Google Search Appliance Connectors

Deploying the Connector for SharePoint

Google Search Appliance Connector for SharePoint software version 4.1.0
Google Search Appliance software versions 7.2 and 7.4

June 2015

About this guide

This guide is intended for anyone who needs to deploy the Google Search Appliance Connector 4.1.0 for SharePoint. The guide assumes that you are familiar with Windows or Linux operating systems and configuring the Google Search Appliance by using the Admin Console.

See the Google Search Appliance Connectors Administration Guide 4.1.0 for general information about the connectors, including:

- What's new in Connectors 4?
- General information about the connectors, including the configuration properties file, supported ACL (Access Control List) features, and other topics
- Connector security
- Connector logs
- Connector Dashboard
- Connector troubleshooting

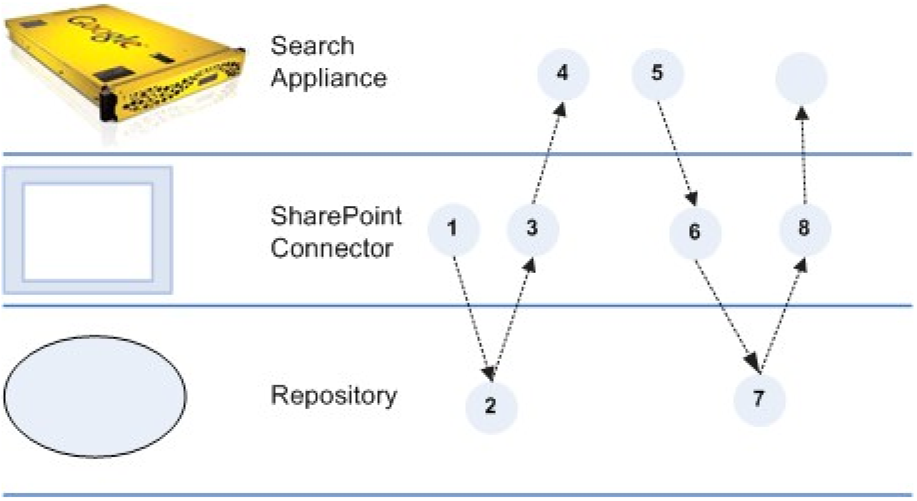
For information about using the Admin Console, see the Google Search Appliance Help Center.

For information about previous versions of connectors, see the Connector documentation page in the Google Search Appliance Help Center.

Overview of the GSA Connector for SharePoint

The Connector for SharePoint 4.1 enables the Google Search Appliance to crawl and index content from Microsoft SharePoint. Each connector instance can support only one SharePoint Web Application. If you have more than one Web Application, you need to deploy one connector instance for each one.

The following diagram provides an overview of how the search appliance gets content from SharePoint through the connector. For explanations of the numbers in the process, see the steps following the diagram.



1. The Connector for SharePoint starts communicating with the repository by presenting authentication credentials.
2. The repository sends a limited number of Doc IDs of documents in the repository to the connector.
3. The connector constructs URLs from the Doc IDs and pushes them to the search appliance in a metadata-and-URL feed. Take note that this feed does not include the document contents.
4. The search appliance gets the URLs to crawl from the feed.
5. The search appliance crawls the repository according to its own crawl schedule, as specified in the GSA Admin Console. It crawls the content by sending GET requests for content to the connector.
6. The connector requests the content from the repository.
7. The repository sends the content to the connector.

8. The connector provides the content to the search appliance for indexing. If the content is in HTML format, the search appliance will follow links within the page and send more GET requests for the linked content to the connector.

Automatic updates every 15 minutes

After the initial process completes, the connector periodically informs the search appliance of new documents and changed documents, according to the value set in the connector configuration option `adaptor.incrementalPollPeriodSecs`. The default interval value is 15 minutes, but you can configure it to suit your needs. For more information, see “Common configuration options” in the Administration Guide.

