Get Your Data Ready in Less Time

Solution Brief
September 2015

Cisco Data Preparation on Cisco UCS Integrated Infrastructure



Highlights

Integrated Infrastructure
- Deploy an industry-leading platform for enterprise big data deployments.

Self-Service Big Data Preparation
- Automate data integration and data-quality optimization with a platform built for governance, collaboration, and large scale using operations analytics, predictive modeling, and packaged analytics tools
- Deploy a self-service application that allows nontechnical business analysts to collect, explore, clean, combine, and enrich the data that fuels analytics.

Instant Verification of Data
- Quickly process data preparation requests so that users can validate operations in real time.

Improved Data Quality
- Identify data patterns, errors, duplications, and omissions with full-text search, interactive text and numeric filters, histograms, and data-quality heat maps.

Automated Data Shaping Platform
- Ingest, clean, and combine data using pivots and splits and quickly make data sets available for analytics.

Scalability for Big Data Workloads
- Scale single-rack and multiple-rack deployments without adding complex layers of switching infrastructure.

Instead of spending a lot of time preparing data for analysis, deploy big data solutions that get you the answers you need—fast.

As your organization uses more self-service business intelligence and analytics applications, you're likely spending more time to get the data ready for your analysts to use. You need to bring together data sets from different sources, identify duplicate data and blank fields, fix misspellings, split and reshape columns, and add data to provide context. But with powerful business intelligence tools available today, every minute wasted on data preparation is a minute you are not asking questions and making decisions. Cisco Unified Computing System™ (Cisco UCS®) running Cisco® Data Preparation can help your teams increase their analytics productivity on ever-increasing data volumes, reduce the risk of data chaos, and get greater value from data insights.

Cisco Data Preparation

Cisco Data Preparation is an end-user application that allows nontechnical business analysts to easily gather, clean, combine, and enrich the raw data that fuels analytics (Figure 1). With this solution, your analysts can:

- **Add:** Include data regardless of location, including data from Hadoop Distributed File System (HDFS) files, relational databases, spreadsheets, and flat files.
- **Explore:** Identify problems with data quality with impromptu interactive exploration using full-text searches; interactive text and numeric filters and histograms; and visual data quality heat maps that highlight patterns, errors, duplicates, and sparse or missing data.
- **Clean:** Use sophisticated algorithms that work across sections or entire data sets, without coding or scripting. Because the solution highlights inconsistencies, gaps, and duplicate data, your analysts can fill in blanks, remove or rename duplicates,

fix inconsistent capitalization, and perform other tasks to improve data quality.

- **Shape:** Pivot or depivot data, split columns, and aggregate data with a single click to quickly make data sets more suitable for the required analytics exercise.
- **Enrich:** Add data when your original data set lacks the context needed for analytics. An example of enrichment is the addition of four-digit extensions to standard five-digit U.S. zip codes.
- **Combine:** Bring multiple data sets together in less time so that you can identify the optimal fields to use to merge data. Cisco Data Preparation automatically detects common attributes across multiple data sets and provides best-match options so that analysts can select the combination that best matches the analysis to be performed. Data sets are assembled into a single answer set. Overlapping references are merged into deduplicated, trusted entities without scripting, SQL, or complex spreadsheet functions such as pivot tables and macros.
- **Publish:** Make answer sets available directly through open database connectivity (ODBC) live queries to QlikView, Tableau, Microsoft Excel, and other ODBC-compliant analytics tools and applications.

Designed for Scalability

The solution uses a four-layer architecture designed for interactive, self-service data preparation at scale (Figure 2). It combines Cisco Data Preparation, built on Apache Spark, and Cisco UCS to deliver a self-service solution with automated data integration capabilities on infrastructure that can scale easily to support growing volumes of data.

- **User-interface layer:** The application front end is a visually dynamic, multiuser interface designed with HTML5 and web-socket technology, making it as interactive and intuitive as popular consumer applications.
- **Web services:** A lightweight Java layer translates and mediates actions from the user interface into commands to the underlying platform layer. This layer handles critical capabilities for rules for tenants, users, projects, and

Figure 1. Cisco Data Preparation Allows You to Easily Gather, Process, and Enrich Data for Improved Analysis

cell-level modifications, creating a comprehensive backbone for governance capabilities.

- **Parallel in-memory pipelined engine:** The engine uses proprietary machine learning, latent semantic indexing, statistical pattern recognition, and text analytics techniques. Data is handled in a model-free environment that accelerates performance with a vector query processor and operates over large and diverse volumes of structured and unstructured data in real time.
- **File management and storage:** All data sets are stored and accessed through a library that runs on top of HDFS.

