EMC²
where information lives®

Using EMC ControlCenter UNC File Level Reporting on CIFS Shares

EMC Proven Professional Knowledge Sharing 2009

Chad DeMatteis
Sr. Storage Operations Specialist
EMC Corp.
Dematteis_chad@emc.com

Michael Horvath
Data Storage Administrator
Fifth Third Bancorp
michael_horvath@yahoo.com

EMC² PROVEN PROFESSIONAL

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect EMC Corporation's views, processes or methodologies.

Introduction

Controlling storage costs is a recurring theme of modern Information Technology Asset Management. As a result, most storage professionals and system administrators are required to provide an accounting of the data stored on Enterprise Devices. The toolset offered by many client operating systems is insufficient to quickly and accurately provide *timely* reporting on most aspects of this data.

Collecting file and folder statistics within a NAS CIFS share is one aspect of data accounting. It can be daunting to provide detailed reporting of a large CIFS share with deep directory trees. It is especially difficult when using Windows native tools that can have poor enumeration performance. In an environment with tens of thousands of folders and millions of files, these activities can consume a large amount of time.

This article discusses how EMC ControlCenter's® Network File System Assisted Discovery feature, introduced in v6.0, can provide Celerra CIFS administrators with file and folder level reporting(FLR); covering UNC(Universal Naming Convention) FLR configuration considerations, and lessons learned during an EMC ControlCenter deployment. It also provides practical examples of how you can use CIFS FLR reports to quickly determine file age and type distribution, top storage users, and utilization trending. These reports provide administrators with the information needed to maximize storage utilization and address CIFS storage consumption issues before they impact end users.

Audience

This article assumes the reader is knowledgeable about EMC ControlCenter Console, EMC ControlCenter Storage Scope™, and Crystal Reports. It focuses specifically on using EMC ControlCenter's Network File System feature for data collection and reporting.

Understanding the Challenge

We found there was a need for reports on Celerra CIFS file and folder utilization during an EMC ControlCenter upgrade and deployment. The Windows administrators tried using Windows native tools to manually gather folder sizes via properties, storage used by certain file types, and the top CIFS users. They encountered long enumeration times (see figure 1) and the "flashlight" wait icon while traversing large folder trees; forcing them to map and remap to deeper paths within the tree.

There was an obvious need for a reporting tool given the need to collect data that covered multiple CIFS shares, each with up to 5000 folders and over 1 million files.

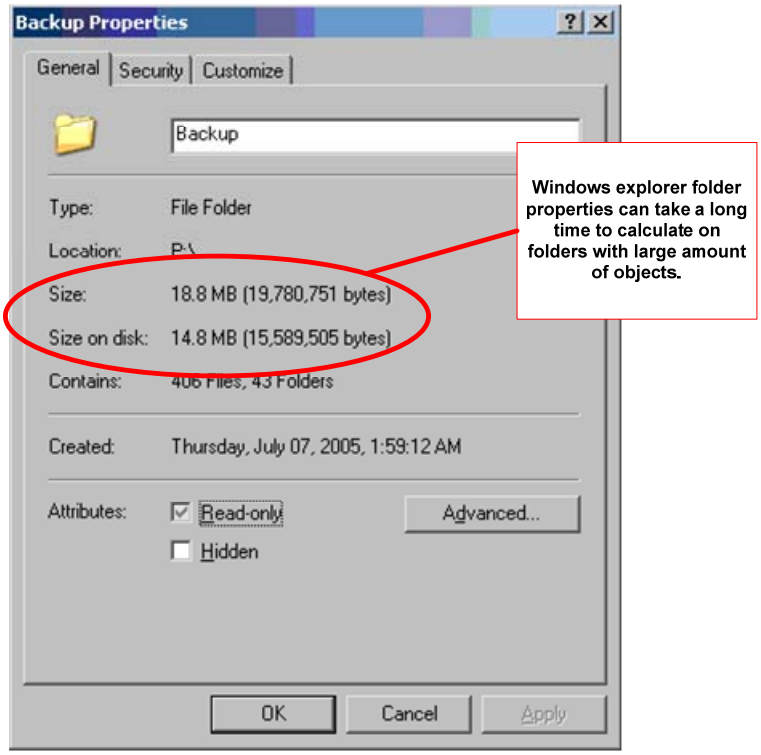


Figure 1: Folder Properties

The Solution

There were a few ways to provide the reports needed: command line/windows explorer, scripting (perl or vbscript), and EMC ControlCenter. Command line was straight forward, but time consuming and labor intensive. We could write a script to gather the file and folder properties, but that would require time to write and also a database to store the gathered data. The third choice was EMC ControlCenter 6.0. It provided a new feature to discover a Network File System using UNC along with UNC FLR data collection policies. This feature is capable of collecting data from a Network File System hosted on a Celerra®. Since EMC ControlCenter is a proven platform with facilities to manage and store data collections and provided reporting capabilities, it was the most logical choice.

