Endeca® Content Acquisition System

CAS Console Help

Version 3.1.1 • December 2012

ORACLE

ENDECA

Contents

Copyright and disclaimer

Preface

The Oracle Endeca Commerce solution enables your company to deliver a personalized, consistent customer buying experience across all channels — online, in-store, mobile, or social. Whenever and wherever customers engage with your business, the Oracle Endeca Web commerce solution delivers, analyzes, and targets just the right content to just the right customer to encourage clicks and drive business results.

Oracle Endeca Commerce is the most effective way for your customers to dynamically explore your storefront and find relevant and desired items quickly. An industry-leading faceted search and Guided Navigation solution, Oracle Endeca Commerce enables businesses to help guide and influence customers in each step of their search experience. At the core of Oracle Endeca Commerce is the MDEX Engine™, a hybrid search-analytical database specifically designed for high-performance exploration and discovery. The Endeca Content Acquisition System provides a set of extensible mechanisms to bring both structured data and unstructured content into the MDEX Engine from a variety of source systems. Endeca Assembler dynamically assembles content from any resource and seamlessly combines it with results from the MDEX Engine.

Oracle Endeca Experience Manager is a single, flexible solution that enables you to create, deliver, and manage content-rich, cross-channel customer experiences. It also enables non-technical business users to deliver targeted, user-centric online experiences in a scalable way — creating always-relevant customer interactions that increase conversion rates and accelerate cross-channel sales. Non-technical users can control how, where, when, and what type of content is presented in response to any search, category selection, or facet refinement.

These components — along with additional modules for SEO, Social, and Mobile channel support — make up the core of Oracle Endeca Experience Manager, a customer experience management platform focused on delivering the most relevant, targeted, and optimized experience for every customer, at every step, across all customer touch points.

About this guide

This guide describes how to create, configure, crawl, and monitor data sources using CAS Console for Oracle Endeca Workbench.

Who should use this guide

This guide is intended for application developers who are building applications using the Endeca Content Acquisition System, and are responsible for crawling source data in different source formats to transform the data into Endeca records.

Conventions used in this guide

This guide uses the following typographical conventions:

Code examples, inline references to code elements, file names, and user input are set in `monospace` font. In the case of long lines of code, or when inline monospace text occurs at the end of a line, the following symbol is used to show that the content continues on to the next line: ¬

When copying and pasting such examples, ensure that any occurrences of the symbol and the corresponding line break are deleted and any remaining space is closed up.

Contacting Oracle Support

Oracle Support provides registered users with important information regarding Oracle Endeca software, implementation questions, product and solution help, as well as overall news and updates.

You can contact Oracle Support through Oracle's Support portal, My Oracle Support at *https://support.oracle.com*.

Chapter 1

Getting started with the Endeca CAS Console

This section provides an introduction to crawling data sources with the Endeca CAS Console for Oracle Endeca Workbench.

About the Endeca CAS Console for Oracle Endeca Workbench

The CAS Console for Oracle Endeca Workbench is a Web-based application used to crawl various data sources including file systems, content management systems, and custom data source extensions. During the Content Acquisition System installation, the CAS Console is installed as an extension to Oracle Endeca Workbench.

In a typical implementation, Endeca developers use the CAS Console for Oracle Endeca Workbench to add one or more data sources to an application. The CAS Server crawls the data source and converts the source data into Endeca records. These records are subsequently processed by Forge and the Indexer in an Endeca pipeline for use in an MDEX Engine.

Overview of working with a data source

The following steps provide an overview of the procedure to add a data source to CAS Console, configure filtering criteria, and acquire data from the data source.

Each of the following steps is described in its own help topic:

- Log in to Oracle Endeca Workbench and select the Data Sources page.
- Add a data source to CAS Console (either a file system data source, CMS data source, or other type of custom data source extension).
- Configure filtering criteria for a data source if necessary, and other configuration options.
- Acquire data from a data source.
- Monitor a data source as it is being crawled.

Chapter 2

Working with data sources and manipulators

This section describes how to add data sources and manipulators to CAS Console and how to configure filters and options for data sources. You have to purchase and enable a CMS data source for it to appear in CAS Console. Some data sources listed in this section maybe not be available in your application.

Overview of the default CAS data sources and manipulators

The Content Acquisition System ships with a set of default data sources and manipulators. Each is described here:

Data Source	Description
Delimited File	Crawls records in delimited text files, including .csv files.
Documentum Content Server	Crawls Documentum Content Server repositories (docbases).
Documentum eRoom	Crawls Documentum eRoom repositories.
Endeca Record File	Crawls Endeca record files including .xml, .xml.gz, .bin, .bin.gz, .binary, and .binary.gz.
File System	Crawls folders and files on both local drives and network drives.
FileNet Document and Image Services	Crawls Content and Image Services repositories.
Filenet P8	Crawls FileNet P8 repositories.
Interwoven TeamSite	Crawls Interwoven TeamSite repositories.
JDBC	Crawls a JDBC-accessible database.
JSR-170 Compliant	Crawls Java content repositories (JCR).
Lotus Notes	Crawls Lotus Notes repositories.

Data Source	Description
Microsoft SharePoint Object Model	Crawls SharePoint repositories using a custom Web Service that is implemented with the SharePoint Object Model API.
Microsoft SharePoint Web Services	Crawls SharePoint repositories.
Open Text Livelink	Crawls Open Text Livelink repositories.
Record Store Merger	Crawls CAS record store instances.

For information about version support for a particular repository, see the connector's chapter in *CAS Developer's Guide*.

Manipulator	Description
Filtering Script	This manipulator runs an inline BeanShell script that filters Endeca records from crawl output.
Modifying Script	This manipulator runs an inline BeanShell script that modifies Endeca records.

For information about configuring a manipulator, see the *CAS Console Help* or use the `cas-cmd` utility with the `getModuleSpec` task.

Adding a Delimited File data source

You add a Delimited File data source by specifying delimited text files to crawl, delimiter information, and optional information about column names and whether columns contain multi-assign values.

Each new line in a delimited text file results in one corresponding Endeca record after crawling. Columns and column fields become Endeca properties and property values.

