Microsoft

Windows Server 2008 R2

File Classification Infrastructure

Technical White Paper

Published: May 2009

For the latest information, please see
http://www.microsoft.com/windowsserver2008/en/us/default.aspx

Contents

Introduction 1
FCI Out-of-the-Box Capabilities 1
Extensibility 2
SharePoint Integration 2

File Classification Infrastructure 3
FCI and Policy-Based File Management 3
Storing Classification Properties 4
Extensibility Points 4
APIs for External Applications 4

FCI Usage Scenarios 5
Expiring Files 5
Detecting Sensitive Information on File Shares 5
Reducing the Cost and Time of File Backup and Restoration 5

Conclusion 7

Appendix: Details of File Classification Infrastructure in Windows Server 2008 R2 8
Installing File Classification Infrastructure 8
Defining Classification Properties 8
Creating Automatic Classification Rules 9
Scheduling Automatic File Management Tasks 9
Generating Classification-Based Storage Reports 10
Configuring Global Classification Settings 11
Managing File Classification Infrastructure on Remote Computers 11
Backing Up File Classification Infrastructure Configurations 12

Introduction

Data and data storage continue to grow in importance for most businesses and organizations. But the very reliance of business operations on electronic data adds new dimensions of complexity to managing this vital resource. Companies also bear increasing regulatory pressure over their data, while at the same time facing increasing risks from data leakage.

IT departments must manage increasingly large and unwieldy storage infrastructures. According to IDC's report "Workloads 2008: Understanding Server and Storage System Deployment," file server workload storage terabyte growth will have a 51 percent compound annual growth rate for calendar years 2008 through 2012. Storage is no longer simply a marginal expense, and the cost of storage-related security breaches can be hundreds of dollars *per record*. Additionally, any reductions in the cost of storage hardware are quickly outpaced by the demand for more storage.

Managing storage is no longer simply about volume and availability; organizations need to manage their data more effectively and more efficiently. Only by gaining insight into their data can companies reduce the cost of storing, maintaining, and managing data. Only by enforcing company policies and knowing how storage is utilized can administrators efficiently use their storage and mitigate the risks of data leakage. The next frontier for administrators is to be able to manage data based on business value.

Windows Server® 2008 R2 File Classification Infrastructure (FCI), a built-in solution for file classification, enables manual processes to be automated with predefined policies based on the data's business value. File Classification Infrastructure provides you with insight into your data to help you manage your data more effectively, reduce costs, and mitigate risks. FCI also provides an extensible infrastructure upon which independent software vendors (ISVs) and IT departments can build end-to-end solutions that enable organizations to perform advanced classification and data management, thus helping enable Microsoft partners to deliver rich classification solutions.

FCI Out-of-the-Box Capabilities

FCI's out-of-the-box functionality includes the ability to define classification properties, automatically classify files based on location and content, apply file management tasks such as file expiration and custom commands based on classification, and produce reports that show the distribution of a classification property on the file server.

Classification includes:

- **Automatic classification**—Using automatic classification rules, FCI can classify files according to the folder in which the file is located or based on the contents of the file.
- **Manual classification**—An end user can manually classify a file using the file properties interface built into the Microsoft® Office system files, and FCI will recognize these properties.
- **Line of Business (LOB) applications and IT scripts**—Using an API, LOB applications and IT scripts can set classification properties to files.

FCI also provides the following data management functionality with no additional third-party applications:

- **File expiration**—Dealing with stale, unused data can be a paramount data management issue for organizations. Expiring files based on usage and business value can reduce both the cost of storage and management and the risk of information leakage on file servers. The out-of-the-box FCI solution provides automatically scheduled tasks that expire files based on age, location, or other classification categories.

