Content Manager

Software Version 24.3

Elasticsearch Integration Installation and Configuration Guide

opentext™

Document Release Date: March 2023
Software Release Date: July 2024

Legal notices

Documentation updates

The title page of this document contains the following identifying information:

- Software Version number, which indicates the software version.
- Document Release Date, which changes each time the document is updated.
- Software Release Date, which indicates the release date of this version of the software.

To check for updated documentation, visit https://www.microfocus.com/support-and-services/documentation/.

Support

Visit the MySupport portal to access contact information and details about the products, services, and support that OpenText offers.

This portal also provides customer self-solve capabilities. It gives you a fast and efficient way to access interactive technical support tools needed to manage your business. As a valued support customer, you can benefit by using the MySupport portal to:

- View information about all services that Support offers
- Submit and track service requests
- Contact customer support
- Search for knowledge documents of interest
- View software vulnerability alerts
- Enter into discussions with other software customers
- Download software patches
- Manage software licenses, downloads, and support contracts

Many areas of the portal require you to sign in. If you need an account, you can create one when prompted to sign in.

Contents

Getting started with Elasticsearch

Content Manager 24.3 provides support for integration with Elasticsearch for document content indexing. Customers have the option to choose either IDOL or Elasticsearch for this functionality.

Elasticsearch is an open source search and content analytics engine built on top of Apache Lucene™.

Unlike IDOL, Elasticsearch is a third-party product not shipped with Content Manager (i.e. non-OEM). Instead, organisations choosing to use Elasticsearch will be required to download it from elastic.co.

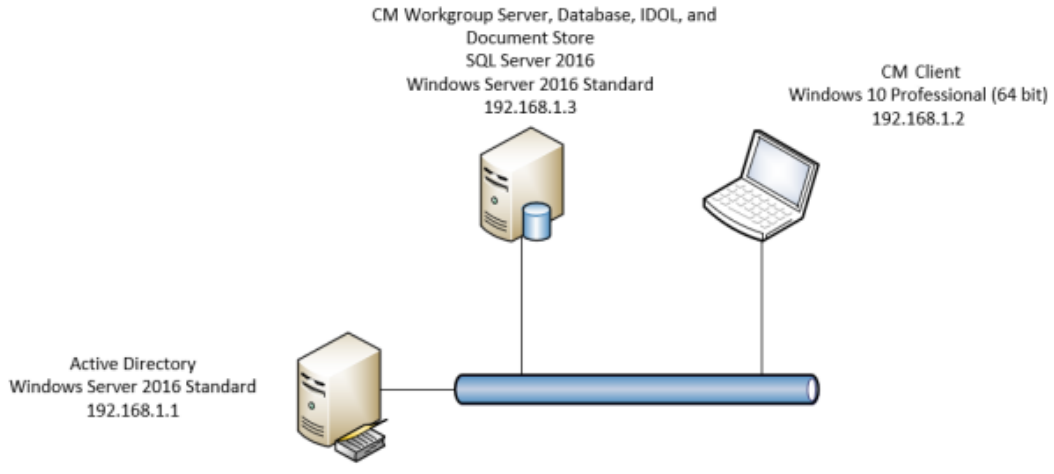
This guide will help you understand what's required to establish a basic Content Manager 24.3 and Elasticsearch instance after upgrading from a previously supported release or installing Content Manager 24.3 for the first time.

Before you begin, it's important to note Elasticsearch can be configured and scaled in many different ways. Specialist Elasticsearch configuration advice and support cannot be provided by the Content Manager team. If, after reading this guide, you believe your organization requires specialist Elasticsearch configuration advice and support, you should consult the Elasticsearch team directly regarding your requirements.

Environment overview

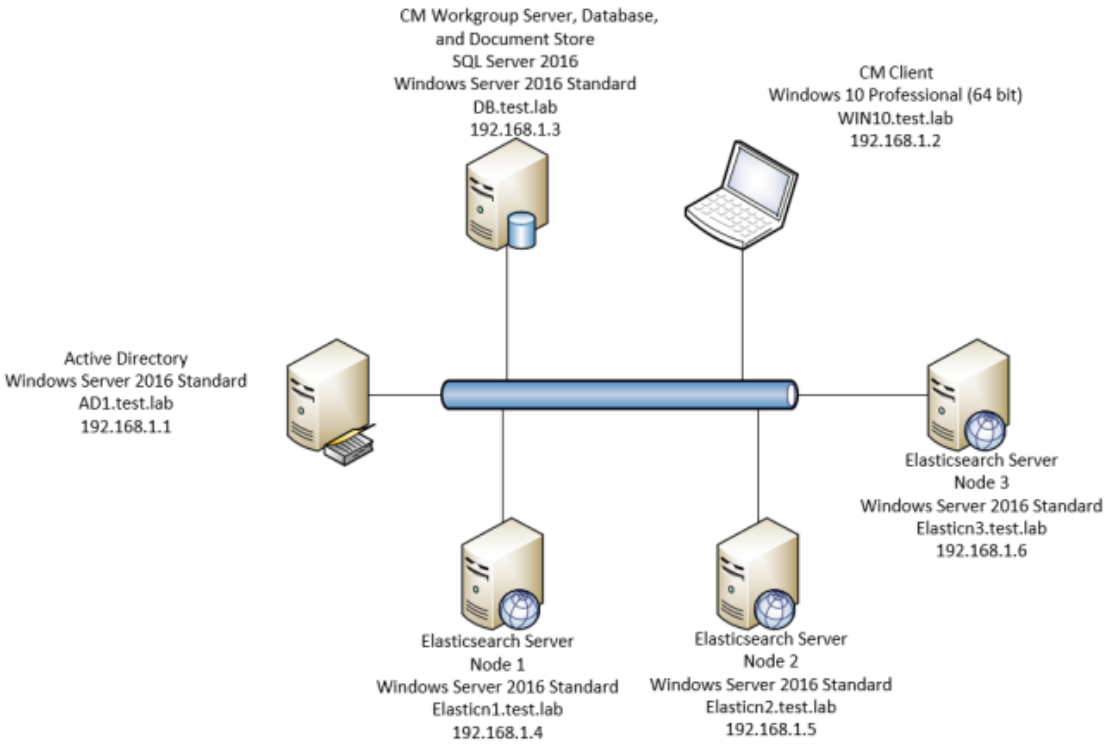
This guide will begin by running you through a scenario where a customer is upgrading from Content Manager 9.4 and IDOL to Content Manager 24.3 and Elasticsearch.

The starting environment will be a simple Content Manager 9.4 and IDOL implementation. We assume you already know how to install Content Manager 24.3 as outlined in Installation Guide (**CM24.3_ Install.pdf**), for example:



Whether you're upgrading or doing a new installation the resulting Content Manager 24.3 and Elasticsearch environment, in this example scenario, will include three additional "node" servers in an

Elasticsearch cluster. This is inline with the Elasticsearch team's best practice guidelines regarding the minimum number of nodes you should have in a production cluster:



Install Content Manager

Upgrading from earlier versions

If you are running an earlier version of Content Manager we recommend upgrading to Content Manager 24.3 and IDOL environment before attempting to install and configure Elasticsearch. This will allow you to confirm your environment and content searches are working as expected prior to reconfiguring your environment to use Elasticsearch integration instead of IDOL.

Follow the instructions outlined in the Installation and Setup Guide (**CM24.3_Install.pdf**) to upgrade your environment to Content Manager 24.3.

New Installations

If you're installing Content Manager 24.3 for the first time you should do so as outlined in the Installation and Setup Guide (**CM24.3_Install.pdf**) before you implement Elasticsearch.

Install Elasticsearch

When you have upgraded to or installed Content Manager 24.3 you can proceed to installing Elasticsearch on each of the servers that will make up the cluster.

We do not recommend installing Elasticsearch on a Workgroup Server as it can use ports that are used by Content Manager or IDOL, and adds unnecessary complexity when you're trying to troubleshoot a problem in the future. The Elasticsearch team are very clear on keeping it simple. We suggest you follow this approach when designing and setting up your environment.

Elastic.co recommend you dedicate a separate server for each Elasticsearch node as we've done in the above network diagram.

The Elasticsearch team recommend a minimum of three nodes for production environments so that if any single node fails no data will be lost.

Install Elasticsearch on your nominated servers

Download Elasticsearch 7.x or 8.x from elastic.co and unzip it to your preferred directory.

In this example, we downloaded the zip package and unzipped it to `C:\Elasticsearch\` on Elasticn1, Elasticn2, and Elasticn3.

Edit the Elasticsearch configuration file

Elasticsearch is configured via a configuration file (`elasticsearch.yml`) located at `\config` folder in the directory you've just unzipped it to.

You can open, edit, and save this file using a text editor like TextPad or Notepad.

The most common settings are already in this file and you'll find they're commented out via the # symbol at the start of each line. The following table lists the ones that are most commonly used:

node.name - Name of the server node. This name is just for reference. If you leave this blank Elasticsearch will choose a random default name for you.

cluster.name - For multi node systems, each node needs to have the same cluster name. All nodes with the same cluster will use a network discovery protocol to find the other nodes in the cluster, and automatically configure themselves. The default name is *elasticsearch*, but you should change it to an appropriate name which describes the purpose of the cluster. A cluster does not become active until this is set.

path.data - Path where Elasticsearch stores the index data. The default is the 'data' directory in the Elasticsearch folder.

path.logs - Path where Elasticsearch stores the log files. The default is the 'log' directory in the Elasticsearch folder.

network.host - The node will bind to this hostname or IP address and publish (advertise) this host to other nodes in the cluster. Accepts an IP address, hostname, a special value, or an array of any combination of these.

discovery.seed_hosts - Out of the box, without any network configuration, Elasticsearch will bind to the available loopback addresses and will scan local ports 9300 to 9305 to try to connect to other nodes running on the same server. This provides an auto- clustering experience without having to do any configuration. When you want to form a cluster with nodes on other hosts, you must use the discovery.seed_hosts setting to provide a list of other nodes in the cluster that are master-eligible and likely to be live and contactable in order to seed the discovery process. This setting should normally contain the addresses of all the master-eligible nodes in the cluster. This setting contains either an array of hosts or a comma-delimited string.

cluster.initial_master_nodes - When you start a brand new Elasticsearch cluster for the very first time, there is a cluster bootstrapping step, which determines the set of master-eligible nodes whose votes are counted in the very first election. In development mode, with no discovery settings configured, this step is automatically performed by the nodes themselves. As this auto-bootstrapping is inherently unsafe, when you start a brand new cluster in production mode, you must explicitly list the master-eligible nodes whose votes should be counted in the very first election. This list is set using the cluster.initial_master_nodes setting.

