Pentaho High-Performance Big Data Reference Configurations using Cisco Unified Computing System

By Jake Cornelius
Senior Vice President of Products
Pentaho

June 1, 2012

Pentaho Delivers High-Performance Big Data Configurations Using Cisco Unified Computing System

Pentaho, together with the Cisco Unified Computing System provides companies with Big Data Platform that delivers high performance, robust data integration, and advanced analytics features that expedite the implementation of end-to-end big data analytic solutions.

Next-Generation Big Data Solution

The combination of world-leading Cisco Unified Computing System™ (Cisco UCS™) and Pentaho Business Analytics enables companies to significantly reduce time-to-value and the operating expenses associated with Big Data.

Pentaho Business Analytics: Rapidly Design and Deploy Big Data Solutions

By tightly coupling data integration with business analytics, Pentaho brings together IT and business users to easily access, integrate, visualize, explore and mine all data that impacts business results. Pentaho's open source heritage drives continued innovation in a modern, unified, embeddable analytics platform that works with any data including big data and diverse data types. Pentaho Business Analytics (BA) provides a complete solution, is fast to deploy, easy to use, and extremely cost-effective — in short, delivering business analytics that work. The unified suite includes data integration, data discovery and exploration, and data mining.

Cisco UCS and Pentaho BA can help businesses manage many different data integration and analytics use cases. Table 1 provides examples of how the Pentaho Reference Configurations can accelerate big data initiatives

Table 1. Sample Use Cases for Cisco UCS and Pentaho Business Analytics

Scenario	Pentaho Reference Configuration Analytics
Data Acquisition	Easily collect and store structured, semi-structured and unstructured data in a fault-resilient, scalable store that can be organized and sorted for indexing and analysis.
Data Preparation	Design powerful ETL jobs in an easy-to-use, graphical environment to process (batch or real-time) large quantities of structured, semi-structured and unstructured data.
Orchestration	Graphically design workflows for data acquisition, data processing and analytics which can be executed on a scheduled basis or in real-time by integrating with your existing IT infrastructure.
Analytic Solutions	Agilely design and generate new analytic solutions; Visually explore and analyze data; Share and distribute results (examples: online dashboards, interactive analytics, bursted reports)
Big Data Solutions	Deliver scalable, end-to-end solutions for a broad spectrum of big data use cases: for example, sentiment analysis, customer risk analysis, trade analytics, credit scoring, and fraud detection.
Data Fabric	Holistically create and manage solutions in a hybrid data environment: example, collect data from a mix of cloud and on-premise sources, store raw data in Hadoop/NoSQL, spin off processed data to an analytic data mart (relational, columnar, in-memory) for interactive analysis.

As shown in Figure 1, the major components of Pentaho BA include:

Data Integration

Data is everywhere and the volume and variety of data is growing by the minute. With Pentaho Data Integration organizations can extract data from complex and heterogeneous sources and diverse data types to produce consistent, high quality ready-to-analyze data for powering business analytics. With a rich graphical user interface and a parallel processing engine, Pentaho Data Integration offers high performance ETL (extract, transform and load). Tight integration with the Pentaho Business Analytics platform further provides the fastest path to delivering rich reporting, dashboards, data discovery and predictive analytics solutions.

Highlights:

- Rich, graphical designer
- Enterprise scalability and performance
- Big data integration and job orchestration for Hadoop, NoSQL and analytic databases
- Integrated, interactive reporting and data analysis

Big Data

Pentaho Business Analytics for big data dramatically lowers the technical barriers and shortens the time it takes to help companies pragmatically operationalize the promise of big data by delivering an integrated analytics solution.

