Powering the Internet of Everything with Big Data Analytics

How Cisco IT defined an information architecture that streamlines access to big data and analytics tools

By Piyush Bhargava

Distinguished Engineer, Cisco IT

Abstract

The Internet of Everything (IoE) means billions of new network connections generating a deluge of data. Enterprises can convert this deluge into accessible big data with the right kind of technology, organization, and architecture for analytics.

This article discusses how Cisco IT is preparing to deliver the analytics capabilities necessary to make IoE data a useful and valuable business asset. It is the second article in a two-part series about the impact of IoE-generated big data in the enterprise. The first article, The Internet of Everything and Big Data for the Connected Enterprise, describes the IoE opportunity and its impact on several key business functions.

Defining an Information Architecture for the Connected Enterprise

Connected enterprises are already capturing and generating more data from more sources—at a growth rate that is increasing exponentially. This growth in data is coinciding with a growing demand for analytics and a growing number of enterprise systems and users looking to consume that data.

To meet the data demands of today and to realize the promise of IoE and big data, enterprises need a clear strategy to evolve their existing information architecture. For Cisco IT, developing this strategy meant addressing several questions:

- What are the drivers for creating a big data and analytics architecture?
- How are we approaching the architecture challenge today?
- How do we match the diverse business use cases with the right technologies for big data and analytics?
- What changes will be required in employee skill sets and the IT organization?
- What challenges cannot be addressed with our current IT architecture?
- How will our IT infrastructure need to evolve in order to truly capture the IoE opportunity?

The IT strategy should also consider how to manage growth in the volume, variety, and velocity of data. Given that growth, both IT and business users seek new solutions for managing and making sense of data. Typically, the IT department assumes this solution will be delivered as an IT service that is built from a combination of commercial products and internally developed software. But commercial Software as a Service (SaaS) and cloud offerings also make it easy for users and business leaders to build their own point solutions for data management and analytics. And users may do so if the IT department doesn't offer credible alternatives at a reasonable cost and with an aggressive delivery timeframe.

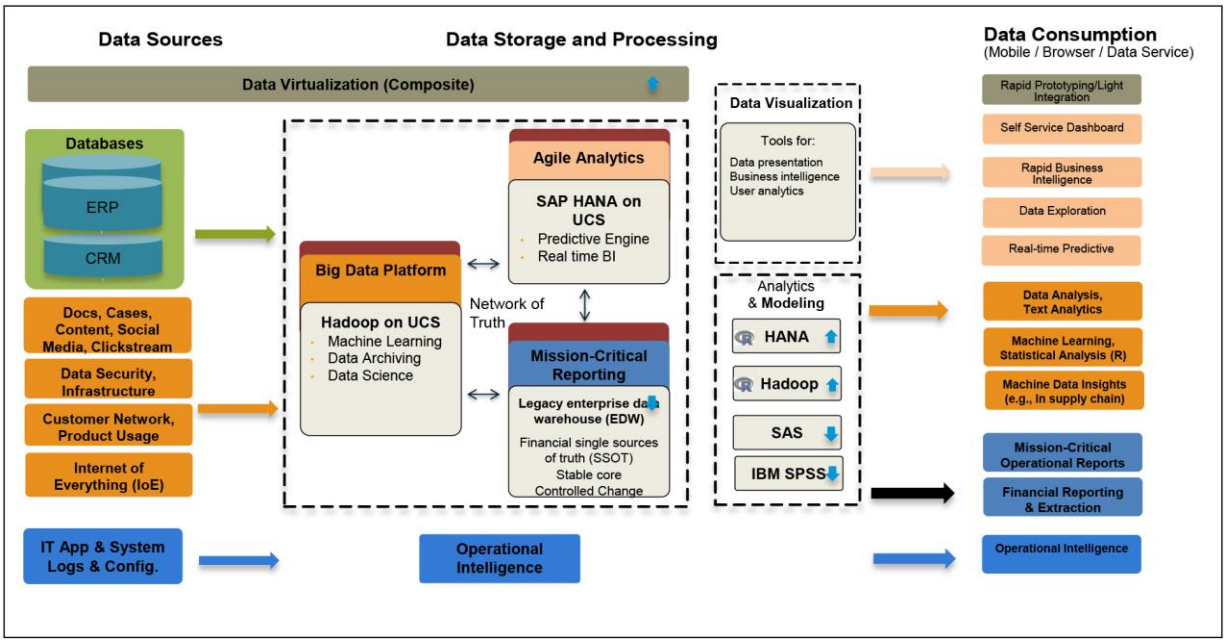
However, the IT challenge posed by the demands of IoE and big data isn't merely about implementing new technologies within the enterprise. A larger challenge is to integrate the new technologies with the existing architecture and have a coherent strategy for matching the right technology to each business driver. (Table 1)

Table 1. Technologies Used for Delivering IT Services

Business Drivers	IT Technology Strategy
New data types and processing needs	Deploy new data processing technologies, e.g., Hadoop
User experience: Performance, visualization, bring your own device (BYOD)	Precompute selected data on Hadoop or with in-memory technologies
Business velocity	Support user self-service, agile analytics, BYOD
Demand for analytics	Use machine learning, statistical computing with R language, predictive analytics
Data governance	Maintain metadata management, network of truth

An information architecture for the IoE and big data must meet two requirements. First, the architecture must support information as it flows seamlessly across different systems and technologies. Second, it must provide a blueprint for introducing new technologies while also making use of legacy investments when appropriate. Cisco IT addressed these requirements by defining a reference architecture for data platforms in a connected enterprise. (Figure 1)

Figure 1. Cisco IT Data Platforms Reference Architecture for the Connected Enterprise



As shown in Figure 1, data sources at Cisco include enterprise applications using traditional database technologies, application data from SaaS providers, clickstream data from the Cisco.com web platform, system log files, documents, and content of various types. The key change in data sources due to IoE is the addition of unstructured data such as documents, web content, and logs.

Cisco IT data storage and processing platforms consist of a Massive Parallel Processing (MPP) platform, a SAP HANA in-memory analytics platform, a Hadoop big data platform, and an operational intelligence platform. Prior to 2012, we used only the MPP platform; the IoE has generated the need for the additional data platforms.

We also use multiple platforms to meet the requirements of different data use cases and consumption patterns across the company. These platforms include the Cisco® Data Visualization Suite, Apache Hadoop, and SAP HANA.

Cisco Data Virtualization Suite

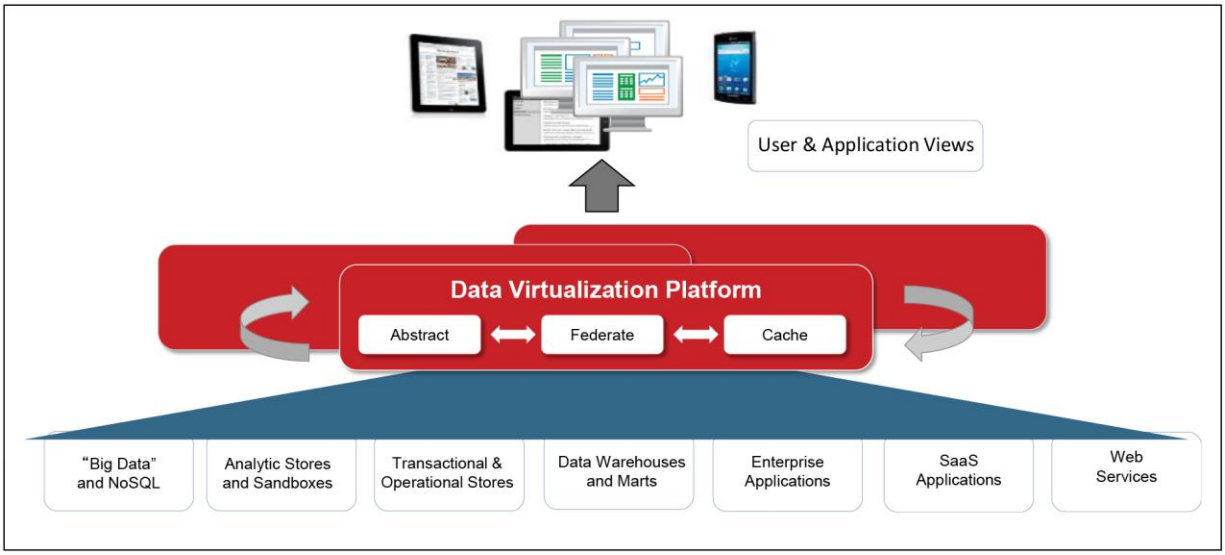
A constant challenge facing IT and business teams working with data is the amount of time and resources required to finalize architecture requirements around data processing, analytics, and consumption patterns. The reason for this challenge is that the business needs to work with the data to identify the relevant nuggets of information, yet the architecture requirements often change after the business sees that processed data.

Another challenge is that the IT process - from finding initial requirements to delivering a prototype analytics tool—can be lengthy and expensive. Time delays can mean that an internal analytics project has a high chance of failing or not being completed because it no longer meets business needs.

From the IT department's perspective, the reason the virtualization process takes so long is that data needs to be extracted from multiple applications, each with its own data model, into a common data model in the reporting application. This extraction is a manual and cumbersome process that requires significant expertise in data modeling, data integration, and business processes. Several iterations of the requirements and modeling effort may be necessary in order to produce a data model that is suitable for user applications. Often so much investment goes into this data preparation phase that the business objectives of delivering data to applications for user consumption are left unrealized.

We use the Cisco Data Virtualization Suite as a data integration technology that creates a logical abstraction for data collected from multiple, disparate, internal and external sources. The Cisco Data Virtualization Suite connects to data sources, executes queries to retrieve requested data, combines or federates data from these sources, and delivers the result to consuming applications. User applications have a single, logical place to go for retrieving data that has already been prepared for viewing from a business perspective. (Figure 2)

Figure 2. Logical Abstraction in a Data Virtualization Platform



Data virtualization offers distinct advantages for data modeling and access:

- Data isn't moved or replicated so it preserves a high degree of consistency.
- Logical data modeling can help create an abstraction that combines data from multiple sources in a single model.
- The data supply chain shrinks dramatically. The focus shifts from moving data between disparate data models and technologies to logically integrating data.
- Latency issues are limited because data at the consuming application reflects the latest data from the source.

We are applying the Cisco Data Virtualization Suite for the following use cases:

- **Application development.** Instead of bulky processes and change controls, data virtualization enables iterations in data integration and consumption patterns for applications. In turn, this iterative approach enables developers to rapidly identify data requirements and create prototype applications.
- **Low-volume data.** Data virtualization is the ideal way to meet business needs if an application integrates or analyzes only a small volume of data.
- **Cost control.** We want users and applications to migrate to lower-cost platforms as the performance and capabilities of those platforms improve. Yet the change management and coordination with users for that migration can be a daunting and expensive effort. Providing continuous data access via data virtualization can help isolate users from changes in the underlying application platform.

Enterprise Hadoop Platform

Apache Hadoop is an open-source framework for developing distributed, data-intensive applications based on the concepts of Google Map Reduce and Google File System (GFS). Hadoop is one of the most disruptive data technologies today and it offers enterprises a new and alternative way to store, process, and manage enterprise data.

Cisco IT chose Hadoop because of the following benefits:

- **Low costs.** Hadoop runs on relatively low-cost servers with local storage. We have implemented the Hadoop platform on Cisco Unified Computing System™ (Cisco UCS) rack servers with built-in storage.
- **Scalability.** A simple programming model and a very efficient framework design allows Hadoop to scale linearly as additional nodes are added to the server cluster. Hadoop abstracts distributed computing aspects from the developer so no programming changes are required as the amount of data grows. The IT department simply adds nodes to the cluster.
- **Flexibility.** Unlike traditional databases, Hadoop doesn't require a pre-defined schema to store the data. Data can be stored in any format and a schema can be applied at the time of processing. This flexibility makes Hadoop suitable for storing and processing unstructured data. It also allows for multiple uses of the same data without requiring any changes to how the data was originally stored.
- **Throughput performance.** Hadoop was designed for maximizing the data throughput of batch processing. Large data processing that may take hours or days on traditional databases can be processed in significantly less time on a Hadoop cluster.

These benefits have also prompted the open-source community and commercial vendors to innovate on Hadoop and build an associated ecosystem of developer tools. Along with this ecosystem, Hadoop can be used for data storage, data processing, structured query language (SQL) processing, machine learning, graph processing, transactional processing on NoSQL databases, and other processing requirements that can be executed as a software program. (Table 2)

Table 2. Hadoop Evolution to Support Many Types of Data and Processing

Data Processing Function	Tools and Applications
Application logic	Application code
Stream and publish/subscribe	Storm, Kafka, Flume
Data integration	Sqoop and third-party applications
Graph processing	Giraph, Titan, GraphX
Machine learning	Mahout, R language
SQL support	Low-latency SQL: Impala, Presto, Drill, Tez; SQL Hive
NoSQL database	HBase
Processing framework	Hadoop YARN and Map Reduce
Data platform	Hadoop infrastructure implementation on Cisco UCS servers for management, distributed file system, security

Due to the flexible nature and potential benefits of Hadoop, almost every IT organization is seeking to build a Hadoop infrastructure for big data. If enterprises do not take this type of architectural approach, they face the risk of managing multiple Hadoop infrastructures along with creating fragmented data collections and insights. Using a defined architecture, Cisco IT created a multi-tenant enterprise Hadoop platform in 2012 to support use cases for multiple company functions.

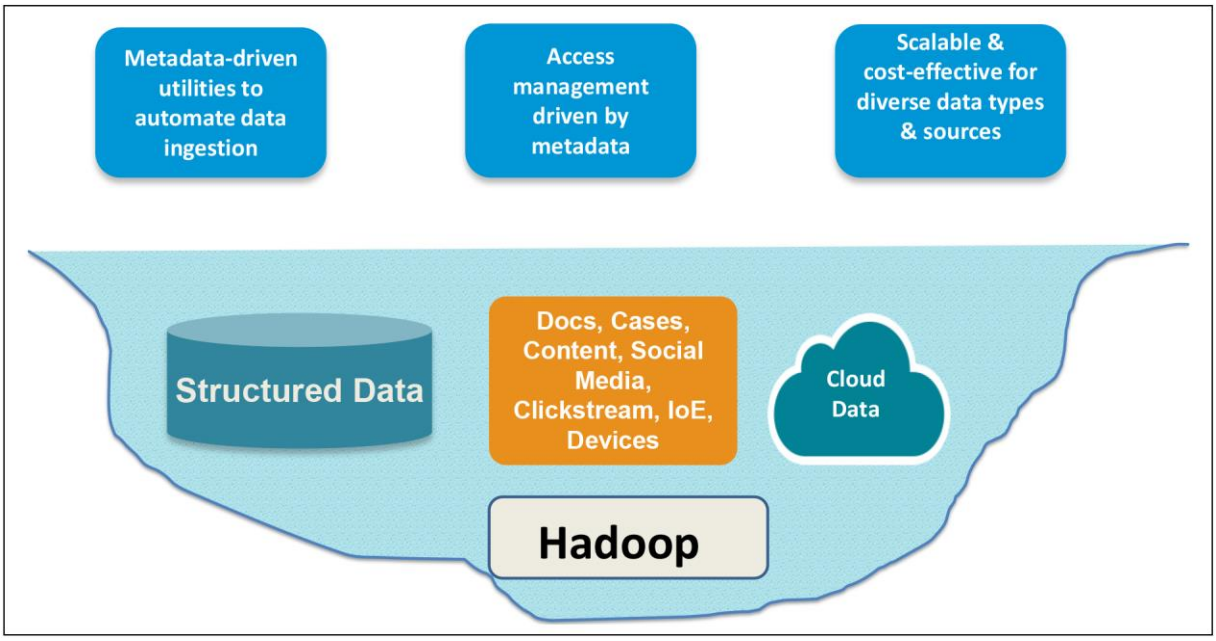
A primary use case for Hadoop within Cisco IT is extract/transform/load (ETL) offload from our legacy MPP platform. ETL data processing is required to prepare data for analytics. This process is hidden from the user and takes up 30-50 percent of the capacity in a data warehouse. Moving ETL processing to Hadoop reduces the time required and frees precious capacity on our legacy MPP platform. More details about this implementation is presented in the Big Data Analytics case study.

Initially, our enterprise Hadoop platform provided a shared server infrastructure and operations model for users. However, each use case had its own data repository and was responsible for bringing data into Hadoop and managing its access and processing. This architecture led to duplication, redundancy of common data elements, and precluded any opportunities to share data. We quickly realized that our vision for shared data and insights would not be realized unless we had a better way to manage how data is ingested into the Hadoop platform, how it is cataloged, and how the user and application access is provisioned.

The solution for shared data management was establishing an enterprise data lake. (Figure 3) A data lake is defined as a large repository that stores data from multiple sources in its native raw format, using three key components:

- Automated, metadata-driven ingestion of data that eliminates the need for each application to code its own way of bringing in data
- Policy and workflow-driven access management
- Support for a variety of workloads and access patterns on this shared data

Figure 3. Enterprise Data Lake on Hadoop



In addition to the use case for shared data management, other Cisco IT use cases for Hadoop include:

- **Data Archiving.** Old data from our enterprise data warehouse is archived to Hadoop to simplify access.
- **Unstructured data.** Any data from a non-database source that requires processing uses the Hadoop platform. Data that is considered unstructured often in fact has a well-defined structure; the issue is that it is not a structure defined by a data modeler. Common examples of unstructured data are documents, web content, system logs such as clickstream logs, emails, social media, and source code.
- **Machine learning or statistical processing.** These use cases often require significant data processing and Hadoop provides the scale necessary to meet application objectives.
- **Analytics.** Network, collaboration, and security analytics currently leverage the Hadoop platform.

SAP HANA on Cisco UCS Servers

The combination of SAP HANA and Cisco UCS servers provides high-performance, in-memory analytics tools. This combination operates as a MPP database system with columnar technologies, advanced data compression, and a built-in analytics engine.

Traditional database systems were designed for optimizing performance by working around the constraints of disk I/O bottlenecks and limited memory on the hardware. By running fully in the server's main memory, SAP HANA eliminates the disk I/O constraint, an important factor when working with huge, continually changing data sets.

The high-performance SAP HANA architecture eliminates the need to create performance structures like indexes and aggregate tables. This architecture also gives Cisco several distinct advantages:

- Business users can directly access hundreds of millions of data rows and get sub-second response time without the need for IT to optimize access.
- Real-time analytics are possible because the steps required to aggregate data at a summary level are no longer needed.
- Fewer materialized tables translate into the need for fewer IT resources and improved agility to respond faster to business information demands.

Cisco IT has positioned SAP HANA for use cases that meet the following characteristics:

- High-value data uses that demand very fast performance, aggregation of large datasets across multiple dimensions, and a high degree of change in business requirements.
- Use cases that need near-real-time analytics where data is frequently changing and the business depends on access to the latest data.

The first Cisco IT use case that went live on SAP HANA was the Dynamic Insights for Sales Executives (DISE) platform. DISE provides field sales executives with near-real-time access to sales bookings, opportunities, and forecast data. The solution improves sales productivity by streamlining the weekly sales forecasting process. Cisco sales executives also credit DISE with improving the accuracy of sales forecasts and providing better quality data.

The other prominent Cisco IT use case on SAP HANA is for the supply chain, providing near-real-time data access for improving sales order visibility across the entire supply chain process. This use case requires connecting data sets from contract manufacturers, logistics partners, and multiple internal systems to provide accurate visibility into the latest status on every order. It also requires the ability to initiate a workflow based on predictive analytics in the event of manufacturing anomalies and delays.

Operational Intelligence Platform

Running an effective IT operation with a focus on minimizing service downtime and improving the experience of the business is a top priority for every IT organization. Health dashboards are examples of monitoring tools that help IT manage the availability and performance of applications, services, and systems. These monitoring tools have been implemented either as part of individual, siloed technology deployments or as part of comprehensive, third-party monitoring solutions. In both cases, these tools have almost always failed to keep up with the changing landscape of IT.

Big data technologies have also disrupted these critical monitoring capabilities by providing an alternative way to collect the intelligence that improves IT operations. Error messages and alerts are often available in the log files written by different physical or virtual devices, applications, or other elements in the data center. Multiple technologies are available to collect, index, and search these logs to identify issues and provide the intelligence that helps improve IT operations.

For delivering this type of critical intelligence to our IT operations teams, we use an operational intelligence platform to index a broad range of system logs for networking devices, operating systems, storage resources, databases, unified communications and video events, and applications. Based on the indexing, the platform correlates events and generates two types of alerts: exceptions and warnings. These alerts are routed to the appropriate teams by email and appear on our internally developed EventPro dashboard for immediate analysis, research, and action. The dashboard presents both virtual and physical stacks in a single view, giving Cisco IT operations teams a single place to look for issues and to work together on resolution.

The operational intelligence platform gives us the ability to:

- Implement self-servicing and self-healing capabilities for infrastructure elements, which increases the availability of IT services
- Centrally monitor, alert, report, and analyze metrics, logs, and events in real time across all physical, virtual, and cloud resources to reduce Mean Time to Detect (MTTD) for active or potential problems
- Correlate and connect events across every level and technology in a Cisco UCS domain
- Proactively predict and detect infrastructure performance problems and prevent them from affecting users
- Determine the root causes of outages and performance problems to reduce Mean Time to Repair (MTTR)
- Facilitate real-time reporting when changes are made to mission-critical systems
- Provide transparency for system issues across all IT teams

Conclusion

For an IT department, big data will be perhaps the largest and most complex challenge that comes with the Internet of Everything. To deliver true business value, IT must be able to manage, process, and make that data available in a reliable and relevant form to many users and applications. As we have found, IT can meet this challenge by defining a new information architecture that supports the ever more connected enterprise.

For More Information

- To read the first article in this series visit The Internet of Everything and Big Data for the Connected Enterprise
- Cisco Data Virtualization Suite
- Cisco UCS servers
- Read the Cisco IT case study on the SAP HANA deployment and the DISE business intelligence platform.
- Read the Cisco IT case study on Big Data Analytics
- To read additional Cisco IT case studies on a variety of business solutions, visit Cisco on Cisco: Inside Cisco IT.
- To view Cisco IT webinars and events about related topics, visit Cisco on Cisco Webinars & Events.

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.