By default, the Delimited File data source reads the header row of a file and uses the header values as column names. In cases where a delimited file does not have a header row of column names, you can specify column names manually using the **Column Names** option.

The data source supports both single-assign and multi-assign values. (See the configuration options below.)

To add a new Delimited File data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Delimited File** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from others in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **Path to Input File(s)**, specify an absolute path to the delimited files you want to crawl.

   Wildcards may be used in the filename but not in the path preceding the filename.

   Example of local folders on Windows:

   - `C:\Endeca\apps\test\data\incoming\records.txt`

Example of syntax for network drives:

- `\\abchost.endeca.com\documents\customers.csv`

5. In **Record Id Column**, specify the name of the column that you want to map to the record ID property in the generated records.

   The values of this column must be unique across all files being crawled.
6. In **Delimited Character**, specify a single character that delimits the fields in the records. The default delimiter is a comma (,).
7. In **Quote Character**, specify a single character that escapes occurrences of the delimited character within a field. The default quote character is a quote (").
8. Optionally, in **Column Names**, click **Add** for each column in the file and name it as appropriate for the column value. Specify column names in the order in which they appear in a delimited text file. This optional configuration is typically only necessary in cases where a delimited file does not contain a header row. If **Column Names** are unspecified, the data source treats the first row of the file as the header row and uses the column names as Endeca property names.
9. Optionally, in **Multi-Assign Delimiter Character**, specify a single character that delimits multi-assign values within a multi-assign column. If you specify a value and omit adding any **Multi-Assign Columns**, the data source parses all columns in the file as if they may contain multi-assign values. In this example, the pipe character (|) delimits multi-assign values named `Value2a`, `Value2b`, and `Value2c` within a multi-assign column named `Header2`:

   ```
   Header1,Header2,Header3
   Value1,Value2a|Value2b|Value2c,Value3
   ```
10. Optionally, in **Multi-Assign Columns**, click **Add** for each column in the file that contains multi-assign values and name it as appropriate for the column value.
11. Optionally, in **Trim Whitespace**, select **true** to trim the leading or trailing whitespace from the data stored in columns of the delimited file. The default value is **true**.
12. Optionally, in **Character Encoding**, specify the character encoding of the delimited file that is being crawled. If unspecified, the default value is **UTF-8**.
13. Click **Save**.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Documentum Content Server data source

You add a Documentum Content Server data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Documentum Content Server data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Documentum Content Server** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

4. In **User Name**, specify a valid user account for the Documentum data source and then specify values for **Password** and **Confirm Password**.
5. Optionally, in **Domain**, specify the domain of the user name from the previous step.
6. In **Docbase Name**, select the Documentum repository name.
7. Optionally, in **Webtop URL**, specify the base URL of the local Documentum StringWebtop installation. For example: `http://myhost:8080/webtop/`.
8. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Note that seeds are not required for CMS data sources. If unspecified, CAS Server crawls all folders.

   You can specify multiple seeds. The path should start with a forward slash followed by the cabinet name and the folder name inside the cabinet. For example, to search the folder 2007reports in the Sales cabinet, specify `/Sales/2007reports`.

9. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Documentum eRoom data source

You add a Documentum eRoom data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Documentum eRoom data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Documentum eRoom** from the list and click **Add**.

   The **Data Source** tab displays.

3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **eRoom Server**, specify the DNS name of the eRoom server.
6. In **Facility Name**, specify the name of the eRoom facility. In an eRoom page url, the facility name is the first item following `eRoom`.
7. Click **Show optional configuration**.
8. Optionally, in **Protocol**, select either **HTTP** or **HTTPS** protocol.
9. Optionally, in **Port Number**, specify the network port to use. The default is 80 if you are using **HTTP** or 443 if you are using **HTTPS**.
10. Optionally, in **Web Context**, specify the Web context of the eRoom Web services. The default is `/eRoomXML/`.

11. Optionally, in **Enable Cache**, select **true** to enable caching for users and groups or **false** to disable caching. This is a per connector instance cache without any expiration time. The default value is **true**.
12. Optionally, in **Check SSL Certificate**, specify whether all SSL certificates are accepted, including self-signed certificates. Select **true** to accept only trusted SSL certificates, or select **false** to accept all SSL certificates including self-signed certificates. The default value is **false**.
13. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Specify everything in the path to the folder following `My eRooms` (as it appears at the top of your browser window when using eRoom), substituting each arrow (>) in the path with a forward slash (/). Seed names are case sensitive.
14. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding an Endeca Record File data source

You add an Endeca Record File data source to CAS Console by specifying one or more Endeca record files and a record ID property. Valid file types are .xml, .xml.gz, .bin, .bin.gz, .binary, and .binary.gz. Wildcard characters may be used to specify multiple files in a given directory, but wildcards cannot be used in directory syntax to specify multiple directories.

To add a new Endeca Record File data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Endeca Record File** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from others in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **Path to Input File(s)**, specify an absolute path to the files you want to crawl.

   Wildcards may be used in the filename but not in the path preceding the filename.

   Examples of local folders on Windows:

   - `C:\Endeca\apps\atgtest\data\incoming\*.xml`
   - `C:\tmp\records.bin.gz`

   Examples of syntax for network drives:

   - `\\abchost.endeca.com\documents\*.xml`
   - `\\xyzhost\home\smith\*.binary`
5. In **Record Id Property**, specify the name of the source property that you want to map to the record ID property in the generated records.

   This property must be unique across all files being crawled.
6. Click **Save**.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a File System data source

You add a File System data source by specifying folders and files to crawl (seeds), optionally specifying filters that include or exclude files, and specifying options for Endeca records that result from crawling the data source.

If you want to crawl network drives, map the drive on the machine running CAS before specifying it as a seed.

To add a new File System data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **File System** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from others in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. Under **Seeds**, specify at least one seed to crawl.

   For crawling file systems, a seed is an absolute path to a folder you want to crawl. Seeds may be local folders or network drives. Note that for Windows, you should specify network drives by universal naming convention (UNC) syntax rather than by using the letter of a mapped drive. For UNIX, you can specify mounted or local drives using standard file path syntax.

   Examples of local folders on Windows:

   - `D:\documents\reports`
   - `C:\tmp`
   - `C:\Documents and Settings\username\My Documents`

   Examples of syntax for network drives:

   - `\\abchost.endeca.com\documents`
   - `\\xyzhost\home\smith`
5. If necessary, click **Add Another** to create additional seeds.
6. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a FileNet Document and Image Services data source

You add a FileNet Document and Image Services data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Filenet Document and Image Services data source:

1. On the **Data Sources** page, click **Add Data Source...**.

2. Select **Filenet Document and Image Services** from the list and click **Add**.

   The **Data Source** tab displays.

3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **IDM Web Services URL**, specify the complete URL to the Web Services for FileNet Panagon, using the following format: `http://Panagon Web Services server:port/directory` , where `Panagon Web Services server` is the name of the FileNet Panagon Web Services server, `port` is the port number for the Default Web Site (usually 80), and `directory` is the directory where you installed the Web services component.
6. In **System Type**, select **Content Services** or **Image Services**.
7. In **Library Name**, specify one of the following:

   - For Document Services, specify the name of the FileNet Document Services library, according to this format: *Host*^*System*. For example: `FILENETDSHOST^FILENETDSSYSTEM`
   - For Image Services, specify the name of the FileNet Image Services library, according to this format: `Library^Domain^Organization`. For example, `IMS^panagonis2^MyCompany`

8. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Specify the syntax: `/folder1`. To find folders to set as seeds:

   - In Windows Explorer, navigate to FileNet Neighborhood.
   - Select the desired Image or Document Services instance.
   - You are prompted for username and password.
   - Navigate to folders you wish to crawl in this FileNet instance.

   Seed names are case sensitive.

9. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Filenet P8 data source

You add a Filenet P8 data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Filenet P8 data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Filenet P8** from the list and click **Add**.

   The **Data Source** tab displays.

3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **Object Store**, specify the name of the FileNet P8 Object Store.
6. In **Remote Server URL**, specify the URL of the Content Engine web service, for example, `http://fnetserver:9080/wsi/FNCEWS40SOAP`.
7. Optionally, in **Workplace URL**, specify the base URL of the FileNet Workplace or WorkplaceXT application. This is used to generate URLs to documents and folders. For example: `http://fnetserver:9080/app_engine/`. (The URLs are stored in `Endeca.CMS.Uri`.)
8. Optionally, in **Remote file transfer URL**, specify the URL of the Content Engine file transfer web service, for example: `http://fnetserver:9080/wsi/FNCEWS40MTOM`.
9. Optionally, in **Check SSL Certificate**, specify whether all SSL certificates are accepted, including self-signed certificates. Select **true** to accept only trusted SSL certificates, or select **false** to accept all SSL certificates including self-signed certificates. The default value is **false**.
10. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Specify the path to the folder in the Object Store as displayed in the FileNet/Workplace browser, replacing each arrow (>) with a forward slash (/). For example, if the current browse state is `Object Stores > Store2 > MyFolder > dir1 > directory3`, then the corresponding path is `/MyFolder/dir1/directory3`. Seed names are case sensitive.
11. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding an Interwoven TeamSite data source

You add an Interwoven TeamSite data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add an Interwoven TeamSite data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Interwoven TeamSite** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **Area Path**, specify the server relative area path.
6. In **Service URL**, specify the base URL. The syntax of this property is `http://`*`server name`*`:`*`port`*`/iw/services/cm/2.0.`

7. Optionally, in **Check SSL Certificate**, specify whether all SSL certificates should be accepted, including self-signed certificates. If set to `true`, only trusted SSL certificates are accepted. The default value is `false`.
8. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Unless otherwise indicated, the syntax for a CMS seed begins with a forward slash, specifies the root, and then any level of folder depth, for example, `/CMSRoot/folderA/folderB`.

   Seed names are case sensitive.
9. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a JDBC data source

You add a JDBC data source to CAS Console by specifying connection information, a JDBC driver class, and a SQL query to execute against the database. The data source executes the SQL query against the database, and each row in the result set becomes an Endeca record. Each column in a row becomes an Endeca property on the record.

Before using the JDBC data source, you must install the JDBC driver into CAS. For details, see "Installing a JDBC driver into CAS" in the *CAS Developer's Guide*.

To add a JDBC data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **JDBC** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **Driver class**, specify the fully qualified Java class name for the JDBC driver.
   For example: `com.mysql.jdbc.Driver`
5. In **JDBC URL**, specify the connection string that includes, at a minimum, the database vendor, the host and port, and the database instance name. If desired, you can also specify the **Username** and **Password** as part of the connection string.
   For example: `jdbc:mysql://localhost:3306/my_database`
6. Optionally, in **Username**, specify a valid user account for the data base and then specify a corresponding value for **Password**.
7. Optionally, in **Connection Properties**, specify any additional connection properties that your database may require. Specify properties in the format: `name=value`.
   For example: `MaxPoolSize=10`
8. If necessary, click **Add Another** to create additional connection properties.
9. In **SQL/Record Generation**, specify a SQL query to execute against the database.
   For example: `SELECT * FROM my_table;`
10. In **Key Column**, specify the name of the column in the database that you want to map to the record ID property in the generated records.

The values of this column must be unique across all records being returned in the result set of the SQL query.

11. Click **Save** or select the **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a JSR-170 compliant data source

You add a JSR-170 compliant data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a JSR-170 compliant data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **JSR-170 Compliant Connector** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **JNDI name or RMI address**, specify the JNDI name or RMI address that CAS uses to retrieve the repository implementation.
6. In **JCR repository retrieval method**, select **JNDI** if you specified a JNDI name above or select **RMI** if you specified an RMI address.
7. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Unless otherwise indicated, the syntax for a CMS seed begins with a forward slash, specifies the root, and then any level of folder depth, for example, `/CMSRoot/folderA/folderB`.

   Seed names are case sensitive.
8. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Lotus Notes data source

You add a Lotus Notes data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Lotus Notes data source:

1. On the **Data Sources** page, click **Add Data Source...**.

2. Select **Lotus Notes Connector** from the list and click **Add**.

   The **Data Source** tab displays.