Background of Environment

This article is based on the following scenario:

- Distributed EMC ControlCenter infrastructure: StorageScope and EMC ControlCenter server running on separate servers each with Windows 2003 sp2; 4 CPU-3GHz; 4GB Memory.
- Host for FLR scans: windows 2003 sp2; 8 CPU-3GHz; 4GB Memory
- NAS: NS702G and NS80
- File systems: 8 CIFS file systems totaling 22,000 folders and 13,000,000 Files

EMC ControlCenter UNC Discovery Steps

Overview

The discovery of UNC file systems within EMC ControlCenter is done through the Assisted Discovery (AD) interface using the Network File system Discover Type (see figure 2). The actual discovery of the file system is handled through a designated Windows host agent in which the host agent scans the file system via UNC (it does not map the file system to a drive letter). Once discovered, the Network file system is considered a managed object of type “NetworkShare.”

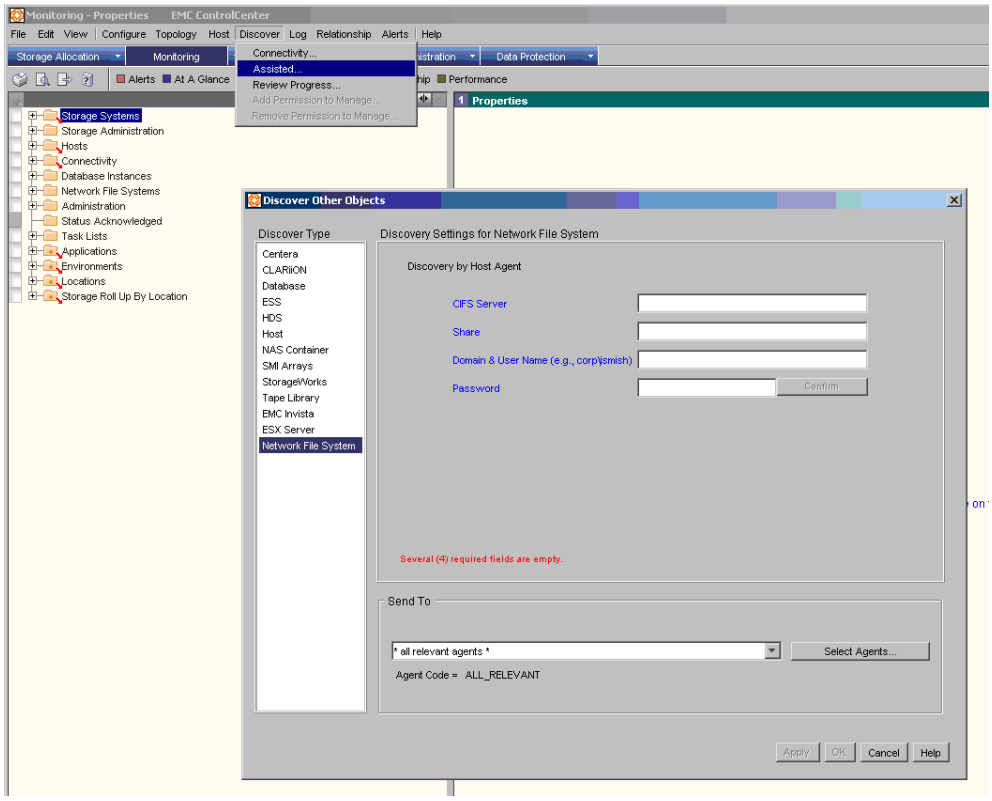


Figure 2: Assisted Discovery

Requirements/Considerations

The discovery of a Network File system through assisted discovery requires:

- Celerra/Server CIFS interface name used to access Network File System; referred to as “CIFS Server” in AD.
- Name of Network File system referred to as “Share” in the AD.
- Domain user ID that has permissions to access desired share on CIFS server.
- Server running a Windows agent with network access to the CIFS Server.
- StorageScope file level reporting license.