- **Custom tasks**—FCI empowers administrators to run custom commands to automate management tasks based on file name, age, location, or other classification categories. For example, IT administrators can automatically move data based on policies for either centralizing the location of sensitive data or for moving data to a less expensive storage facility.
- **Reporting**—Reports can provide administrators with a powerful tool to assess the risk of files being in the wrong place on their servers. Using the built-in capabilities of FCI, administrators can create reports in a variety of formats that contain details—including location—about files that have a particular classification. The FCI reporting infrastructure can also be used to generate information that can be used by another application.

Extensibility

There are many solutions on the market that do a good job of dealing with specific aspects of the challenges of ballooning storage requirements. FCI provides an extensible infrastructure to allow otherwise siloed solutions to work with one another and to empower companies to craft rich, end-to-end data management solutions that meet their organization’s specific business objectives.

FCI persists file classification between different ISV offerings so that products that classify files can work with products that consume file classifications. For example, if a data leakage-prevention product classifies files as containing personal information, then a backup product can back the files up to an encrypted store rather than the regular store. Moreover, IT administrators can build in-house solutions that plug into the classification infrastructure and interoperate with ISV product offerings.

SharePoint Integration

FCI integrates with Microsoft® Office SharePoint® Server 2007 so that file classification properties defined for Microsoft Office files on a file server persist with those files when they are uploaded into SharePoint. FCI-classified, third-party file types can also achieve this level of SharePoint integration through SharePoint file parsers. These parsers enable you to manage in SharePoint the metadata of file types for which SharePoint does not have built-in parsing support.

This white paper explains how the file classification infrastructure in Windows Server 2008 R2 works with organizations’ broader data management processes. It also examines FCI functionality in some specific deployment scenarios.

File Classification Infrastructure

Windows Server 2008 R2 File Classification Infrastructure fits into broader policy-based file management processes and can have its out-of-the-box capabilities augmented by third-party solutions.

FCI and Policy-Based File Management

While specific implementation of FCI will vary based on business needs, FCI fits into every phase of the process that makes up policy-based file management.

- **Planning classification properties**—It is a good practice for IT and business units to first consider which data management actions they would like to perform (for example, expiration, location of sensitive files, or backup). Based on that, they can then determine which classification properties should be assigned to files (for example, Secrecy = Top Secret, Classified, or Unclassified; PersonalInformation = Yes or No; and BusinessImpact = High, Medium, or Low). FCI provides a variety of property types that can be used to define a classification taxonomy that meets an organization's business needs. Available property types include Boolean, date, number, ordered list, and string value, among others.
- **Identifying files to be classified**—Based on their needs, organizations may choose to classify files on specific servers and shares. FCI comes with a built-in file system scanner that can scan files targeted for classification. Initial and ongoing scans can be scheduled to take place during low-utilization hours for large file systems. FCI can also discover file data at the request of other applications.
- **Classifying files**—There are several ways in which files can be classified.
  - Manual classification: Organizations can define Microsoft Office templates so that users can manually classify files.
  - Line-of-business (LOB) application classification: LOB applications can use FCI to set classification properties on files. For example, a human resource application can set the property *PersonalInformation* = *Yes* on a file when it saves a file to the file server.
  - Automatic classification: Using FCI, IT administrators can create automatic classification rules that classify files according to the location or content of the files. This functionality is provided right out of the box. IT administrators can also classify files according to any other mechanism supported by third-party classification plug-ins.
- **Managing files based on classification**—There are a number of management tasks based on classification that can be performed.
  - Out-of-the-box file management tasks: IT administrators can use scheduled file management tasks in FCI to perform custom commands based on classification properties, age, and file name wildcards. IT administrators can also configure file management tasks to send automated notifications to content owners when an automated file management task runs. For example, when files become old enough to be automatically expired, content owners can be notified in advance and given the opportunity to evaluate the proposed expiring content.
  - Data management applications: Partner solutions can use FCI to provide data management such as backup and archival based on classification. For example, to save backup costs, the administrator can define that important information is

backed up once a day, while less important information is backed up once a week.