Supported SharePoint versions

The Connector for SharePoint 4.1 supports the following versions:

- SharePoint 2010
- SharePoint Foundation 2010
- SharePoint 2013
- SharePoint Foundation 2013

Supported operating systems for the connector

The Connector for SharePoint 4.1.0 must be installed on one of the following supported operating systems:

- Windows Server 2012
- Windows Server 2008 (32 and 64 bit)
- Windows Server 2003 (32 and 64 bit)
- Ubuntu
- Red Hat Enterprise Linux 5.0
- SUSE Enterprise Linux 10 (64 bit)

Indexing unpublished docs

The Connector for SharePoint 4.1.0 always honors the Search Visibility setting on SharePoint (you cannot override this). For draft documents, indexing depends on the permissions that are given to the connector user account. If the connector user has only "Full Read" permissions, the connector will honor all "Draft item visibility" settings on SharePoint.

Supported authentication mechanisms

The Connector for SharePoint 4.1.0 supports any authentication mechanism where a verified user ID is rendered.

Known connector limitations

- Only one connector instance is allowed per Virtual Server / SharePoint Web Application.
- The number of content databases will affect document change detection latency.
- The number of unique users and groups used in ACLs for each site collection will affect memory consumption.

CPU and memory recommendations

The following table contains CPU and memory recommendations for the Connector for SharePoint. Take note that the value for the GSA host load is just for reference and not a recommended value. You should use a host load value that SharePoint can handle.

Number of Site Collections	Number of Users per Site Collection	Number of SharePoint Local Group Memberships	Number of Role assignments	GSA Host Load	Recommended RAM	Recommended CPU
200	50K	50K	5K	4	4GB	4 core
2000	50K	50K	5K	4	8GB	4 core
2000	5K	5K	5K	4	4GB	4 core
200	50K	50K	5K	8	8GB	4 core
2000	50K	50K	5K	8	12GB	4 core
2000	5K	5K	5K	8	8GB	4 core

To discover the numbers in your SharePoint installation, run the pre-deployment powershell script `diagnose_sp.ps1`, which outputs the following information:

- Number of site collections
- Number of users per site collection
- Total number of SharePoint local group memberships
- Number of role assignments (permissions given to users and groups)

Before you deploy the Connector for SharePoint

Before you deploy the Connector for SharePoint, ensure that your environment has all of the following required components:

- GSA software version 7.2.0.G.264 or higher
  To download GSA software, visit the Google for Work Support Portal (password required).
- Java JRE 1.7u6 or higher installed on the computer that runs the connector
- Connector for SharePoint 4.1.0 JAR executable
  For information about finding the JAR executable, see Step 2 Install the Connector for SharePoint.
- User account for the connector, with Full Read permissions to SharePoint Web Application in the User Policy
- If running the connector on Linux, the user account used for running the connector should belong to same domain as the SharePoint server. A user from a child domain or from the same forest is not sufficient.
- If there are any write-locked site collections, run the PrepareWriteLockedSitesForAdaptor.ps1 script on SharePoint using an account that has Admin privileges before installing the connector.
- To gather additional information about your SharePoint environment which can be handy in configuring SharePoint connector, run diagnose_sp.ps1 on the SharePoint server using an account that has farm administration privileges.

  While it is not a mandatory step to run the script before deploying SharePoint connector, output of the script, which includes information such as number of web applications, authentication mechanism, number of documents and user group membership count, is very helpful in estimating number of connector instances required, memory requirements , as well as expected document count.

- Optionally, configure the search appliance for the authentication method in use (typically LDAP for Active Directory). For detailed information about configuring authentication, see Managing Search for Controlled-Access Content.

Deploy the Connector for SharePoint

Because the Connector for SharePoint is installed on a host that is separate from the GSA, you must establish a relationship between the connector and the search appliance.