Given these requirements, there are some important considerations prior to discovery of a Network File system.

First, the domain ID should be a dedicated functional account intended strictly for discovery and scanning of network file systems. Ideally, the password for the account should remain static (within security policy guidelines) as a password change would require a rediscover of the network file systems. You might be tempted to use your personal account or share a functional ID with another application, but if the password changes the next scheduled discover/scan of the network file system will cause the ID to be locked out. This will impact other business functions.

The second, and most important, consideration is the selection of the Windows host agent for the discovery. Running FLR scans against a network file system results in a CPU performance cost on the Windows agent host. *Note: The performance impact is described later in the performance section of this paper.* A best practice is to use a dedicated host or one that has free CPU cycles to manage the file system**. Do not select** the default: "*all Relevant agents*." If all relevant agents are selected and the EMC ControlCenter environment contains a large number of Windows agents, the ownership and the performance impact of the FLR scans will bounce around from server to server as agents are restarted. This makes it difficult to coordinate data collection schedules and can have unpredictable performance impacts if a server is heavily utilized when a scan is performed.

If additional management hosts need to be added or removed after the initial discovery, the management permissions of each Network File System can be controlled through the assisted discover manage permissions interface.

Example Network File System Discover

The following example shows the discovery of a Network file system hosted on a Celerra.

Share Name: \\cifs01\drive01
Celerra CIFS interface: cifs01
Domain ID: corp\eccflr
Windows Host agent server: flrscan01

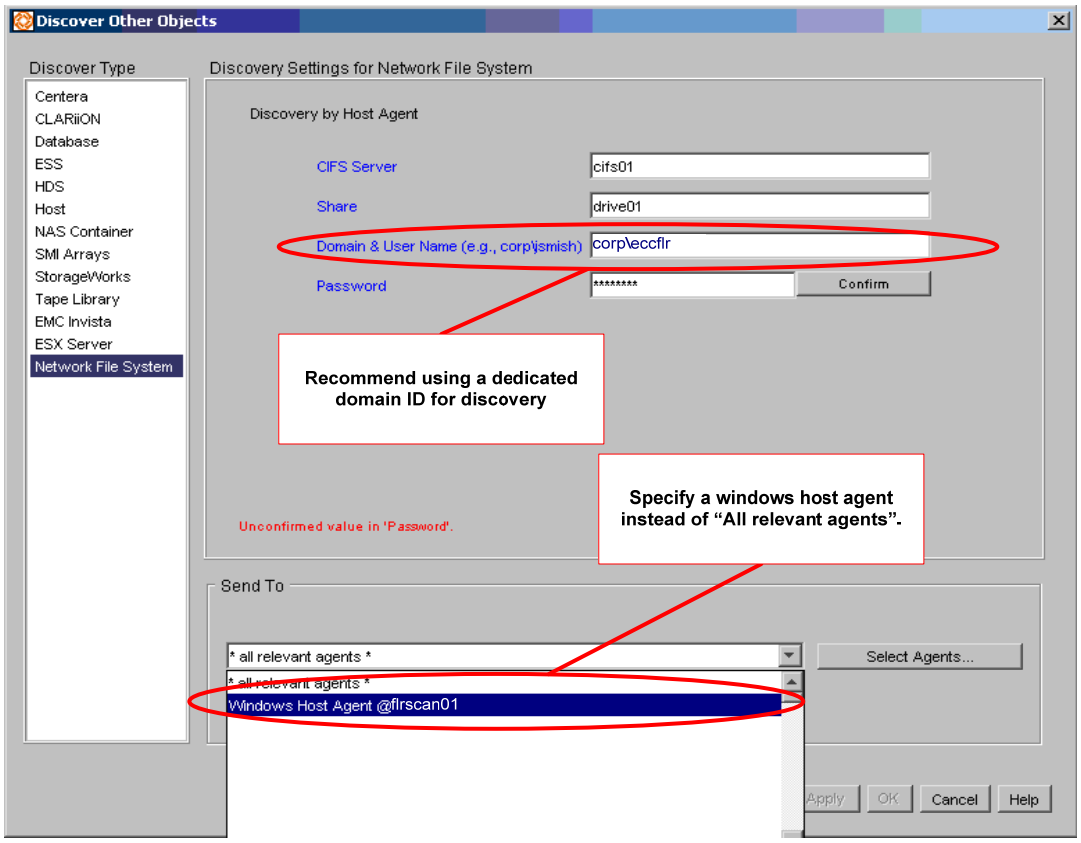


Figure 3: example of Network file system Discovery

Progress of the discovery can be monitored through AD's "review progress" option:

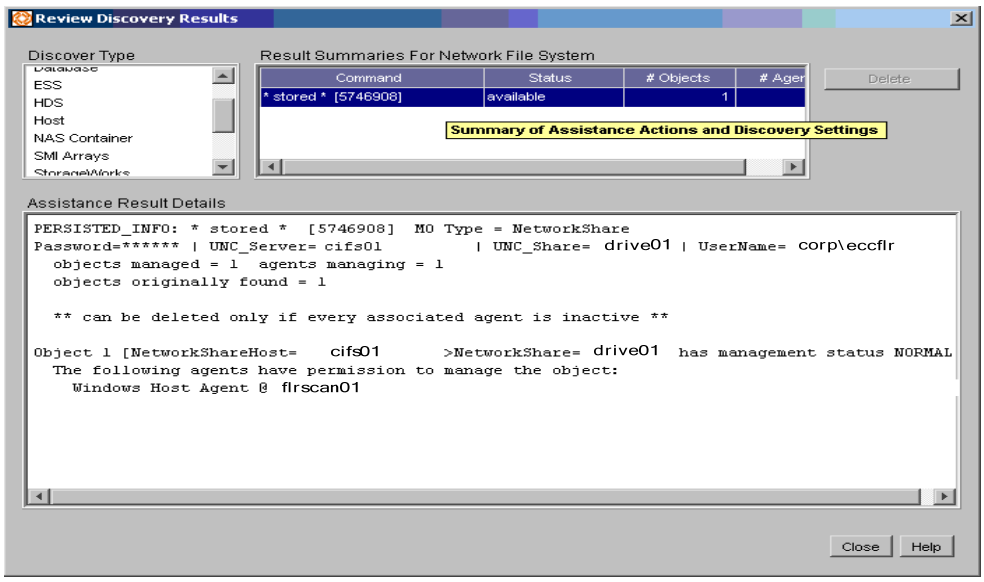


Figure 4: review progress of assisted discovery

Once discovered, the share will show up as a managed object within the EMC ControlCenter console under the “Network File Systems” folder:

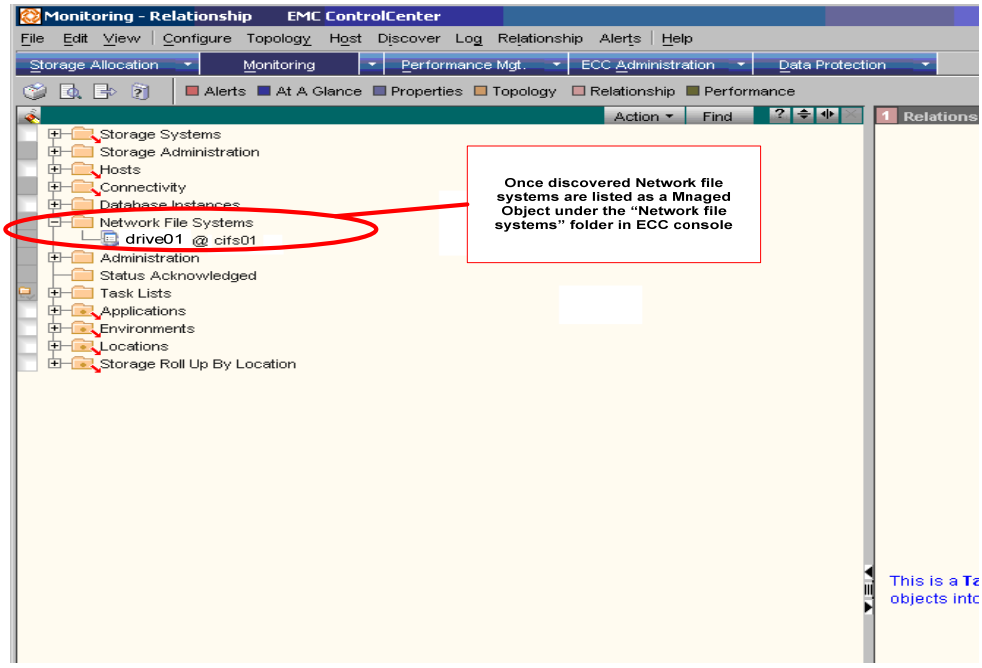


Figure 5: Network File system Managed Object

Network File System Data Collection Policies

As with any other managed object within EMC ControlCenter, a collection policy is required to maintain active discovery and to gather information about an object. EMC ControlCenter provides a specific Data Collection Policy (DCP) template for both discovery and File level collection of Network file systems. The templates are part of the Host agent for windows, called “Discovery UNC” and “File Level Collection for UNC Connections.”

