

# Cisco Unified Data Center Solutions for MapR: Deliver Automated, High-Performance Hadoop Workloads

## What You Will Learn

MapR Hadoop clusters on Cisco Unified Computing System™ (Cisco UCS®) computing and fabric solutions and Cisco® Tidal Enterprise Scheduler (TES) workload automation solutions provide excellent performance, efficiency, scalability, and automation for your enterprise big data environments.

### Highlights

- Optimized for performance: Cisco® Unified Data Center solutions combined with MapR deliver integrated and highly scalable Hadoop implementations that are specifically engineered to handle the most demanding MapReduce and HBase workloads.
- Advanced Hadoop distribution: MapR brings leading innovation to make Hadoop easy, dependable, fast, and ready for all big data analytics.
- Ease of infrastructure management: Cisco UCS® Manager provides unified, embedded management of all computing, networking, and storage-access resources.
- Enterprise-ready workload automation: Cisco TES delivers transparent and powerful workload automation for MapR Hadoop implementations, providing automated load balancing, data exchange, and advanced event-based scheduling.
- Choice of configurations: The solution provides a choice of Cisco configurations, letting organizations select performance and capacity as their needs dictate.
- Enterprise-class support and services: Reference configurations from Cisco help accelerate implementation of Hadoop deployments and promote their success with worldwide enterprise-class support from Cisco and MapR.

## Challenges

Big data technologies - Apache Hadoop in particular - are being used in numerous applications and evaluated and adopted by enterprises of all kinds. As this important technology helps transform large volumes of data into actionable information, many organizations are struggling to deploy effective and reliable Hadoop infrastructure that performs and scales appropriately for mission-critical applications in the enterprise.

Deployed as part of a comprehensive data center architecture, the Cisco Unified Data Center solution for MapR Hadoop delivers a powerful and flexible infrastructure that increases business and IT agility, reduces total cost of ownership (TCO), and delivers exceptional return on investment (ROI) at scale, while fundamentally transforming the way that organizations do business with Hadoop technology.

## Cisco Unified Data Center Overview

Cisco provides a complete data center solution for big data environments. The combination of Cisco TES and Cisco UCS C-Series Rack Servers addresses the needs of high-performance MapR Hadoop environments, as well as the business and operational requirements of evolving data centers.

This approach offers:

- Infrastructure simplicity and a building-block method to reduce TCO.
- Enhanced business resilience with greater operational continuity.
- Capability to use existing operational models and administrative domains for easy deployment.

- Transparent integration of MapR Hadoop data workloads with Oracle e-Business Suite, Oracle PeopleSoft, and Oracle 10 and 11; Microsoft SQL Server; SAP Enterprise Resource Planning (ERP), Business Information Warehouse, and BusinessObjects; Informatica; Cognos; and other industry-standard integrations, such as web services, Secure Shell (SSH), Simple Network Management Protocol (SNMP), Java Database Connectivity (JDBC), and FTP and Secure FTP (SFTP) file transfer automation.

### **Cisco Unified Data Center Benefits**

- Tested and certified building blocks to reduce cost and risk: Cisco offers infrastructure building blocks that help organizations deploy MapR Hadoop distributions quickly, while scaling configurations rapidly and predictably as demand dictates.
- Powerful and cost-effective Cisco UCS C-Series Rack Servers: You get a choice of Cisco UCS servers that deliver a best-in-class, next-generation platform, which can be deployed quickly and scaled to meet data-processing needs.
- Workload automation with Cisco TES: With an intuitive GUI, Cisco TES makes the process of scheduling advanced workloads easy for operations staff by supporting:
  - Detailed, dependency-based event processing
  - Point-and-click dynamic variables and parameters
  - Scalable, extensible architecture
  - Detailed logging, alerts, and automated remediation
  - Granular security and access-rights management

### **MapR: An Enterprise-Ready Hadoop Platform**

As the technology leader in Hadoop, MapR provides enterprise-class Hadoop solutions that can be developed quickly and administered easily. With significant investment in critical technologies, MapR offers the industry's most comprehensive Hadoop platform, fully optimized for performance scalability. MapR's distribution delivers more than a dozen tested and validated Hadoop software modules over a fortified data platform, offering exceptional ease of use, reliability, and performance for Hadoop solutions.

