Coveo Platform 7.0

Amazon S3 Connector Guide

Notice

The content in this document represents the current view of Coveo as of the date of publication. Because Coveo continually responds to changing market conditions, information in this document is subject to change without notice. For the latest documentation, visit our website at www.coveo.com.

Document part number: PM-150604-EN

Publication date: 1/3/2019

Table of Contents

1. Amazon S3 Connector

CES 7.0.7711+ (June 2015)

Amazon simple storage service (S3) is a cloud based object storage, designed to store, distribute and manage a large quantity of data worldwide.

The Coveo connector for Amazon S3 allows Coveo administrators to index and integrate the Amazon S3 content into the Coveo unified index and make it easily searchable by end-users.

Note: An access key is needed to connect to the Amazon Web Services (AWS) service through the software development kit (SDK). The access key is a way to authenticate from the SDK as an Identity and Access Management (IAM) account. The number of requests is unlimited, but every request to your Amazon S3 bucket(s) has a charge (see Request Pricing).

1.1 Connector Features Summary

Features		Supported	Additional information
Amazon S3 version		Latest cloud version	Following available Amazon S3 releases
Searchable content types[1]		✔	Buckets[2] and objects (folders and files)
Content update	Incremental refresh	✖	
	Full refresh	✔	
	Rebuild	✔	
Document-level security		✖	Permissions must be manually defined on the source [more]

1- An access key is needed to connect to the Amazon Web Services (AWS) service through the software development kit (SDK). The access key is a way to authenticate from the SDK as an Identity and Access Management (IAM) account. The number of requests is unlimited, but every request to your Amazon S3 bucket(s) has a charge (see Request Pricing).

2- Amazon S3 Requester Pays buckets are not supported.

1.2 Features

The Amazon S3 connector features are:

Content indexing

The connector can retrieve and index exclusively the following Amazon S3 types of items:

- Buckets
- Objects (folders and files)

Note: The connector supports Amazon S3 server-side encryption on object data.

Pause/Resume

When indexing Amazon S3 content, the connector can be paused and resumed.

1.3 Limitations

- The Amazon S3 connector does not support the Amazon S3 security model. Thus, permissions must be manually defined on the source (see Permissions).
- The Requester Pays feature in Amazon S3 is not supported (see Requester Pays Buckets).

What's Next?

Review the steps to deploy the Amazon S3 connector (see "Amazon S3 Connector Deployment Overview" on page 3).

2. Amazon S3 Connector Deployment Overview

The following procedure outlines the steps needed to deploy the Amazon S3 connector. The steps indicate the order in which you must perform configuration tasks on the Coveo server.

To deploy the Amazon S3 connector

1. Validate that your environment meets the requirements (see "Amazon S3 Connector Requirements" on page 4).
2. On the Coveo server, in the Coveo Administration Tool:
   a. Create an Amazon S3 field set to take advantage of the available Amazon S3 metadata.
      i. It is recommended to start by importing the default Amazon S3 field set file (`[CES_Path]\Bin\Coveo.CES.CustomCrawlers.AmazonS3.FieldSet.xml`) to create fields for all the metadata available by default from Amazon S3 documents.
      ii. When you created custom metadata for your Amazon S3 documents, add corresponding fields to the field set.
   b. Configure and index an Amazon S3 source.

      The connector must know details about the authorized access to the Amazon S3 bucket(s) to index its content (see "Configuring and Indexing an Amazon S3 Source" on page 5).
   c. If you encounter issues, verify if modifying the default value of hidden source parameters can help resolve the problems (see "Modifying Hidden Amazon S3 Source Parameters" on page 11).

3. Amazon S3 Connector Requirements

Your environment must meet the following requirements to be able to use the Amazon S3 connector:

- CES 7.0.7711+ (June 2015)
- Coveo license for the Amazon S3 connector

  Your Coveo license must include support for the Amazon S3 connector to be able to use this connector.
- A valid IAM Account

  You need an IAM account with at least the Read permission on the bucket to be crawled.

