How Cisco IT Built Big Data Platform to Transform Data Management

EXECUTIVE SUMMARY

CHALLENGE

- Unlock the business value of large data sets, including structured and unstructured information
- Provide service-level agreements (SLAs) for internal customers using big data analytics services
- Support multiple internal users on same platform

SOLUTION

- Implemented enterprise Hadoop platform on Cisco UCS CPA for Big Data - a complete infrastructure solution including compute, storage, connectivity and unified management
- Automated job scheduling and process orchestration using Cisco Tidal Enterprise Scheduler as alternative to Oozie

RESULTS

- Analyzed service sales opportunities in one-tenth the time, at one-tenth the cost
- $40 million in incremental service bookings in the current fiscal year as a result of this initiative
- Implemented a multi-tenant enterprise platform while delivering immediate business value

LESSONS LEARNED

- Cisco UCS can reduce complexity, improves agility, and radically improves cost of ownership for Hadoop based applications
- Library of Hive and Pig user-defined functions (UDF) increases developer productivity.
- Cisco TES simplifies job scheduling and process orchestration
- Build internal Hadoop skills
- Educate internal users about opportunities to use big data analytics to improve data processing and decision making

NEXT STEPS

- Enable NoSQL Database and advanced analytics capabilities on the same platform.
- Adoption of the platform across different business functions.

Enterprise Hadoop architecture, built on Cisco UCS Common Platform Architecture (CPA) for Big Data, unlocks hidden business intelligence.

Challenge

Cisco is the worldwide leader in networking that transforms how people connect, communicate and collaborate. Cisco IT manages 38 global data centers comprising 334,000 square feet. Approximately 85 percent of applications in newer data centers are virtualized and IT is working toward a goal of 95 percent virtualization.

At Cisco, very large datasets about customers, products, and network activity represent hidden business intelligence. The same is true of terabytes of unstructured data such as web logs, video, email, documents, and images.

To unlock the business intelligence hidden in globally distributed big data, Cisco IT chose Hadoop, an open-source software framework that supports data-intensive, distributed applications. “Hadoop behaves like an affordable supercomputing platform,” says Piyush Bhargava, a Cisco IT distinguished engineer who focuses on big data programs. “It moves compute to where the data is stored, which mitigates the disk I/O bottleneck and provides almost linear scalability. Hadoop would enable us to consolidate the islands of data scattered throughout the enterprise.”

To offer big data analytics services to Cisco business teams, Cisco IT first needed to design and implement an enterprise platform that could support appropriate service-level agreements (SLAs) for availability and performance. “Our challenge was adapting the open-source Hadoop platform for the enterprise,” says Bhargava.

Technical requirements for the big data architecture included:

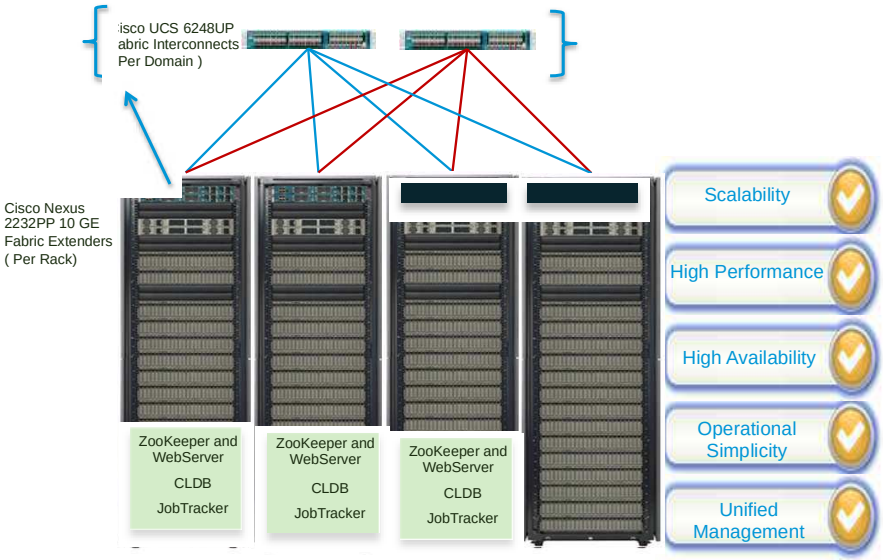
- Open-source components

- Scalability and enterprise-class availability
- Multitenant support so that multiple Cisco teams could use the same platform at the same time
- Overcoming disk I/O speed limitations to accelerate performance
- Integration with IT support processes

Solution

Cisco IT Hadoop platform is build using Cisco® UCS Common Platform Architecture (CPA) for Big Data. High level architecture of the solution is shown in Figure 1.

Figure 1. Cisco IT Hadoop Platform



Cisco IT Hadoop Platform is designed to provide high performance in a multitenant environment, anticipating that internal users will continually find more use cases for big data analytics. “Cisco UCS CPA for Big Data provides the capabilities we need to use big data analytics for business advantage, including high-performance, scalability, and ease of management,” says Jag Kahlon, Cisco IT architect.

Linearly Scalable Hardware with Very Large Onboard Storage Capacity

The compute building block of the Cisco IT Hadoop Platform is the Cisco UCS C240 M3 Rack Servers, 2-RU server power by two Intel Xeon E5-2600 series processors, 256 GB of RAM, and 24 TB of local storage. Out of the 24 TB, Hadoop Distributed File System (HDFS) can use 22 TB, and the remaining 2 TB is available for the operating system.

“Cisco UCS C-Series Servers provide high performance access to local storage, the biggest factor in Hadoop

performance" says Virendra Singh, Cisco IT architect.

> "Cisco UCS C-Series Servers provide high performance access to local storage, the biggest factor in Hadoop performance"
>
> **Virendra Singh, Cisco IT Architect**

The current architecture comprises of four racks, each containing 16 server nodes supporting 384 TB of raw storage per rack. "This configuration can scale to 160 servers in a single management domain supporting 3.8 petabytes of raw storage capacity," says Kahlon.

Low-Latency, Lossless Network Connectivity

Cisco UCS 6200 Series Fabric Interconnects provides high speed, low latency connectivity for servers and centralized management for all connected devices with UCS Manager. Deployed in redundant pairs offers the full redundancy, performance (active-active), and exceptional scalability for large number of nodes typical in big data clusters.

Each rack connects to the fabric interconnects through a redundant pair of Cisco Nexus® 2232PP Fabric Extenders, which behave like remote line cards.

Simple Management

Cisco IT server administrators manage all elements of the Cisco UCS including servers, storage access, networking, and virtualization from a single Cisco UCS Manager interface. "Cisco UCS Manager significantly simplifies management of our Hadoop platform" says Kahlon. "UCS Manager will help us manage larger clusters as our platform grows without increasing staffing." Using Cisco UCS Manager service profiles saved time and effort for server administrators by making it unnecessary to manually configure each server. Service profiles also eliminated configuration errors that could cause downtime.

> "Cisco UCS Manager significantly simplifies management of our Hadoop platform. UCS Manager will help us manage larger clusters as our platform grows without increasing staffing"
>
> **Jag Kahlon, Cisco IT Architect**

Open, Enterprise-Class Hadoop Distribution

Cisco IT uses MapR Distribution for Apache Hadoop, which speeds up MapReduce jobs with an optimized shuffle algorithm, direct access to the disk, built-in compression, and code written in advanced C++ rather than Java.

"Hadoop complements rather than replaces Cisco IT's traditional data processing tools, such as Oracle and Teradata," Singh says. "Its unique value is to process unstructured data and very large data sets far more quickly and at far less cost."