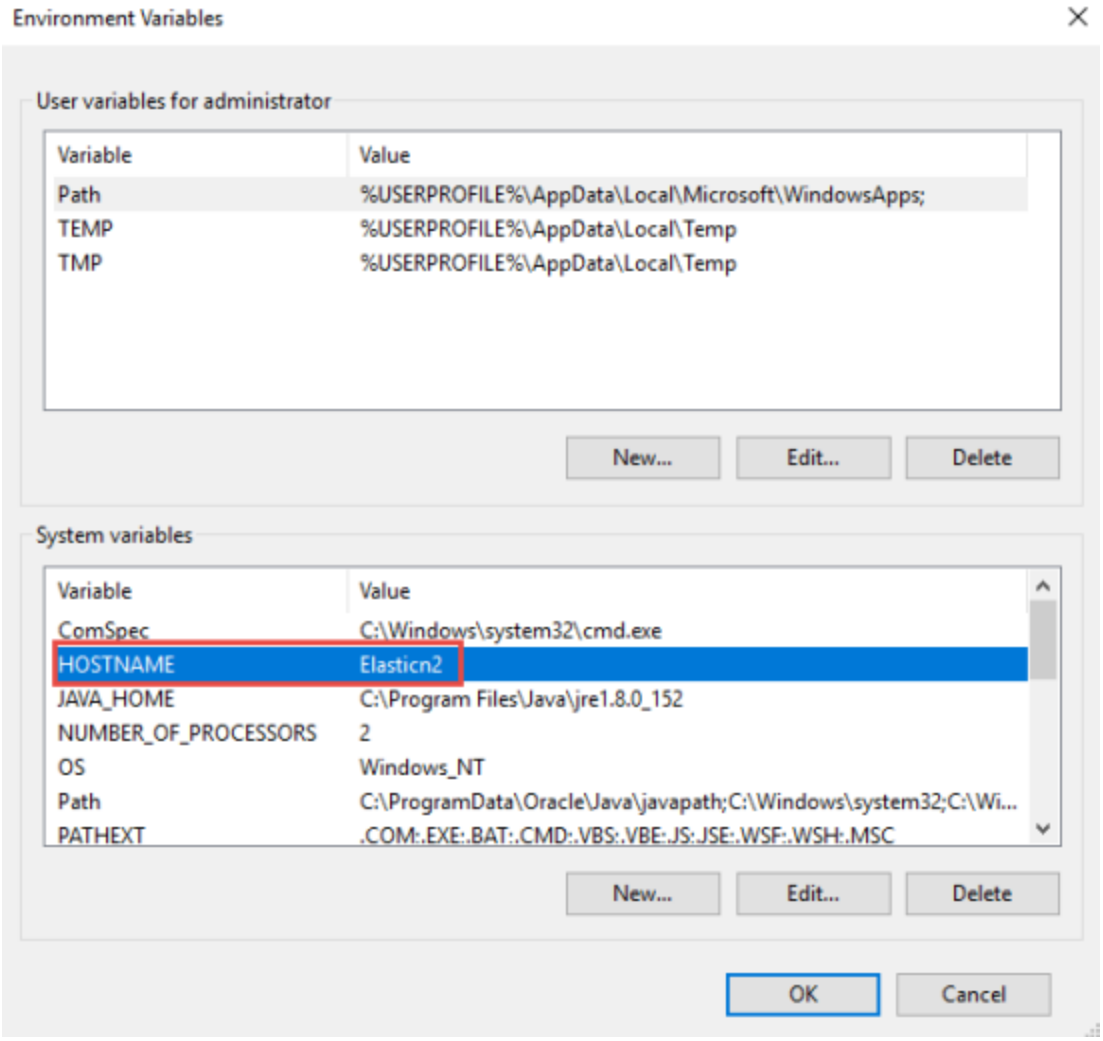
To setup the three-node cluster in our environment we'll be changing these settings in the configuration file on each server:

node.name - We'll set this to be the server's hostname by removing the # and using the entry ${HOSTNAME}

```
# ---------------------------------- Node -
#
# Use a descriptive name for the node:
#
node.name: ${HOSTNAME}
#
# Add custom attributes to the node:
#
#node.attr.rack: r1
#
```

To use this entry, you'll also have to enter the HOSTNAME environment variable like you did for JAVA_HOME.

This will be Elasticn1, Elasticn2, and Elasticn3 on each of the node servers.



cluster.name - We'll set this by removing the # and using the entry `contentmanager`

```
# ----------------------------Cluster--------
#
# Use a descriptive name for your cluster:
#
cluster.name: contentmanager
xpack.security.enabled: true
#
```

network.host - We'll set this by removing the # and using the entry `_site_` which will set the value to any site-local addresses on the system, for example 192.168.1.4

```
# ---------------------------------- Network -----------------------------------
#
# Set the bind address to a specific IP (IPv4 or IPv6):
#
network.host: _site_
#
# Set a custom port for HTTP:
#
#http.port: 9200
#
# For more information, consult the network module documentation.
```

discovery.seed_hosts - We'll set this by removing the # and replacing the host entries with the IP addresses of the three node servers that will make up the cluster (192.168.1.4, 192.168.1.5, and 192.168.1.6). We could have also used hostnames instead of IP addresses (Elasticn1, Elasticn2, and Elasticn3).

```
# ---------------------------------- Discovery -----------------------------------
#
# Pass an initial list of hosts to perform discovery when this node is started:
# The default list of hosts is ["127.0.0.1", "[::1]"]
#
discovery.seed_hosts: ["192.168.1.4", "192.168.1.5", "192.168.1.6"]
#
```

> **TIP:** If you change these settings after your environment is operational you'll have to stop and start the Elasticsearch service to implement the changes.

Reset password and update authentication (Only for Elasticsearch 8.x)

The Elasticsearch 8.x includes built-in user `elastic` with auto-generated password. You can reset the password and configure this user to work with Content Manager.

To reset the password, see the Elasticsearch documentation available at https://www.elastic.co/guide/en/elasticsearch/reference/current/zip-windows.html.

To configure this user to work with Content Manager, perform the following steps:

1. In Content Manager Enterprise Studio, right-click the dataset and navigate to **Content Index > Properties**.
2. In the **Authentication** tab.
3. Check the **Enable X-Pack authentication** check box.
4. Enter user name as `elastic`.
5. Click **Set Password**. The **Set Password** dialog is displayed.
6. Enter the same password you reset in the above steps for **Enter a new value for password** field.
7. Click **OK**.
8. In the **General** tab, click **Test** and make sure that the connectivity to the Elasticsearch cluster is successful. Click **OK** in the pop up dialog.

Memory Management and configuring the Java Virtual Machine (JVM)

Getting your memory allocation configured correctly is critical to the successful operation of Elasticsearch. You must size your hardware and configure your environment appropriately to handle the indexing load your organization will be generating.

There are three memory variables you have to consider when setting up Elasticsearch:

- **Installed RAM**

The Elasticsearch team advise a server that has 64 GB of RAM is the ideal amount in a production environment. The Elasticsearch team do not recommend having less than 8 GB of RAM as it can lead to a need for many small servers.

- **Heap memory**

This is the memory allocated to the JVM in which Elasticsearch runs. The Elasticsearch team recommend giving 50% of the available memory to Elasticsearch heap, while leaving the other 50% free.

When setting the heap memory, it is also recommended you do not allocate more than 26 GB. This is due to a Java limitation that becomes apparent at 32 GB and can have a negative impact on performance. This limitation is described in detail in the **Heap: Sizing and Swapping** section of **Elasticsearch: The Definitive Guide** on the elastic.co web page.

- **Off heap memory**

This is the portion of memory used by the operating system and Lucene (the underlying engine used by Elasticsearch) outside of the JVM heap.

When configuring the JVM heap you should ensure there's enough memory left to accommodate off heap processes. In line with the above recommendation, this should be around 50% of the available memory.

Elasticsearch provides a way to manage JVM settings via the `jvm.options` file which has the default heap size set to 1 GB. You can open, edit, and save this file using text editor like TextPad or Notepad. You'll find this file in:

```
<Drive>:\<Elasticsearch installation directory>\config\
```

In our test environment, we'll be setting the initial and maximum heap size to 1g (1 GB).

```
################################################################
## IMPORTANT: JVM heap size
################################################################
##
## You should always set the min and max JVM heap
## size to the same value. For example, to set
## the heap to 4 GB, set:
##
## -Xms4g
## -Xmx4g
##
## See https://www.elastic.co/guide/en/elasticsearch/reference/current/heap-size.html
## for more information
##
################################################################

# Xms represents the initial size of total heap space
# Xmx represents the maximum size of total heap space

-Xms1g
-Xmx1g
```

Because of a bug in Elasticsearch 7.2 that doesn't set up the Java configuration properly when installing Elasticsearch as a service at the next step, we'll also explicitly specify a path for the Java IO temp location. Also in the jvm.options file set the following parameter to your choice of temporary location, for example

Change this line from the default value

```
-Djava.io.tmpdir=${ES_TMPDIR}
```

To

```
-Djava.io.tmpdir=C:\Windows\Temp
```

> **NOTE:** Path cannot contain spaces.

```
# log4j 2
-Dlog4j.shutdownHookEnabled=false
-Dlog4j2.disable.jmx=true

-Djava.io.tmpdir=C:\Windows\Temp

## heap dumps
```

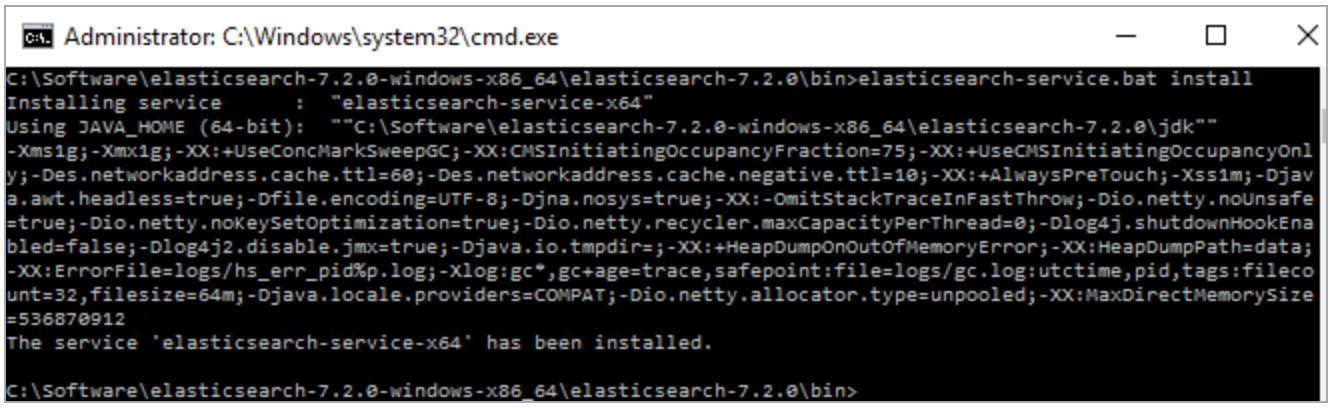
Save your changes to `jvm.options` before proceeding.

Configure Elasticsearch to run as a service

You can use the supplied batch script `<Drive>:\<Elasticsearch installation directory>\bin\elasticsearch-service.bat` to install Elasticsearch as a Windows service.