Storing Classification Properties

FCI can store the classification information about a file in a number of locations. By default, FCI stores the properties in an NTFS alternate data stream (ADS), so that all file types can be classified. The alternate data stream persists as long as the file is on an NTFS file system. For Microsoft Office files, FCI stores the file properties in the file itself in addition to the ADS. This allows Microsoft Office file properties to be maintained if the files are uploaded to SharePoint or if they are e-mailed.

Lastly, FCI provides an extensibility model for storing properties in files such that third-party format owners can provide a plug-in that stores classification properties within their file formats.

Extensibility Points

FCI is extensible by third-party plug-ins at the point where files are classified (classifier plug-in) and the point where properties get read/stored for files (property storage module plug-in).

- **File classification (Classifier plug-in)**—Products that classify data can hook into the FCI automatic classification rules by providing a classification plug-in.
- **File property storage (Property Storage Module plug-in)**—The method for storing and reading classification properties is extensible, so that for different file formats, properties can be stored in various locations by means of third-party plug-ins. For example, a third-party video file plug-in can provide a way to store and extract properties from a video file. In another example, classification properties can be stored in a database or in the cloud.

In addition, the custom file management tasks can be extended by applications and custom scripts.

- **Custom file management tasks**—Custom file management tasks run on a scheduled basis and invoke a custom script or application based on a condition. Custom scripts, such as moving a file and leaving a symbolic link in the original location or changing an access control list (ACL) on a sensitive file, can be provided by IT departments to apply automatic data management to files based on their classification.

APIs for External Applications

FCI includes Distributed Component Object Model (DCOM)-based application programming interfaces (APIs) to support integration with third-party data classification and management applications. These APIs allow third-party applications to do the following:

- **Get and set all of the individual properties of a given file**—For example, a third-party backup application can use the APIs to *get* properties from files and to back up files classified as containing personal information to an encrypted file store. A LOB finance application can set the “Project” property in files to “Finance.”
- **Automatically manage the FCI configuration on a file server**—For example, FCI configuration APIs enable a central management solution to configure FCI rules and tasks across multiple servers.

Moreover, because the APIs are based on Microsoft DCOM technology, they are remotely scriptable by tools such as Windows PowerShell™ and can be used across multiple servers.

FCI Usage Scenarios

Windows Server 2008 R2 File Classification Infrastructure provides end-to-end data management solutions to organizations both out of the box and coupled with IT department custom scripts and Microsoft partner applications. These range from data grooming to helping secure personal information and improving file backup service level agreements (SLAs).

Expiring Files

Suppose IT administrators at a specialty software company, A. Datum, have noticed that more than 60 percent of their storage is taken up by files that network users create and then never touch again. This represents both large cost and potential risk for A. Datum. To handle this problem, A. Datum IT management decides to institute a process of file expiration.

The IT management staff at A. Datum classifies files on the network file shares based on their business criticality (high, medium, or low) on an ongoing basis. The IT managers then use management tasks in FCI to expire files that are three years old or older, that have not been modified in the last year, and that are of only medium or low importance. Before expiring the files to an archive, IT management has FCI automatically notify the file owners that their files are about to be expired so that the owners can take action if necessary. This saves management time, frees up storage space on the A. Datum network, and intelligently reduces the growth in demand for more storage while decreasing backup costs. It also reduces risks arising from stale data—for example, the risk of outdated (and thus out-of-context) data being brought into litigation through e-discovery is reduced. This functionality is available right out of the box with Windows Server 2008 R2.

Detecting Sensitive Information on File Shares

As another example, suppose legal staff and IT administrators at Trey Research, a pharmaceutical firm, are worried about confidential information from participants in clinical trials being kept on improperly accessible file shares.

Trey administrators can use the functionality built into FCI to address this concern. Trey legal counsel determines that participants' personal records are the primary concern, and Trey IT staff creates a classification property for files on Trey shares that flags the files as confidential.