What's Next?

Create an Amazon S3 field set (see Amazon S3 Connector Deployment Overview).

4. Configuring and Indexing an Amazon S3 Source

A source defines a set of configuration parameters for indexing the content of a specific Amazon S3 site.

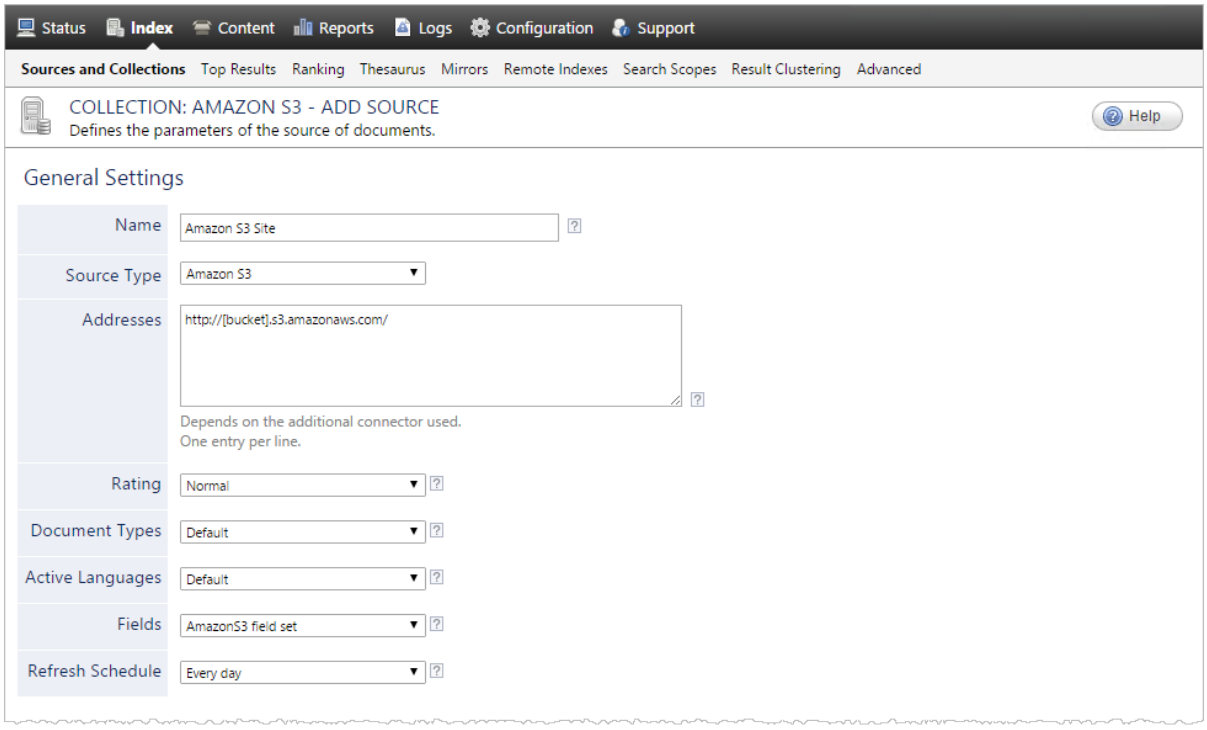
To configure and index an Amazon S3 source

1. On the Coveo server, access the Administration Tool.
2. Select **Index** > **Sources and Collections**.
3. In the **Collections** section:
   a. Select an existing collection in which you want to add the new source.

      OR

   b. Click **Add** to create a new collection.
4. In the **Sources** section, click **Add**.

   The **Add Source** page that appears is organized in three sections.
5. In the **General Settings** section of the **Add Source** page:



   a. Enter the appropriate value for the following required parameters:

      Name

      A descriptive name of your choice for the connector source.