3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

4. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
5. In **Domino Server**, specify the name of the Domino server where the database is located.
6. In **Database Path**, specify the absolute or relative path of the Notes database.
7. Click **Show optional configuration**.
8. Optionally, in **HTTP Port Number**, specify the port of the Domino web server or DIIOP server. If no value is added the default of 80 is used.
9. Optionally, in **QuickPlace**, indicate whether the Notes database being accessed is an IBM Lotus QuickPlace or IBM Lotus Quickr database. If set to **true**, the settings for Content strategy and RFC 822 strategy are ignored. The default is **false**.
10. Optionally, in **Content Strategy**, select the strategy to use for document content. The strategies are:

    - Text Body
    - Text Body + Attachments (This is the default).
    - Attachments
    - HTML Body + Attachments

11. Optionally, in **RFC 822 Strategy**, select `true` for the server to attempt to reconstruct the original RFC 822 message and return it as content. If the document is not in RFC 822 format, the strategy specified by `contentStrategy` is used instead.
12. Optionally, in **Keep Rich Text**, select `true` to return rich-text fields as properties, or select `false` to ignore them.
13. Optionally, in **Display Views**, specify comma-separated names of views/folders that should be made available by the connector even if they are normally hidden.
14. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Unless otherwise indicated, the syntax for a CMS seed begins with a forward slash, specifies the root, and then any level of folder depth, for example, `/CMSRoot/folderA/folderB`.

    Seed names are case sensitive.

15. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Microsoft SharePoint Object Model data source

You add a Microsoft SharePoint Object Model data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Microsoft SharePoint Object Model data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Microsoft SharePoint Object Model** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **User Name**, set a user name for the data source.
5. Specify values for **Password** and **Confirm Password**.
6. If you are using NTLM authentication, specify the domain of the SharePoint server in **Domain**.

   In order to authenticate to SharePoint using NTLM, the domain name must be specified to log on to the server. Alternatively, encode the domain name in the user name as in DOMAIN\username. In that case, do not specify a domain name in the credentials domain.

   For non-NTLM authentication, this is a convenience property for prepending the value of this property to the username property. The domain will be appended with a backslash separating it from the username. Oracle recommends specifying the domain only in the username property and not adding this property, for clarity.
7. In **SharePoint Site URL**, specify a value for the SharePoint server name and port, such as `http://sharepoint:10000`. The **SharePoint Site URL** can only be set to the repository site or the home SharePoint site collection. The **SharePoint Site URL** cannot be set to a Document Library.
8. Click **Show optional configuration**.
9. Optionally, in **Enable HTTP Chunking**, select **true** to use chunked encoding for HTTP messages. The default value is **false**.
10. Optionally, in **Check SSL Certificate**, specify whether all SSL certificates are accepted, including self-signed certificates. Select **true** to accept only trusted SSL certificates, or select **false** to accept all SSL certificates including self-signed certificates. The default value is **false**.
11. Optionally, in **Socket timeout**, specify a different timeout value (in milliseconds) for content retrieval. The default value is 15 seconds (15000 milliseconds).
12. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Specify the relative path of the URL in relation to the siteUrl, such as `/Shared Documents/Word Docs`. Do not use `"%20"` to denote spaces; use spaces if needed in the URL. When crawling lists, do not specify the seed as the navigation url (such as `/Lists/ListName`); simply specify `/ListName`.

    Seed names are case sensitive.
13. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Microsoft SharePoint Web Services data source

You add a Microsoft SharePoint Web Services data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add a Microsoft SharePoint Web Services data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Microsoft SharePoint Web Services** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **User Name**, set a user name for the data source. For SharePoint 2003, this username must match the SharePoint user's Account as it appears in **Central Administration** > **People and Groups** > **User Information**
5. If you are using NTLM authentication, specify the domain of the SharePoint server in **Domain**.

   In order to authenticate to SharePoint using NTLM, the domain name must be specified to log on to the server. Alternatively, encode the domain name in the user name as in DOMAIN\username. In that case, do not specify a domain name in the credentials domain.

   For non-NTLM authentication, this is a convenience property for prepending the value of this property to the username property. The domain will be appended with a backslash separating it from the username. Oracle recommends specifying the domain only in the username property and not adding this property, for clarity.
6. Specify values for **Password** and **Confirm Password**.
7. In **SharePoint Site URL**, specify a value for the SharePoint server name and port, such as `http://sharepoint:10000`. The **SharePoint Site URL** can only be set to the repository site or the home SharePoint site collection. The **SharePoint Site URL** cannot be set to a Document Library.
8. Click **Show optional configuration**.
9. Optionally, in **Generic Lists Support**, select **true** to support additional Sharepoint lists such as Issues, Wiki, Surveys, custom lists. By default, the connector manages document libraries. The default is **true**.
10. Optionally, in **Enable HTTP Chunking**, select **true** to use chunked encoding for HTTP messages. The default value is **false**.
11. Optionally, in **Check SSL Certificate**, specify whether all SSL certificates are accepted, including self-signed certificates. Select **true** to accept only trusted SSL certificates, or select **false** to accept all SSL certificates including self-signed certificates. The default value is **false**.
12. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Specify the relative path of the URL in relation to the siteUrl, such as `/Shared Documents/Word Docs`. Do not use `"%20"` to denote spaces; use spaces if needed in the URL. When crawling lists, do not specify the seed as the navigation url (such as `/Lists/ListName`); simply specify `/ListName`.

    Seed names are case sensitive.
13. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding an Open Text Livelink data source

You add an Open Text Livelink data source to CAS Console by specifying connection information, seeds, filters that include or exclude files, and advanced settings.