IT configures FCI to find sensitive words in Trey Research files. IT administrators at Trey then create a management task that runs nightly. When files that are flagged as containing confidential information are detected in the targeted file shares, the task changes the access rights to the file and sends a notification to administrators of the location and name of the file or files.

Trey's processes for securing the personal information of clinical trial participants can be improved further by integrating a third-party classification application with FCI. Accordingly, Trey Research purchases a third-party classification application approved by the Health Insurance Portability and Accountability Act (HIPPA) that plugs into FCI. Trey administrators add classification rules that use the new plug-in to detect confidential information, and the file management task that is already in place will report on the files found by the new plug-in. This saves administrative time and money and enables Trey IT administrators to more finely search for confidential information.

Reducing the Cost and Time of File Backup and Restoration

Contoso is a logistics firm struggling with the cost of backing up data. Suppose Contoso management knows that all of the data they regularly back up is not of the same business value, but because they cannot positively differentiate between files of varying value, they have to back up all data nightly. As the number of files generated in the course of business operations continues to explode, Contoso's backup costs are soaring.

The CIO and IT management at Contoso have decided to couple FCI with a third-party backup application to address this challenge. Working with division heads, Contoso IT management has devised a taxonomy and a process for classifying files as having either high, medium, or low business impact. They then use FCI to apply this classification to all files on Contoso file shares. When backups are run, the backup application queries files based on their properties. The backup application backs up high business impact files nightly, medium business impact files weekly, and low business impact files monthly. Moreover, Contoso is able to restore its high business impact files first in cases of disaster, getting the company's business solutions back up and running again faster and minimizing the business cost of missing data.

The result for Contoso is a dramatic savings in backup costs and time without degradation to Contoso's backup SLA. Contoso is able to differentiate its data based on business operations and then backup and restore data accordingly.

Conclusion

Reliance on data and storage resources continues to grow in importance for almost all organizations. CIOs face increasing regulations and concerns about data leakage, while IT administrators face the steadily growing challenge of overseeing larger and more complex storage infrastructures. Simultaneously, IT departments are being tasked with maintaining the total cost of ownership of storage at reasonable levels. Managing storage resources is thus no longer just about storage volume or data availability—it is also about the enforcement of company policies and knowing how storage is consumed to enable efficient utilization and compliance to mitigate risk.

Windows Server 2008 R2 File Classification Infrastructure (FCI) provides you with insight into your data by automating the classification and file management processes so that you can manage your data more effectively. FCI enables this by providing an extensible, automatic classification mechanism that partners can plug into and by providing interfaces to query file classification in order to apply actions to files based on classification. These mechanisms include built-in, out-of-the-box functionality in Windows® and interfaces for partners to build rich, end-to-end solutions for classifying and applying policy based on classification. Through FCI, Microsoft customers can save money and reduce risk by managing files based on their business value and business impact.

Appendix: Details of File Classification Infrastructure in Windows Server 2008 R2

Out of the box, File Classification Infrastructure in Windows Server 2008 R2 provides the functionality to automatically classify files, expire and run custom commands based on classifications, and create detailed reports.

This appendix familiarizes you with installing, configuring, and using FCI for this functionality and provides some examples of this functionality in use.

Installing File Classification Infrastructure

FCI can be controlled through File Server Resource Manager (FSRM). FSRM is a feature of the File Services role in Windows Server 2008 R2 that can be installed as part of the File Services role by using Server Manager. After you install the file server role with FSRM, you can use the FSRM Microsoft® Management Console (MMC) snap-in.

Membership in the local administrators group or equivalent is the minimum required to use FSRM.

Defining Classification Properties

FCI enables organizations to define classification properties for their files based on the organization's policies for managing data and business operations. Different organizations might use different properties; FCI enables organizations to define the property set that makes sense for their business needs.