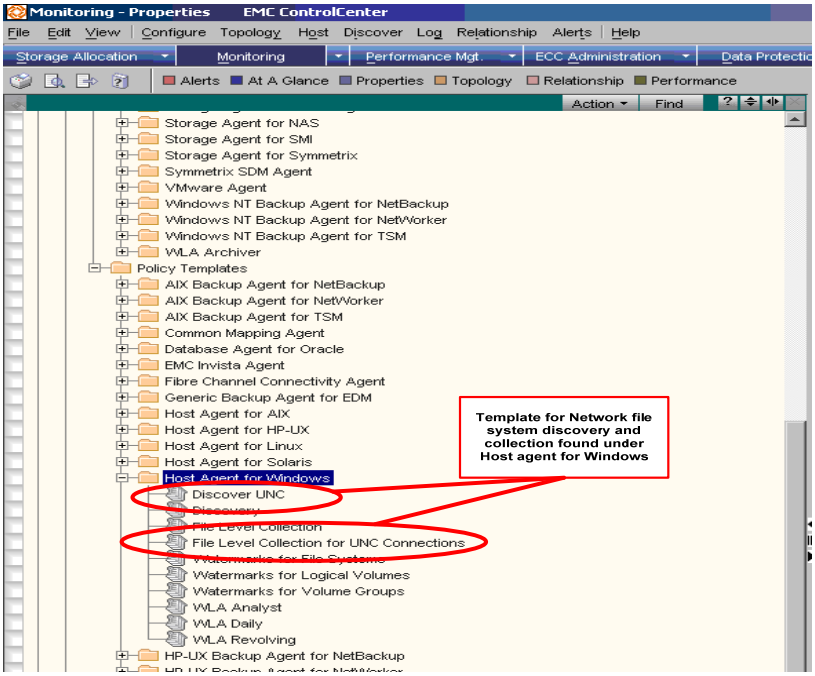


Figure 6: Data Collection policy template for UNC

Defining a Collection Policy

The DCP creation process requires decisions about the type of statistics to gather and the frequency to gather those statistics. Base these decisions on the type of reporting required.

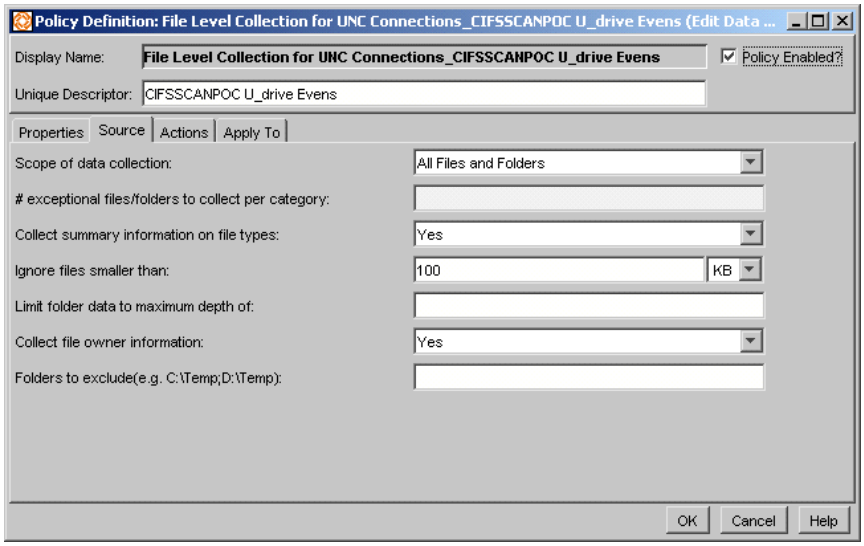


Figure 7: example of data collection policy

The “source” dialog box in the UNC DCP has several options to consider (see figure7):

Scope of data collection: defines what level of scan to be run.

