Endeca Content Acquisition System

Version 3.0.0

Release Announcement

May 2011

Endeca Content Acquisition System

The Endeca Content Acquisition System (CAS) provides core data integration needs by combining file system and Web crawlers, a portfolio of connectors to various content management systems, and an extension API, enabling connectivity to a wide range of enterprise source systems. CAS includes the following capabilities:

- CAS Server - Robust, multi-threaded crawling is included as part of the core Endeca IAP. The CAS Server also supports CMS data sources when the specific connector is licensed separately and custom data sources when built with the CAS Extension API.

- CAS Console - A Web-based application used to manage crawls run by the CAS Server. During the Content Acquisition System installation, the CAS Console is installed as an extension to Endeca Workbench.

- Web Crawler - Web crawling is included as part of the core Endeca IAP. The Web Crawler delivers improved performance and extensibility compared to previous crawling solutions.

- CMS Connectivity - Connectors to numerous content management systems such as Microsoft SharePoint, IBM Lotus Notes, and EMC Documentum are available as separately licensed additions to the Endeca IAP. For the latest list of CMS connectors, please contact your account representative.

- Record Store – A persistent generational storage of data to simplify the operational environment for data updates.

- CAS Extension API – A Java-based API that provides a set of interfaces and classes to build data source and manipulator extensions to CAS. Using this API, it is possible to build connectors to Enterprise Resource Planning (ERP) systems, Product Lifecycle Management (PLM) systems, Enterprise Content Management (ECM) systems, Ratings & Review services, Site Analytics services, Social Networking sites, or other systems and services that are not accessible by Endeca's out-of-the-box connectors.

CAS Server

The CAS Server can acquire data from file systems, content management systems, and custom data sources. The CAS Server is highly configurable and features integrated security mechanisms.

The features of the CAS Server include:

- *High Performance*. Multi-threaded crawling and integrated text extraction processes result in high performance robust crawls. Obtaining content from archive files is also supported with file system and CMS data sources. The CAS Server is capable of differential crawling when run in incremental mode, pulling content only when files have been modified, added or deleted since the last crawl.

- *Integrated Security.* A crawl automatically acquires the Access Control Lists (ACLs) associated with a file or folder. Group memberships are determined at query time via the Endeca Security Framework (which includes security adapters for systems such as LDAP and Active Directory), and results returned by the Endeca MDEX Engine respect the acquired ACLs. This leads to no information leakage, and no need to post-process results

at the application layer. If the CMS does not take advantage of LDAP or Active Directory, APIs are provided that allow users to write their own custom security adapter to determine group membership information in the application layer.

- *Integrated Document Conversion.* During the crawling process, text and metadata can be extracted automatically from approximately 500 file formats, including commonly-used formats such as PDF and Microsoft Office documents, including Microsoft Office 2007 formats. In the early stages of the application development process, the crawler can be run without the document conversion process enabled, facilitating a very fast crawl for initial data inspection.

- *Highly configurable.* The CAS Server is easily configured to set up data sources, manipulators, and filters in CAS Console or via the CAS APIs.

CAS Console

The CAS Console is a Web application that serves as the user interface to the CAS Server. This tool enables users to create, execute, and monitor multiple crawls of file system, CMS data sources, or any data sources for which a connector has been created using the CAS Extension API. The CAS Console executes commands on the CAS Server via Web service calls.

The features of the CAS Console include:

- *Data Source Configuration.* An administrator can create or edit a data source with an easy to use interface. Each data source's configuration tabs indicates required and optional settings.

- *Manipulator Configuration.* An administrator can add and configure manipulators with an easy to use interface. Each manipulator's configuration tab indicates required and optional settings.

- *Crawl Control*. The CAS Console may be used for starting and stopping acquisition from data sources. During normal operation, the system will automatically determine the correct acquisition (full or incremental) mode.

- *Crawl Monitoring.* Progress of a running crawl displays on the CAS Console Status page. Statistics including acquisition rate, number of files and directories crawled, and number of records created. For stopped crawls, a status flag indicates whether the crawl ran to completion, failed due to a fatal error, or was aborted by an administrator.

CAS Web Crawler

Content stored on Web servers can be accessed using the CAS Web Crawler. The Endeca CAS Web Crawler is highly configurable and features integrated security mechanisms.

The features of the CAS Web Crawler include:

- *High Performance*. The multi-threaded Web Crawler can crawl multiple URLs in parallel, allowing it to traverse a high number of sites and pages.

- *Fine-tunable Behavior*. The Web Crawler supports crawler directives such as robots.txt. It is capable of throttling to avoid stressing servers. The Web Crawler also supports

Enterprise Security Standards and can crawl through secure sites, supporting HTTPS, Basic, NTLM and Form-Based. The details of Form-Based authentication can vary significantly from site to site; therefore, this functionality is highly customizable.

- *JavaScript Support.* The Web Crawler can leverage custom logic to deal with specific JavaScript routines and capture URLs within the scripting language.

- *Integrated Document Conversion:* During the crawling process, text and metadata can be extracted automatically from approximately 500 file formats, including commonly-used formats such as PDF and Microsoft Office documents, including Microsoft Office 2007 formats.

- *Highly Extensible.* Based upon the highly extensible open-source Nutch project, the CAS Web Crawler can be extended easily via custom plug-ins. The Web Crawler can also seamlessly accept Nutch plug-ins produced by the open source community, allowing for endless custom data manipulation opportunities.

Record Store

The Record Store provides persistent generational storage of Endeca records for use in the ITL process. It is most commonly used to simplify the operational processes surrounding update-focused applications.

The features of the Record Store include:

- *Baseline and Incremental Input.* The Record Store provides a Web services interface that accepts both baseline and incremental data as input. For example, a file system data source may write incremental output to an instance of the Record Store while a Web crawl writes baseline data.

- *Baseline and Incremental Output.* Regardless of the mode of input to the Record Store, the system provides a Web services interface for baseline and incremental reading of data to the rest of the ITL process. For example, a Web crawl may write only baseline data but a Forge pipeline can read only the changed data from the Record Store for processing.

- *Operational Simplicity.* The Record Store supports a standards-compliant Web services interface to integrate with any operational environment. It also supports transactions to allow for simultaneous reading and writing of data without concern for data loss.

CAS Extension API

The CAS Extension API provides a Java API for developing extensions to CAS. Extensions include data source connectors and record manipulators. Once installed into CAS, the extensions are available and configurable using the CAS Console, the CAS Server API, and the CAS Server Command-line Utility.

The features of the CAS Extension API include:

- *Support for data source extensions*. Data source extensions can access any type of data source that you want to include in the Content Acquisition System. For example, data source extensions might access flat files, databases, content management repositories (that do not already have a corresponding CMS Connector), and so on.

- *Support for manipulator extensions*. Manipulator extensions transform Endeca records as part of data processing in a CAS acquisition. In a typical usage, manipulators run in a CAS acquisition to provide record pre-processing before a Forge pipeline runs.

- *Leverage CAS operational infrastructure.* New data source extensions and manipulators are available alongside Endeca's native functionality for use by administrators. The CAS Console, for example, can be used to configure and monitor crawls for sources that are accessed via data source extensions. The resulting records from a crawl are stored in a Record Store instance.

Release Information

CAS 3.0.0 (May 2011)

Endeca CAS release 3.0.0 introduces a number of feature enhancements and resolved issues.

Manipulators configurable in CAS Console

You can now add and configure any number of manipulators to a data source and re-order them in the CAS Console.

Disabling of Manipulators

To improve debugging and iterative development, any manipulator can be disabled so that it does not run during data acquisition. It can later be re-enabled to run during data acquisition. Manipulators can be enabled and disabled through the CAS Console, CAS Server API, or CAS Server Command-line Utility.

CAS Document Conversion module packaging

The CAS 3.0.0 release installs its own instance of the Document Conversion Module. If you have purchased the Document Conversion Module, you no longer need to download and install the standalone Document Conversion installation. You can now enable the module in CAS by providing a license key. For enablement details, see the document titled "Enabling the Document Conversion Module in CAS" that is available on EDeN

CAS Document Conversion Module upgrade

The 3rd party libraries in Endeca's Document Conversion module were upgraded. This module, which can be used within CAS to extract metadata and text from binary file formats such as PDF and Microsoft Word, supports approximately 500 different file formats.

Performance Logging Enhancements

The cas-service.log includes additional performance metrics for each completed crawl. The new metrics include the time each manipulator spends processing records in a crawl.

Performance Improvements

Performance enhancements were made to Record Store output, Crawl History tracking, and Document Conversion. Crawls that use these features are likely to see performance improvements between previous versions of CAS and CAS 3.0.0.

Platform Support

The CAS 3.0.0 release adds support for SUSE Enterprise Linux 11, VMware vSphere 4 and 4.1, and Amazon Elastic Compute Cloud (EC2).

Early Access features for Data Integration

CAS 3.0.0 introduces early access features to support eBusiness use cases. These features are compatible with the MDEX 6.2.0 release. The early access features include:

- A Record Store Merger data source that supports merging product data and taxonomies that have been loaded into Record Stores by crawls from other data sources.
- A Dimension Value Id Manager component that runs in the Component Instance Manager. With the Dimension Value Id manager, dimension value ID's can be generated and later retrieved based on a dimension value spec. This capability keeps configuration stable across environments.
- An MDEX output option for configuring a crawl to output MDEX compatible data that can be directly processed by Dgidx, bypassing Forge.

These early access features currently enable a limited number of use cases. Please contact your Endeca services representative for more information.

Contact Us

Kristen Fortino
Director, Product Management
T +1 617.674.6224
E KFORTINO@ENDECA.COM