To add an Open Text Livelink data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Open Text Livelink Connector** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from other data sources in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **Domain**, specify the domain of the repository.
5. In **User Name**, specify a valid user account for the data source and then specify values for **Password** and **Confirm Password**.
6. In **Server Name**, specify the DNS name of the Livelink server (the host).
7. Click **Show optional configuration**.
8. Optionally, in **Port Number**, specify the network port of the Livelink server. If no value is entered, the default value of `2099` is used.
9. Optionally, in **Search Broker**, specify the name of the search broker to use. If no value is entered, the default value of `Enterprise` is used.
10. Optionally, in **Search Broker (All Versions)**, specify the name of the search broker to use if searching across versions. If no value is entered, the default value of `Enterprise [All versions]` is used.
11. Optionally, in **Livelink CGI Path**, specify the server-relative path of the Livelink CGI if you are using http/https tunneling. For example: `/Livelink/livelink.exe`.
12. Optionally, in **Use HTTPS**, specify whether to use an unsecure (http) or secure (https) connection channel. In order to use https connections, the Livelink ECM Secure Connect module must be installed and this value must be set to `true`. The default is `false`.
13. Optionally, in **HTTP User Name**, specify the user name to use to authenticate against the tunneling web server. The default is to use the current user name. This parameter is ignored if the web server doesn't require authentication.
14. Optionally, in **HTTP Password**, specify the password to use to authenticate against the tunneling web server. The default is to use the current user password. This parameter is ignored if the web server doesn't require authentication.
15. Optionally, in **Display URL**, specify the full URL to the Livelink CGI. For example, `http://<host-name>[:port][/<context>]/livelink.exe`. This URL is used to build a URL to items. If none is specified then the URL property of items will not be populated.
16. Optionally, specify one or more seeds by clicking **Add** under **Seeds**. Unless otherwise indicated, the syntax for a CMS seed begins with a forward slash, specifies the root, and then any level of folder depth, for example, `/CMSRoot/folderA/folderB`.

    Seed names are case sensitive.
17. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

At this point, you can add manipulators, acquire data from the data source, and monitor its status.

Adding a Record Store Merger data source

You add a Record Store Merger data source to an acquisition by specifying at least one data record store instance that CAS reads from. You can optionally specify any number of additional data record store instances and taxonomy record store instances to read from.

The Record Store Merger data source reads from the specified record store instances, merges the records (if you specified more than one record store instance), and writes the output to a record store instance.

CAS treats record store instances as either data record stores or as taxonomy record stores depending on where you specify them on the **Data Source** tab. However, when a Record Store Merger data source reads records from a taxonomy record store instance, CAS tags each record with an `Endeca.taxonomy` property during the merge. This property is not added to records that CAS reads from a data record store instance. Later in the processing pipeline, CAS skips the property the `Endeca.taxonomy` property when converting taxonomy records into dimension configuration. This means that if you are using a manipulator to create additional taxonomy records that were not read by the Record Store Merger data source, then be sure to add the `Endeca.taxonomy` property to the new records.

To add a Record Store Merger data source:

1. On the **Data Sources** page, click **Add Data Source...**.
2. Select **Record Store Merger** from the list and click **Add**.

   The **Data Source** tab displays.
3. In **Name**, specify a unique name for the data source to distinguish it from others in the CAS Console.

   You can create a data source name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
4. In **Data Record Stores**, specify the name of a record store instance that contains product data.
5. Optionally, click **Add Another** to add multiple data record store instances.
6. Optionally, in **Taxonomy Record Stores**, click **Add** and specify the name of a record store instance that contains taxonomy data.
7. Optionally, click **Add Another** to add multiple taxonomy record store instances.
8. Click **Show optional configuration**.
9. Optionally, in **Host**, specify the fully qualified name of the computer hosting the record store instances that CAS should read from. If unspecified, the default is the host running the CAS Service.
10. Optionally, in **Port**, specify the network port of the computer hosting the record store instances that CAS should read from. If unspecified, the default is port number for the CAS Service (typically `8500`).
11. Optionally, in **Is port SSL**, specify whether the **Port** is SSL or not. Select **true** to connect to the record store instances using HTTPS, or select **false** to connect using HTTP. Specify **false** if you enabled HTTPS redirects in Oracle Endeca Workbench. The default value is **false**.
12. Click **Save** or select the **Filters** or **Advanced Settings** tab to continue configuring the data source.

The data source displays **Acquisition Steps** where you can add manipulators, revise the data source configuration if necessary, or start acquiring data from the data source.

Editing a data source

You edit a data source by modifying its configuration options or by adding manipulators to the data source. The changes take effect the next time the CAS acquires data from the data source.

You can view data source details while the CAS Server is acquiring data, but the options are unavailable to edit.

To edit a data source:

1. On the **Data Sources** page of CAS Console, locate the data source you want to edit.
2. Click the data source's name to display the **Acquisition Steps** list.
3. Click the **Acquisition** link.
   The **Data Source** tab displays.
4. Edit the options on the **Data Source** tab, and if applicable, the **Filters**, or **Advanced Settings** tabs.
5. When you finish editing the data source, click **Save** to go back to the **Acquisition Steps** list.
6. Optionally, click the Advanced Settings link and modify the **Log Level** or **Maximum Threads** settings.
   - If you modified these options, click **Save** to return to the **Data Sources** page, or click **Apply** to remain on this page and continue editing acquisition steps.
7. When you are finished editing the data source, click **Return to Data Sources** or click **Data Sources**.

Deleting a data source

If a data source is no longer necessary, you can delete it from an application.

To delete a data source:

1. On the **Data Sources** page, locate the name of the data source you want to delete.
2. Click the icon in the **Delete** column for the corresponding row.
3. Click **Yes** to confirm the action.

Specifying data source filters

Filters define what folders and files are included and excluded while acquiring data from a data source.

If you do not specify any filters, all folders and files included in seeds are crawled by default.

Note the following about filters:

- Include filters apply only to files and not folders.
- Exclude filters apply to files and folders.
- Filters are case insensitive except for property names themselves, or unless otherwise specified in a regular expression.
- Custom data source extensions do not support filters in this release.

Common Endeca properties for data source filters

The following Endeca properties are commonly used for data source filters:

- `Endeca.FileSystem.Name`
- `Endeca.FileSystem.Extension`
- `Endeca.File.Size`
- `Endeca.FileSystem.Path`
- `Endeca.FileSystem.ModificationDate`
- `Endeca.CMS.Name`
- `Endeca.CMS.ModificationDate`
- `Endeca.CMS.MimeType`
- `Endeca.CMS.ContentLength`
- `Endeca.CMS.Author`

Specifying string filters for a data source

String filters allow you to include or exclude content based on the values of string properties you specify. String filters support Java regular expressions as well as wildcard expressions.