- **“All Files and Folders”**- scans all files and folders in a particular filesystem. This option provides the most detail but is the most resource intensive.
- **“Exceptional Files and Folders”-** this option scans all files and folders that are in the “top <N>”; where ‘N’ is defined in “# exceptional files/folders to collect per category.” The categories include number of files greater than 1MB and number of files not accessed in more than 60 days.
- **“folders only”-** scans folders and returns summary information.

 # Exceptional files/folders to collect per category:

- number of files to scan for “top N”.

Collect summary information on file types:

- If “yes,” summary information on file types as defined in the “File Type Definitions” will be included in scan. This is required if file type distribution is needed for reporting.

Ignore files smaller than:

- The scan will not collect on files less than a defined maximum size in bytes, KB, MB or GB.

Limit folder data to maximum depth of:

- Defines the number of directories to scan before stopping. Large folder trees require longer time to scan and greater CPU resources on host.

Collect file owner information:

- If “yes” is selected, the owner of information will be gathered.

Folders to exclude:

- A directory path to be excluded from the scan. Must be a full path e.g. \\cifs01\archive.

The “actions” and “apply to” tabs define the schedule and the particular Network File systems that will be scanned with the DCP. The schedule should consider the amount of files and folders to be scanned. For example, a folder summary on a File System with only a few hundred folders can be scheduled on a nightly basis, while a complete scan of millions of files and thousands of folders is best scheduled on a weekly basis.

Note: Take care when scheduling DCPs for both a summary and a detailed file level scan. If a summary scan is scheduled after a detailed scan of a particular file system; the detailed FLR data from that scan will be lost. A summary has to run prior to the detailed scan.

Example Defining a Collection Policy

In this scenario, a detailed file level scan of the file systems was required. We ran multiple DCPs on Sundays because performance and scan times were a concern. Each DCP contained a mix of NS702G and NS80 file systems, and collected all files and folders including owner and file type information with files greater than 100KB.

Chart 1 below summarizes the schedule:

DCP Name	Filesystems	cifs server	#folders	#files(Millions)	Schedule
Cifs FLR Evens	drive02	cifs01 (ns702g)	4000	3.3	9am Sundays
	drive04	cifs01 (ns702g)	5700	2.3	9am Sundays
	drive06	cifs02(ns80)	100	0.6	9am Sundays
	drive08	cifs02(ns80)	826	1	9am Sundays
Cifs FLR Odds	drive03	cifs01 (ns702g)	5500	5.2	Noon Sundays
	drive05	cifs02(ns80)	6000	1.2	Noon Sundays
	drive09	cifs02(ns80)	1000	0.5	Noon Sundays

Chart 1: DCP schedule example

FLR Data Removal Schedule

Purge older collections from the StorageScope repository since FLR data changes constantly. This is done through the “FLR data Removal schedule” in StorageScope.

In this example, since the DCP for the file systems runs on Sundays, the previous scan is removed on Saturdays (see figure 8).

Note: The FLR removal schedule doesn’t remove file trending data.