To do this, open an administrative command prompt, navigate to the bin directory, and run the following command on each Elasticsearch server:

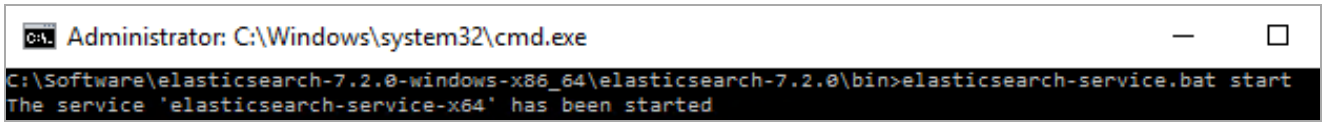
```
elasticsearch-service.bat install
```



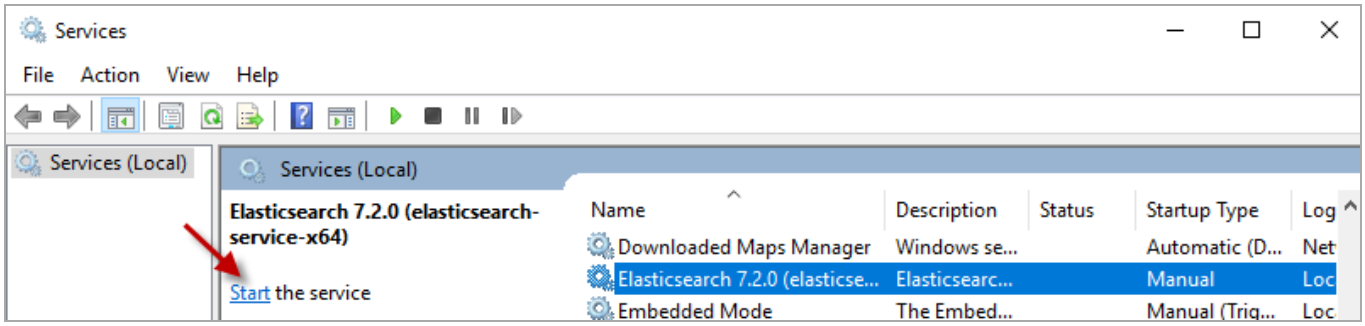
Run Elasticsearch

Start the Elasticsearch service on each server via an administrative command prompt using the following command:

```
elasticsearch-service.bat start
```



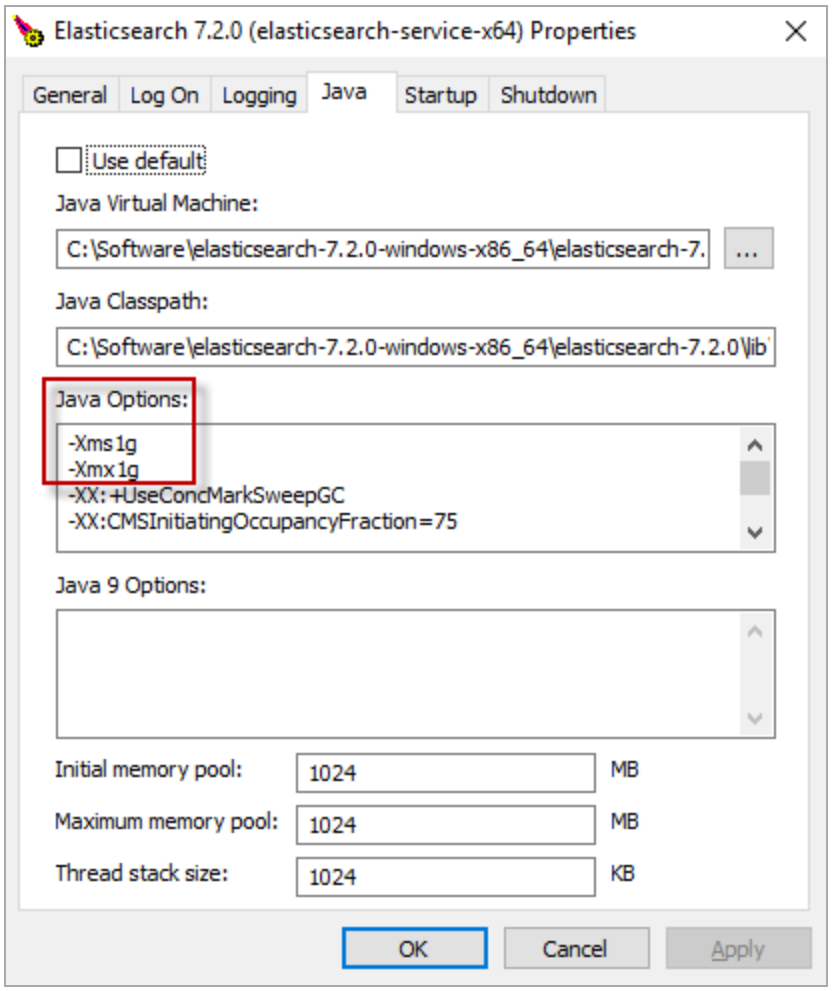
You can also start the Elasticsearch service via the Windows Services Management Console (Start Button > Run > services.msc)



Adjusting the JVM heap size after installing Elasticsearch as a service

To adjust the JVM heap size after installing Elasticsearch as a service you'll need to use the service manager GUI located in the …\bin directory. Editing `jvm.options` will not pass on heap changes to the service once it's installed.

You can open the service manager window by invoking `elasticsearch-service.bat` manager from the command line:



Any changes you make via the service manager will require you to re-start the service which can be done from the **General** tab.

Confirm the Elasticsearch service and cluster are operational

Check the health of your cluster by opening `http://<ip address or hostname>:9200/_cat/health?v` in a browser. The IP address or hostname used can be any Elasticsearch server in your environment.

Elasticsearch uses a traffic-light scheme for server health: green, yellow if there are warnings (e.g. disk space more than 85% full), and red for serious errors.

The Elasticsearch service can take some time to start if you've had a reason to restart it e.g. after changing configuration values. During this period, you may see the health status as **red**. If this occurs, you should wait until the service is finished its start-up sequence and try the health check again.

Configure an Elasticsearch Content Index

Content Manager 24.3 will only allow you to have one type of content indexing engine configured. If you upgrade from earlier versions your existing IDOL content index will continue to be available in 24.3.

However, if you decide to use Elasticsearch as your content index your IDOL content index will no longer be available.

> **NOTE:** If you decide to go back to an IDOL content index, you will need to delete your Elasticsearch content index and create an new IDOL index. See **Setting up document content indexing and searching** in the Content Manager Enterprise Studio help file.

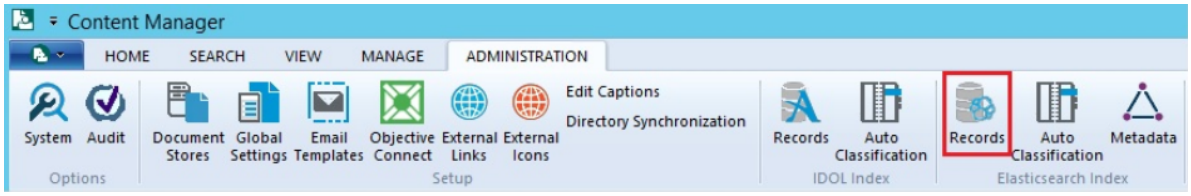
Now we have a working Elasticsearch environment in place we can move on to enabling it within Content Manager 24.3.

Reindex your document store with Elasticsearch

If you're setting up a new instance of Content Manager there's no need to perform this step yet. You can go straight to Create a new Elasticsearch index, on page 23 and follow the steps to setup a new Content Index in Enterprise Studio.

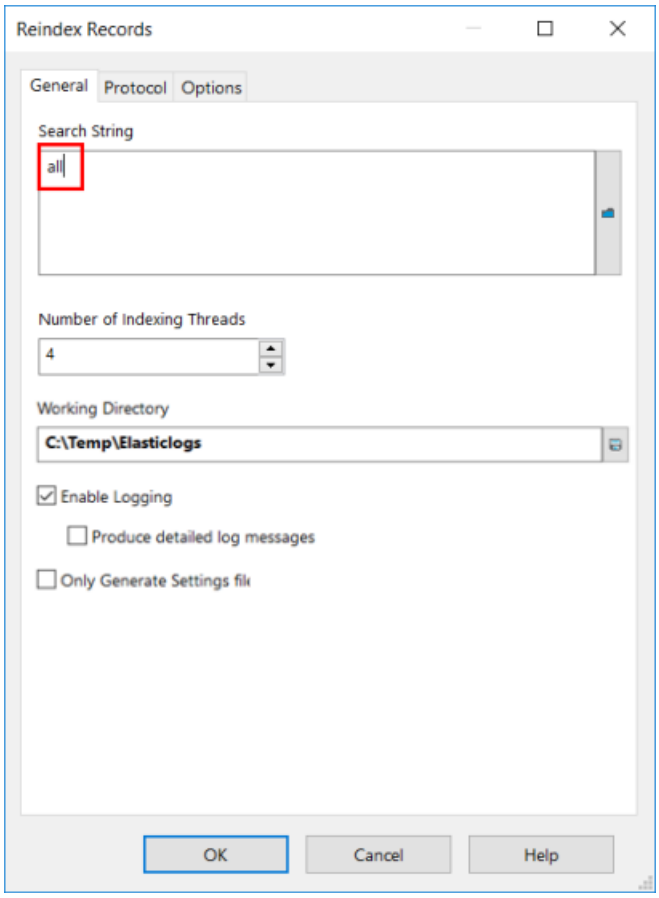
Open the Content Manager client and select the **Administration** tab.

From the **Administration** tab, in the **Elasticsearch Index** group, click **Records**.



The **Reindex Records** dialog will appear.

In the **Search String** field on the **General** tab, type all.