Pentaho is the leading solution for big data analytics and provides numerous benefits including:

- **Improve Productivity** - visual design and management tools providing a 10x productivity improvement over custom development
- **Reduce costs** – empowers organizations to leverage existing skill sets to implement Big Data solutions by providing familiar tools for Data Integration and Business Analytics that hide the complexities of Big Data platforms
- **Prevent Big Data 'Silos'** – prevents Big Data platforms from becoming information silos by providing the ability to easily design sophisticated workflows that orchestrate events that span Big Data and traditional data platforms
- **Freedom to Choose** – broad support for Big Data platforms including Hadoop, NoSQL and MPP Databases allows you to choose the right tool for each use case and ensure solutions can be designed and managed in a single environment
- **End-to-end Analytic Solutions** – provides a clear path to designing complete business analytic solutions from standard reporting and dashboards to data discovery to predictive analytics

Data Discovery and Exploration

Pentaho Business Analytics provides a highly interactive and easy to use web-based interface for business users to access and visualize data, create and interact with reports and dashboards, and analyze data across multiple dimensions, without depending on IT or developers. For IT, Pentaho Business Analytics is built on a modern lightweight high-performance platform and can be flexibly deployed on-premise, in the Cloud, or seamlessly embedded into other software applications.

Data Mining and Predictive Analytics

The powerful, state-of-the-art machine learning algorithms and data processing tools in Pentaho Business Analytics enable users to uncover meaningful patterns and correlations that may otherwise be hidden with standard analysis and reporting. These sophisticated, advanced analytics help plan for future outcomes based on a better understanding of prior business performance. Pentaho's Business Analytics includes:

- Dozens of powerful algorithms including classification, regression, clustering and association
- Support for the whole process of experimental data mining, including:
  - Preparation of input data
  - Statistical evaluation of learning schemes
  - Visualization of input data and the result of learning