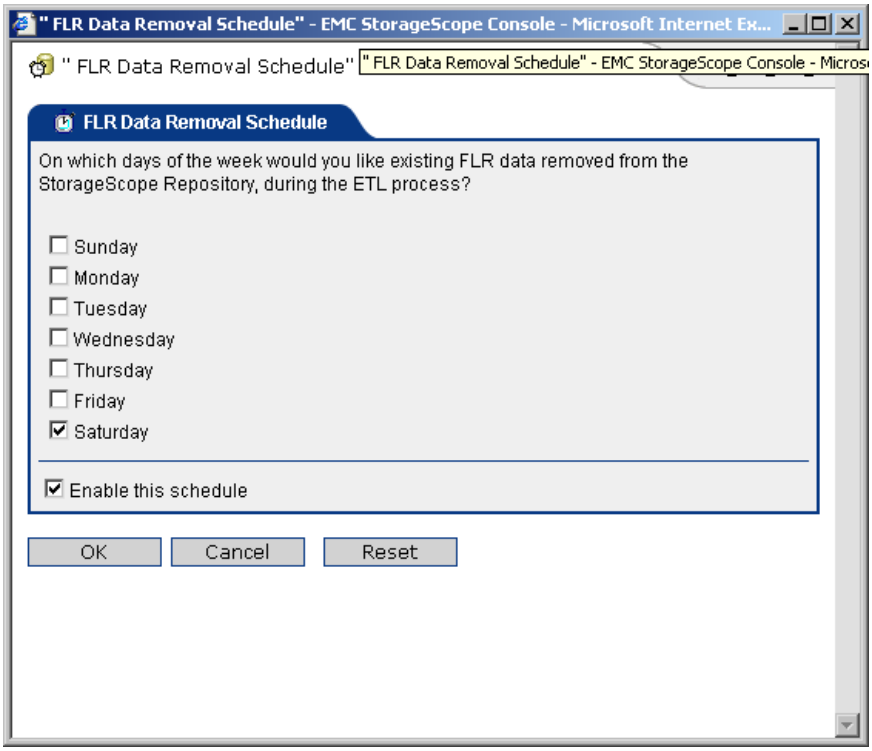


Figure 8: FLR data removal example

Performance Considerations

Performance impact is an important consideration when scanning Network file Systems on a Celerra. There is a direct correlation between the number of files and folders to be scanned and the amount of resources required to run the scan. The greatest performance impact is on CPU utilization of the Windows server that hosts the agent responsible for the FLR scan. That is why careful consideration is required when selecting which hosts have permissions to manage the file systems.

The amount of CIFS and SMB calls generated per second during a scan is another performance consideration. Although in our experience the scans did not generate more calls than an average business day, due to the amount of folders and files the scans did contribute a noticeable amount. Therefore, consider Celerra performance when scheduling a scan, particularly if the Celerra datamover is heavily utilized.

Example Performance Charts

The following figures show an example of scans based on the DCPs found in Chart 1. We collected the data using EMC ControlCenter workload analyzer.

Figure 9 shows the CPU utilization of the windows server running the scan. The noticeable CPU utilization is fairly sustained over the period of the scans.

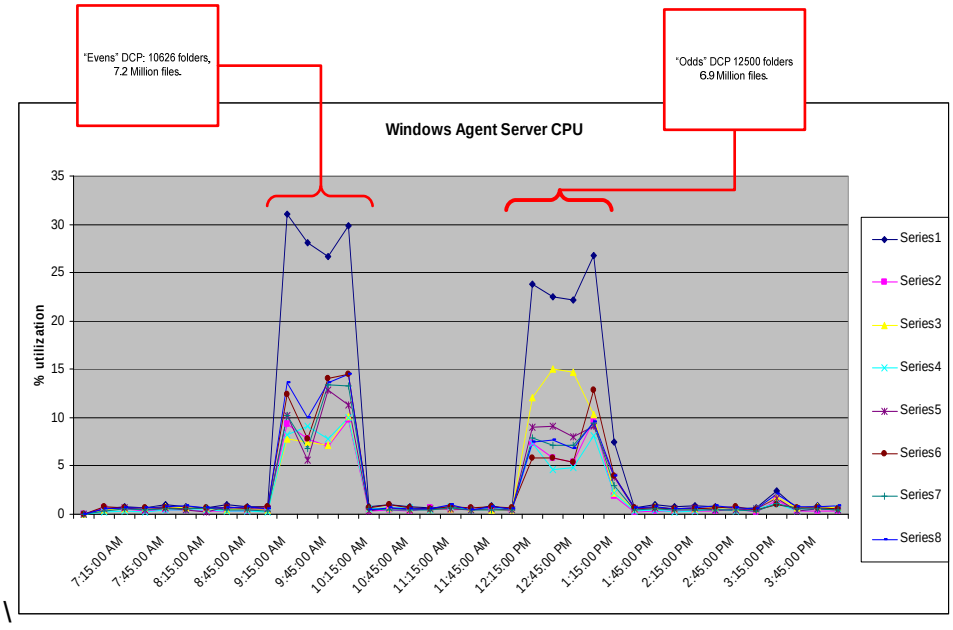


Figure 9: host agent server CPU Utilization

Figures 10a and 10b show the CIFS and SMB calls on a NS702G and NS80 during the two scans. Since these are scans of file/folder properties, there isn't a large amount of data transferred from the Celerra; but every file and folder query generates a call represented from the spikes on the graphs.