### **MapR Benefits**

MapR offers:

- Ease of use: The MapR 100-percent Portable Operating System Interface (POSIX)-compliant system allows users to access the Hadoop cluster through industry-standard APIs, such as Network File System (NFS), Open Database Connectivity (ODBC), Linux Pluggable Authentication Modules (PAM), and Representational State Transfer (REST). MapR also provides multitenancy, data-placement control, and hardware-level monitoring of the cluster.
- Reliability: The MapR distribution provides automated "lights-out" data center capabilities for Hadoop. Features include self-healing of the critical services that maintain the cluster nodes and jobs, snapshots that allow point-in-time data recovery, mirroring that allows wide-area intercluster replication, and rolling upgrades that prevent service disruption.

- Performance: To provide superior and exceptional performance over other Hadoop distributions, MapR uses an optimized shuffle algorithm, direct access to the disk, built-in compression, and code written in advanced C++ rather than Java. As a result, the MapR distribution provides better hardware utilization when compared to any other distribution. MapR technology innovations bring these capabilities onto a single data platform built for all big data analytics.
- Broad platform support: The platform supports a wide range of applications that process structured and unstructured data stored in files, as well as NoSQL databases. By using a single data platform for all Hadoop workloads, MapR further solidifies its value proposition - providing excellent ROI for hardware utilization.

## Cisco UCS Solution for MapR

The Cisco UCS solution for MapR is based on the Cisco Common Platform Architecture (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 supplied by the following components:

- Cisco UCS 6200 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management provided for all connected devices 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 typically found in clusters serving big data applications. Cisco UCS Manager helps enable rapid and consistent server configuration using service profiles, which can automate ongoing 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 display alarms and send notifications about the health of the entire cluster.
- Cisco UCS 2200 Series Fabric Extenders extend the network into each rack, acting as remote line cards for fabric interconnects and providing highly scalable and extremely cost-effective connectivity for a large number of nodes.
- Cisco UCS C240 M3 Rack Servers are designed for a wide range of computing, I/O, and storage-capacity demands in a compact two-rack-unit (2RU) design. Cisco UCS C240 M3 servers are powered by dual Intel Xeon processor E5-2600 series CPUs and support up to 768 GB of main memory.

These servers support a range of disk-drive options as well as Cisco UCS virtual interface cards (VICs) optimized for high-bandwidth and low-latency cluster connectivity, with support for up to 256 virtual devices.

High-capacity and high-performance options are available, and reference architectures are available in single-rack and multirack configurations, with considerable capacity built into the Cisco Unified Fabric provided in each rack.

## Managing Big Data Workloads Using Cisco TES

With so much data coming in from so many sources and in such a diversity of formats, any workload management weaknesses can become magnified in a big data environment. As a result, a successful big data solution requires end-to-end business process visibility and transparent data transitions into and out of applications, databases, and processing environments. Big data workload solutions must therefore manage data input and output at every point.

---

Whether the data comes from log files, telemetry feeds, reservation agents, retail consignments, or partner spreadsheet files, one solution, with the following attributes, should handle every piece of incoming data:

- End-to-end visibility: A true enterprise workload automation solution that can scale to meet the demands of a big data environment must be able to see the entire business-processing environment, including activity by business partners using FTP and data exchange mechanisms, ERP processes and databases, data warehouse solutions, data integration activities, and business information solutions. All need to be accessible through a single management console.
- Comprehensive API integration: Enterprise-class workload solutions need a high degree of visibility; they must also be able to interact with each application or database in a comprehensive and easy-to-administer way. Each process step in the entire workload has to be smoothly integrated with the previous one, allowing administrators to quickly define, test, run, and customize the entire job stream.
- Capacity awareness: Data center administrators need to understand what percentage of their resources each business process requires, and which application the data passes through as it enters and exits each phase of business processing. Understanding the volume of data that will emerge from every input source is critical because there is less opportunity to adjust to capacity fluctuations as the amount of data grows. A lack of capacity awareness can increase operating costs and processing errors, while decreasing resource utilization.
- Latency-free throughput: The capability to run business processes in real time is an important complement to capacity awareness because latency is critical to the delivery of big data analytics. As each process is completed, another must start and continue the business-process flow. This flow management traditionally is a labor-intensive, manual process, but it must be automated in a big data environment.
- Predictive workload analytics: Workload analytics give administrators the necessary tools to manage big data business-processing environments error free. Operators must be able to predict and plan for the impact of big data workloads on their environments. Working without this information results in a higher error rate due to disconnected processes, unforeseen effects on resources, and missed or degraded service levels.