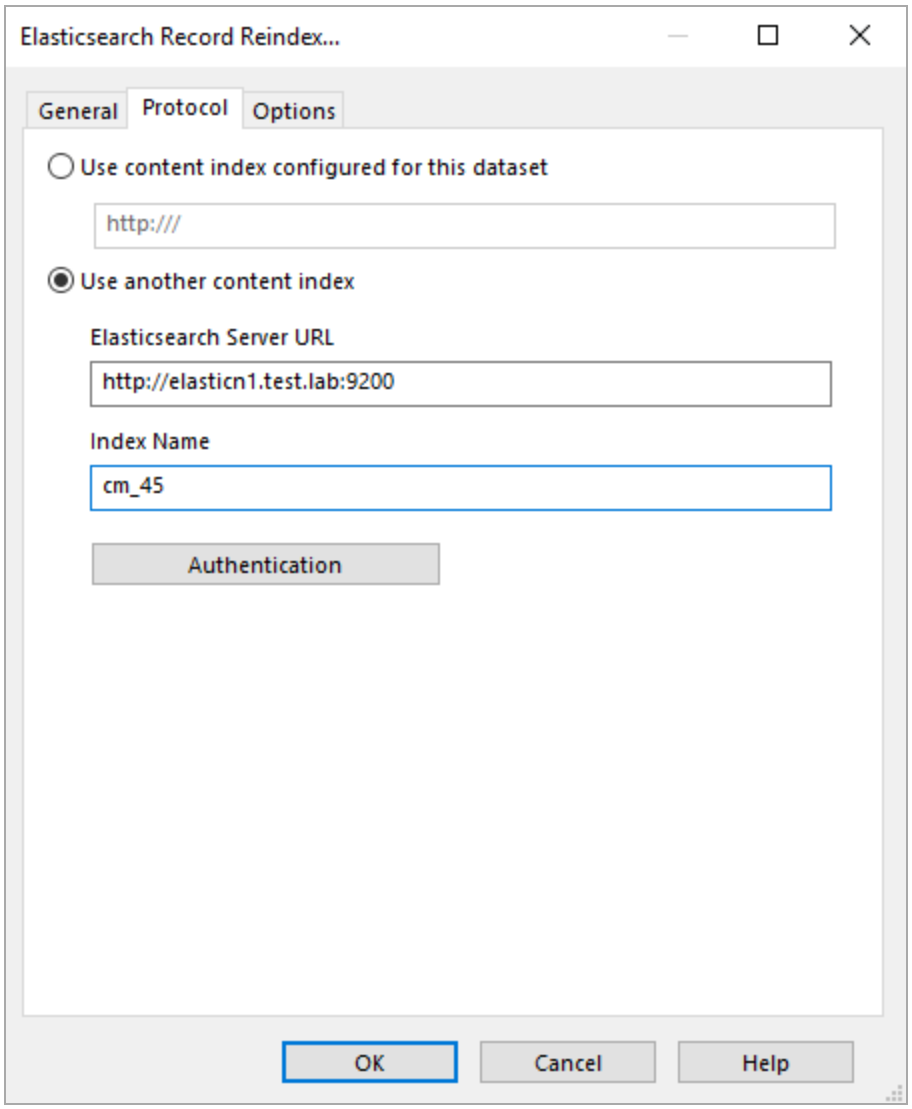
IMPORTANT: To successfully complete a full re-index, make sure you are using a Content Manager location profile that has sufficient access to bypass all current security settings you may have in place.

You also have the opportunity to potentially improve performance by increasing the **Number of Indexing Threads**. This is largely dependent on your hardware. You'll have to trial a number of different thread settings to find out exactly how many threads it's capable of handling. For this scenario we'll use the default value of 4.

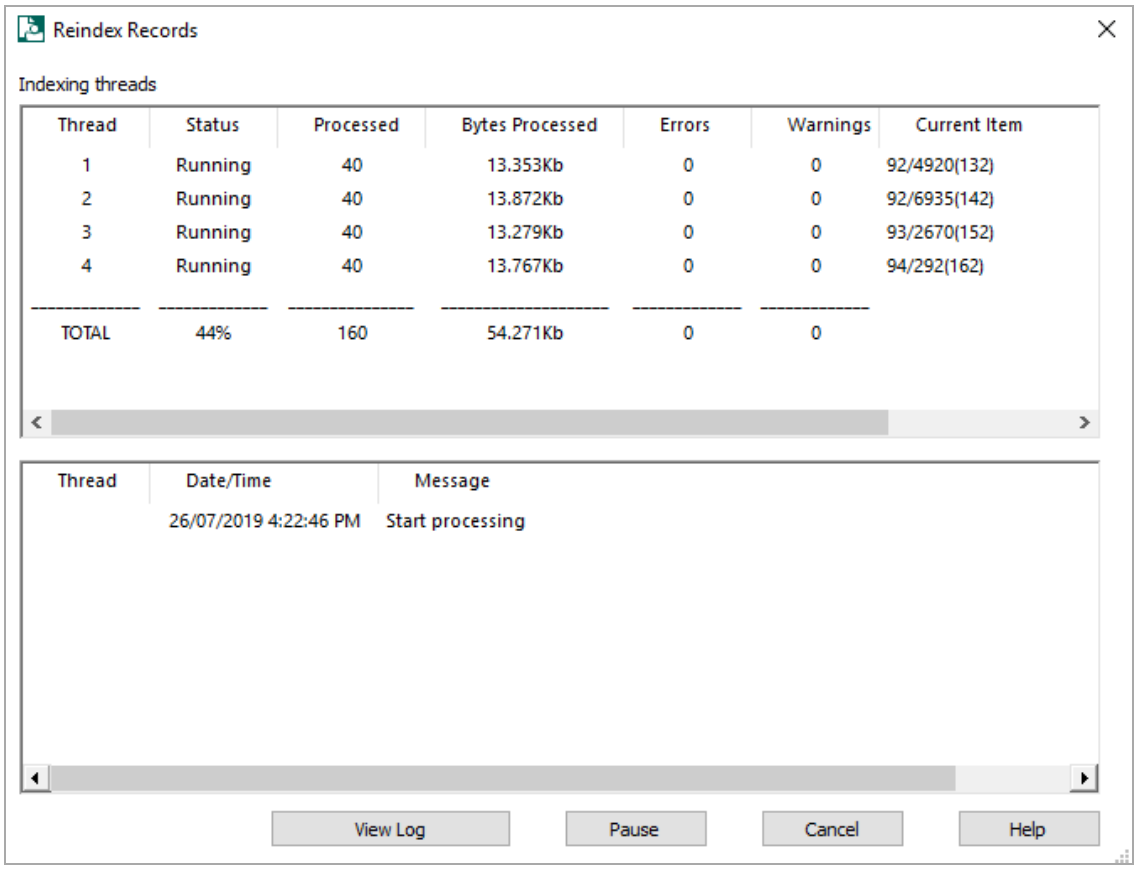
There is a checkbox for logging, **Enable Logging and Produce detailed log messages**, which allows verbose logging. Also, by not checking the checkbox, logging is disabled entirely, apart from a header entry in the log file which shows the settings and date.

The checkbox, **Only Generate Settings file**, when selected generates a settings file for use with the Content Manager command line tool and doesn't perform re-indexing. It is saved to the **Working Directory**.

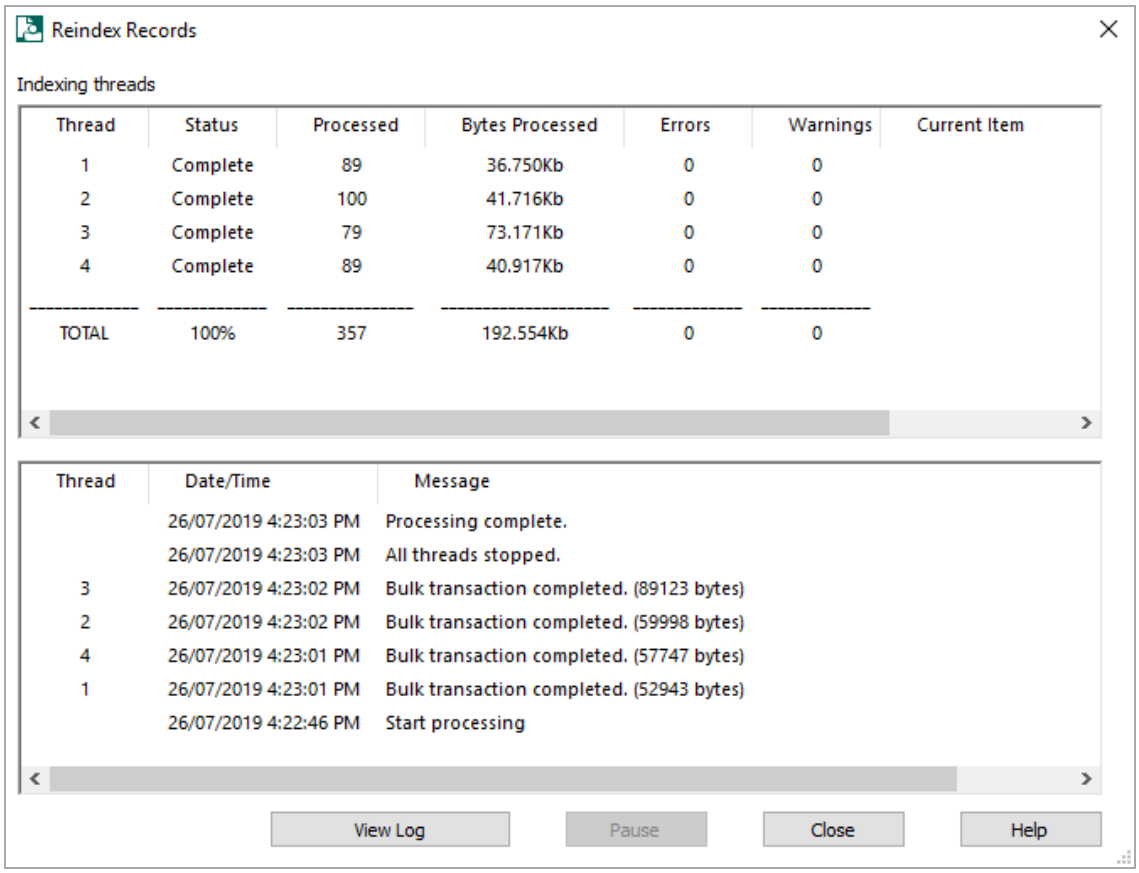
Because we haven't configured an Elasticsearch index yet in Content Manager Enterprise Studio, select **Use another content index** on the **Protocol tab** and in the **Elasticsearch server URL** field type the URL for the Elasticsearch server. This can be any Elasticsearch server in the cluster.



Click **OK** and the re-indexing process will start.



