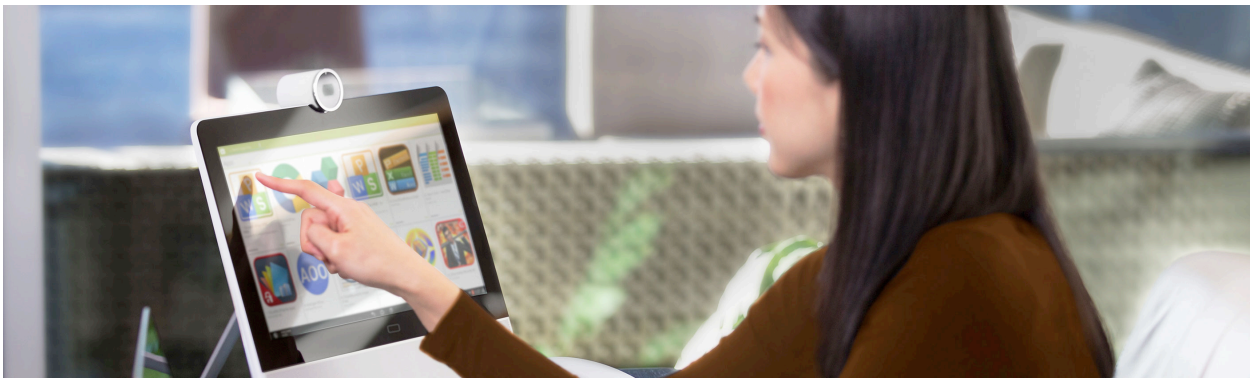
Cisco Data Preparation



Clean, Complete, Ready-to-use Data For Analysis

Business analysts are adopting self-service data preparation tools to more quickly address inconsistent, incomplete, inaccurate and exponentially increasing data. But they are frustrated when the tools they buy have limited reach or cumbersome coding requirements and may not deliver the clean, complete and ready-to-use data their analytics require.

IT leaders support the migration of ad hoc data prep workloads out of their organization. However, they are growing concerned about the increasing inefficiency, cost and risks associated with the unmanaged purchase of lightweight data prep tools, and the proliferation of ungoverned data sets retained across multiple sandboxes and desktops, without consideration of core IT factors such as scalability, security and governance.

Cisco Data Preparation helps business analysts deliver analytic business impacts sooner, working the way business analysts want to work, while achieving business and IT leaders' efficiency, cost and risk mandates.

- **Fast** – Speeds difficult and time-consuming data gathering, exploration cleansing, enrichment and more to accelerate business impact from your analytics.
- **Flexible** – Turns inconsistent, incomplete and inaccurate data from myriad data sources into the clean, complete and ready-to-use Answer Sets your analytics require.
- **Easy to Use** – Your analysts comfortably interact with the actual data using familiar Excel-like functions rather than forcing them to understand abstracted data models and adopt IT-oriented tools.
- **Scalable** – Your data is stored and prepared within a Spark-based analytic sandbox environment provisioned and managed by IT; so scale and performance are never an issue and sampling is an option rather than requirement.
- **Governed** – To ensure the security and governance you require, every step is tracked and auditable; so there is never a question about where the data came from, how it was transformed and where it was delivered.

Unlike lightweight or IT-specialist self-service data prep tools that often result in a proliferation of ungoverned data sets and multiple sandboxes, Cisco Data Preparation is enterprise-grade, self-service data prep software for both business analysts and IT.

Features

Cisco Data Preparation is a self-service application that supports the complete data preparation lifecycle, enabling business analysts to quickly transform inconsistent, incomplete and inaccurate data from myriad data sources into the clean, complete and ready-to-use Answer Sets your analytics require.