## Managing MapR Hadoop Workloads

By configuring, managing, and automating MapR workloads using Cisco TES, the burden of Day 2 operations administration is significantly reduced for Hadoop administrative staff, database administrators, data architects and scientists, and data center operations staff. After the logical data flows are designed and configured and the specific connections are made, the Hadoop team can concentrate on developing MapReduce algorithms (the Hadoop data analysis methodology's key value) and entering the appropriate algorithm into the MapReduce job definition dialog box. Then Cisco TES does the rest. Cisco TES eliminates the bottleneck of scripted workflows that rely on "hope-and-pray" data throughput and uncertainty about the effects of data traffic on the network.

Cisco TES aids prevalidation, sizing, and performance optimization, reducing integration and deployment risk. It also facilitates integration with the existing operational processes of new big data analytics as part of a managed business process - and provides the following advantages:

- Cisco TES allows connection to each MapR instance, creating a true lights-out, end-to-end, policy-based big data analytics engine. All other data source connections and job definitions are created using the same interface, delivering a transparent single-pane-of-glass view for complex workload automation. Managing multiple MapR instances is as simple as defining a new connection. Multiple clusters can be managed in parallel or as parent-children relationships, providing increased flexibility.

- You can define Data Mover jobs as you would define any other job in your job stream, with the added benefit that MapR can store and access data to NFS, HBase, and standard Hadoop Distributed Files System (HDFS) data nodes. The Hadoop job tracker and task tracker process the jobs as they would any other Hadoop job, eliminating the management overhead required for handling static scripts or manually running complex workflows.
- MapReduce job definition and processing are simplified with reduced managed switch nodes and business-processing visibility and automation. As you implement a scaled-out solution, the value of MapReduce automation is magnified. Cisco TES simplifies MapReduce job configuration with an easy-to-use interface, which allows administrators access to cluster-health and job-output information.
- Running Cisco TES for MapR workloads delivers a powerful and flexible Day 2 operations solution that increases IT flexibility and reduces TCO through ease of use and true automation, while fundamentally transforming the way that IT departments manage Hadoop technology.

## Conclusion

Despite the compelling power of big data technology, deployment of a successful, reliable, and high-performance infrastructure for Apache Hadoop can be daunting. The MapR Hadoop distribution provides critical technology advances that help make Hadoop implementation easy, dependable, and fast for enterprise-ready deployments. MapR has made crucial investments to improve the underlying technology, while maintaining the Apache Hadoop API and application compatibility for broad applicability. When MapR and the Cisco Unified Data Center are combined, they bring the power of the MapR platform to a dependable deployment model. This model can be implemented rapidly and customized for either high performance or high capacity, using Cisco Unified Fabric, powerful and efficient Cisco UCS rack servers, and Cisco TES for end-to-end big data workload automation. Whether you are deploying a large data center or buying single racks through the Cisco SmartPlay program, Cisco Unified Data Center solutions for MapR can be sized to meet the challenges of any Hadoop implementation.

## For More Information

For more information, please visit <http://www.cisco.com/go/smartplay>, <http://www.cisco.com/go/bigdata>, <http://blogs.cisco.com/datacenter/cpa/>, and <http://www.cisco.com/go/workloadautomation>.



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](http://www.cisco.com/go/offices).

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: [www.cisco.com/go/trademarks](http://www.cisco.com/go/trademarks). Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)