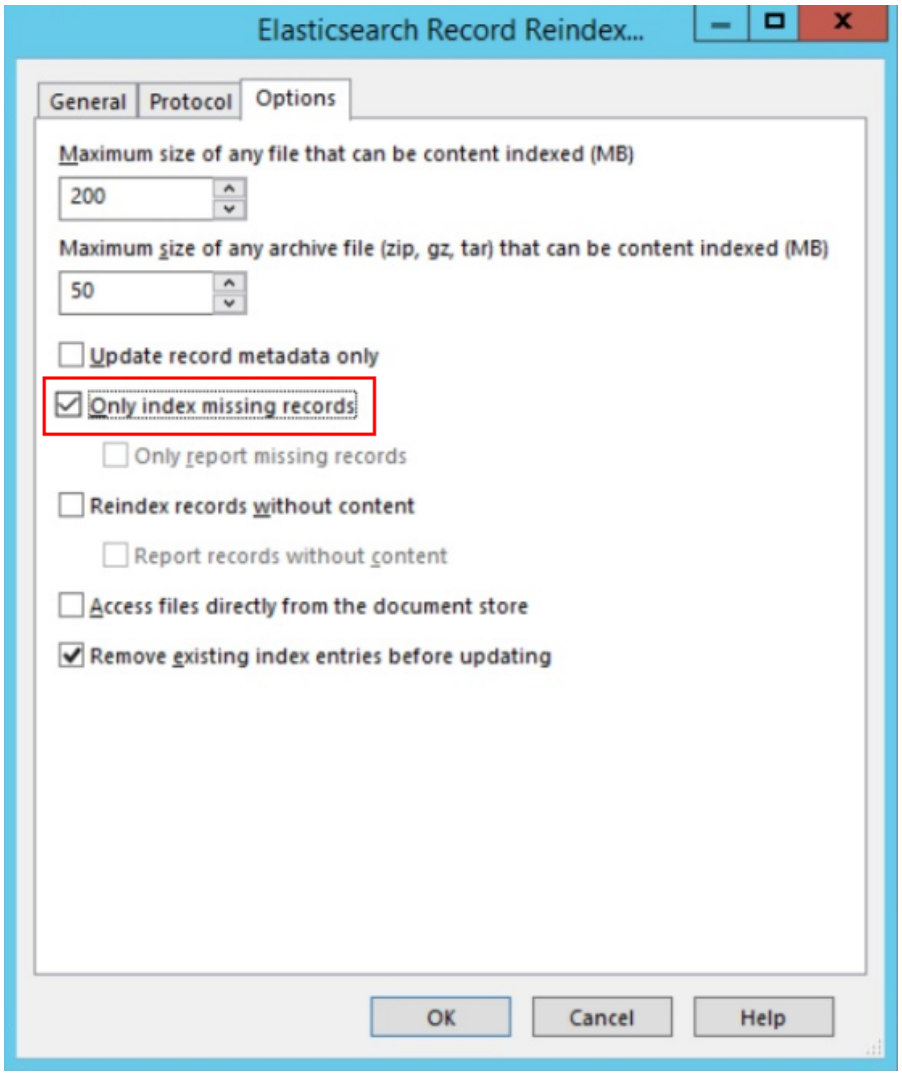
When the re-indexing process is complete the Complete status will be displayed along with any errors it encountered.



If there are errors, click **View Log** to get additional information.

To rectify this, we can re-run the re-index but this time only index those records that are missing by selecting the **Only index missing items** on the **Options** tab.

The option **Only report missing items** is available when the option **Only index missing items** is selected. Select the option **Only report missing items** to merely find the records that have not been indexed before and then report them in the log files, but not reindex them.



The option **Reindex records without content**, checks the Elasticsearch index for records that have electronic documents associated with them but do not have their content in the index. When this option is set only records with missing content will be reindexed.

When the option **Report records without content** is checked, it reports on which records are missing content and writes this information to the log file.

> **NOTE:** The **Report missing records** and **Report records without content** can only be run with a single thread. If the thread count is greater than 1, a warning will be displayed.

The default indexing behaviour is to transfer the document from the document store to the client's working directory using the Content Manager Workgroup server and then extract the text there. When the option **Access files directly from the document store** is selected the document content is extracted directly from the document store, as long as the file can be read. This will improve indexing performance if the indexing client is located close to the document store i.e. on the same network space, and also reduce the amount of disk space required. This option is not recommended over slow network connections.

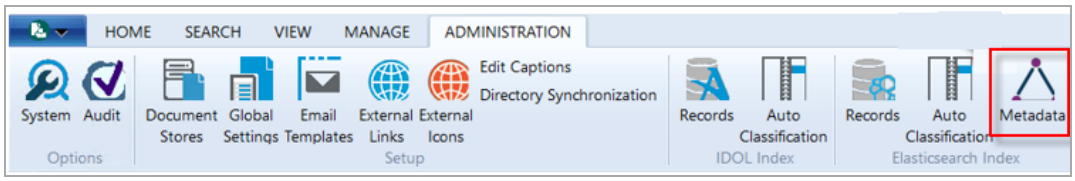
Remove existing index entries before updating option

When a record's electronic document is indexed into Elasticsearch, and there is text content that can be indexed, at least 2 entries are created in the Elasticsearch index. The first entry is the parent item (or 'document' in Elasticsearch terminology) that contains all the metadata for the record, and the second is a child item that contains the content of the electronic document. Depending on the content index setting in Content Manager Enterprise Studio for "Elasticsearch Document Content Field size limit", and the amount of text to be indexed, there may be many more child items(documents) created. Additionally, if the electronic document is a compound file (eg. a zip file), then at least one child item(document) is created for each file found in the compound file.

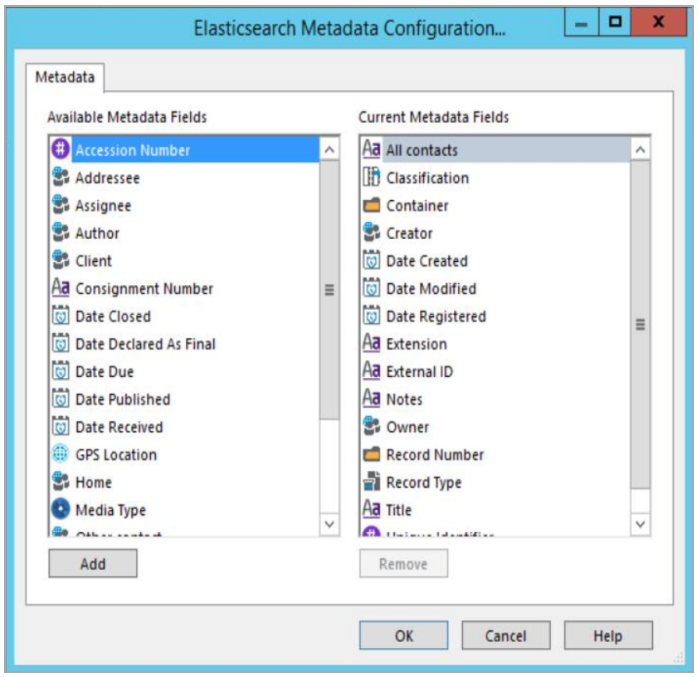
If the electronic document is updated, then we have no way of knowing if the number of child items will change, so we need to do a search for all child items and remove them before updating the entry for this record. This is what happens for a content index event update for the Record. However, there is the overhead of searching for, and then removing these child items. When performing a reindex operation on a large dataset, this has the potential to significantly slow down the operation. For a new index that has no data, or for the case where the user knows that the electronic documents have not changed to a large degree, then disabling this option will allow the reindex to complete sooner.

Configure Elasticsearch Metadata

A **Metadata** option has been added to the **Adminstration** tab in the client that allows you to add metadata fields that will be indexed by Elasticsearch.



Click **Metadata** to display **Elastic metadata configuration** dialog.



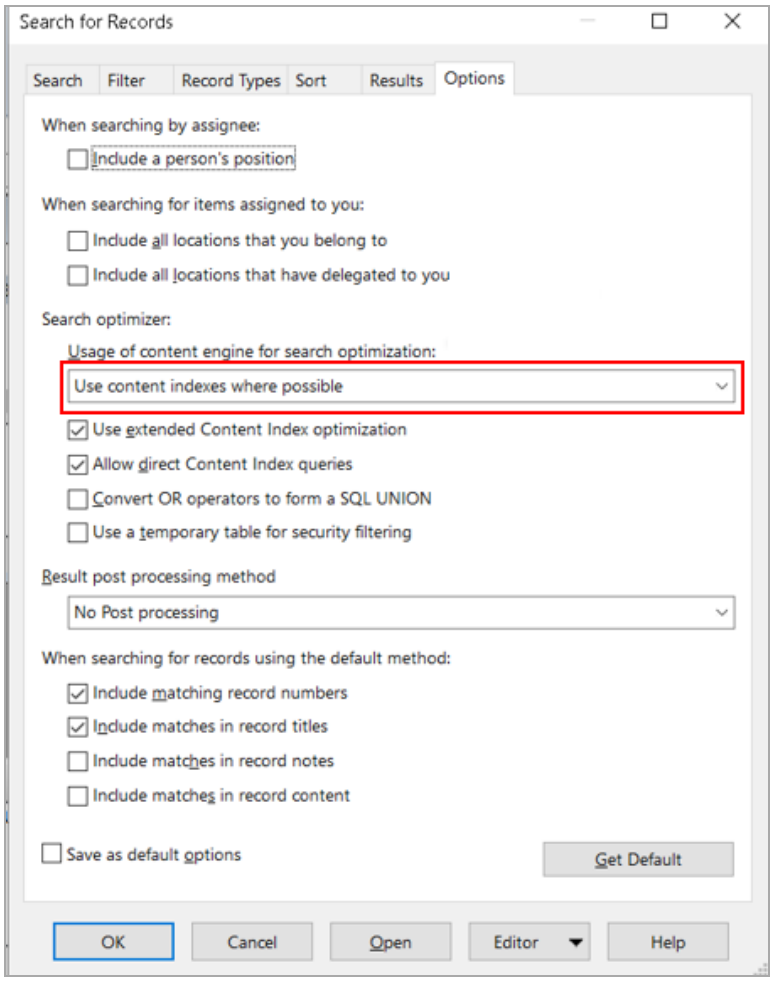
The **Current Metadata Fields** displays the default list of metadata fields that will be indexed by Elasticsearch. You can modify the current list at any time to include additional metadata fields from the **Available Metadata Fields** that will be indexed by Elasticsearch.

Custom or User defined fields will show up in the available list after creation.

If the current metadata fields list has new fields added after index creation, the new metadata will only be indexed from that point forward. If you want to include the new metadata field for all previously indexed content, a reindex of the metadata is required.

Metadata fields cannot be removed after they are added to the **Current Metadata Fields** list. To remove metadata from the Elasticsearch index, remove the index and recreate it using Content Manager Enterprise Studio, followed by a reindex.

To make use of the metadata fields in Elasticsearch, the settings for the search optimiser should be set to **Use content indexes where possible**.

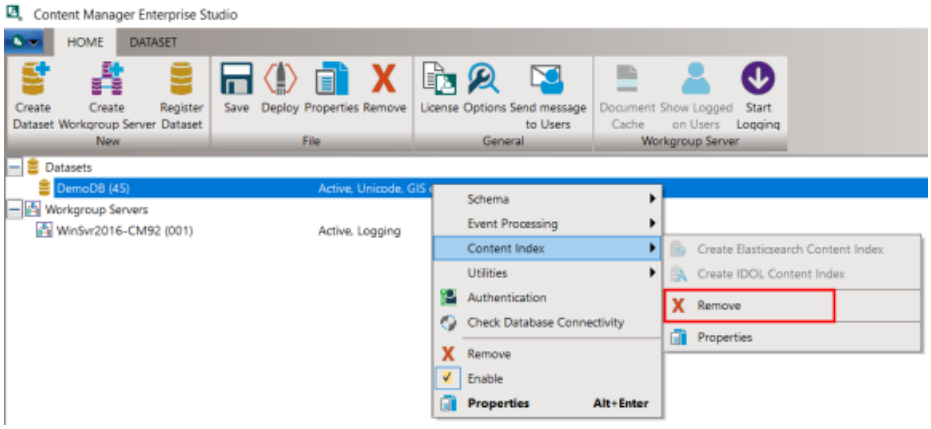


Configure Content Manager Enterprise Studio to use Elasticsearch

Remove existing IDOL content indexes

Now we've successfully created an Elasticsearch content index the existing IDOL index can be removed.