To deploy the Connector for SharePoint, perform the following tasks:

1. Configure the search appliance
2. Install the Connector for SharePoint
3. Optionally, configure adaptor-config.properties variables
4. Run the Connector for SharePoint

Step 1 Configure the search appliance

For the search appliance to work with the Connector for SharePoint, the search appliance needs to be able to crawl SharePoint content and accept feeds from the connector. To set up these capabilities, perform the following tasks by using the search appliance Admin Console:

1. Add the URL pattern provided by the connector to the search appliance's crawl configuration follow patterns.
2. Add the IP address of the computer that hosts the connector to the list of Trusted IP addresses so that the search appliance will accept feeds from this address.
3. Set up connector security.

Add the URL pattern

To add the URLs provided by the connector to the search appliance's crawl configuration follow patterns:

1. In the search appliance Admin Console, click **Content Sources > Web Crawl > Start and Block URLs**.
2. Under **Follow Patterns**, add the URL that contains the hostname of the machine that hosts the connector and the port where the connector runs.
   For example, you might enter http://connector.example.com:5678/doc/
   where connector.example.com is the hostname of the machine that hosts the connector.
   By default the connector runs on port 5678.
3. Click **Save**.

Add the IP address

To add the IP address of the computer that hosts the connector to the list of trusted IP addresses:

1. In the search appliance Admin Console, click **Content Sources > Feeds**.
2. Under **List of Trusted IP Addresses**, select **Only trust feeds from these IP addresses**.
3. Add the IP address for the connector to the list.
4. Click **Save**.

Set up security

For information about setting up security, see “Enable connector security” in the Administration Guide.

Step 2 Install the Connector for SharePoint

This section describes the installation process for the Google Search Appliance Connector for SharePoint on the connector host computer. This connector version does not support installing the connector on the Google Search Appliance.

You can install the Connector for SharePoint on any host running one of the supported operating systems, however, the host must be in the same domain as the SharePoint installation.

As part of the installation procedure, you need to edit some configuration variables in the configuration file. Take note that you can encrypt the value for `sharepoint.password` before adding it to the file by using the Connector Dashboard or standalone command line executable, as described in “Encode sensitive values” in the Administration Guide.

NTLM and **Kerberos:** (**Windows**): When SharePoint’s and the current user’s domain are the same or are from the same domain hierarchy, Windows operating systems automatically use the credentials of the person currently signed on to Windows.

If they are not from the same domain hierarchy, you need to specify a username and password. For example, suppose that you run the connector from a coprorate domain such as @mycompany.com against a SharePoint instance from another domain, such as GSA-CONNECTORS. In this case, you need to specify user credentials for the GSA-CONNECTORS domain.

Take note that the specified username and password from the connector configuration file are used only when the computer that hosts the connector is not part of the same domain forest as the SharePoint server.

Forms Authentication (**Linux** and **Windows**): Always specify the username and password.

Windows installation

To install the Connector for SharePoint on Windows:

1. Log in to the computer that will host the connector by using an account with sufficient privileges to install the software.
2. Start a web browser.
3. Visit the connector 4.1.0 software downloads page at http://googlegsa.github.io/adaptor/index.html.
   Download the `exe` file by clicking on **Microsoft SharePoint** in the Windows Installer table.
   You are prompted to save the single binary file, `sp-install-4.1.0.exe`.
4. Start installing the file by double clicking `sp-install-4.1.0`.
5. On the **Introduction** page, click **Next**.
6. On the **GSA Hostname and other required configuration values** page, enter values for the following options:
   a. **GSA Hostname or IP address** of the GSA that will use the connector.
      For example, enter **`yourgsa.example.com`**
   b. **SharePoint Server**.
      For example, enter **`http://yoursharepoint.example.com/`**
   c. **Adaptor port number** for any crawlable documents this connector serves.
      Each instance of a Connector on same machine requires a unique port. The default is 5678.
   d. **Server dashboard port** for the Connector Dashboard.
      The value is the port on which to view web page showing information and diagnostics about the connector. The default is 5679.
   e. **Maximum Java Heap size (in megabytes).**
      Refer to the entry in the table in GSA host load, CPU and memory recommendations that shows a suggested value for "Recommended RAM." Specify that value as the **Maximum Java Heap size**.
   f. Whether or not to run the connector after the installer finishes.
7. Click **Next**.
8. On the **Choose Install Folder** page, accept the default folder or navigate to the location where you want to install the connector files.
9. Click **Next**.
10. On the **Choose Shortcut Folder**, accept the default folder or select the locations where you want to create product icons.