For example, at Fabrikam, a custom engineering firm, the CIO works with the legal department and the IT department to define properties with which all files should be labeled. These labels include "Confidentiality," "Business Impact," "Project," and "Personal Information." Classification properties are the means by which IT administrators apply these labels to files with FCI. The IT department then defines the following properties to apply to Fabrikam's files:

- **Project**—String value
- **BusinessImpact**—Ordered list (*High*, *Medium*, or *Low)*
- **PersonallyIdentifiableInformation**—Boolean (*Yes* or *No)*
- **Confidentiality**—Ordered list (*BusinessNeedOnly*, *Confidential*, or *Public)*

FCI classification properties are defined on a per-file server basis:

- If a property is not defined on the server, it cannot be set on or applied to a file on that server.
- However, if a property was applied to a file on another server but is not defined on this server, it can still be queried from the file on this server.

FCI currently supports the following property types:

- **Ordered List**—A list of values, of which only one can be selected at a time. Entries early in the list have a higher priority during conflicts than later entries. (For example, *Confidentiality = High, Medium, or Low*.)
- **Boolean**—A Yes/No value. (For example, *PersonalInformation=Yes/No.*)
- **String**—A simple text label. (For example, this may be the project for this file.)

- **Number**—A number value.
- **Date/Time**—A date and time value.
- **Multi-choice**—A list of values, of which several can be selected at a time. (For example, a list of data types that this file contains.)
- **Multi-string**—A series of text labels. (For example, multiple projects for a file.)

Note that defining a classification property does not alter any files. Property values are assigned to files by either automatic classification, set by LOB applications, or manually set by users.

Creating Automatic Classification Rules

Automatic classification rules are a mechanism that assigns properties to files based on a specific classification mechanism. Classification rules are used to evaluate which values should be assigned to properties for files on the server.

To return to the example of Fabrikam, the records manager and the legal department have determined classification rules for Fabrikam documents as laid out in Table 1.

Table 1: Fabrikam Document Classification Rules

Engineering	Finance	All others (unless shown otherwise)
• Medium business impact • Confidential	• High business impact • Accessible only with a business need	• Low business impact • Publicly available

The **Project** property is set on a per-project basis within each department.

Automatic classification rules are configured to assign values to the *BusinessImpact* and *Confidentiality* classification properties based on the location of Fabrikam files.

Fabrikam administrators can also add a further rule based on the content of files. Administrators can create a property named *PersonalRecord* and have FCI flag files that contain sensitive, private information such as Social Security numbers or telephone numbers. (Using custom file management tasks, files with *PersonalRecord = Yes* could then automatically be moved to file shares with controlled access to better secure personal information.)

The classification engine uses a scheduled task to perform this operation and evaluate any defined rules. By default, there is no schedule for automatic classification rules; administrators must set this schedule.

Scheduling Automatic File Management Tasks

The File Management Tasks feature enables administrators to automate scheduled file management tasks such as file pruning, protection of sensitive information, and so forth. These tasks can be scheduled to occur periodically. Target files for processing by File Management Tasks can be defined through the following values:

- Location
- Classification properties

- Creation time
- Modification time
- Last time accessed
- File name/extension wildcards

File management tasks can also be configured to notify file owners of any impending policy that will be applied to their files.

Consider Fabrikam once again. Now that files are being classified on the server, Fabrikam needs to ensure that policy is maintained correctly. The biggest challenges facing Fabrikam are:

- Users are sharing out files that should only be available to those with a genuine business need. Fabrikam managers fear that this could end up costing the company money if the confidentiality of the information on these shares is breached.
- Employees are creating large numbers of files that they do not subsequently use. Fabrikam's costs for procuring more disks and the continuous cost of backing up all that data are spiraling out of control.
- The legal department has asked that all documents with a high business impact be mirrored to another location for disaster recovery purposes.

FCI makes it easier for Fabrikam to address each of these challenges. Classification-based reports delivered weekly to Fabrikam administrators will identify any files in the share folders that do not meet respective policies. IT managers can configure file expiration to automatically expire (move) files to a specified directory based on certain criteria; from there, they can back up those files and delete them. Moreover, managers can set up a similar mechanism to automatically copy all files with high business impact to a mirroring location.