On your Workgroup Server open Enterprise Studio -> right mouse click on your Dataset -> **Content Index -> Remove**.



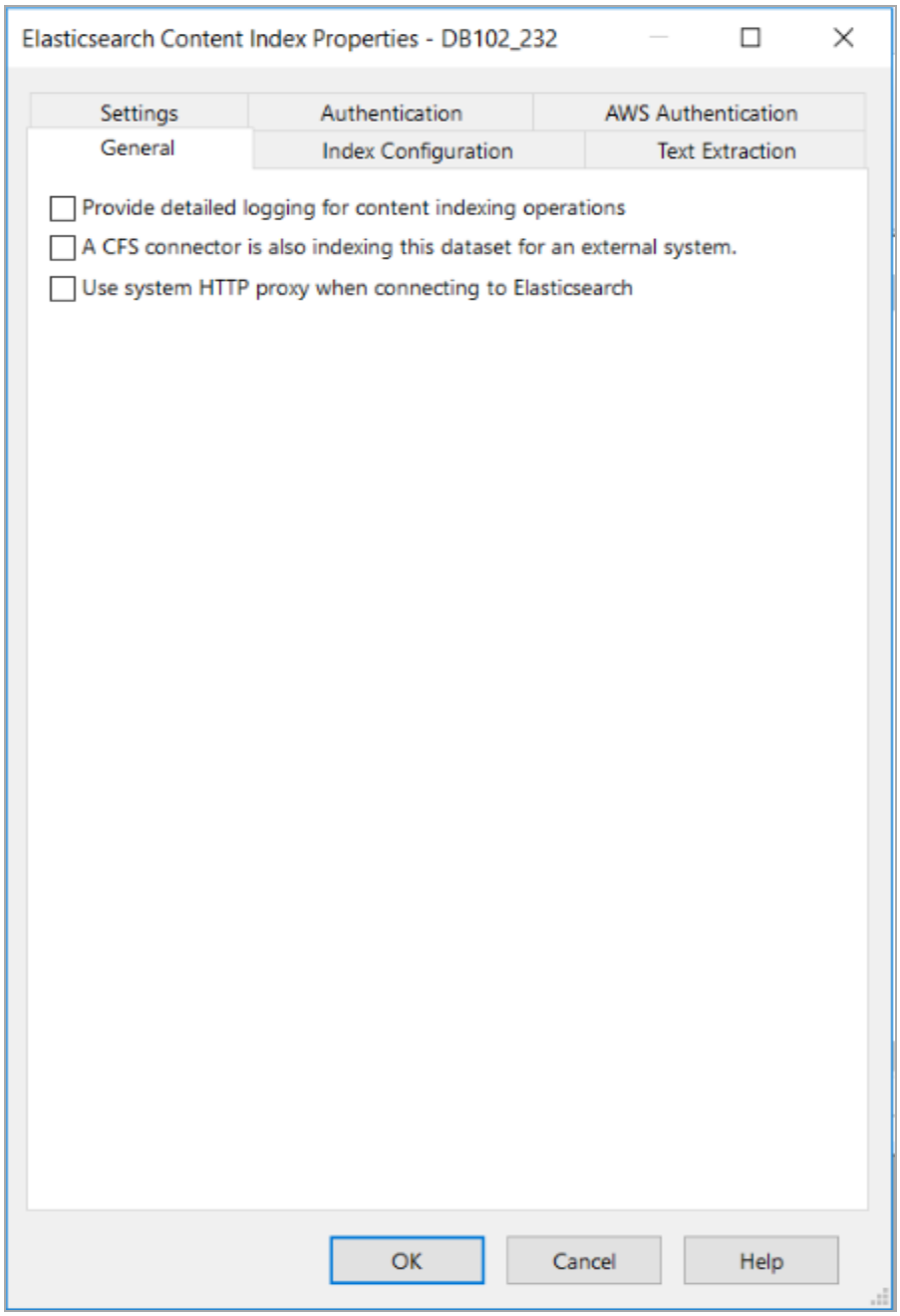
Select **Delete Content Index from Server** then select **OK**.

Create a new Elasticsearch index

On your Workgroup Server open Enterprise Studio -> right mouse click on your Dataset -> **Content Index -> Create Elasticsearch**.

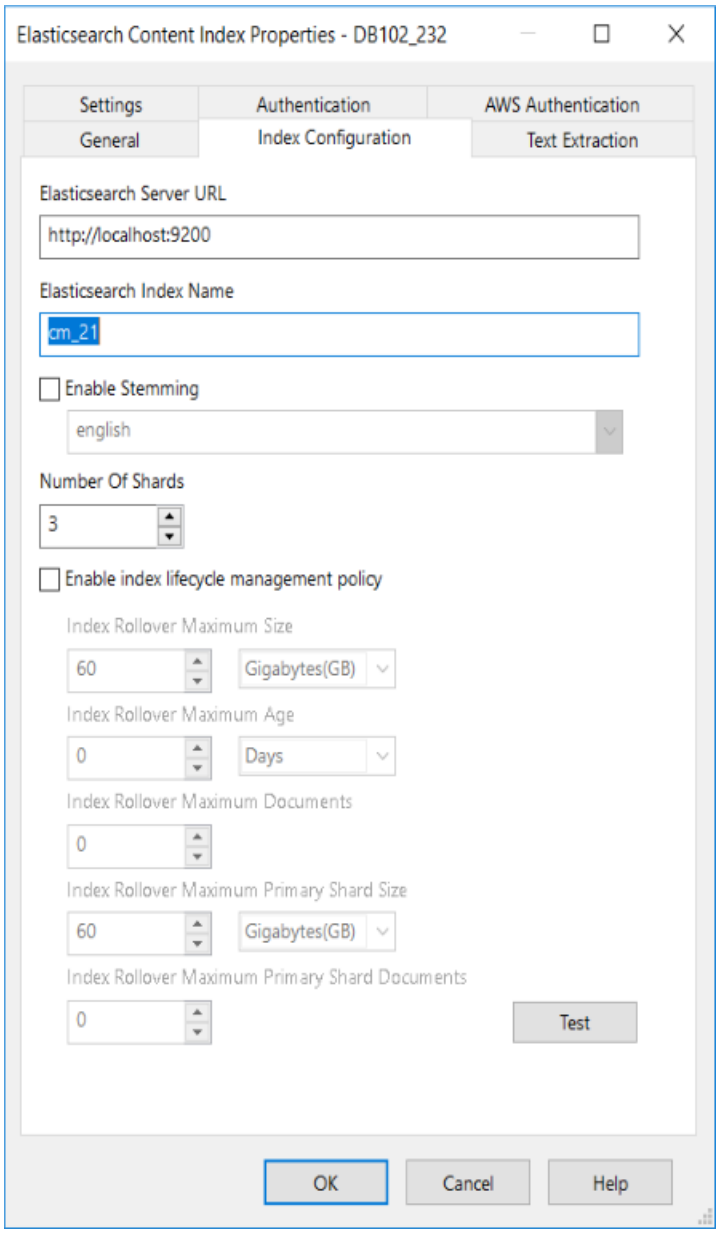
Fill in the fields on the Content index properties dialog:

On the **General** tab:



- **Provide detailed logging for content indexing operations** - select this option to generate detailed log files.
- **A CFS connector is also indexing this dataset** - select this option if you have also installed and configured the CFS connector. Content Manager will send messages to the CFS connector to ensure that the Elasticsearch index is up to date with changes to the Content Manger database.
- **Use system HTTP proxy when connecting to Elasticsearch** - select this option if your organization needs to connect to Elasticsearch using a proxy.

On the **Index Configuration** tab:



- **Elasticsearch Server URL** - enter the Elasticsearch Server URL. If using the X-Pack authentication, specify the https-based URL for the ES server here.
- **Elasticsearch Index Name** - Name of the Elasticsearch index to use for this Content Manager dataset. The characters in the name must be alphanumeric characters or underscores. If you performed a reindex at an earlier step this would be the same value used for the Index name during that process.
- **Index text that was deleted from a document with revision tracking** - select this option to index content that was deleted from a document that was edited using revision tracking/ "tracked changes".
- **Index hidden text from Microsoft Office files** - select this option to index content that is marked 'hidden' in documents, for example, hidden cells in Microsoft Excel.

- **Elasticsearch Stemming** - Allows you to select a stemmer that is appropriate for your index. If no stemmer is chosen, the content will be indexed without stemming. The default stemmer is “English”. Selecting a stemmer that is going to match the language used in the majority of records and documents will ensure the return of more relevant search results.

  A custom stemmer can be entered by typing in its name the option box. However, you must ensure the stemmer you enter is supported by version of Elasticsearch you are using.

  This option will be greyed out after the index is created and cannot be changed without deleting the index and re-creating it.

- **Number of shards** - Sets the number of shards contained in each Elasticsearch index. The default is three. Avoid setting many shards per index as it can have an impact on performance. It is recommended to avoid a shard size of more than 50GB.

  For example, if the index size is set at 60 GB and the number of shards to three, this will lead to an index that has three shards each having a size of about 20GB. However, the final index size will vary depending on how you have configured shard replication on your Elasticsearch cluster. In this example replication is turned off.

  The numbers of shards cannot be changed after index creation and will be disabled.

- **Enable lifecycle management policy and Index rollover size** - When this option is enabled an index will “rollover” to another index when the index size is reached. This is useful in large environments that are ingesting lots of new content.

  If this option is not set and an index size is not specified a customer’s index has the potential to reach a theoretical maximum at which point the Elasticsearch index may experience issues. e.g. shard size exceeds 50GB.

  Using a lifecycle policy based on index size allows you to keep the shard size under 50GB and grow your content index over time.

