views with BigSheets from within the cluster. No data extraction is required.

For data scientists, the BigInsights Data Scientist module includes Big R and advanced text analytics capabilities.

Big R

Big R enables data scientists to run native R functions to explore, view, transform, and model big data from within the R environment. Data scientists can now run scalable machine-learning algorithms with a wide class of algorithms and R-like syntax for new algorithms and customize existing algorithms. BigInsights for Apache Hadoop running Big R can use the entire cluster memory, spill to disk, and run thousands of models in parallel.

Benefits of Big R includes:

- End-to-end integration with open source R
- Transparent processing on Hadoop
- Transparent access to comprehensive and scalable machine-learning algorithms provided in Big R
- Text analytics capabilities to extract meaningful information from unstructured data

Text Analytics

A sophisticated text analytics capability unique to BigInsights allows developers to easily build high-quality applications that can process text in multiple written languages and derive insights from large amounts of native textual data in various formats.

For administrators, the BigInsights Enterprise Management module provides a management console and built-in security features.

Management Console

A comprehensive web-based interface included in BigInsights simplifies cluster management, service management, job management, and file management.

Administrators and users can share the same interface, launching applications and viewing a variety of configurable reports and dashboards.

Built-in Security

BigInsights was designed with security in mind, supporting Kerberos authentication and providing data privacy, masking, and detailed access controls with auditing and monitoring functions to help ensure that the environment stays secure.

Reference Architecture

The current version of the Cisco UCS Integrated Infrastructure for all big data offers the following configuration based on the computing and storage requirements for IBM BigInsights (see Table 1).

Table 1:

Cisco UCS Integrated Infrastructure for Big Data with IBM BigInsights

Connectivity:

2 Cisco UCS 6296UP 96 Port Fabric Interconnects

Scaling:

- Up to 80 servers per domain
- Scaling to thousands of servers through the use of Cisco Nexus® 7000 and 9000 Series Switches

Management nodes for IBM BigInsights (per cluster):

3 Cisco UCS C220 M4 Rack Servers, each with:

- 2 Intel Xeon processor E5-2680 v3 CPUs
- 256 GB of memory
- 8 x 600-GB 10,000-rpm SAS HDDs
- Cisco 12-Gbps SAS Modular Raid Controller with 2-GB flash-based write cache (FBWC)
- Cisco UCS VIC 1227 with 2 x 10 Gigabit Ethernet

Note: For IBM BigInsights 4.x, there will be 6 Management nodes for running the management services (namenode, resource manager, etc) in High-Availability mode

Data nodes (per rack):

16 Cisco UCS C240 M4 Rack Servers (LFF), each with:

- 2 Intel Xeon processor E5-2680 v3 CPUs
- 128 GB of memory
- Cisco 12-Gbps SAS Modular Raid Controller with 2-GB FBWC
- 12 x 6-TB 7200-rpm LFF SAS drives (768 TB total)
- 2 x 120-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot
- Cisco UCS VIC 1227 with 2 x 10 Gigabit Ethernet SFP+ ports

For More information

Cisco Big Data Portal: **www.cisco.com/go/bigdata**
Cisco UCS Design for Big Data: **http://www.cisco.com/go/bigdata_design**
Cisco UCS Integrated Infrastructure for Big Data with IBM BigInsights:

- Cisco Validated Design: **http://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/Cisco_UCS_Integrated_Infrastructure_for_Big_Data_with_IBM.html**
- Cisco Data Center Blog: **http://blogs.cisco.com/datacenter/biginsights**

IBM BigInsights for Apache Hadoop: **http://www-01.ibm.com/software/data/infosphere/hadoop/enterprise.html**

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**.