Example: `Amazon S3 Site`

Source Type

The connector used by this source. In this case, select **Amazon S3**.

Addresses

The address of the Amazon S3 bucket site in one of the following types:

- Virtual-host style

  Examples:

  - http://[bucket].s3.amazonaws.com/
  - http://[bucket].s3-[aws-region].amazonaws.com/

  where you replace `[bucket]` by your actual bucket name and `[aws-region]` with your region-specific endpoint.

- Path style

  Examples:

  - http://s3.amazonaws.com/[bucket]
  - http://s3-[aws-region].amazonaws.com/[bucket]

  where you replace `[bucket]` by your actual bucket name and `[aws-region]` with your region-specific endpoint.

You can enter more than one bucket address on separate lines, but you must ensure that all source parameters apply to all Amazon S3 buckets. Otherwise, create other sources for other buckets.

Notes:

- The starting address must specify one bucket with its region. URLs that do not specify any region are using the US Standard (us-east-1) region endpoint.
- When the URL point to a folder inside a bucket, only keys starting with that prefix will be crawled.
- You can index more than one bucket.

Fields

If you defined an Amazon S3 field set, select it (see Amazon S3 Connector Deployment Overview).

Refresh Schedule

Time interval at which the index is automatically refreshed to keep the index content up-to-date. By default, the **Every day** option instructs CES to refresh the source everyday at 12 AM.

Note: You can create a new or modify an existing source refresh schedule.

b. Review the value for the following parameters that often do not need to be modified:

Rating

Change this value only when you want to globally change the ranking associated with all items in this source relative to the rating of other sources.

Example: If this source is for a legacy PLM, you may want to set this parameter to **Low**, so that in the search interface, results from this source appear later in the list compared to those from other sources.

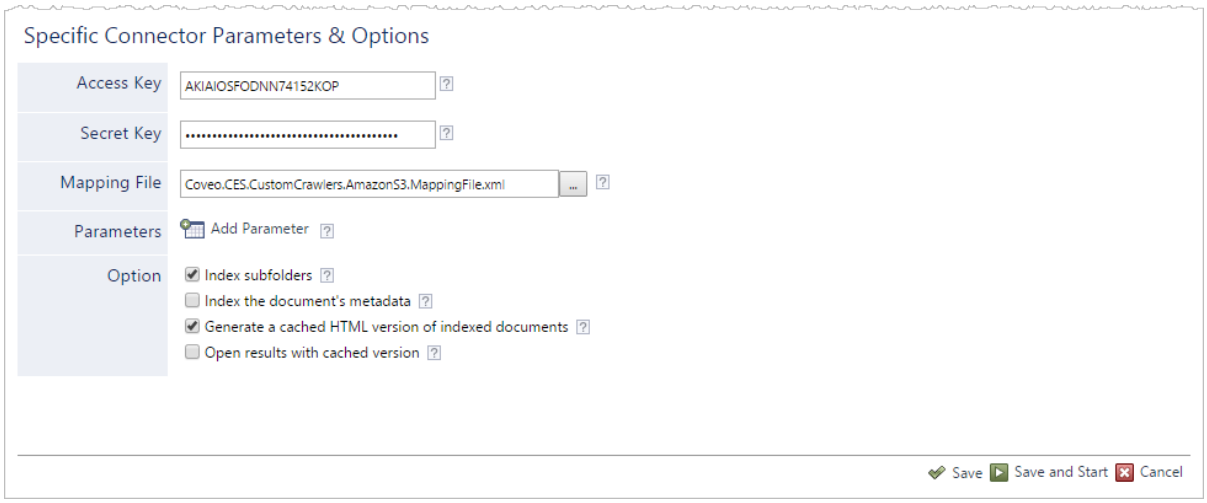
Document Types

If you created a custom document type set for this source, select it. Otherwise, leave **Default**.