Excellent for Many Use Cases

Ease of use and a fast in-memory processing architecture make Cisco Data Preparation an excellent environment for solving many types of business problems. Because the solution is not industry specific or limited in functional scope, it can help your organization:

- Turn data insights into revenue
- Merge and reconcile client records
- Access online web retail information, such as user transactions and product transit time
- Identify supply-chain congestion points
- Manage inventory
- Establish product value bundles for sales organizations
- Assess data quality
- Evaluate compliance
- Migrate data between applications

Built on Cisco UCS Integrated Infrastructure for Big Data

Cisco Data Preparation is based on Cisco UCS Integrated Infrastructure for Big Data, a highly scalable architecture designed to meet scale-out application demands. Offered in a complete and easy-to-order package, this infrastructure includes computing, storage, connectivity, and unified management capabilities.

Cisco UCS 6200 Series Fabric Interconnects

Fabric interconnects establish a single point of connectivity and management for the entire system. Deployed in redundant pairs, Cisco UCS fabric interconnects offer the high-bandwidth and low-latency connectivity, active-active redundancy, high performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications.

The system integrates and unifies management for all connected infrastructure components. Cisco UCS Manager supports rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities, such as firmware updates, across the entire cluster as a single operation. It also offers advanced monitoring capabilities with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C-Series Rack Servers

Cisco UCS C240 and C220 M4 Rack Servers support a wide range of computing, I/O, and storage-capacity demands in a compact design. Based on the Intel® Xeon® processor E5-2600 v3 family and supporting 12-Gbps serial-attached SCSI (SAS) throughput, these rack servers deliver significant performance and efficiency gains over previous generations of servers.

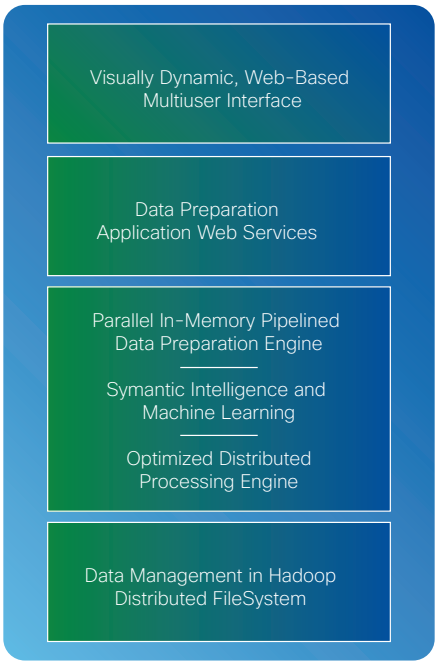

Figure 2. Cisco Data Preparation Architecture

Table 1. Cisco Data Preparation Reference Architecture

Component	Description
Connectivity	• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects • Up to 80 servers with no additional switching infrastructure

Spark pipeline servers	• 8 Cisco UCS C240 M4 Rack Servers, each with:
	  • 2 Intel Xeon processor E5-2680 v3 family CPUs
	  • 512 GB of memory
	  • 2 x 120-GB internal SSD boot drives
	  • 12 x 1.6-TB SSD drives for data storage

Application server	• 4 small-form-factor (SFF) Cisco UCS C220 M4 Rack Servers, each with:
	  • 2 Intel Xeon processor E5-2680 v3 family CPUs
	  • 256 GB of memory
	  • 2 x 1.2-TB SAS SFF hard-disk drives (HDDs

Metadata management servers	• 3 Cisco UCS C220 M4 Rack Servers, each with:
	  • 2 Intel Xeon processor E5-2680 v3 family CPUs
	  • 256 GB of memory
	  • 2 x 1.2-TB SAS SFF HDDs
	  • 2 x 1.6-TB SSDs

Data library servers	• Cisco UCS Integrated Infrastructure for Big Data • Cloudera 5.4 or later

The servers use dual CPUs, support up to 768 GB of main memory (128 to 256 GB is typical for big data applications), and support a range of disk and solid-state-disk (SSD) drive options. Cisco UCS C220 M4 servers offer industry-leading computing density, and Cisco UCS C240 M4 servers offer a balance of computing and storage resources.

Table 1 lists the recommended components for the Data Preparation reference architecture.

Conclusion

If your organization needs to get data ready more quickly so that answers can be found in time to address your pressing needs, consider Cisco Data Preparation. This innovative solution provides excellent data preparation flexibility and responsiveness, accelerating time to value and enhancing the productivity of your data analysts. With the power of intelligent, self-service data preparation capabilities, you can empower your organization to identify relevant data and make it worth analyzing.

For More Information

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

For more information about the Cisco Smart Play Program, visit http://www.cisco.com/go/smartplay.

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

For more information about the Cisco UCS Common Platform Architecture (CPA) for big data, visit http://blogs.cisco.com/datacenter/cpav3/.

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**.

LE-53502-00 09/15