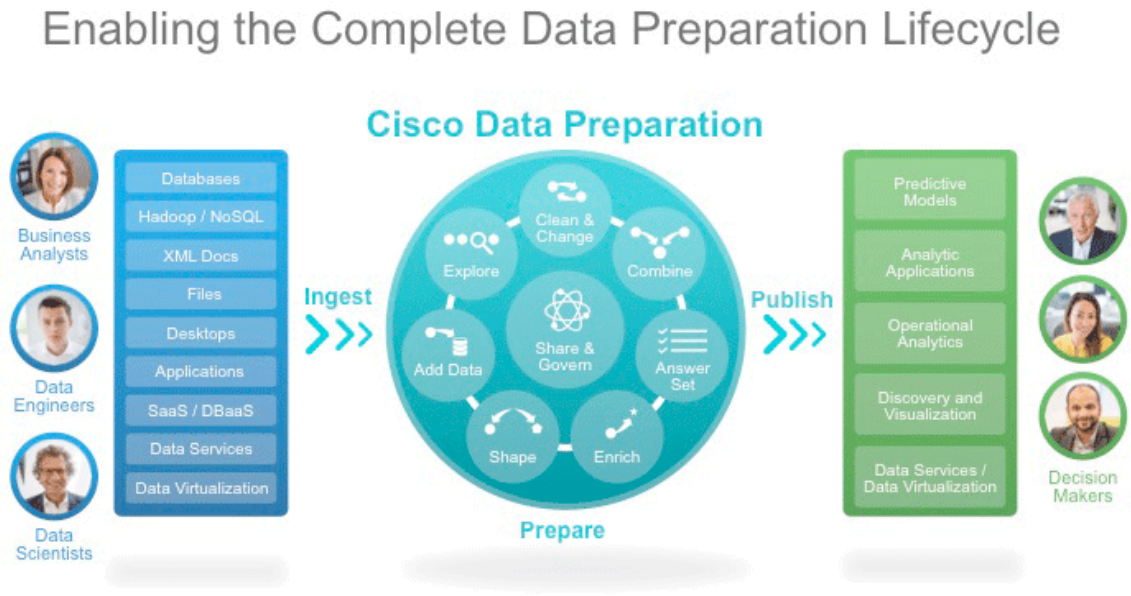


Table 1. Data Preparation Features

Features	Benefits
Ingest	Build a scalable, controlled analytic sandbox environment. Sources include flat files, semi-structured big data, relational, SaaS and enterprise apps, data services, Cisco Data Virtualization and more.
Add data	Work with all the data you need. Use a billion rows if you want, sampling is not required.
Explore	Quickly understand your data. Built-in full text search, interactive text and numeric filters and histograms, and visual data quality heat maps highlight patterns, errors, duplicates and sparse or missing data to speed the exploration process.
Clean & Change	Fix inconsistent, incomplete and inaccurate data quality issues without coding, SQL or scripting. Automatically normalize similar values using natural language processing (NLP), split columns, concatenate columns, de-duplicate, detect and remediate blanks, nulls, and whitespace on the fly.
Shape	Organize your data in minutes, not hours. In a single click, you can pivot or de-pivot data sets, create aggregations, and more to quickly make your data sets more suitable for the required analysis.
Combine	Bring multiple data sets together with ease. CDP automatically detects common attributes and provides best-match options. Then with one click, CDP assembles a unified data set, resolving and removing overlapping entity references, so you avoid difficult scripting, SQL or complex Excel functionality like VLOOKUPS, pivot tables and macros.
Enrich	Easily add context to your data. You can add industry data, append 5-digit zip codes with +4, or integrate information from third-party data providers, and more.
Share & Govern	Automatically capture the work you perform so you can collaboratively recreate, reuse and refine these steps anytime. CDP's authentication, authorization, versioning and auditing capabilities ensure that your work is safe, secure and shared appropriately.
Publish	Visualize your clean, complete and ready-to-use Answer Sets in your favorite analytic tool as well as Cisco Data Virtualization. Connect directly via Hive or Impala. Export data in one of many supported file formats. Or use ClicktoPrep bi-directional access.

Data Preparation with Data Virtualization

Combining Cisco Data Preparation and Cisco Data Virtualization further speeds time-to-solution for new analysis, while enhancing scalability, security and governance.

Table 2. Data Virtualization Integration Areas

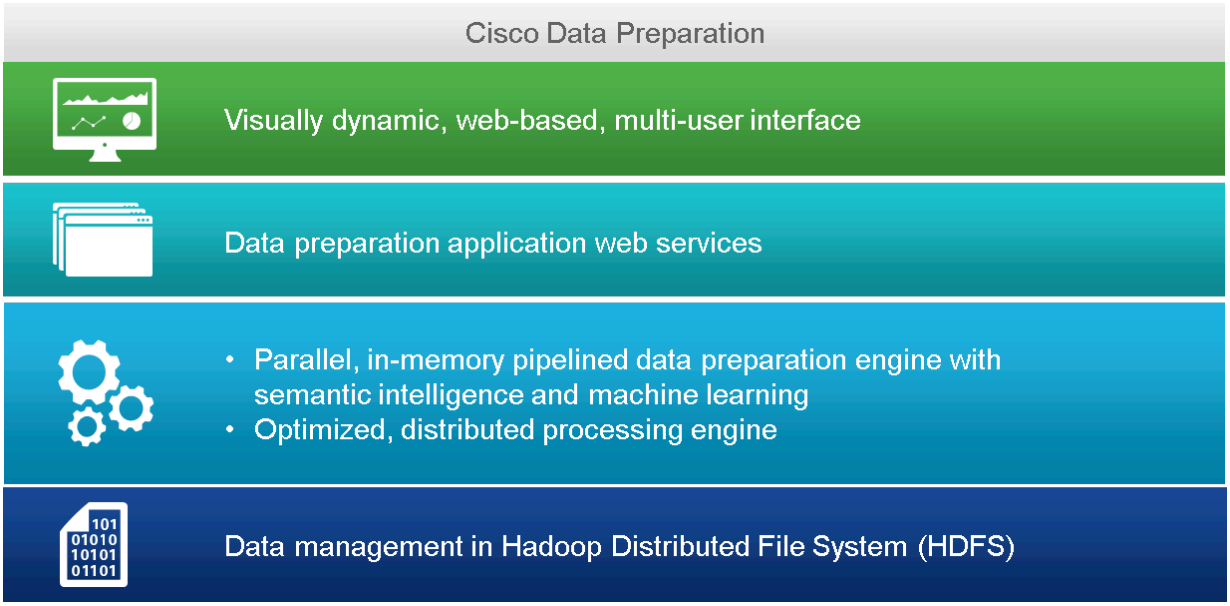
Integration Area	Description
Data Virtualization data sources	Find Data in Business Directory – Connect to business directory, to share and reuse curated data from one or more instances of Cisco Information Server. The curated data has been vetted by IT, annotated by end-users, and gains value from repeated use. Find Data in Cisco Information Server (CIS) – Connect to CIS and gain access to an even broader range of virtualized data sets managed by IT. This virtualized data is integrated from numerous sources, ranging from databases to packaged apps to cloud sources. Ingest Data via Cisco Information Server – Use CIS to load virtualized data sets into Data Preparation where analysts can further refine and drive analytic value from them.
Data Virtualization deployments	Promote Answer Sets into CIS – Data sets prepared by business users in Data Preparation can be made available for broader organizational via CIS and business directory.

Technology

Cisco Data Preparation runs on an enterprise-scale platform built on Hadoop and powered by Spark. It is built on a four-layer architecture so you can store and interactively prepare data at limitless scale. (See Figure 1.)