Active Languages

If you defined custom active language sets, ensure to select the most appropriate for this source.

6. In the **Specific Connector Parameters & Options** section of the **Add Source** page:



   a. When the Amazon S3 content is private, enter the appropriate value for the following parameters. Otherwise (for public data set) leave them empty:

Notes:

- The Access Key and Secret Key are accessible in the IAM console (see Understanding and Getting Your Security Credentials).
- CES 7.0.7914+ (October 2015) The Access Key and Secret Key parameters are optional.

Access Key

The ID of the IAM account access key used to request data from the Amazon S3 servers.

Example: `AKIAIOSFODNN74152KOP`

Secret Key

The IAM account secret access key used to request data from the Amazon S3 servers.

Example: `wJalrXUtnFEMI/K7MDENG/bPxRfiCYifc51AYQQf`

b. In the **Mapping File** box, leave the default mapping file name (`Coveo.CES.CustomCrawlers.AmazonS3.MappingFile.xml`) unless you created a custom mapping file, in which case, enter the full path of your valid mapping file.

c. Click **Add Parameter** when you want to show and change the value of hidden source parameters (see "Modifying Hidden Amazon S3 Source Parameters" on page 11).

d. In the **Option** section, the state of check boxes generally does not need to be changed:

Index Subfolders

Check to index all subfolders below the specified starting addresses.

Index the document's metadata

When selected, CES indexes all the document metadata, even metadata that are not associated with a field. The orphan metadata are added to the body of the document so that they can be searched using free text queries.

When cleared (default), only the values of system and custom fields that have the **Free Text Queries** attribute selected will be searchable without using a field query.

Example: A document has two metadata:

- `LastEditedBy` containing the value `Hector Smith`
- `Department` containing the value `RH`

In CES, the custom field `CorpDepartment` is bound to the metadata `Department` and its **Free Text Queries** attribute is selected.

When the **Index the document's metadata** option is cleared, searching for `RH` returns the document because a field is indexing this value. Searching for `hector` does not return the document because no field is indexing this value.

When the **Index the document's metadata** option is selected, searching for `hector` also returns the document because CES indexed orphan metadata.

Generate a cached HTML version of indexed documents

When you select this check box (recommended), at indexing time CES creates HTML versions of indexed documents and saves them in the unified index. In the search interfaces, users can then more rapidly review the content by clicking the **Quick View** link to open the HTML version of the item rather than opening the original document with the original application.

Consider clearing this check box only if you do not want to use **Quick View** links or to save resources when building the source.

Open results with cached version

Leave this check box cleared (recommended) so that in the search interfaces, the main search result link opens the original document with the original application. Consider selecting this check box only when you do not want users to be able to open the original document but only see the HTML version of the document as a Quick View. When this option is selected, you must also select the **Generate a cached HTML version of indexed documents** check box.

e. Click **Save** to save the source configuration.

