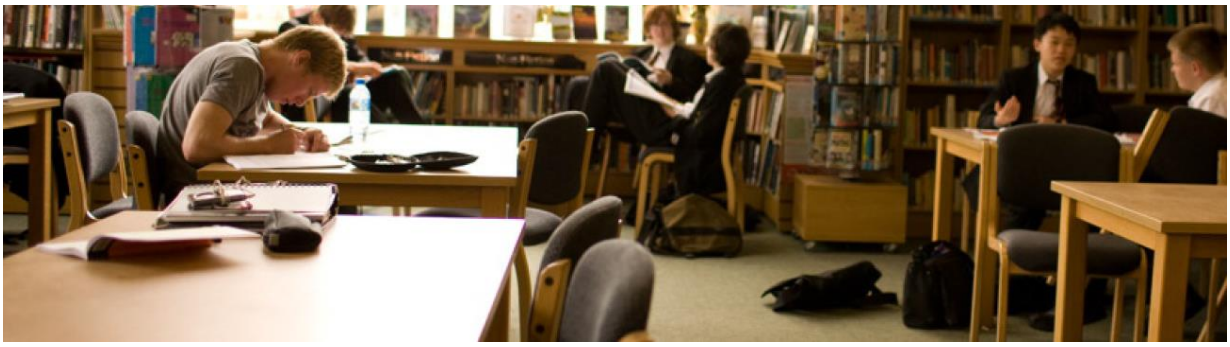
White Paper

Optimize Your Data Warehouse with a Joint Solution from Cisco and Informatica



Today's information-based business culture challenges organizations to integrate data from a wide variety of sources to improve customer acquisition and retention, increase operation efficiency, strengthen product and service delivery, and enter new markets.

Today's information-based business culture challenges organizations to integrate data from a wide variety of sources to improve customer acquisition and retention, increase operation efficiency, strengthen product and service delivery, and enter new markets.

To meet these goals, enterprises demand clean, secure, accessible, timely, actionable data—plus analytical environments that can scale to accommodate the tremendous growth in data volume, variety, and speed while also handling diverse types of enterprise data processing workloads. Big data platforms such as Apache Hadoop have emerged to complement the data warehouse and address new requirements for scalability. In contrast to the traditional "scale up and scale out" approach—adding infrastructure using high-end servers or appliances with expensive shared storage (for example, SAN or network-attached storage [NAS])—newer data architectures are enabling organizations to ingest and process greater volumes of data more cost effectively.

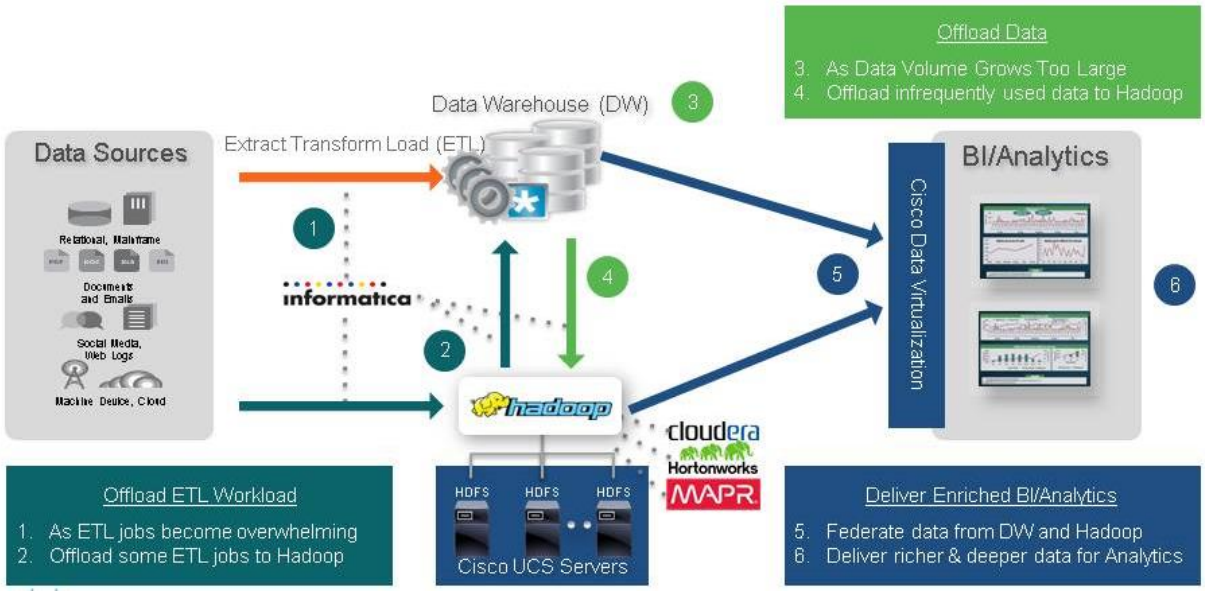
By using Hadoop and related tools to optimize the data warehouse, organizations can gain the following benefits:

- Infrequently used data can be offloaded to more cost-effective storage so that data warehouse storage is optimized.
- A greater variety and volume of data can be ingested and stored to derive new business insights. In contrast to traditional approaches, which provide data this is summarized or aggregated, data can be stored in raw form for more precise insights.
- Data transformation (that is, extract, load, and transform [ELT]) and data quality workloads can be transitioned off so that data warehouse CPU capacity is optimized.
- Network performance is no longer a bottleneck with a distributed environment that pushes processing to the data.
- Unstructured and semi-structured data can be easily stored and manipulated.

Solution Overview

To help customers quickly capture the value of Hadoop, Cisco and Informatica are offering a joint data warehouse optimization solution (Figure 1). This solution provides a single platform for offloading processing and storage from data warehouses to Hadoop. It consists of the Cisco Unified Computing System™ (Cisco UCS®), Informatica Big Data Edition, a Hadoop distribution of the customer's choice, and Cisco® Data Virtualization. The joint offering provides the services to help customers analyze their current data warehouse utilization and assess opportunities for offloading. Where potential savings are identified, Informatica Big Data Edition enables customers to run data transformation and data quality processes using a simple visual development environment on Hadoop clusters installed on Cisco UCS servers. The distributed data environments can be federated using Cisco Data Virtualization to provide business intelligence and analytics with a single point of access to all data.

Figure 1. Data Warehouse Optimization Architecture



Products

Cisco UCS Intergrated Infrastucture for Big Data

Cisco UCS changes the way organizations do business through policy-based automation and standardization of IT processes. Cisco UCS combines industry-standard x86-architecture servers, networking, and enterprise-class management into a single system. The system's configuration is entirely programmable using unified, model-based management to simplify and accelerate deployment. A unified I/O infrastructure uses a high-bandwidth, low-latency unified fabric to support networking, storage I/O, and management traffic. The system uses a wire-once architecture with a self-aware, self-integrating, intelligent infrastructure that eliminates the time-consuming, manual, error-prone assembly of components into systems.

The Cisco UCS Intergrated Infrastucture for Big Data, the third geneartion of Cisco Common Platform Architecture (CPA) for Big Data, is a highly scalable architecture designed to meet a variety of scale-out application demands. The solution has been widely adopted for agriculture, education, finance, healthcare, service provider, entertainment, insurance, and public-sector environments. The Cisco UCS

Integrated Infrastructure solution improves performance and capacity. It also offers additional complete solutions with industry-leading partnerships.

With complete, easy-to-order packages that include compute, storage, connectivity, and unified management features, Cisco UCS Integrated Infrastructure for Big Data accelerates deployment, delivers predictable performance, and reduces total cost of ownership (TCO).

Informatica Big Data Edition

Informatica Big Data Edition and Hadoop enable organizations to integrate and prepare greater volume and variety of data faster and more economically. Informatica simplifies the ingestion of all types of data, including customer transactions and interactions, mainframe data, server logs, and sensor data, in batches and in real time, using hundreds of prebuilt connectors. Informatica supports high-speed data ingestion to Hadoop using native connectivity to applications, databases, machines, social media, and prebuilt parsers for industry formats including FIX, SWIFT, HL7, HIPAA, EDI, and ASN.1. Informatica also supports multiple styles of data ingestion, including batch processing, log-based replication, change-data capture, and real-time streaming.

Informatica developers can then profile, parse, integrate, cleanse, and refine the data up to five times faster using a simple visual development environment, prebuilt transformations, and reusable business rules. Because data pipelines are optimized and abstracted from underlying processing technology, developers can adopt new technology innovations without the risk of needing to redo work as Hadoop continues to change and evolve. With more than 100,000 developers skilled in Informatica technology, organizations can use existing skills and accelerate project staffing.

Visual development environment: Most Hadoop development today is performed manually in a manner similar to the way that ETL code was developed a decade ago, before ETL tools such as Informatica PowerCenter were created. Graphical codeless development has already proven to reduce development time by as much as 500 percent while identifying data errors not caught by manual Hadoop coding.

- Universal data access: Organizations use Hadoop to store and process a variety of diverse data sources and often face challenges in combining and processing all relevant data from their traditional data sources along with new types of data. Informatica Big Data Edition helps organizations easily and reliability perform preprocessing and postprocessing of data into and out of Hadoop.
- High-speed data ingestion: Access, load, transform, and extract big data to and from source and target systems or directly into Hadoop or your data warehouse using native connectivity to applications, databases, machines, social media, and prebuilt parsers for industry-standard formats. Collect log files and machine and sensor data in real time and reliably stream data at scale directly into Hadoop.
- Comprehensive data transformation: Hadoop excels at storing diverse data, but the capability to derive meaning and make sense of it across all relevant data types is a major challenge. Informatica Big Data Edition provides an extensive library of prebuilt transformation capabilities on Hadoop for data integration, data parsing, and data quality. Integrate, cleanse, and prepare all types of structured and unstructured data natively in Hadoop.
- Holistic data governance: Profile data on Hadoop to understand the data, identify data quality issues, and collaborate on data pipelines. Automate the discovery of data domains and relationships on Hadoop such as sensitive data that needs to be protected. Help ensure complete transparency of data pipelines with

end-to-end data lineage of all data movement from source data, through Hadoop, and to target applications.

Cisco Data Virtualization

Cisco Data Virtualization federates disparate data, abstracts and simplifies complex data, and delivers the results as data services or relational views (that is, logical business views) to consuming applications such as business intelligence and analytics tools and other information dashboards. With advanced query optimization technology, it delivers extremely high performance.

After the selected warehouse data has been offloaded to Hadoop, Cisco Data Virtualization is used to federate both data sources and offer a single view of the data. Analytics and business intelligence reports are enriched because they now have access to more data from the warehouse as well as the from Hadoop big data store.

The main features of Cisco Data Virtualization include:

- Data access: Connect and expose data from diverse sources.
- Data federation: Process and optimize queries across the data warehouse, Hadoop big data stores, and more. Optimized algorithms accelerate queries across disparate data sources.
- Data delivery: Deliver data to diverse consuming applications through data services, including analytics and business intelligence tools.

Services

As part of the solution, Cisco and Informatica offer onsite professional services to help you complete the project successfully. We follow an implementation methodology with documented and proven tasks to get the job done right and on time. The services are summarized here:

- Assess: Before you deploy any products, our team will examine your data warehouse and IT environments. We develop a picture of your data and transformation workload. Then we create a plan of action that describes what needs to be offloaded and the best way to do so.
- Virtualize: Deploy a data abstraction layer above your data warehouse and big data cluster. This layer hides any complexity in the physical data layer. Data and ETL migration can take place without disrupting the business intelligence and analytics applications that consume the data.
- Migrate: Map data and ELT and ETL workloads to be transitioned. Determine the data movement approach and migrate identified data and workloads from the warehouse to the big data target.
- Operate: We help you establish an operating plan to enable you to run the solution after it is live. Standard operating procedures are documented for administrators to reference as needed.

In addition, we offer services to plan and design the physical infrastructure, as well as the software configurations to help you deploy a predictable and scalable environment. We can work with you to:

- Provision a network to handle the multiple streams of data traffic between the data warehouse, the Informatica server. and the Hadoop nodes
- Deploy a network infrastructure that can handle the consumption by business intelligence and analytics applications that federate data warehouse and Hadoop data through the Cisco Data Virtualization layer

- Absorb large bursts of I/O traffic that occur during initial data load as well as internodal traffic during Hadoop processing

Solution Benefits

Business Benefits

The Informatica and Cisco Data Warehouse Optimization solution delivers a high impact on net income, with a substantial return on investment. Business benefits include:

- Cost control: Offload data and computation processing to get more from your technology investments.
- Enhanced analytics: Access not just current and recent history but extended historical data that typically is archived and not accessible.
- Competitive advantage: Use your company's massive data assets with effective analytics to increase productivity and address business change.
- Reduced risk: Use proven software, networking, and computing infrastructure to adopt big data and logical data warehousing.

Technical Benefits

The offering also provides a number of technical benefits that are unique to the solution's combination of best-in-class Cisco and Informatica products and services. Technical benefits include:

- Linear scalability: The platform is designed to scale with simplicity as your data volume increases. It can grow to up to 160 Hadoop nodes without the need to add any new switching components or redesign the system's connectivity in any way. Furthermore, it can scale up to 10,000 servers, all managed within a single management domain with Cisco UCS Central Software.
- Simplified data migration: Simplify the processing of data transformation and data quality jobs with Informatica Big Data Edition. Using an intuitive visual development environment with hundreds of prebuilt connectors and transformations, developers are up to five times more productive than when they use manual coding.
- Query optimization: Cisco Data Virtualization contains a comprehensive set of techniques and algorithms for federating disparate data. This intellectual property is the result of hundreds of person-years of research and development.
- High performance: Cisco UCS delivers the adaptive performance needed to handle the diverse workloads for both ETL offload and regular Hadoop jobs. The solution's low-latency, lossless 10-Gbps unified fabric is fully redundant and, through its active-active configuration, delivers higher performance compared to other vendors' solutions. It handles the peak I/O load from clients while maintaining a quick response time.
- Deployment and management simplicity: The Cisco UCS platform can host big data and enterprise applications in the same management and connectivity domains, further simplifying data center management. Informatica Big Data Edition uses the full power of the Hadoop framework, helping ensure that all data pipelines built are production ready for high performance, scalability, availability, and maintainability.

Conclusion

Cisco UCS and Informatica, along with Hadoop vendors, bring the power of big data to radically transform the existing enterprise data warehouse landscape with improved response time as well as lower total cost of ownership (TCO) to meet the challenges of growing data volumes and new types of data. This approach also reduces operating costs with improved productivity and simplified deployment of Informatica Big Data Edition along with Cisco UCS CPA for Big Data and Cisco Data Virtualization, offering a solution that can be implemented rapidly and customized for either best performance or high capacity.

For More Information

For more information about Cisco UCS big data solutions, please visit http://www.cisco.com/go/bigdata.

For more information about Cisco UCS Integrated Infrastructure for Big Data, please visit http://blogs.cisco.com/datacenter/cpav3.

For more information about Cisco Data Virtualization, please visit http://www.cisco.com/go/datavirtualization.

For more information about Cisco Validated Designs for big data, please visit http://www.cisco.com/c/en/us/solutions/enterprise/data-center-designs-cloud-computing/bigdata.html.

For more information about Informatica Big Data Edition, please visit http://www.informatica.com/us/products/big-data/informatica-big-data-edition

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.**

Printed in USA CXX-XXXXXX-XX 10/11