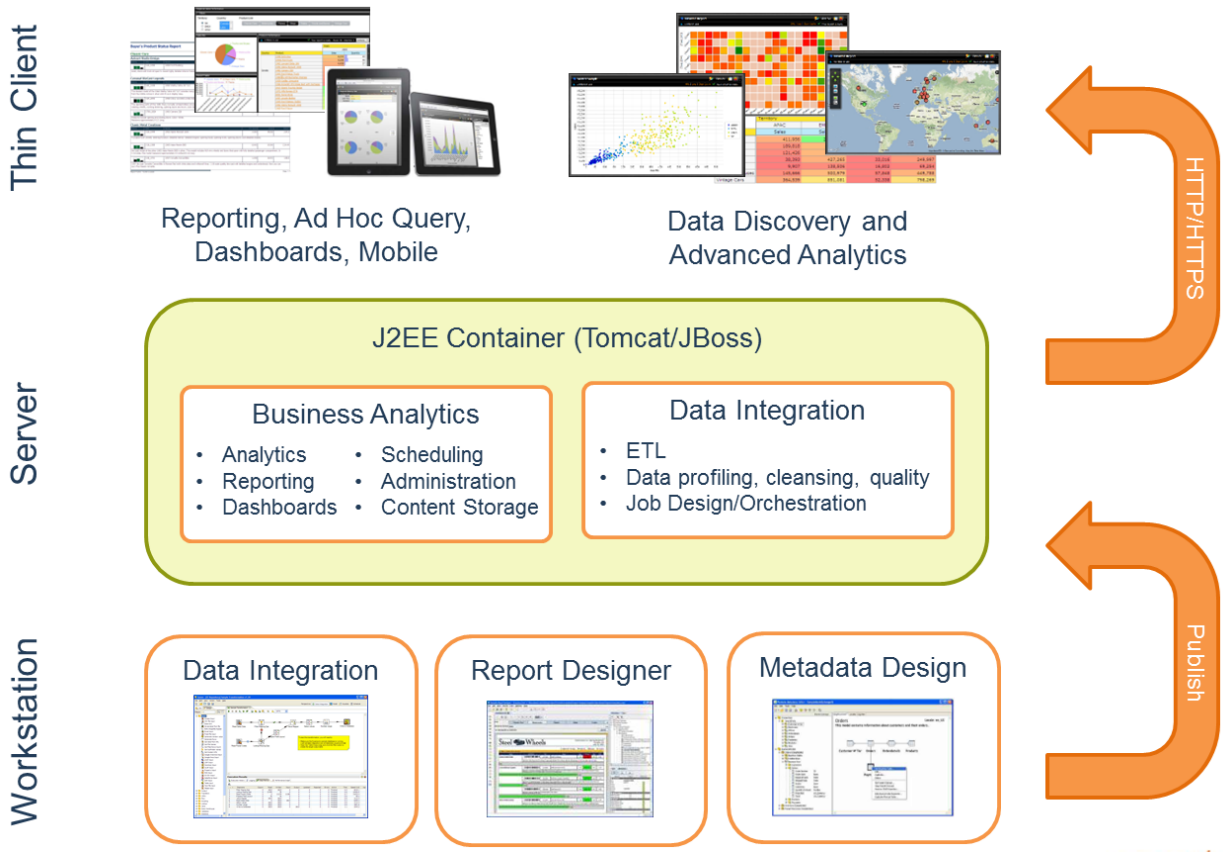

Figure 1. Pentaho Business Architecture and Components

Cisco UCS: The Ideal Analytics Platform

Cisco UCS is the ideal platform for Pentaho Business Analytics. It is the outcome of a thorough testing and development process between Pentaho and Cisco. Cisco UCS innovations combine industry-standard, x86-architecture servers with networking and storage access into a single converged system (Figure 2). The system is entirely programmable using unified, model- based management to simplify and accelerate the deployment of enterprise-class applications and services running in bare-metal, virtualized, and cloud-computing environments.

Big Data implementations can present a number of challenges to enterprise environments. Many of these challenges arise from the dichotomy between the introduction of innovative new technology and the enterprise-class performance, reliability, and support demanded by mission-critical systems. The joint Cisco and Pentaho solution is designed to provide a solution to these challenges and offers radically simplified deployment, management and system monitoring capabilities, high availability, exceptional performance and scalability, and enterprise-class service and support.

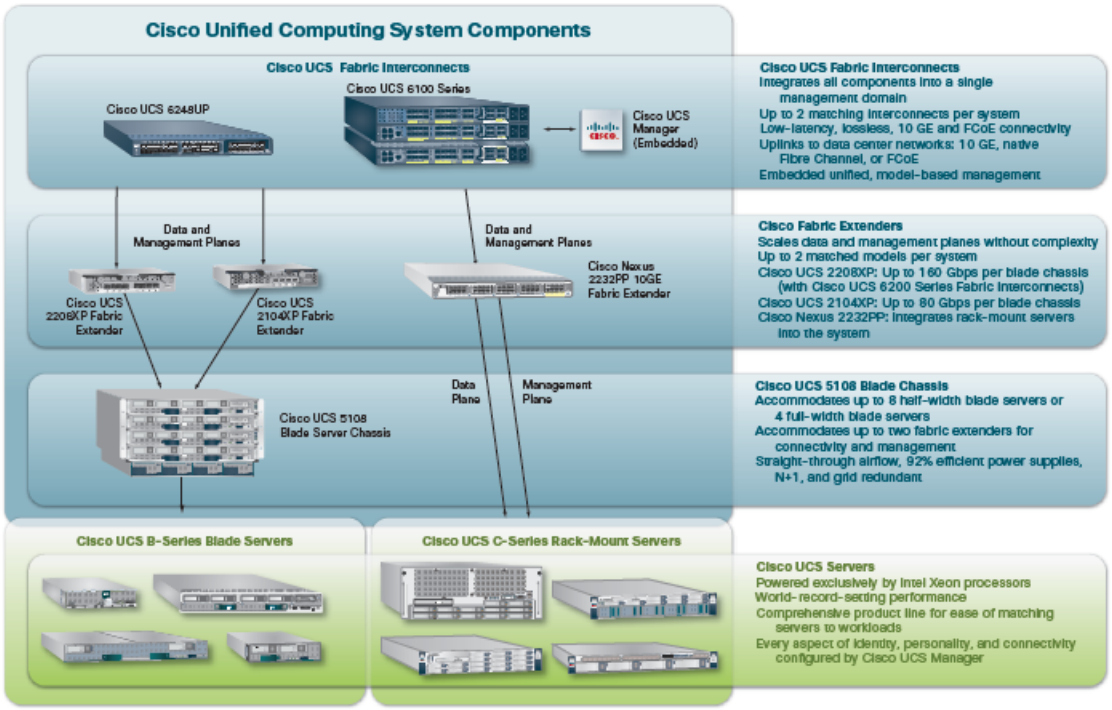

Figure 2. Cisco UCS Is a Single Unified System

Reference Configuration

The reference configuration is built using the Cisco Big Data Common Platform following components:

- Cisco UCS 6200 Series Fabric Interconnects: The Cisco UCS 6200 Series Fabric Interconnects are a core part of Cisco UCS, providing both network connectivity and management capabilities across Cisco UCS 5100 Series Blade Server Chassis as well as Cisco UCS C-Series Rack-Mount Servers. Typically deployed in redundant pairs, the fabric Interconnects offer line-rate, low-latency, lossless 10 Gigabit Ethernet connectivity and unified management with Cisco UCS Manager in a highly available management domain.
- Cisco UCS 2200 Series Fabric Extenders: Cisco UCS 2200 Series Fabric Extenders behave as remote line cards for a parent switch and provide a highly scalable and extremely cost-effective unified server-access platform.
- Cisco UCS C240 M3 Rack-Mount Servers: Cisco UCS C240 M3 Rack-Mount Servers are general-purpose 2-socket platforms based on Intel® Xeon® E-2600 series processors. These servers support up to 768 GB of main memory and 24 small factor (high performance) or 12 large form factor (high capacity) internal front-accessible, hot-swappable to provide data performance, capacity and data protection.
- Cisco UCS P81E Virtual Interface Card (VIC): Unique to Cisco UCS is a dualport PCI Express (PCIe) 2.0 x8 10-Gbps adapter designed for use with Cisco UCS C-Series Rack-Mount Servers.
- Cisco UCS Manager: Cisco UCS Manager resides within the Cisco UCS 6200 Series Fabric Interconnects. It makes the system self-aware and self-integrating, managing all of the system components as a single logical

entity. Cisco UCS Manager can be accessed through an intuitive GUI, a command-line interface (CLI), or an XML API. Cisco UCS Manager uses service profiles to define the personality, configuration, and connectivity of all resources within Cisco UCS, radically simplifying provisioning of resources so that the process takes minutes instead of days. This simplification allows IT departments to shift their focus from constant maintenance to strategic business initiatives.

The single-rack configuration consists of two fully redundant Cisco UCS 6248UP 48-Port Fabric Interconnects and two Cisco Nexus® 2232PP 10GE Fabric Extenders, as depicted in Figure 3. Each node in the configuration connects to the unified fabric through two active-active 10-Gbps links using a Cisco UCS P81E VIC (data traffic) and Cisco Integrated Management Controller (IMC; management traffic). Multi-rack configurations include components from a single rack and two Cisco Nexus 2232PP fabric extenders for every additional rack.

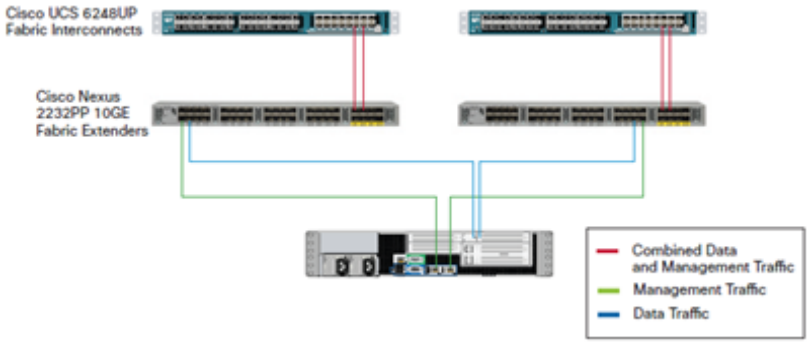

Figure 3. UCS Fabric Architecture

The high performance cluster node is a Cisco UCS C240 M2 Rack-Mount Server with two Intel Xeon E5-2665 processors, 256 GB of memory, a Cisco UCS VIC 1225 , an LSI 6G MegaRAID 9266-8i card, and 24 1-TB SATA SFF internal disk drives for a total of 24 TB of storage. The high performance cluster node is a Cisco UCS C240 M2 Rack-Mount Server with two Intel Xeon E5-2640 processors, 128 GB of memory, a Cisco UCS VIC 1225 , an LSI 6G MegaRAID 9266-8i card, and 12 3-TB SAS LFF internal disk drives for a total of 36 TB of storage. The high performance and high capacity reference configurations are depicted in Figure 4.