To add a string filter to a data source:

1. On the **Data Sources** page:
   - Click **Add Data Source** and select a new data source from the drop-down menu, or
   - Click the name of a saved data source to edit its configuration.
2. Select the **Filters** tab.
3. Determine whether you want to apply this filter to files or folders. Then, under the appropriate heading, click **Add**.
4. Select **String Filter**.
5. In the **Specify Property for String Filter** dialog, enter the **Property Name** to filter against and click **Ok**.
6. Select either **Include** or **Exclude** for this filter. For example: to exclude files with an extension of `exe` and `dll`, add a string filter for the property Endeca.FileSystem.Extension, select **Exclude**, and enter the following in the text field, separating string entries with commas:

   exe, dll

    **Note:** Include filters apply only to files and not folders.

7. If desired, check **Use a regular expression** and specify a Java regular expression to match against. See Sun Microsystems' documentation for additional information about syntax and usage of Java regular expressions.

    **Note:** You cannot have multiple entries separated by commas in this case.

8. Click **Save**.

Specifying date filters for a data source

Date filters allow you to include or exclude content based on properties with date values occurring either before or after values that you specify.

To add a date filter to a data source:

1. On the **Data Sources** page:
   - Click **Add Data Source** and select a new data source from the drop-down menu, or
   - Click the name of a saved data source to edit its configuration.
2. Select the **Filters** tab.
3. Determine whether you want to apply this filter to files or folders. Then, under the appropriate heading, click **Add**.
4. Select **Date Filter**.
5. In the **Specify Property for Date Filter** dialog, enter the **Property Name** to filter against and click **Ok**.
6. Select either **Include** or **Exclude** for this filter.

    **Note:** Include filters apply only to files and not folders.

7. Specify whether this filter applies **Before** or **After** the specified date.
8. Enter the date in mm/dd/yyyy format, or click the calendar icon to select a date. The filter defaults to 12 AM on the date you select.
9. Click **Save**.

Specifying numeric filters for a data source

Numeric filters allow you to include or exclude content based on properties with numeric values that are evaluated against expressions that you specify.

To add a numeric filter to a data source:

1. On the **Data Sources** page:
   - Click **Add Data Source** and select a new data source from the drop-down menu, or
   - Click the name of a saved data source to edit its configuration.
2. Select the **Filters** tab.
3. Determine whether you want to apply this filter to files or folders. Then, under the appropriate heading, click **Add**.
4. Select **Numeric Filter**.
5. In the **Specify Property for Numeric Filter** dialog, enter the **Property Name** to filter against and click **Ok**.
6. Select either **Include** or **Exclude** for this filter.

    **Note:** Include filters apply only to files and not folders.

7. Select a comparison operator from the drop-down menu.
8. Enter a number in the input field, without commas or decimal points. The filter defaults to bytes as the unit of measure.
9. Click **Save**.

Removing data source filters

If a filter is no longer necessary, you can delete it from a data source configuration.

To remove a data source filter:

1. On the **Data Sources** page, click on the data source you want to edit.
2. Select the **Filters** tab.
3. Locate the filter you want to remove and click the delete icon in the upper-right corner of this filter's configuration section.

Specifying advanced setting for an acquisition

Advanced settings for an acquisition control document conversion, caching, and ACL property retrieval. Settings you configure on this page over ride the default settings for this acquisition.

To specify advanced settings for an acquisition:

1. On the **Data Sources** page, do one of the following:
   - Click **Add Data Source** and select a new data source from the drop-down menu.
   - Click the name of a saved data source to edit its configuration.
2. On the **Edit** page, click the **Show Advanced Settings**.
3. Select a **Log level** from the drop-down menu:

Log level	Description
`Use Server Default`	Logs messages based on the default log level on the CAS Server.
`ALL`	Logs all levels of messages on the CAS Server.
`TRACE`	Logs TRACE-level messages on the CAS Server.
`DEBUG`	Logs DEBUG-level messages on the CAS Server.
`INFO`	Logs INFO-level messages on the CAS Server.
`WARN`	Logs WARN-level messages on the CAS Server.
`ERROR`	Logs ERROR-level messages on the CAS Server.
`FATAL`	Logs FATAL-level messages on the CAS Server.
`OFF`	Disables logging on the CAS Server.

   In this release, the **Log level** setting does not apply to the acquisition processing of data sources built with the CAS Extension API. Logging does apply to starting and stopping acquisition from such data sources, maintaining crawl history, and other internal CAS operations, but not record processing.
4. Specify a value for **Maximum threads** if you wish to modify the number of threads available to the CAS Service. The default number of threads is one more than the number of CPUs of the machine running the CAS Service. For example, if the CAS Service is running on a machine with four CPUs, the default number of threads is five.
5. Click **Save**.

Specifying advanced settings for a data source

Advanced settings control the number of threads available to CAS Service and the CAS Service log level. Settings you configure on this page over ride the default settings for this data source.

To specify advanced settings for a data source:

1. On the **Data Sources** page, do one of the following:
   - Click **Add Data Source** and select a new data source from the drop-down menu.
   - Click the name of a saved data source to edit its configuration.
2. On the **Edit** page, click the **Acquisition** step.
3. Select the **Advanced Settings** tab.
4. Optionally, under Document Conversion, uncheck **Enable document conversion** to disable document conversion. This option is enabled by default.

   

   Note: This option is available to CMS connectors, file system data sources, and may be available to custom data sources. For a full list of file formats that the CAS Document Conversion Module Supports, see Appendix B of the *CAS Developer's Guide*.

5. Optionally, check **Cache files locally** to cache files in the CAS temporary directory.

   

   Note: This option is available to file system data sources and may be available to custom data sources. It is not available to CMS connectors.

6. Optionally, check **Expand archive files**. Enabling this option creates a record for each archived entry and populates the record's properties.

   

   Note: This option is available to CMS connectors and file system data sources. It is not available to custom data sources.

7. Optionally, uncheck **Retrieve ACLs** if you do not want ACL properties for the records.

   

   Note: This option is available to file system data sources. It is not available to CMS connectors and custom data sources.

8. Click **Save**.

Modifying records with CAS manipulators

CAS manipulators modify Endeca records as part of data processing in a CAS acquisition. CAS manipulators typically modify records by adding or deleting properties on records, or by filtering records from the output of a CAS acquisition. Manipulators are optional.

By default, CAS provides a Modifying Script manipulator and a Filtering Script manipulator. These manipulators provide a light-weight solution to perform record manipulation using the BeanShell Scripting Language. Each manipulator has a small example BeanShell script in the default value of the **Script Source** field. You can modify the default BeanShell scripts as necessary for your acquisition.

If the Modifying Script manipulator or the Filtering Script manipulator do not meet your needs, you can implement a custom manipulator using the CAS Extension API. For details about implementation and installation, see the *CAS Extension API Guide*.

Adding a custom manipulator to a data source

You add a custom manipulator to a data source on the **Edit** page of CAS Console. You then specify configuration properties and processing order of the manipulator within the acquisition process. Custom manipulators are extensions to CAS created with the CAS Extension API. For implementation and installation information, see the *CAS Extension API Guide*.

A manipulator must be installed into CAS before you can add it to a data source.

To add a manipulator:

1. On the **Data Sources** page, click a data source name to access its acquisition steps.
2. Click **Add Manipulator**...
3. Select a manipulator and click **Add**.
   The **Manipulator Settings** page displays.
4. Specify configuration properties as necessary for the manipulator.

   To determine the configuration property values, you may have to coordinate with the extension developer who created the manipulator.
5. Click **Save**.

The manipulator displays on the **Acquisition Steps** list.

Adding a Filtering Script manipulator to a data source

You add a **Filtering Script** manipulator to a data source on the **Edit** page of CAS Console. The manipulator runs an inline BeanShell script that filters Endeca records from acquisition output.

The example script, provided in the **Script Source** field, evaluates a record and then returns a Boolean to indicate whether a particular record is included or excluded from the acquisition output. You can modify the inline script as appropriate for your acquisition. A return value of `true` includes a record and `false` excludes the record from the acquisition output. If the script does not return a Boolean, the data source fails.

The manipulator has access to the methods in the `Record` and `PropertyValue` classes (i.e. `com.endeca.itl.record.Record` and `com.endeca.itl.record.PropertyValue`). For details about the methods in these classes, see the CAS Record Store API Reference (Javadoc) installed in `CAS\<version>\doc\recordstore-javadoc`. The manipulator also has access to the methods in the `Logger` class in `org.slf4j.Logger`. Other classes may be imported as necessary.

To add a Filtering Script manipulator to a data source:

1. On the **Data Sources** page, click a data source name to access its acquisition steps.
2. Click **Add Manipulator...** and select **Filtering Script**.
3. Click **Add**.
   The **Manipulator Settings** page displays.
4. In **Manipulator Name**, specify a unique name for the manipulator to distinguish it from other manipulators in the data source.

You can specify a name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.

5. In **Script Source**, modify the inline BeanShell script as appropriate for your data source.
6. Click **Save**.

The manipulator displays on the **Acquisition Steps** list.

Adding a Modifying Script manipulator to a data source

You add a **Modifying Script** manipulator to a data source on the **Edit** page of CAS Console. The manipulator runs an inline BeanShell script that modifies Endeca records.

The example script, provided in the **Script Source** field, adds a property to each record being crawled and then logs that addition. You can modify the inline script as appropriate for your acquisition.

The manipulator has access to the methods in the `Record` and `PropertyValue` classes (i.e. `com.endeca.itl.record.Record` and `com.endeca.itl.record.PropertyValue`). For details about the methods in these classes, see the CAS Record Store API Reference (Javadoc) installed in `CAS\<version>\doc\recordstore-javadoc`. The manipulator also has access to the methods in the `Logger` class in `org.slf4j.Logger`. Other classes may be imported as necessary.

To add a Modifying Script manipulator to a data source:

1. On the **Data Sources** page, click a data source name to access its acquisition steps.
2. Click **Add Manipulator...** and select **Modifying Script**.
3. Click **Add**.
   The **Manipulator Settings** page displays.
4. In **Manipulator Name**, specify a unique name for the manipulator to distinguish it from other manipulators in the data source.

   You can specify a name with alphanumeric characters, underscores, dashes, and periods. All other characters are invalid for a name.
5. In **Script Source**, modify the inline BeanShell script as appropriate for your data source.
6. Click **Save**.

The manipulator displays on the **Acquisition Steps** list.

Editing a manipulator

You edit a manipulator by modifying its configuration options in CAS Console. The configuration changes take effect the next time CAS acquires data from the data source.

You can view manipulator details while an acquisition is running but not edit the configuration options until the acquisition completes.

To edit a manipulator:

1. On the **Data Sources** page of CAS Console, locate the data source that contains the manipulator you want to edit.
2. Click the data source's name to open the **Edit** page for the manipulator.
3. In **Acquisition Steps**, click the name of the manipulator you want to edit.
4. Edit the manipulator settings as necessary.

To determine the configuration property values, you may have to coordinate with the extension developer who created the manipulator, or you can run the `getModuleSpec` task of `cas-cmd` to retrieve the configuration properties of a manipulator.

5. When you finish editing the manipulator, click **Save**.

Disabling a manipulator

You can disable a manipulator on the **Edit** page of CAS Console. By default, a manipulator is enabled after you save it in a data source. A manipulator runs as part of an acquisition process unless you explicitly choose to disable it.

To disable a manipulator:

1. On the **Data Sources** page, click a data source name to access its acquisition steps.
2. Locate the manipulator you want to disable and click the **Disable** link.
3. Click **Save**.

The manipulator displays as Disabled on the **Acquisition Steps** list. You can re-enable the manipulator by clicking the **Enable** link.

Deleting a manipulator from a data source

If a manipulator is no longer necessary, you can delete it from a data source.

To delete a manipulator:

1. On the **Data Sources** page, select the data source that contains the manipulator you want to delete.
2. Under **Acquisition Steps**, locate the manipulator you want to delete and click the icon in the **Delete** column.
3. Click **Yes** to confirm the action.
4. Click **Save**.