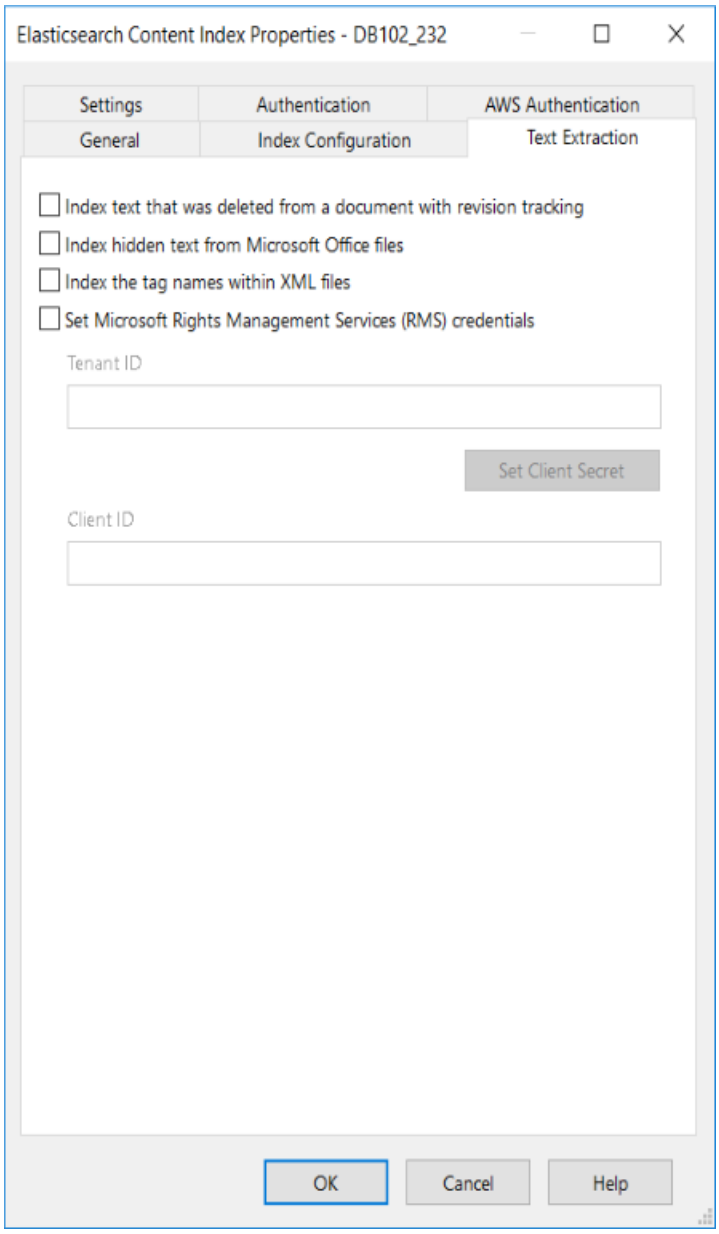
  Depending on organisational requirements, indices can also be rolled over based on index age (days) or the number of Elasticsearch JSON documents.

  The most common option is to rollover indices based on size as it allows organisations to predict and provision future storage requirements.

  The lifecycle management policy and index rollover size cannot be changed after index creation and will be disabled.

  - **Index Rollover Maximum Size** - set the maximum size of the index before it is rolled over to another index.
  - **Index Rollover Maximum Age** - set the maximum age of the index before it is rolled over to a another index.
  - **Index Rollover Maximum Documents** - set the maximum number of Elasticsearch JSON documents that can be created before it is rolled over to another index.
  - **Index Rollover Maximum Primary Shard Size** - triggers a rollover when the size of the largest primary shard reaches this value.
  - **Index Rollover Maximum Primary Documents** - triggers a rollover when the number of documents in a primary shard reaches this value.

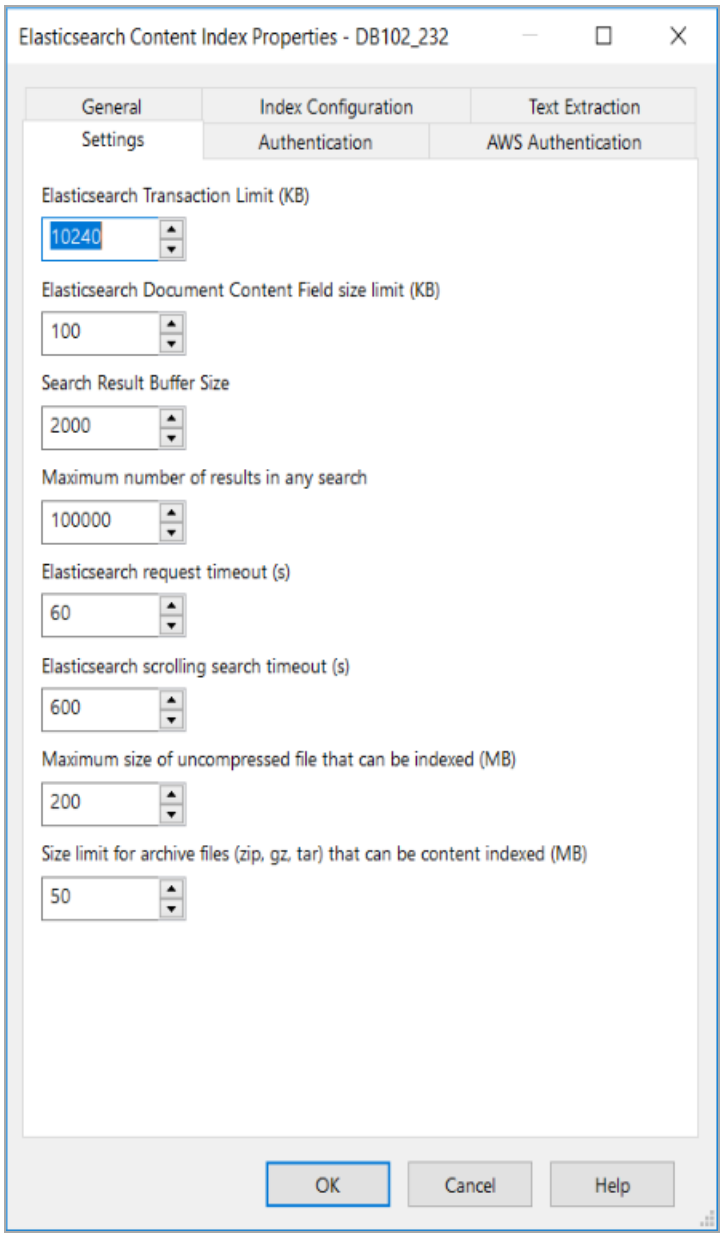
On the **Text Extraction** tab:



- **Index text that was deleted from a document with revision tracking** - select this option to index content that was deleted from a document that was edited using revision tracking/ "tracked changes".
- **Index hidden text from Microsoft Office files** - select this option to index content that is marked 'hidden' in documents, for example, hidden cells in Microsoft Excel.
- **Index the tag names within XML files** - select this option to index XML Format text files.
- **Set Microsoft Rights Management Services (RMS) credentials** - enter the credentials for the Microsoft Rights Management Services (RMS) to allow Microsoft Information Protection (MIP) encrypted documents to be indexed.
  - **Tenant ID** - enter the RMS Tenant ID.
    - **Set Client Secret** - if required, click the **Set Client Secret** option to set the password.

- **Client ID** - enter the RMS Client ID.

On the **Settings** tab:

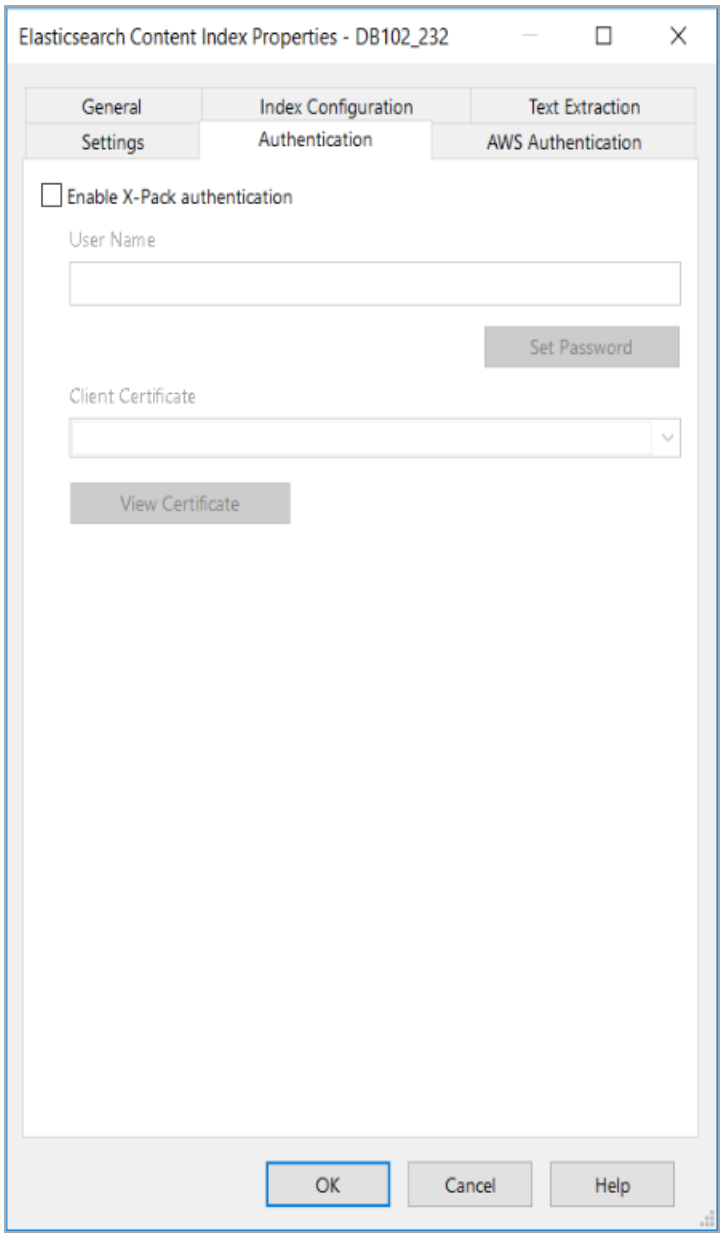


- **Elasticsearch Transaction Limit (KB)** - sets the size limit for content to be sent to Elasticsearch for indexing. The transaction limit can be set between 1KB and 1 000 000 KB.
  • Documents that are indexed via the Content Index event processor are queued and processed in a batch based on a timer event. If the amount of text extracted for the queued documents is greater than the transaction limit it is split up and sent in as many transactions as required to complete the indexing. If it's less than the transaction size, it is sent as a part of the timer event.
  • Documents that are reindexed on the client, via the Reindex option in the Elasticsearch group on the Administration tab in the Content Manager client, the transactions are sent to

Elasticsearch by the client. If all the extracted text is smaller than the transaction limit, the client sends that, otherwise it is split into batches based on the size of the transaction limit until all documents have been indexed.

- **Elastic Document Content Field size limit (KB)** - set the maximum size of the Content field in an Elasticsearch document that is created as a part of the text extraction process. If there is a large file being indexed the extraction process will create multiple Elasticsearch documents and link them together by URI so they are all referenced to the same record. The limits for the document content field size are 1KB to 100 000KB.
- **Search Result Buffer Size** - default value of 2000. For customers running searches that retrieve many results, increasing the buffer size will reduce the number of service calls to the Elasticsearch server to retrieve all the results.
- **Maximum number of results in any search** - the maximum number of results Content Manager will get back from Elasticsearch, by default, this is set to 100,000. If the number of results found by a search exceeds this number, a warning displays and the user only gets this number. The maximum value that can be set is 10,000,000.
- **Elasticsearch Request timeout (s)** - Limits how long Content Manager will wait for Elasticsearch to process a single request.
- **Elasticsearch scrolling search timeout (s)** - default value is 600 - Limits how long Content Manager Elasticsearch should keep the search context alive. It accepts a value between 60 and 86400. When scroll timeout occurs during an Elasticsearch query, a warning message is displayed *Content Manager Workgroup Server on 'local' reported an error. Could not serialize the server-side recordset. The search context has expired for this Elasticsearch query. You will need to reissue the request.*
- **Maximum size of uncompressed file that can be content indexed (MB)** - default: 2048. Content Manager does not index the content of documents whose file size is greater than this number in megabytes.
- **Set limit for archive file (zip, gz, tar) that can be content indexed (MB)** - default: 2048. Content Manager does not index the content of archive files whose file size is greater than this number in megabytes.