Hadoop Distributed File System (HDFS) aggregates the storage on all Cisco UCS C240 M3 servers in the cluster to create one large logical unit. Then, it splits data into smaller chunks to accelerate processing by eliminating time-consuming extract, transform, and load (ETL) operations. “Processing can continue even if a node fails because Hadoop makes multiple copies of every data element, distributing them across several servers in the cluster,” says Hari Shankar, Cisco IT architect. “Even if a node fails, there is no data loss.” Hadoop senses the node failure and automatically creates another copy of the data, distributing it across the remaining servers. Total data volume is no larger than it would be without replication because HDFS automatically compresses the data.

Cisco Tidal Enterprise Scheduler (TES)

For job scheduling and process orchestration, Cisco IT uses Cisco TES as a friendlier alternative to Oozie, the native Hadoop scheduler. Built-in Cisco TES connectors to Hadoop components eliminate manual steps such as writing Sqoop code to download data to HDFS and executing a command to load data to Hive. “Using Cisco TES for job scheduling saves hours on each job compared to Oozie because reducing the number of programming steps means less time needed for debugging,” says Singh.

Another advantage of Cisco TES is that it operates on mobile devices, enabling Cisco end users to execute and manage big data jobs from anywhere.

> “Using Cisco TES for job scheduling saves hours on each job compared to Oozie because reducing the number of programming steps means less time needed for debugging.”
>
> **Virendra Singh, Cisco IT Architect**

Looking Inside a Cisco IT Hadoop Job

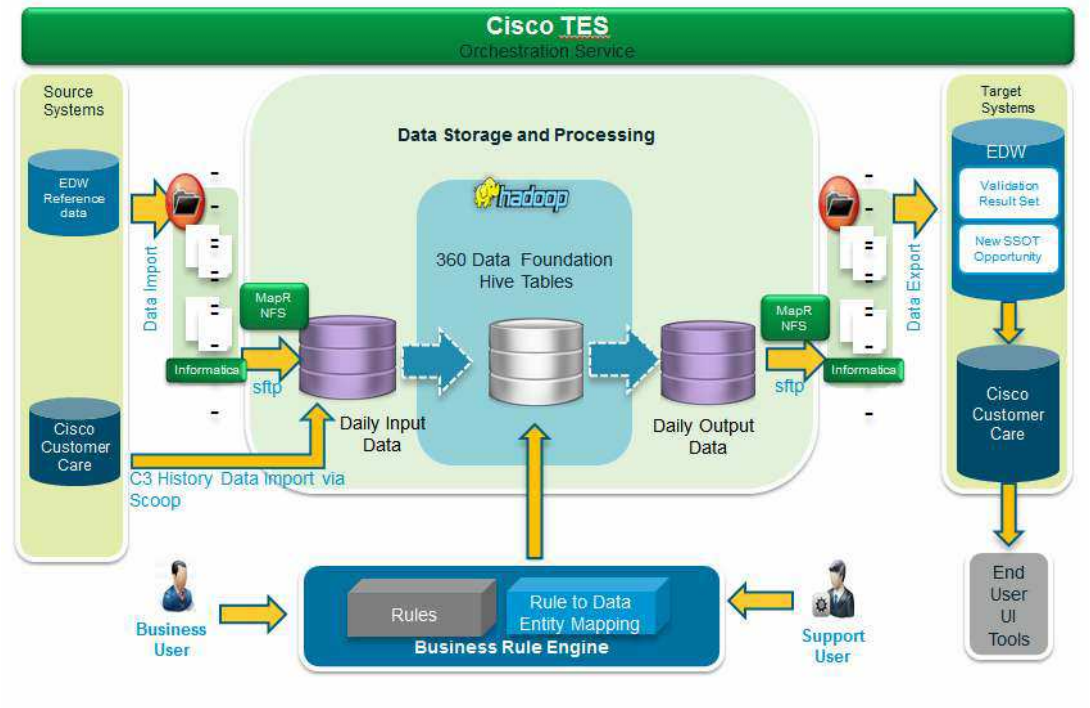
The first big data analytics program in production at Cisco helps to increase revenues by identifying hidden opportunities for partners to sell services. “Previously, we used traditional data warehousing techniques to analyze the install base that identified opportunities for the next four quarters,” says Srini Nagapuri, Cisco IT project manager. “But analysis took 50 hours, so we could only generate reports once a week.” The other limitation of the old architecture was the lack of a single source of truth for opportunity data. Instead, service opportunity information was spread out across multiple data stores, causing confusion for partners and the Cisco partner support organization.