Acquiring from a data source

This section covers how to acquire data from a data source that you have configured, and how to abort the acquisition process.

Starting acquisition from a data source

After you create and configure a data source, you can acquire data from it and, if desired, you can monitor its progress or stop acquiring from the data source.

When you acquire from a data source, the CAS Server automatically determines which acquisition mode is necessary. By default, the CAS Server attempts incremental acquisition, and it switches to full acquisition if any of the following conditions are true:

- A data source has not been acquired before, which means no crawl history exists.

- A Record Store instance that stores record output does not contain at least one record generation. This applies to the default case in which the CAS Server is configured to output to a Record Store instance rather than a file on disk.
- Seeds have been removed from the data source configuration (adding seeds does not require full acquisition).
- The document conversion setting has changed.
- Filters have been added, modified, or removed in the data source configuration.
- Repository properties have changed, such as the `username` property setting for CMS data sources.

In all other cases, the CAS Server acquires incrementally. However, you may force full acquisition of a data source. For more information on the difference between full and incremental acquisition, see the *Endeca CAS Developer's Guide*.

Also, a data source many contain one or more manipulators as part of its configuration. In this release, manipulators do not display in CAS Console. You can view, add, and configure manipulators using the CAS Server Command-line Utility. When you start an acquisition, any manipulators included with the data source also run and perform record manipulation.

To acquire from a data source:

1. On the **Data Sources** page, locate the data source from which to acquire data.
2. In the Acquire Data column, Click **Start** or click the icon next to the **Start** button to force full acquisition.

The message in the Status column reads "Acquiring" for this data source. A timestamp appears for the Start Time, and Duration shows the time elapsed since the CAS Server started acquiring from this data source.

Aborting acquisition from a data source

After you start acquiring from a data source, you can abort the process at any time.

Note: When you abort acquisition from a data source, the following occurs:

- The CAS Server produces no record output and all Record Store transactions roll back.
- Metadata history returns to its state before the CAS Server started acquiring from the data source.
- The statistics shown on the detailed status page are not returned to the state before the CAS Server started acquiring from the data source. This page shows information from the aborted acquisition process.

To abort acquisition from a data source:

1. On the **Data Sources** page, locate the data source from which to stop acquiring data. The status message reads "Acquiring."
2. Click **Abort**.

The data source status changes to "Stopping" and then to "Aborted" to indicate that you intentionally stopped acquiring from the data source, and that you can begin acquisition again.

Monitoring a data source

This section covers how to monitor the status of a data source that is being crawled, or view statistics about the most recent results of crawling that data source.

Viewing the status of a data source

The **Acquisition Status** column of the CAS Console page indicates the current status of each data source. In addition to showing status, the table on the Data Sources page displays the start time, end time, and duration of each crawling process.

You can also see the detailed status of a data source by clicking in the **Acquisition Status** field for a data source. Detailed status information includes crawl mode, crawl rate, files and folders crawled, conversion rates, and so on.

Note: Custom data sources created with the CAS Extension API may not display all of the metrics listed below.

Table 1: Definitions of acquisition status messages

Status label	Description
Never Started	The CAS Server has not yet crawled the data source.
Acquiring	The CAS Server is crawling the data source. It may be crawling either fully or incrementally. Refer to the Mode field in the data source status view.
Stopping	The data source is in the process of stopping.
Aborted	A user aborted acquisition from the data source.
Failed	The CAS Server did not successfully crawl the data source. An error message describing the cause of the failure appears below the status label on the data source status view.
Completed	The CAS Server successfully finished crawling the data source.

Table 2: Definitions of detailed acquisition status messages

Status label	Description
Mode	The mode in which the CAS Server crawled the data source. This value can be either full or incremental.
Start time	The time at which crawling of the data source started according to the clock of the machine running the CAS Server.
End Time	The time at which crawling of the data source ended according to the clock of the machine running the CAS Server. This value is empty if the data source is Acquiring.
Duration	The length of time in which the CAS Server crawled the data source.
Crawl Rate	This value is expressed in records per second. It is the total number of Files crawled and Folders crawled, divided by the crawl Duration.

Status label	Description
Files	This is the number of files crawled after filters are applied to the data source. The value is calculated by subtracting the Excluded Files value from the total number of files crawled.
Folders	This is the number of folders included in the data source. This value is calculated by subtracting the Excluded Folders value from the total number of folders crawled.
Excluded Files	This is the number of files excluded from the data source configuration based on your filtering criteria.
Excluded Folders	This is the number of directories excluded from the data source configuration based on your filtering criteria.
Conversion Rate	This value is expressed in records per second. It is calculated by adding the values of Documents Converted plus Conversion failures and dividing by the Crawl Duration.
Documents Converted	The number of documents that the Document Conversion Module successfully converted to text.
Conversion Failures	The number of documents that the Document Conversion Module failed to convert to text.
Endeca Records Created/Updated	The number of new or updated Endeca records.
Endeca Records Deleted	The number of deleted Endeca records.
Endeca Records Failed	The number of Endeca records that failed during the crawl. CAS Server logs each failed record in the `cas-service.log` file.

The CAS Console updates the status of a data source every five seconds.

To view the status of a data source:

1. Select the CAS Console page.
   The **Acquisition Status** column displays the basic status for each data source.
2. If you want to examine detailed status, click the **Acquisition Status** field for a specific data source.
3. Click **Close** when you are done examining the status.

Viewing data source logs

If the status tab for a data source shows that there are failures, you can find further information in the log file.

To view the log file for a data source configuration:

1. On the CAS host, navigate to the log directory of the CAS Server. This directory is located in `<install path>\CAS\workspace\logs` under the root directory of CAS.
2. Open the `cas-service.log` file. You can identify log messages relating to a specific crawl because they contain the name of the data source within brackets. For example, a log message for a data source configuration named myDataSource might look like this:

   ```
   08-06-12 11:48:47 DEBUG [myDataSource] Starting work: Processing c:\My
   Documents\Test.doc (WorkExecutor$WorkRunnable)
   ```

For information on troubleshooting data source crawling errors, see the *Endeca CAS Developer's Guide.*

Index