11. To create icons for all users of the Windows machine where you are installing the connector, check **Create Icons for All Users** and click **Next**.
12. On the **Pre-Installation Summary page**, review the information and click **Install**. The connector Installation process runs.
13. On the **Install Complete** page, click **Done**. If you selected the option to run the connector after the installer finishes, the connector starts up in a separate window.
14. If the connector is running from host machine which is not in same forest as SharePoint server, then add these additional configuration options to `adaptor-config.properties`:

    ```
    sharepoint.username=YOURDOMAIN\\ConnectorUser
    sharepoint.password=user_password
    ```

15. If the connector needs to use Live Authentication to connect to SharePoint, then add these additional configuration options to `adaptor-config.properties`:

    ```
    sharepoint.username=AdaptorUser Live Authentication Id
    sharepoint.password=uS3R_passWoRD
    sharepoint.useLiveAuthentication=true
    ```

16. If the connector needs to use ADFS Authentication to connect to SharePoint, then add these additional configuration options to `adaptor-config.properties`:

    ```
    sharepoint.username=AdaptorUser@yourdomain.com
    sharepoint.password=uS3R_passWoRD
    sharepoint.sts.endpoint=https://adfs.example.com/adfs/services/trust/2005/usernamemixed
    ```

    `sharepoint.sts.realm=urn:myserver:sharepoint` or `https://yoursharepoint.example.com/_trust`

Command-line installation for Linux or Windows

The following procedure gives the steps for installing the Connector for SharePoint on Linux. Take note that if you prefer not to use the Windows installer, you can also follow this procedure to install the Connector on Windows.

To install the connector:

1. Download the Connector for SharePoint JAR executable (`adaptor-sharepoint-4.1.0-withlib.jar`) from http://googlegsa.github.io/adaptor/index.html.
2. Create a directory on the host where the connector will reside. For example, create a directory called sharepoint_connector_410.
3. Copy the Connector for SharePoint 4.1.0 JAR executable to the directory.
4. Create an ASCII or UTF-8 file named `adaptor-config.properties` in the directory that contains the connector binary.
5. Provide the following configuration (replacing bolded items with your real configuration) within the file:

   ```
   gsa.hostname=yourgsa.example.com or IP address
   sharepoint.server=http://yoursharepoint.example.com/
   ```

   where **yoursharepoint.example.com** is a fully-qualified domain name. If it is not a fully-qualified domain name, then you must set DNS override on the connector host.

6. **Linux**: Add these additional configuration options to `adaptor-config.properties`:

   ```
   sharepoint.username=YOURDOMAIN\\ConnectorUser
   sharepoint.password=user_password
   ```

   Forms Authentication (**Linux** and **Windows**): Always specify the username and password.

7. Create a folder named **logs** in the same directory.
   Create an ASCII or UTF-8 file named **logging.properties** in the same directory that contains the connector binary and add the following content:

```
handlers = java.util.logging.FileHandler
.level = WARNING
com.google.enterprise.adaptor.level = INFO
com.google.enterprise.adaptor.sharepoint.level = INFO
java.util.logging.FileHandler.formatter=com.google.enterprise.ada
ptor.CustomFormatter
java.util.logging.FileHandler.pattern=logs/sp-adaptor.%g.log
java.util.logging.FileHandler.limit=104857600
java.util.logging.FileHandler.count=20
```

HTTPS configuration

If SharePoint is configured to use HTTPS, get a SharePoint certificate to add it as a trusted host for the connector by performing the following steps:

1. Navigate to SharePoint in a browser.
   A warning page appears with a message such as "This Connection is Untrusted." This message appears because the certificate is self-signed and not signed by a trusted Certificate Authority. Click, "I Understand the Risks" and "Add Exception."
2. Wait until the "View..." button is clickable, then click it.
3. Change to the "Details" tab and click "Export...".
4. Save the certificate in your connector's directory with the name "`sharepoint.crt`".
5. Click Close and Cancel to close the windows.
6. To allow the connector to trust SharePoint, enter the following command:

```
keytool -importcert -keystore cacerts.jks -storepass changeit
-file sharepoint.crt -alias sharepoint
```

7. When prompted Trust this certificate?, answer yes.

Step 3 Configure adaptor-config.properties variables

Optionally, you can edit or add additional configuration variables to the `adaptor-config.properties` file. The following table lists the most important variables that pertain to the Connector for SharePoint, as well as their default values. See also “Common configuration options” in the the Administration Guide.

Variables	**Description**	**Default**
`server.dashboardPort`	Port on which to view web page showing information and diagnostics. The Windows installer prompts for this information.	5679
`adaptor.namespace`	Namespace used for ACLs sent to GSA	Default
`sharepoint.xmlValidation`	Whether to enable strict checking of XML responses using the expected schema.	False
`sharepoint.maxIndexableSize`	Number of bytes of a document that GSA indexes.	2097152
`sharepoint.siteCollectionOnly`	Whether sharepoint.server is a site collection, instead of a virtual server. If true then adaptor will index `sharepoint.server` as a site collection.	Auto-detected value

Step 4 Run the Connector for SharePoint

After you install the Connector for SharePoint, you can run it by using `cmd.exe` on the host machine:

```
java -Djava.util.logging.config.file=logging.properties -jar adaptor-
sharepoint-4.1.0-withlib.jar
```

To Verify that the connector has started and is running, navigate to the Connector Dashboard at `http://<CONNECTOR_HOST>:<nnnn>/dashboard` or `https://<CONNECTOR_HOST>:<nnnn>/dashboard`

where `<nnnn>` is the number you specified as the value for the `server.dashboardPort` in the configuration file.

To run the connector as a service, use the Windows service management tool or run the `prunsrv` command, as described in "Run a connector as a service on Windows" in the Administration Guide.

Uninstall the Google Search Appliance Connector for SharePoint

To uninstall the Connector for SharePoint on Windows:

1. Navigate to the SharePoint connector installation folder, **_GSA_SP_Adaptor_installation**.
2. Click `Uninstall_GSA_SP_Adaptor.exe`.
   The **Uninstall GSA_SP_Adaptor** page appears.
3. Click **Uninstall**.
   Files are uninstalled.
4. Click **Done**.

Multi-Tenant configurations

Multi-Tenant SharePoint deployments typically host multiple customer sites under same Web application. Customers gets permissions only for their respective site collections. In such scenario it is not possible to get Full Read permissions on SharePoint web application as required SharePoint Adaptor.

Such multi-tenant configurations are supported in Connectors 4.1 via Site Collection Only mode. To support a multi-tenant configuration, site collection mode must be enabled by using the `sharepoint.siteCollectionOnly` configuration option in the `adaptor-config.properties` file. To index site collection at root level in site collection only mode, you need to set `sharepoint.siteCollectionOnly` to true explicitly.

The connector will index a site collection and its child items. For this reason the connector user account on SharePoint needs site collection administrator permissions.

If you have multiple site collections to index in a multi-tenant environment, you need to configure one connector instance for each of the site collections.

To configure SharePoint Adaptor in site collection only mode:

1. Specify sharepoint.server as site collection URL, for example, http://sharepoint.example.com/sites/sitecollection
2. If Site collection URL is root site collection (for example, http://sharepoint.example.com), explicitly set `sharepoint.siteCollectionOnly = true`
3. Add the adaptor host URL to include a pattern on GSA, for example, `http://adaptorhost.example.com:5678/doc/`

   Note: With Connector 4.1, you don't need to specify start URL or more restrictive include patterns on GSA.

Non-canonical URLs in site collection only mode

The SharePoint connector 4.1.0 allows non-canonical URLs in site collection only mode. That is, the connector URL as specified by the `sharepoint.server` configuration option in the `adaptor-config.properties` file need not be in exactly the same case as on SharePoint.

Advanced Topics

The information in this section extends beyond basic SharePoint connector configuration.

Override Content-Type for Microsoft Outlook .msg files

If the connector encounters Outlook .msg files when crawling content, it overrides the Content-Type for the files and indexes them as as “application/vnd.ms-outlook.”

Perform SID lookup

With v4.1, you can configure the connector to perform SID (security identifier) based lookup for domain group principals to resolve group names from Active Directory (AD) instead of using the principal display names available on SharePoint. This is necessary only when SharePoint is configured to use NTLM claims and display names for domain groups on SharePoint are different than actual group names in AD.

By default, this functionality is disabled and generally not required in most common SharePoint environments where DisplayName for domain group is same as actual principal name in AD.

To configure SID Lookup, you need to specify the options described in the following table in `adaptor-config.properties`.

Option	**Description**	**Required or Optional**
`sidLookup.host`	IP Address or hostname for AD Server.	Required
`sidLookup.port`	Default value is 3268. If you want to use SSL for connecting to AD, use port 3269. By default, the connector makes a query against global catalog (required for multi-domain support). You can restrict the SID lookup to single domain by using port 389 or port 636 (for SSL).	Optional
`sidLookup.username`	Username for user to connect to AD for SID lookup. This user can be same as AdaptorUser.	Required
`sidLookup.password`	Password for user to connect to AD for SID lookup.	Required

`sidLookup.method`	Specify as “SSL” if you want to make secure connection to AD for SID lookup.	Optional

Browser leniencies

When the connector crawls content, it might encounter unsupported characters in URLs. By default, the connector encodes URLs and is lenient in handling unsupported characters in URLs, same as most web browsers.

This behavior is controlled by the `adaptor.lenientUrlRulesAndCustomRedirect` configuration option, where the default value is `true`.

You can turn this behavior off by setting

```
adaptor.lenientUrlRulesAndCustomRedirect = false
```

Set number of maximum URL redirects to follow

When the connector is downloading document content, it might need to follow a number of URL redirects to get to the actual content. By default, the connector allows a maximum of 20 redirects.

This behavior is controlled by the `adaptor.maxRedirectsToFollow` configuration option.

You can change the number of maximum allowed redirects by setting the value to a positive integer, for example:

```
adaptor.maxRedirectsToFollow=50
```

Note: This property is applicable only when the connector is configured to handle custom redirects, that is,

```
 adaptor.lenientUrlRulesAndCustomRedirect = true
```

Allow all claims types in ADFS environment

In v4.1, the connector allows all ADFS (Active Directory Federation Services) claim types to be used in ACLs. Previous versions of the connector allowed only user and domain group claims to be part of ACLs.

Note: Even though connector v4.1 allows additional ADFS claims types in ACLs, none of the built-in authentication mechanism available with GSA resolve these claims for the search user. To support these additional claims, you need to write a custom cookie cracker to resolve additional claims for search users. For information about writing a custom cookie cracker, see "Using Cookie Cracking" in Managing Search for Controlled-Access Content.

Configure host load

If you are running multiple SharePoint connectors on the same host, Google recommends configuring the host load to maximize performance. For more information on this topic, refer to the the Administration Guide.

Increase log level for debugging connector Issues

You can increase the log level for the connector to get additional information that can help you resolve connector issues. For more information on this topic, refer to the the Administration Guide.

Troubleshoot the Connector for SharePoint

For information about troubleshooting the Connector for SharePoint, see “Troubleshoot Connectors,” in the Administration Guide.