The new big data analytics solution harnesses the power of Hadoop on the Cisco UCS CPA for Big Data to process 25 percent more data in 10 percent of the time. The data foundation includes the following:

- Cisco Technical Services contracts that will be ready for renewal or will expire within five calendar quarters
- Opportunities to activate, upgrade, or upsell software subscriptions within five quarters
- Business rules and a management interface to identify new types of opportunity data
- Partner performance measurements, expressed as the opportunity-to-bookings conversion ratio

Figure 2 and Table 1 show the physical architecture, and Figure 3 shows the logical architecture.

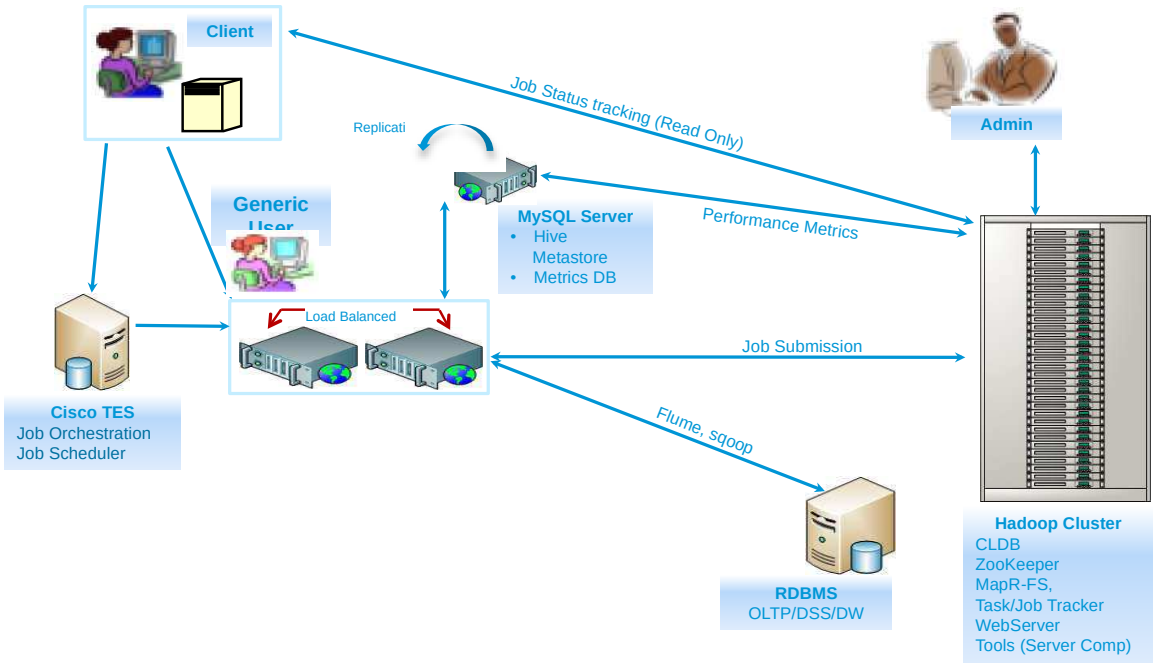
Figure 2. Physical Architecture for Hadoop Platform to Identify Partner Sales Opportunities



Software Components in Cisco IT's Big Data Analytics Platform

Component	Function
MapReduce	Distributed computing framework. Data is processed on the same Cisco UCS server where it resides, avoiding latency while data is accessed over the network.
Hive	SQL-like interface to support analysis of large data sets and data summarization.
Cisco Tidal Enterprise Scheduler (TES)	Workload automation, job scheduling, event management and process orchestration.
Pig	Data flow language. Enables Cisco IT to process big data without writing MapReduce programs.
HBase	Columnar database built on top of HDFS for low-latency read and write operations.
Sqoop	Tool to import and export data between traditional databases and HDFS.
Flume	Tool to capture and load log data to HDFS in real time.
ZooKeeper	Coordination of distributed processes to provide high availability.