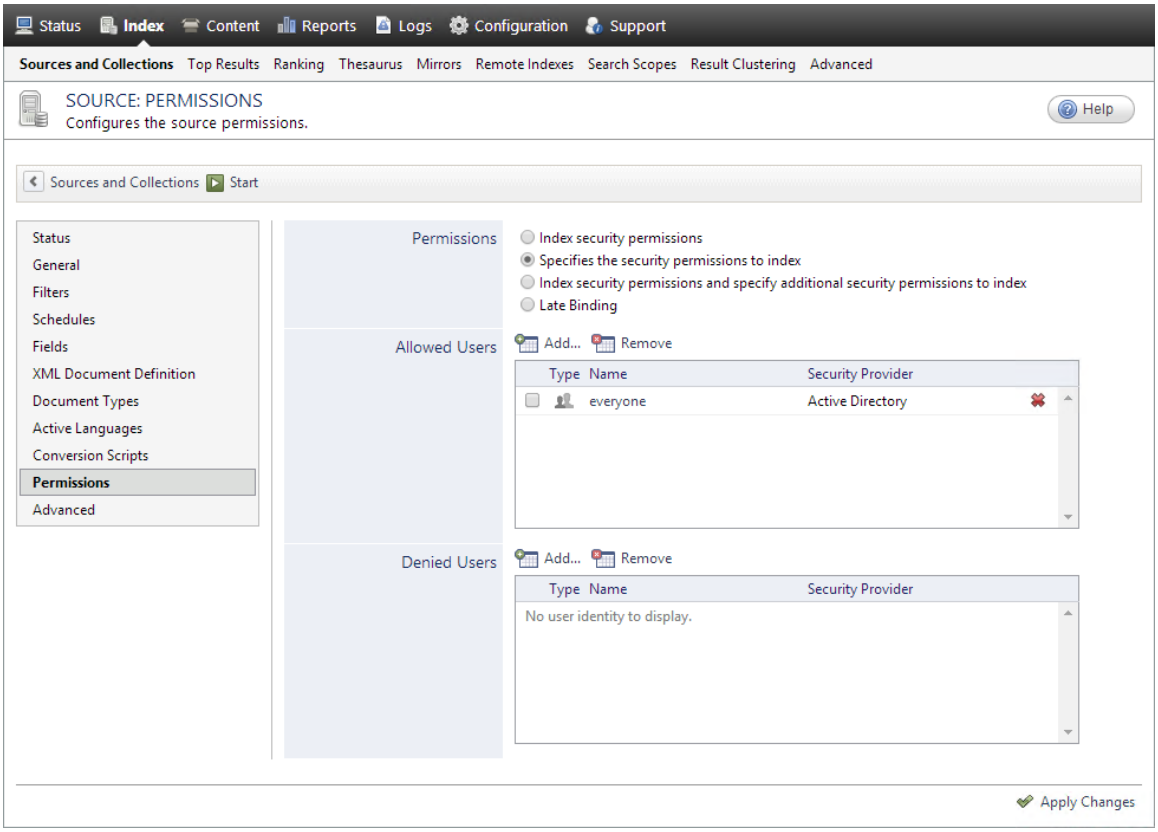
7. Because Amazon S3 security model is not yet supported, the Amazon S3 connector does not index permissions and you must change the default **Permissions** option to set the permissions globally on the source:

Note: You get the following error message in the CES Console when the **Index security permissions** option is selected by default:

```
Permissions indexing is not provided by AmazonS3Crawler.
```

a. In the navigation panel on the left, select **Permissions**.

b. In the **Permissions** page:



i. Select the **Specifies the security permissions to index** option.

ii. In the **Allowed Users** list, add or remove users or groups to precisely specify who has access to the content from this source.

By default, the Active Directory **everyone** group specifies that any Active Directory user can see all the content from this source.

iii. Optionally, in the **Denied Users** list, add users or groups to specify who has not access to the content from this source.

iv. Click **Apply Changes**.

8. On the toolbar, click **Start/Rebuild** to start indexing your source.
9. Validate that the source building process is executed without errors:

- In the navigation panel on the left, click **Status**, and then validate that the indexing proceeds without errors.

OR

- Open the CES Console to monitor the source building activities.

5. Modifying Hidden Amazon S3 Source Parameters

The **Add Source** and **Source: ... General** pages of the Administration Tool present the parameters with which you can configure the connector for most Amazon S3 sites. More advanced and more rarely used parameters are hidden. You can choose to make one or more of these parameters appear in the **Add Source** and **Source: ... General** pages of the Administration Tool so that you can change their default value.

The following list describes the available advanced hidden parameters for Amazon S3 sources. The parameter type (integer, string…) appears between parentheses following the parameter name.

BatchSize (Integer)

The number of objects to retrieve by request (between 1 and 1000). The default value is `100`.