2 x Cisco UCS 6200 Fabric Interconnects
2 x Cisco Nexus® 2232PP 10GE Fabric Extenders

High Performance Configuration

16 x Cisco UCS C240 M3 Rack-Mount Servers
2x Intel® Xeon® E5-2665 processors
256 GB Memory
1 x Cisco UCS VIC 1225
1 x LSI MegaRAID SAS 9226CV-8i Card
24 x 1-TB SATA 7200 RPM SFF Disk Drive

High Capacity Configuration

16 x Cisco UCS C240 M3 Rack-Mount Servers
2x Intel® Xeon® E5-2640 processors
128 GB Memory
1 x Cisco UCS VIC 1225
1 x LSI MegaRAID SAS 9226CV-8i Card
12 x 3-TB SAS 7200 RPM LFF Disk

Figure 4. High Performance and High Capacity reference configuration

The performance and capacity characteristics of high performance and high capacity configurations are shown in table 2 and table 3. Node recommendations for Pentaho Business Analytics is shown in table 4.

Table 2. High Performance Reference Configurations

	Component	Single Rack	Multi-Rack
Network Fabric	Fabric Interconnects	**2**	2 per cluster
	Fabric extenders	**2**	2 per rack
Computing	Servers	**16**	16 per rack
	Computer processor cores	**256**	256 per rack
	Memory	**2TB (up to 12 TB supported)**	2TB (up to 12 TB supported)
	Unformatted storage capacity	**384 TB**	348 TB per rack

Table 3. High Capacity Reference Configurations

	Component	Single Rack	Multi-rack
Network Fabric	Fabric interconnects	**2**	2 per cluster
	Fabric extenders	**2**	2 per rack
Computing	Servers	**16**	16 per rack
	Computer processor cores	**2 TB (up to 12 TB supported)**	2TB (up to 12 TB supported)
	Memory	**576 TB**	576 per rack

Table 4. Node Recommendations for Pentaho Business Analytics

Service	Number of Nodes
Data Integration Server	**Most or all nodes**
Business Analytics Server	**1 to 3**

Figure 6 below represents the combined platform of Cisco UCS and Pentaho Business Analytics

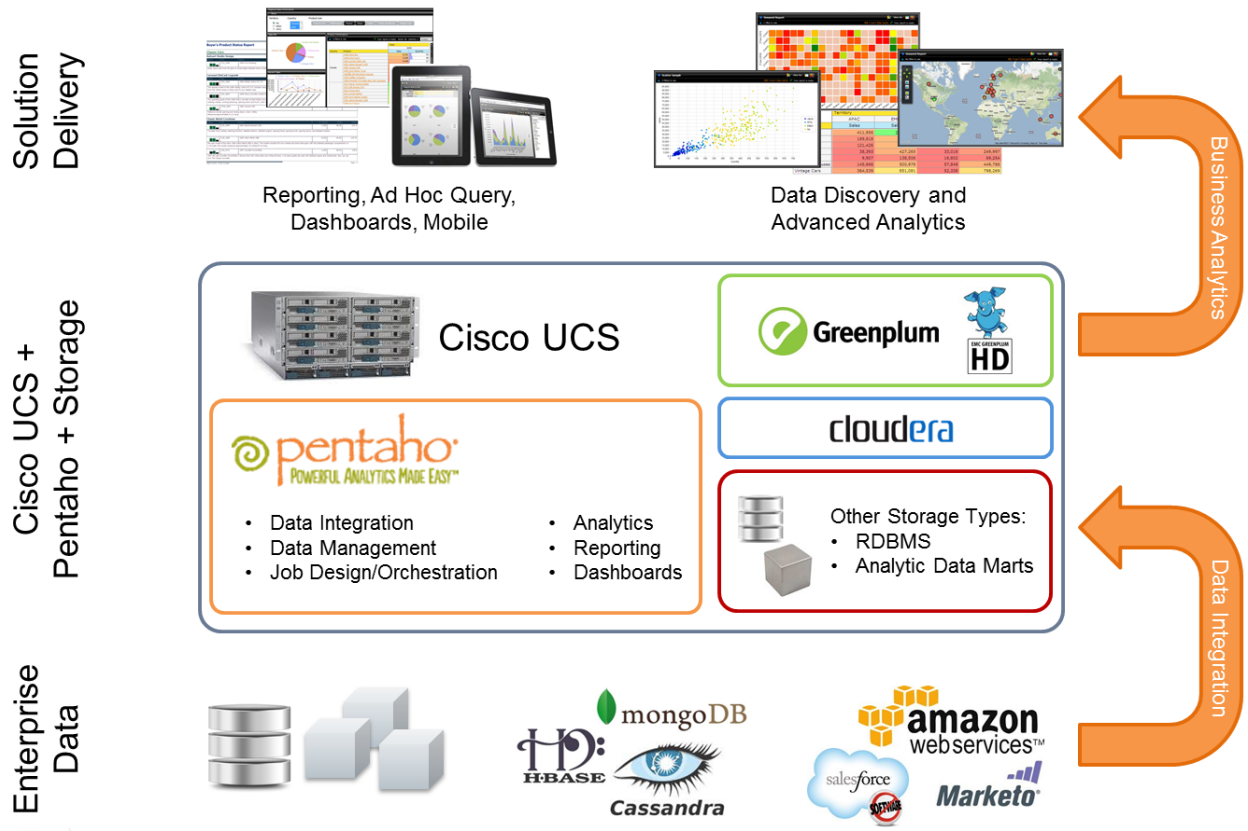

Figure 6. Reference Architecture for Cisco UCS, Pentaho Business Analytics and Big Data platforms

Complete Big Data Analysis Solution

The comprehensive solution from Pentaho built on Cisco UCS Big Data Platform helps organizations deploy big data solutions quickly, with validated configurations that scale easily and predictably, as demand dictates. The reference configurations provide an end-to-end solution that has been tested and validated and that enables enterprise customers to quickly integrate big data initiatives into their existing data center operational models.

High Performance and Exceptional Scalability

Cisco UCS unified fabric architecture provides fully redundant, highly scalable lossless 10-Gbps unified fabric connectivity for big data traffic and can easily scale to support a large number of nodes when required by business demands. The advanced management capabilities of Cisco UCS radically simplify this process with a single point of management that spans all nodes in the cluster.

Simplified Management

Big Data analytics implementations tend to involve large numbers of servers. In traditional environments, it can be challenging to manage these large numbers of servers effectively. Cisco UCS Manager delivers unified, model-based management that applies personality and configures server, network, and storage connectivity resources, making it as easy to deploy large numbers of servers as it is to deploy a single server. Additionally, Cisco UCS Manager can perform system maintenance activates such as firmware updates across the entire cluster as a single operation.

Coexistence with Enterprise Applications

In building Big Data solutions that involve Hadoop and/or NoSQL, organizations need ways to transfer data transparently between their enterprise applications and Big Data platforms. This solution can connect, across the same management plane, to other Cisco UCS deployments running enterprise applications, thereby radically simplifying data center management and connectivity. Pentaho Business Analytics provides a comprehensive platform for designing and managing solutions that cross the boundaries of traditional and Big Data platforms. By providing easy-to-use tools and familiar design concepts for both traditional and Big Data platforms, Pentaho empowers organizations to leverage existing IT skillsets to build Big Data solutions.

Rapid Deployment and Growth

Deployment of large numbers of servers can take time. Systems need to be racked, networked, configured, and provisioned before they can be put into use. Cisco UCS Manager uses a model-based approach to provision servers by applying a desired configuration to physical infrastructure quickly, accurately, and automatically. The ability to create consistent configurations improves business agility and eliminates a major source of errors that can cause downtime. Pentaho Business Analytics' tightly integrated platform demystifies the challenges of building end-to-end solutions that take you from data acquisition and processing to rich analytics solutions.

Enterprise Service and Support

Enterprises want know that the vendors providing a solution have the expertise to help them quickly proceed through the design, deployment, and testing of strategic big data initiatives. Businesses also need to have confidence that if a critical system fails, they will be able to get timely and competent support. The joint Cisco Pentaho solution brings together world-class service and support from Cisco and Pentaho.

For More Information

For complete details about Cisco UCS, visit http://www.cisco.com/go/ucs.
For more information about Pentaho Business Analytics for Big Data, visit http://www.pentaho.com/big-data.

To learn more about Pentaho software and services, **contact Pentaho**:

pentaho.com/contact

+1 (866) 660-7555 (worldwide)