Figure 3. Cisco Hadoop Logical Architecture



Results

Cisco IT has introduced multiple big data analytics programs, all of them operating on the Cisco® UCS Common Platform Architecture (CPA) for Big Data.

Increased Revenues from Partner Sales

The Cisco Partner Annuity Initiative program is in production. The enterprise Hadoop platform accelerated processing time for identifying partner services sales opportunities from 50 hours to 6 hours, and identifies opportunities for the next five calendar quarters instead of four quarters. It also lets partners and Cisco employees dynamically change the criteria for identifying opportunities. “With our Hadoop architecture, analysis of partner sales opportunities completes in approximately one-tenth the time it did on our traditional data analysis architecture, and at one-tenth the cost,” says Bhargava.

> “With our Hadoop architecture, analysis of partner sales opportunities completes in approximately one-tenth the time it did on our traditional data analysis architecture, and at one-tenth the cost.”
>
> **Piyush Bhargava, Cisco IT Distinguished Engineer**

Business benefits of the Cisco Partner Annuity Initiative include:

- Generating an anticipated US$40 million incremental revenue from partners in FY13: “The solution processes 1.5 billion records daily, and we identified new service opportunities the same day we placed

the system in production," says Nagapuri. Cisco is on track to reach the revenue goal.

- Improving the experience for Cisco partners, customers, and employees: Consolidating to a single source of truth helps to avoid the confusion on available service opportunities.
- Creating the foundation for other big data analytics projects: Moving the customer install base and service contracts to the Hadoop platform will provide more value in the future because other Cisco business teams can use it for their own initiatives.

Increased Productivity by Making Intellectual Capital Easier to Find

Many of the 68,000 employees at Cisco are knowledge workers who tend to search for content on company websites throughout the day. But most of the content is not tagged with all relevant keywords, which makes searches take longer. "People relate to content in more ways than the original classification," says Singh. In addition, employees might not realize content already exists, leading them to invest time and money recreating content that is already available.

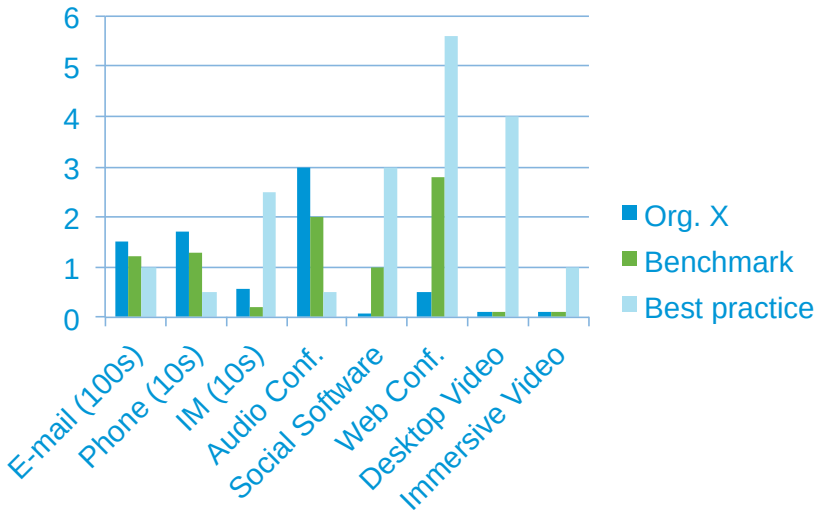
To make intellectual capital easier to find, Cisco IT is replacing static, manual tagging with dynamic tagging based on user feedback. The program uses machine-learning techniques to examine usage patterns and also acts on user suggestions for new tags.

