

Oracle Commerce

Content Acquisition System Console User's Guide

Version 11.2 • October 2015



Contents

Copyright and disclaimer.....	v
Preface.....	7
About this guide.....	7
Who should use this guide.....	7
Conventions used in this guide.....	7
Contacting Oracle Support.....	8
 Chapter 1: Getting started with the Oracle Commerce CAS Console.....	 9
About the Oracle Commerce CAS Console for Oracle Commerce Workbench.....	9
Overview of working with a data source.....	9
 Chapter 2: Working with data sources and manipulators.....	 11
Overview of the default CAS data sources and manipulators.....	11
Adding a Delimited File data source.....	11
Adding an Endeca Record File data source.....	13
Adding a File System data source.....	14
Adding a JDBC data source.....	14
Adding a Record Store Merger data source.....	15
Editing a data source.....	16
Deleting a data source.....	17
Specifying data source filters.....	17
Specifying string filters for a data source.....	17
Specifying date filters for a data source.....	18
Specifying numeric filters for a data source.....	18
Removing data source filters.....	19
Specifying advanced setting for an acquisition.....	19
Specifying advanced settings for a data source.....	20
Modifying records with CAS manipulators.....	21
Enabling script manipulators.....	21
Adding a custom manipulator to a data source.....	21
Adding a Filtering Script manipulator to a data source	22
Adding a Modifying Script manipulator to a data source	23
Editing a manipulator.....	23
Disabling a manipulator.....	24
Deleting a manipulator from a data source.....	24
Acquiring from a data source.....	24
Starting acquisition from a data source.....	24
Aborting acquisition from a data source.....	25
Monitoring a data source.....	25
Viewing the status of a data source.....	25
Viewing data source logs.....	27

Copyright and disclaimer

Copyright © 2003, 2015, Oracle and/or its affiliates. All rights reserved.

This software and related documentation are provided under a license agreement containing restrictions on use and disclosure and are protected by intellectual property laws. Except as expressly permitted in your license agreement or allowed by law, you may not use, copy, reproduce, translate, broadcast, modify, license, transmit, distribute, exhibit, perform, publish, or display any part, in any form, or by any means. Reverse engineering, disassembly, or decompilation of this software, unless required by law for interoperability, is prohibited.

The information contained herein is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

If this is software or related documentation that is delivered to the U.S. Government or anyone licensing it on behalf of the U.S. Government, the following notice is applicable:

U.S. GOVERNMENT END USERS: Oracle programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, delivered to U.S. Government end users are "commercial computer software" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, shall be subject to license terms and license restrictions applicable to the programs. No other rights are granted to the U.S. Government.

This software or hardware is developed for general use in a variety of information management applications. It is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use this software or hardware in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle Corporation and its affiliates disclaim any liability for any damages caused by use of this software or hardware in dangerous applications.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group.

This software or hardware and documentation may provide access to or information on content, products, and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services unless otherwise set forth in an applicable agreement between you and Oracle. Oracle Corporation and its affiliates will not be responsible

for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services, except as set forth in an applicable agreement between you and Oracle.

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at <http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc>.

Oracle customers that have purchased support have access to electronic support through My Oracle Support. For information, visit <http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info> or visit <http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs> if you are hearing impaired.

Preface

Oracle Commerce Guided Search 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, Guided Search enables businesses to influence customers in each step of their search experience. At the core of Guided Search is the MDEX Engine™, a hybrid search-analytical database specifically designed for high-performance exploration and discovery. The Oracle Commerce 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. The Oracle Commerce Assembler dynamically assembles content from any resource and seamlessly combines it into results that can be rendered for display.

Oracle Commerce Experience Manager enables non-technical users to create, manage, and deliver targeted, relevant content to customers. With Experience Manager, you can combine unlimited variations of virtual product and customer data into personalized assortments of relevant products, promotions, and other content and display it to buyers in response to any search or facet refinement. Out-of-the-box templates and experience cartridges are provided for the most common use cases; technical teams can also use a software developer's kit to create custom cartridges.

About this guide

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

Who should use this guide

This guide is intended for application developers who are building applications using the Oracle Commerce 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 answers to implementation questions, product and solution help, and important news and updates about Guided Search software.

You can contact Oracle Support through the My Oracle Support site at <https://support.oracle.com>.

Getting started with the Oracle Commerce CAS Console

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

About the Oracle Commerce CAS Console for Oracle Commerce Workbench

The CAS Console for Oracle Commerce 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 Commerce Workbench.

In a typical implementation, Oracle Commerce developers use the CAS Console for Oracle Commerce 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 Oracle Commerce 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 Commerce 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.
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.
JDBC	Crawls a JDBC-accessible database.
Record Store Merger	Crawls CAS record store instances.

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

Manipulator	Description
Term Extraction	This manipulator extracts terms from Guided Search records and scores them for relevancy.

For more information about configuring a data source or a manipulator run the `cas-cmd` utility with the `get-ModuleSpec` task to return configuration properties.

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 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 MDEX 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 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 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 MDEX 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 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 Commerce 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 properties for data source filters

The following 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.

Log level	Description
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.

- 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.
- 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:

- 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.
- On the **Edit** page, click the **Acquisition** step.
- Select the **Advanced Settings** tab.
- 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*.

- 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.

- 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.

- 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.

- 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.

CAS provides two script manipulators: Filtering Script manipulator and Modifying Script manipulator. To enable them, see [Enabling script manipulators](#) on page 21. 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 these manipulators 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*.

Enabling script manipulators

This section describes how to enable the Filtering Script and Modifying Script manipulators.

Follow these steps:

- Shut down the Oracle Commerce Content Acquisition System Server.
- In the `%CAS_ROOT%\lib\cas-server-plugins\cas\plugin.xml` file. Uncomment the following two plugin extension entries:

```
<!--extension id="com.endeca.cas.manipulator.FilterScriptManipulator"
class="com.endeca.itl.extension.manipulator.script.FilterScriptManipulator"
pointId="com.endeca.cas.extension.Manipulator"/>

<extension id="com.endeca.cas.manipulator.ModifierScriptManipulator"
class="com.endeca.itl.extension.manipulator.script.ModifierScriptManipulator"
pointId="com.endeca.cas.extension.Manipulator"/-->
```

- Start the Oracle Commerce Content Acquisition System Server.

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.

Before you begin, you must enable the script. See [Enabling script manipulators](#) on page 21 for more information.

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.

Before you begin, you must enable the script. See [Enabling script manipulators](#) on page 21 for more information.

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 *CAS Developer's Guide*.

Also, a data source may 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.
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.

Status label	Description
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 <code>cas-service.log</code> 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 *CAS Developer's Guide*.

Index

A

- aborting
 - data source acquisition 25
- adding data sources
 - File System 14
 - JDBC 14
 - Delimited File 12
 - Endeca Record 13
 - Record Store Merger 15
- adding manipulators 22

C

- CAS Console
 - introduced 9
- CAS Server
 - introduced 9
- configuring
 - filters 17
 - advanced settings 20
- crawling
 - process overview 9

D

- deleting
 - data sources 17
 - manipulators 24
- disabling manipulators 24

E

- editing
 - data sources 16, 23

R

- removing
 - filters 19

S

- starting
 - data source acquisition 24

V

- viewing
 - data source logs 27
 - data source status 25