To modify hidden Amazon S3 source parameters

1. Refer to "Adding an Explicit Connector Parameter" on page 11 to add one or more Amazon S3 hidden source parameters.

   For a new Amazon S3 source, access the **Add Source** page of the Administration Tool to modify the value of the newly added advanced parameter:

2.
   a. Select **Index** > **Sources and Collections**.

   b. Under **Collections**, select the collection in which you want to add the source.

   c. Under **Sources**, click **Add**.

   d. In the **Add Source** page, edit the newly added advanced parameter value.

3. For an existing Amazon S3 source, access the **Source: ... General** page of the Administration Tool to modify the value of the newly added advanced parameter:

   a. Select **Index** > **Sources and Collections**.

   b. Under **Collections**, select the collection containing the source you want to modify.

   c. Under **Sources**, click the existing Amazon S3 source in which you want to modify the newly added advanced parameter.

   d. In the **Source: ... General** page, edit the newly added advanced parameter value.

5.1 Adding an Explicit Connector Parameter

Connector parameters applying to all sources indexed using this connector are called explicit parameters.

When you create or configure a source, the Coveo Enterprise Search (CES) 7.0 Administration Tool presents parameters with which you can configure the connector for most setups. For many connectors, more advanced and more rarely used parameters also exist but are hidden by default. CES then uses the default value associated with each of these hidden parameters.

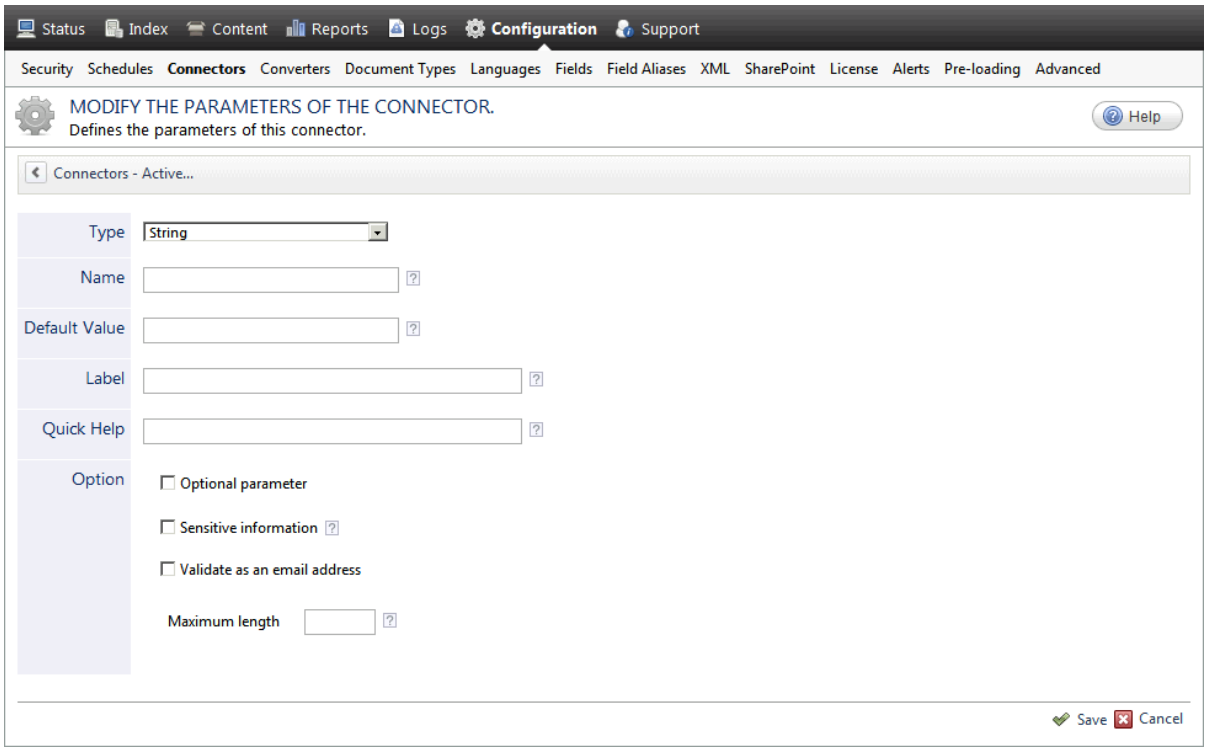
You can however choose to make one or more of these parameters appear in the **Add Source** and **Source: ... General** pages of the Administration Tool so that you can change their default value.

To add an explicit connector parameter

1. On the Coveo server, access the Administration Tool.
2. Select **Configuration** > **Connectors**.
3. In the list on the **Connectors** page, select the connector for which you want to show advanced hidden parameters.
4. In the **Parameters** section of the selected connector page, click **Add Parameter** for each hidden parameter that you want to modify.

   Note: The **Add Parameter** button is present only when hidden parameters are available for the selected connector.

5. In the **Modify the parameters of the connector** page:



   a. In the **Type** list, select the parameter type as specified in the parameter description.

   b. In the **Name** box, type the parameter name exactly as it appears in the parameter description. Parameter names are case sensitive.

   c. In the **Default Value** box, enter the default value specified in the parameter description.

Important: Do not set the value that you want to use for a specific source. The value that you enter here will be used for all sources defined using this connector so it must be set to the recommended default value. You will be able to change the value for each source later, in the **Add Source** and **Source: ... General** pages of the Administration Tool.

d. In the **Label** box, enter the label that you want to see for this parameter.

Example: To easily link the label to the hidden parameter, you can simply use the parameter name, and if applicable, insert spaces between concatenated words. For the **BatchSize** hidden parameter, enter `Batch Size` for the label.

Note: To create multilingual labels and quick help messages, use the following syntax: `<@ln>text</@>`, where *ln* is replaced by the language initials—the languages of the Administration Tool are English (en) and French (fr).

Example: `<@fr>Chemin d'accès du fichier de configuration</@><@en>Configuration File Path</@>` is a label which is displayed differently in the French and English versions of the Administration Tool.

Tip: The language of the Administration Tool can be modified by pressing the following key combination: `Ctrl+Alt+Page Up`.

e. Optionally, in **Quick Help**, enter the help text that you want to see for this parameter when clicking the question mark button that will appear beside the parameter value.

Tip: Copy and paste key elements of the parameter description.

f. When **Predefined values** is selected in the **Type** parameter, in the **Value** box that appears, enter the parameter values that you want to see available in the drop-down parameter that will appear in the Administration Tool interface. Enter one value per line. The entered values must exactly match the values listed in the hidden parameter description.

g. Select the **Optional parameter** check box when you want to identify this parameter as an optional parameter. When cleared, CES does not allow you to save changes when the parameter is empty. This parameter does not appear for **Boolean** and **Predefined values** parameter types.

h. Select the **Sensitive information** check box for password or other sensitive parameter so that, in the Administration Tool pages where the parameter appears, the typed characters appear as dots to mask them. This parameter appears only for the **String** type.

Example: When you select the **Sensitive information** check box for a parameter, the characters typed appear as follows in the text box:

i. Select the **Validate as an email address** check box when you want CES to validate that the text string that a user enters in this parameter respects the format of a valid email address. This parameter appears only for the **String** type.

j. In the **Maximum length** box, enter the maximum number of characters for the string. This parameter

appears only for the **String** type. When you enter `0`, the length of the string is not limited.

k. Click **Save**.

6. Back in the **Connector** page, click **Apply Changes**.

   The hidden parameter now appears in the **Add Source** and **Source: ... General** pages of the Administration Tool for the selected source. You can change the parameter value from these pages. Refer to the documentation for each connector for details.

Note: When you want to modify a hidden source parameter, you must first delete it, and then redefine it with the modified values.