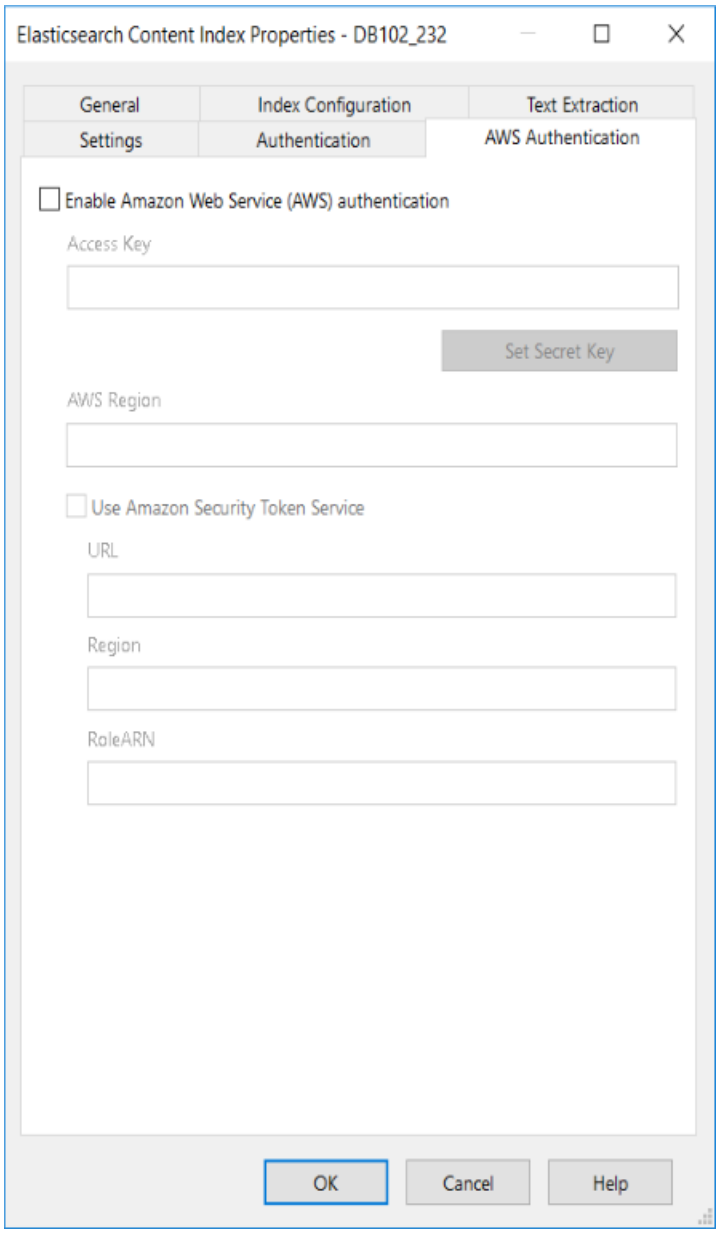
On the **Authentication** tab, fill in the fields:



- **Enable X-Pack authentication** - select this option to enable X-Pack authentication if X-Pack authentication is being used. You can either specify a username and password or a certificate:
  - **User Name** - enter the username to connect to the Elasticsearch server.
  - **Password** - enter the password for the defined user.
  - **Client Certificate** - if using a certificate for the X-Pack authentication, the certificate must be installed to the Personal store of the Local Computer account. Once installed, the certificate will appear in the Client Certificate drop-down list. Select the required certificate from the drop-down list.
  - **View Certificate** - click to view the selected Client Certificate.

> **TIP:** If you get certification validation errors, ensure that the local computer trusts the certificate authority of the certificate that the Elasticsearch server is presenting.

On the **AWS Authentication** tab:



- **Enable Amazon Web Service (AWS) authentication** - select this option to use the Amazon Web Service (AWS) version of the Elasticsearch service.
  - **Access Key** - enter the AWS Access Key. This is the username for the AWS IAM user that has access to the Elasticsearch services inside AWS.
    - **Secret Key** - enter the AWS Secret Key. This is the password for the Access Key user.
  - **AWS Region** - enter the AWS Region name.

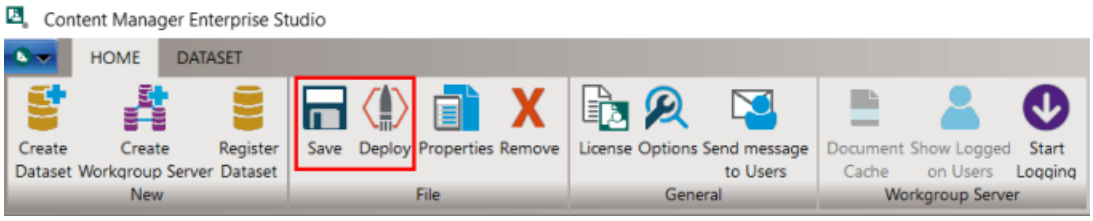
- **Use Amazon Security Token Service (STS)** - To create and provide trusted user with temporary security credentials that can control access to your AWS resources.
  - **URL** – enter the URL for STS endpoint (example: https://sts. ap-southeast-1.amazonaws.com)
  - **Region** - enter the AWS STS Region name (example: ap-southeast-1)
  - **RoleARN** - enter the Role that delegates access to the Amazon AWS resource for the AWS IAM user.

If you've setup your environment correctly the **Test** button should return a **Succeeded** response.

Click **OK** to close the **Content index properties** dialog.

If you encounter the following prompt: **The index 'CM_45' already exists on the Elasticsearch server. Do you want to delete it and create and new one?** Select **No**.

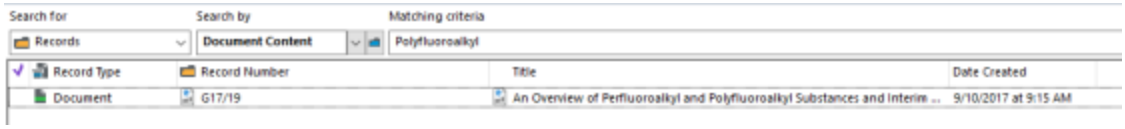
Save and then **Deploy** your new settings in Enterprise Studio.



Confirm your Elasticsearch content index is operational

Performing a simple document content search from a Content Manager client is the best way to confirm your content index is operational.

In this case, we selected a document with unique content so only one result was returned. The use of unique content confirms the searching mechanism can interrogate the entire index and return the correct result.



Troubleshooting common issues

Logging

When it comes to troubleshooting general content indexing issues there are three key logs you should be looking at:

Log file name	Default Log file location	Useful for...
Elasticsearch cluster log	ES_HOME\logs\[cluster_name].log	Identifying issues to do with the health of the Elasticsearch environment. Located on each Elasticsearch server in your cluster.
Workgroup server document content indexing event log	<Drive>\Micro Focus Content Manager\ServerLocalData\TRIM\Log\ DCI_ <DB ID>_<YYYY>_<MM>_<DD>.log	Whenever an electronic record is created, a new version is committed or an electronic document is destroyed, deleted or removed, an event is created to indicate that the document content index should be updated. This log file may help you narrow down where a particular problem is occurring when compared to other logs listed in this table.
Client Reindex log	<Drive>\Users\<username>\AppData\Local\Micro Focus\Content Manager\<DB ID>\Log\ ElasticReindex_ <YYYY>_<MM>_<DD>.log	Identifying problematic documents or timeout issues when reindexing from the Content Manager client.

Diagnosing “operation timed out” errors

The most common issue you’ll probably come across when indexing content is **operation timed out** errors in the client **Reindex** log.

This error doesn’t always mean there’s an issue. Sometimes, it can just mean the client has waited for a response from the Elasticsearch server beyond the **Request Timeout** value you’ve set in Enterprise Studio. Often, you’ll find Elasticsearch has continued to index the content after the client has reported a timeout event. You can check if the content has been indexed by doing a search

based on content. If it doesn't return a result you should then attempt to run a re-index on only missing items.

If you continue to observe timeout errors and content isn't being indexed you can then proceed to examining the logs identified in the previous section. Key areas you should be looking at in this scenario are:

- **Enterprise Studio – Transaction Limit**

Content Manager sends documents to Elasticsearch using a bulk-index API, this value sets the maximum size of the memory buffer before we send the bulk import request to Elasticsearch. When performing re-indexing, using a larger value will speed up indexing. However, you don't want it to be too large as it can result in Elasticsearch taking longer to return a result and/or the JVM running out of heap memory if resources are limited. We recommend you keep the transaction limit set to 5-15 MB (the default transaction limit value is 9 MB).

- **Enterprise Studio – Request Timeout**

The request timeout value sets the time Content Manager will wait for Elasticsearch to respond every time a HTTP request is sent. Every environment will perform differently depending on the resources it has available. If there's adequate resources simply extending the timeout value may resolve many of the timeout errors observed (the default timeout value is 60 seconds).

- **Elasticsearch – available heap and available nodes**

Timeout errors may also indicate there's isn't enough memory allocated to the JVM, your servers aren't large enough to handle the load, or you don't have enough nodes in your cluster. If the cluster log indicates this you should revisit the memory configuration and sizing recommendations outlined in **Memory Management and configuring the Java Virtual Machine (JVM)**.

Proxy servers and firewalls

Content Manager Clients and Workgroup Servers need direct access to your Elasticsearch servers. Make sure your firewall has the appropriate ports open (default port range is 9300-9400) and your proxy server is configured to not cache traffic bound for Elasticsearch hosts.

Useful Information

Kibana

The Elasticsearch team recommend that as your Elasticsearch environment grows to include more nodes and clusters you should consider using Kibana to manage it. Kibana not only allows you to manage your clusters more efficiently it also gives you the ability to analyze and make sense of your data. Visit elastic.co to get more information on Kibana.

cURL

The Elasticsearch team recommend cURL, which is an open source command line utility that can be used to directly query and manage Elasticsearch via its API. If you don't decide to use Kibana you can use cURL commands to perform a large range of tasks e.g. delete an index, get node and cluster health data. The Elasticsearch documentation contains many examples of cURL scripts.

cURL can be downloaded, free of charge, from the official cURL site.

Elasticsearch Reference

The Elasticsearch Reference is a detailed knowledge base of all aspects of Elasticsearch. You should consult this when considering the use of Elasticsearch in your organization.

You can find the Elasticsearch Reference in the Docs section on elastic.co

Hardware requirements

The Content Manager Support team regularly gets asked for assistance to size a customer's environment for content indexing. The Elasticsearch team do a good job of explaining how to approach this problem in the **Hardware** section of **Elasticsearch: The Definitive Guide** on elastic.co

Microsoft also present an excellent overview of Elasticsearch architecture and how to scale it on Azure which can be applied to any environment. You can find this information in the **Elasticsearch on Azure** section of the Azure online documentation.