The content auto-tagging program is currently in proof-of-concept, and Cisco IT is creating a Cisco Smart Business Architecture (SBA) for other companies.

Measured Adoption of Collaboration Applications and Assessed Business Value

Organizations that make significant investments in collaboration technology generally want to measure adoption and assess business value. The Organization Network Analytics program is currently a proof-of-concept. Its goal is to measure collaboration within the Cisco enterprise, develop a benchmark, and present the information on an intuitive executive dashboard (Figure 4). "Our idea is to identify deviations from best practices and to measure effectiveness of change management," Singh says.

Figure 4. Collaboration Benchmark Usage Analysis – Sample Report



The Hadoop platform analyzes logs from collaboration tools such as Cisco Unified Communications, email, Cisco TelePresence®, Cisco WebEx®, Cisco WebEx Social, and Cisco Jabber™ to reveal preferred communications methods and organizational dynamics. When business users studying collaboration enter analysis criteria, the program creates an interactive visualization of the social network.

Next Steps

Cisco IT continues to introduce different big data analytics programs and add more types of analytics. Plans include:

- Identifying root causes of technical issues: The knowledge that support engineers acquire can remain hidden in case notes. “We expect that mining case notes for root causes will unlock the value of this unstructured data, accelerating time to resolution for other customers,” says Nagapuri The same information can contribute to better upstream systems, processes, and products.
- Analyzing service requests, an untapped source of business intelligence about product usage, reliability, and customer sentiment: “A deeper understanding of product usage will help Cisco engineers optimize products,” Singh says. The plan is to mash up data from service requests with quality data and service data. The underlying big data analytics techniques include text analytics, entity extraction, correlation, sentiment analysis, alerting, and machine learning.
- Supporting use cases where NoSQL provides improved performance or scalability compared to traditional relational databases.
- Scaling the architecture by adding another 60 nodes: Cisco IT is deciding whether to add the nodes to the same Cisco UCS cluster or build another cluster that connects to the first using Cisco Nexus switches.

Lesson Learned

Cisco IT shares the following observations with other organizations that are planning big data analytics programs.

Technology

- Hive is best suited for structured data processing, but has limited SQL support.
- Sqoop scales well for large data loads.
- Network File System (NFS) saves time and effort for data loads.
- Cisco TES simplifies job scheduling, process orchestration and accelerates debugging.
- Creating a library of user-defined functions (UDF) for Hive and Pig helps to increase developer productivity.
- Once internal users realize that IT can offer big data analytics, demand tends to grow very quickly.
- Using Cisco® UCS Common Platform Architecture (CPA) for Big Data, Cisco IT built a scalable Hadoop platform that can support up to 160 servers in a single switching domain.

Organization

- Build internal Hadoop skills. “People keep identifying new use cases for big data analytics, and building internal skills to implement the use cases will help us keep up with demand,” says Singh.
- Educate internal users that they can now analyze unstructured data (email, webpages, documents, and so on) in addition to databases.

For More Information

Cisco IT case studies on a variety of business solutions, visit Cisco on Cisco: Inside Cisco IT

- www.cisco.com/go/ciscoit

Cisco® UCS CPA for Big Data:

- blogs.cisco.com/datacenter/cpa/
- http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/wp_greenplum.pdf
- www.cisco.com/go/bigdata

Cisco Tidal Enterprise Scheduler:

- www.cisco.com/go/workloadautomation

Note

This publication describes how Cisco has benefited from the deployment of its own products. Many factors may have contributed to the results and benefits described; Cisco does not guarantee comparable results elsewhere.

CISCO PROVIDES THIS PUBLICATION AS IS WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Some jurisdictions do not allow disclaimer of express or implied warranties, therefore this disclaimer may not apply to you.

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.**