1. User interface layer: Analysts quickly learn and enjoy using the Data Preparation's intuitive, interactive multi-user interface designed using HTML5 and web socket technology.
2. Web services: A lightweight Java layer translates and mediates actions from the user interface into commands to the underlying platform layer. This layer processes critical capabilities for rules for tenants, users, projects, and cell-level modifications, creating a comprehensive governance foundation.
3. Engines: Enabled by proprietary machine learning, latent semantic indexing, statistical pattern recognition, and text analytics techniques. The first engine has parallel in-memory pipelined capabilities that vastly accelerate many of the mundane data prepration functions. The second engine leverages Spark, and operates over a large variety and volumes of structured and unstructured data in real-time, enabling Cisco Data Preparation to scale to billions of rows.
4. File management and storage: Provides a cost effective data management environment. Data sets are stored and accessed through the Cisco Data Preparation library, which resides on top of HDFS.

Figure 1. Cisco Data Preparation Architecture



Data Preparation on Cisco UCS Big Data Infrastructure

Data Preparation installed on Cisco UCS® scales without limits by taking advantage of Cisco's high-performance and easy-to-manage big data infrastructure. Cisco UCS provides a radically simplified architecture with embedded management that makes it easy to scale as your requirements evolve to solve larger problems and explore more complex scenarios. It also reduces your total cost of ownership (TCO) by requiring fewer infrastructure components and reducing operating expenses associated with staff resources. Together you can solve complex analytical problems, improve business performance, and mitigate risk rapidly and confidently.

The recommended configuration for the Cisco Data Preparation platform deployment is based on Cisco UCS C220-M4/C240 M4, with:

- Two Intel Xeon E5-2680 v3 processors
- 256GB RAM
- 10K RPM SAS HDD or SSD drives, which work with an external Hadoop cluster for data storage

Table 3. Cisco UCS Highlights

Highlights	Benefits
Reliable scalability	Cisco Unified Computing System™ delivers reliable scalability of hardware and management to increase business agility, operational efficiency, and help you rapidly respond to changing business requirements.
Reduced TCO and improved staff efficiency	This simplified, intelligent infrastructure reduces your TCO with fewer management points, switches, adapters, cables, and power and cooling components.
Data preparation on Cisco UCS	Cisco Data Preparation on Cisco UCS streamlines customers' ability to prepare their data for analytics at scale, and can be seamlessly integrated into existing enterprise applications environments.

Service and Support

Cisco Services help you gain better visibility, better information, and better understanding to fuel performance, efficiency, and innovation from your software purchases. Cisco Services span three phases of lifecycle management: plan, build, and manage.

- In the plan phase, Cisco assists you to develop your Cisco Data Preparation strategy and plan in support of your business requirements.
- In the build phase, Cisco works with you to validate that your Data Preparation solution is ready for production and then helps you implement it.
- In the manage phase, Cisco assists you to optimize your Data Preparation deployment, focusing on infrastructure, applications, and service management.

Cisco provides around-the-clock Data Preparation product support through Cisco's Technical Assistance Center. Cisco also provides timely access to the latest Cisco Data Preparation releases, service packs and patches.

System Requirements

Cisco Data Preparation has the following system requirements:

Operating System

- 64-bit (x64) operating system
- CentOS Linux, v6.4 and 6.5 for development and testing

Software

- JDK 7 version 1.7 update 67
- Spark 1.3 (prebuilt for CDH 5)
- Cloudera Distribution of Hadoop (CDH) 4.7 and 5.4 / Spark 1.3 (prebuilt for CDH 5)
- Hortonworks (HDP) 2.3.2 / Spark 1.4

Others

- Cisco Information Server 7.0.3 or later

Ordering Information

Cisco Data Preparation is available for ordering. Table 4 lists the product identifiers required for ordering. To place an order, contact your Cisco account representative.

Table 4. Ordering Information

PID	Product Description
CDP-P-T	Data Prep – per core term
CDP-P-1Y	Data Prep – per core term 1 yr
CDP-P-2Y	Data Prep – per core term 2 yr
CDP-P-3Y	Data Prep – per core term 2 yr

For More Information

For more information about Cisco Data Preparation, contact your Cisco account representative.

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 09/15