Defining File Expiration Tasks

FSRM defines the concept of file expiration as follows:

- All files matching certain criteria are moved to a specified directory.
- An administrator can then back those files up and delete them.

Defining Custom File Management Tasks

FCI can perform tasks on files beyond expiration. File Management tasks allow administrators to run custom commands based on classification and other conditions.

Setting Optional File Management Notifications

File management tasks often alter or remove data that a user has placed on a server. It can be important to send out alerts that this procedure will be performed on a set of files. To this end, FSRM can send e-mail messages to administrators or specific users, log an event, and/or run a command or a script at defined intervals before a file management task is triggered. You can configure more than one notification deadline for each file management task. By default, no notifications are generated.

Note that to send e-mail notifications and configure the storage reports with parameters that are appropriate for your server environment, you must first set the general FCI options in FSRM.

Generating Classification-Based Storage Reports

FSRM can generate reports that will help you understand the distribution of classification property values in files on the file server.

From the Storage Reports Management node in the FSRM user interface, you can create report tasks, which are used to schedule one or more periodic reports, or you can generate reports on demand. For on-demand reports, as with scheduled reports, current data is gathered before the report is generated. If you select a “Files by Property” report, the final report will list files grouped by values assigned to classification properties. You can use this report to analyze distribution of values assigned to classification properties and to identify when files have policies applied to them incorrectly.

For more information on how to configure storage reports, see the File Server Resource Manager Step-by-Step Guide for Windows Server 2008 on the TechNet Web site (http://technet.microsoft.com/en-us/library/cc771092.aspx).

Scheduling Files-by-Property Reports

To generate a Files-by-Property report on a regular schedule, you schedule a report task. The report task specifies which reports to generate and what parameters to use, which volumes and folders to report on, how often to generate the reports, and which file formats to save them in.

When you schedule a set of reports, the reports are saved in the report repository. You also have the option of sending the reports to a group of administrators by e-mail.

Configuring Global Classification Settings

You can set FCI configuration options that apply to a whole server in the **File Server Resource Manager Options** dialog box. These options include the following settings:

- The classification schedule specifying when the classification process should start
- Default e-mail notification settings:
  - Who, if anyone, should receive e-mail upon completion of a classification process.
  - The format in which the classification reports should be saved.

Note that to send e-mail notifications to administrators or to users whose files will be affected by a file management task, or to send storage reports over e-mail, you must specify the SMTP server to use and the default e-mail settings in the **E-mail Notifications** tab.

Managing File Classification Infrastructure on Remote Computers

File classification is handled on a per-server basis. To manage FCI on a remote computer (that is, any server other than the local one), you can connect to the computer from FSRM. While you are connected, FSRM will display the objects created on the remote computer, enabling you to manage them in the same way that you manage resources on your local computer.

To manage remote storage resources with FSRM, the following conditions must be met:

- The remote computer must be running Windows Server 2008 R2 with FSRM installed.
- The **Remote File Server Resource Manager Management** firewall exception on the remote computer must be enabled. This exception can be enabled through Windows Firewall in Control Panel.
- The administrator managing the remote computer must be logged on to the local computer with an account that is a member of the **Administrators** group on the remote computer.

It is important to remember that to manage storage resources on a remote computer, the stand-alone FSRM snap-in must be used instead of the Server Manager snap-in. The stand-alone snap-in is available in Administrative Tools.

Backing Up File Classification Infrastructure Configurations

A server’s FCI configuration is backed up and restored when FSRM backs up its system metadata.

To perform a full backup and restoration of FCI configurations, you must use a backup utility such as Windows Server Backup that is compatible with the Volume Shadow Copy Service (VSS) writer infrastructure. System metadata for FCI is backed up and restored as part of the system state.

This is a preliminary document and may be changed substantially prior to final commercial release of the software described herein.

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

The example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious. No association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred.