Google Search Appliance Connectors

Deploying the Connector for File Systems

Google Search Appliance Connector for File Systems software version 4.0.2
Google Search Appliance software version 7.2

July 2014

Table of Contents

About this Guide

This guide is intended for anyone who needs to deploy the Google Search Appliance Connector 4.0.2 for File Systems. The guide assumes that you are familiar with Windows operating systems, file systems, and configuring the Google Search Appliance by using the Admin Console.

See the Google Search Appliance Connectors Administration Guide 4.0.2 for general information about the connectors, including:

- What's new in Connectors 4.0?
- General information about the connectors, including the configuration properties file, supported ACL features, and other topics
- Connector security
- Connector logs
- Connector Dashboard
- Connector troubleshooting

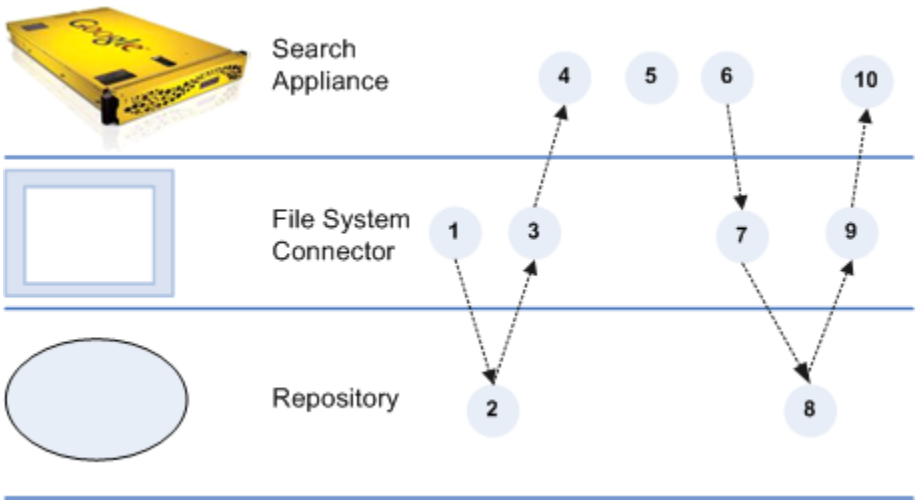
For information about using the Admin Console, see the Google Search Appliance Help Center.

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 File Systems

The Connector for File Systems enables the Google Search Appliance to crawl and index content from Windows shares. A single connector instance can support a single Windows share. The share can be a UNC path or a mapped drive.

The following diagram provides an overview of how the search appliance gets content from the repository through the Connector for File Systems. For explanations of the numbers in the process, see the steps following the diagram.



1. The Connector for File Systems queries the repository for a single DocId.
2. The repository sends the DocId to the connector.
3. The connector constructs a URL from the DocId and pushes it 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 URL 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. If the content is in HTML format, the search appliance follows links within the page.
6. The search appliance requests DocIds that it discovers during the crawl from the connector.
7. The connector queries the repository for the requested DocIds.
8. The repository sends the DocIds to the connector.

9. The connector sends the URLs to the search appliance. For a folder, the connector sends an HTML listing of the contents of the folder to the search appliance.
10. The search appliance continues to crawls the repository.

After the initial process completes, the connector periodically sends files or folders that have been modified, added, or deleted to the search appliance, according to the value set in the connector configuration. The default interval value is 15 minutes. The modification can be renamed content, content that has its attributes changed or had its security permissions (ACLs) changed.

Supported file system protocols

The Connector for File Systems supports the following file system protocols:

- Server Message Block (SMB) 1
- SMB 2
- Common Internet File System (CIFS)
- Samba
- Distributed File System (DFS)

Network File System (NFS) is not supported.

Distributed File System support

For DFS, the Connector for File Systems only supports a DFS link as the root configuration point. DFS links are fully supported by the connector. DFS namespaces are not supported. A Windows share can be a mapped drive, but the connector is not able to read DFS ACLs if a drive maps to a DFS UNC path.

Supported operating systems

The Connector for File Systems must be installed on Windows. The Connector for File Systems 4.0 is compatible with the following Windows operating systems:

- Windows Server 2003
- Windows Server 2008 R2
- Windows Server 2012

Google also supports crawling file shares on these Windows versions, but you might also crawl content on any Windows operating system later than Windows XP.

The Connector for File Systems does not run on Linux.

Before you deploy the Connector for File Systems

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

- GSA software version 7.2.0.G.90 or higher
  To download GSA software, visit the Google Enterprise Support Portal (password required)
- Java JRE 1.7 update 6 or higher installed on computer that runs the connector
- Connector for File Systems 4.0.2 JAR executable
  For information about finding the JAR executable, see Download the connector software
- Ensure that the Windows account has sufficient permissions, as described in the following section

Windows account permissions

The Windows account that the connector is running under must have sufficient permissions to perform the following actions:

- List the content of folders
- Read the content of documents
- Read attributes of files and folders
- Read permissions (ACLs) for both files and folders
- Write basic attributes permissions

  The connector attempts to restore the last access date for documents after it reads the document content during a crawl. In order for the last access date to be restored back to the original value before the content was read, the user account that the connector is running under needs to have write permission. If the account has read-only permission and not write permission for documents, then the last access date for documents will change as the connector reads document content during a crawl.

Membership in one of the following groups grants a Windows account the sufficient permissions needed by the connector:

- Administrators
- Power Users
- Print Operators

- Server Operators

Note: It is not sufficient for the user to be member of one of these groups at the domain level. The user must be a member of one of these groups on the local machine that exports the Windows share.

Download the connector software

The Connector for File Systems must be installed on a host machine. This connector version does not support installing the connector on the Google Search Appliance.

To download the software for Connector for File Systems:

1. Visit https://code.google.com/p/plexi/.
2. Click **Executable** for Microsoft Windows Shares.
   The single binary file, `adaptor-fs-4.0.2-withlib.jar`, is downloaded.

Once you download the connector software, you can copy it to the host and configure it.

Deploy the Connector for File Systems

Because the Connector for File Systems is installed on a separate host, you must establish a relationship between the connector and the search appliance.

To deploy the Connector for File Systems, perform the following tasks:

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

Step 1 Configure the search appliance

For the search appliance to work with the Connector for File Systems, the search appliance needs to be able to crawl file system 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 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.

Add the URL

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.

Step 2 Install the Connector for File Systems

You can install the Connector for File Systems on a host running one of the supported Windows operating systems.

To install the Connector for File Systems:

1. Download the Connector for File Systems JAR executable (`adaptor-fs-4.0.2-withlib.jar`) from https://code.google.com/p/plexi/.
2. Create a directory on the host where the connector will reside. For example, create a directory called filesystem_connector_40.

3. Copy the File System 4.0 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.hostname.com
   filesystemadaptor.src=\\\\host\\share
   ```

   Notes: Backslashes are entered as double backslashes.To represent a single '\' you need to enter '\\'.
   DFS links can be given as filesystemadaptor.src: **\\\\host\\dfsnamespace\\link**

6. Create an ASCII or UTF-8 file named **`logging.properties`** in the same directory that contains the connector binary:

   ```
   .level=INFO
   handlers=java.util.logging.FileHandler,java.util.logging.ConsoleHandler
   java.util.logging.FileHandler.formatter=com.google.enterprise.adaptor.CustomFor
   matter
   java.util.logging.FileHandler.pattern=logs/adaptor.%g.log
   java.util.logging.FileHandler.limit=10485760
   java.util.logging.FileHandler.count=20
   java.util.logging.ConsoleHandler.formatter=com.google.enterprise.adaptor.Custom
   Formatter
   ```

7. Create a folder named logs in the same directory.

Step 3 Configure adaptor-config.properties variables

Optionally, you can add additional configuration variables to the `adaptor-config.properties` file that you created in the previous procedure. The following table lists the most important variables that pertain to the Connector for File Systems, as well as their default values. Variable names are wrapped for readability.

Variable	Description	Default
`server.dashboardPort`	Port on which to view web page showing information and diagnostics.	5679
`filesystemadaptor.supportedAccounts`	Accounts that are in the	`BUILTIN\\Administrators,\\Everyone,BUILTIN\\Users,`

	supportedAccounts will be included in ACLs regardless if they are builtin or not.	`BUILTIN\\Guest,NT AUTHORITY\\INTERACTIVE, NT AUTHORITY\\Authenticated Users`
`filesystemadaptor.builtinGroupPrefix`	Builtin accounts are excluded from the ACLs that are pushed to the GSA. An account that starts with this prefix is considered a builtin account and will be excluded from the ACLs.	BUILTIN\\
`filesystemadaptor.crawlHiddenFiles`	This boolean configuration property allows or disallows indexing of hidden files and folders. The definition of hidden files and folders is platform dependent. On Windows file systems, a file or folder is considered hidden if the DOS hidden attribute is set. By default, hidden files are not indexed and the contents of hidden folders are not indexed. Setting `filesystemadaptor.crawlHiddenFiles` to true will allow hidden files and folders to be crawled by the search appliance.	false
`filesystemadaptor.lastAccessedDate`	Disables crawling of files whose time of last access is earlier than a specific date. The cut-off date is specified in ISO8601 date	disabled

	format, YYYY-MM-DD. Setting `filesystemadaptor.lastAccessedDate` to 2010-01-01 would only crawl content that has been accessed since the beginning of 2010. Only one of `filesystemadaptor.lastAccessedDate` or `filesystemadaptor.lastAccessedDays` may be specified.	
`filesystemadaptor.lastAccessedDays`	Disables crawling of files that have not been accessed within the specified number of days. Unlike the absolute cut-off date used by `filesystemadaptor.lastAccessedDate`, this property can be used to expire previously indexed content if it has not been accessed in a while. The expiration window is specified as a positive integer number of days. Setting `filesystemadaptor.lastAccessedDays` to 365 would only crawl content that has been accessed in the last year. Only one of `filesystemadaptor.lastAccessedDate` or `filesystemadaptor.lastAccessedDays` may be specified.	disabled
`filesystemadaptor.lastModifiedDate`	Disables crawling of files	disabled

	whose time of last access is earlier than a specific date. The cut-off date is specified in ISO8601 date format, YYYY-MM-DD. Setting `filesystemadaptor.lastModifiedDate` to 2010-01-01 would only crawl content that has been modified since the beginning of 2010. Only one of `filesystemadaptor.lastModifiedDate` or `filesystemadaptor.lastModifiedDays` may be specified.	
`filesystemadaptor.lastModifiedDays`	Disables crawling of files that have not been modified within the specified number of days. Unlike the absolute cut-off date used by `filesystemadaptor.lastModifiedDate`, this property can be used to expire previously indexed content if it has not been modified in a while. The expiration window is specified as a positive integer number of days. Setting `filesystemadaptor.lastModifiedDays` to 365 would only crawl content that has been modified in the last year. Only one of `filesystemadaptor.lastModifiedDate` or `filesystemadaptor.last`	disabled

	`ModifiedDays` may be specified.	
`adaptor.incrementalPollPeriodSecs`	Time between incremental crawls.	300 seconds
`adaptor.namespace`	Namespace used for ACLs sent to GSA	Default
`server.port`	Port from which documents are served. GSA crawls this port. Each instance of a Connector on same machine requires a unique port.	5678

Step 4 Run the Connector for File Systems

After you install the Connector for File Systems, you can run it on host machine by using a command like the following example:

```
java -Djava.util.logging.config.file=logging.properties -jar adaptor-fs-4.0.2-withlib.jar
```

To run the connector as a service, use the Windows service management tool or run: `prunsrv start adaptor-fs`

Note: By default the Connector for File Systems service runs using the Windows Local System account. This should be fine in most cases but this can cause issues if access to documents is restricted through ACLs. In cases where the Connector for File Systems service is not able to crawl documents due to ACL restrictions, you would need to specify a user for the Connector for File Systems service through the Windows Service Control Manager that has sufficient access to crawl the documents.

Troubleshoot the Connector for File Systems

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