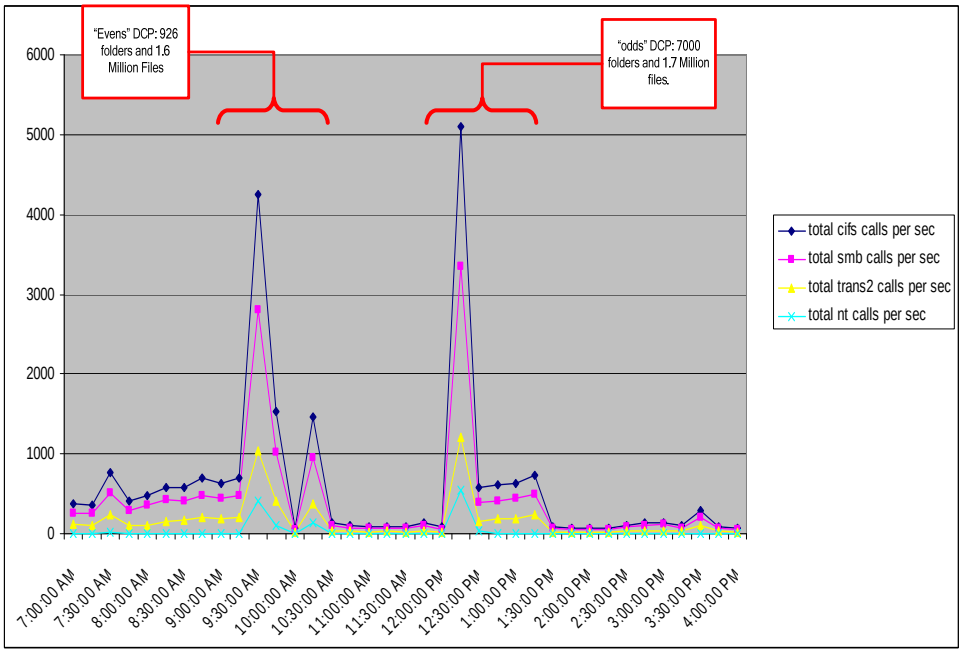


Figure 10a: CIFS calls per second on ns80 "cifs02"

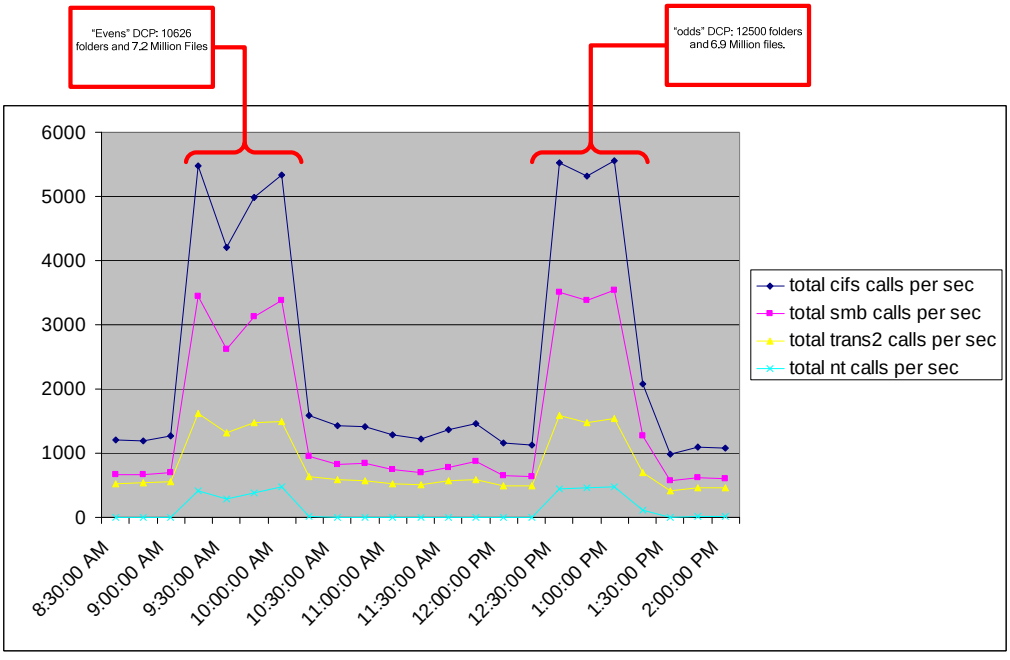


Figure 10b: CIFS calls per second on ns702g "cifs01"

StorageScope Reports for Network File Systems

The metrics gathered by EMC ControlCenter on Network File systems are collected by the Windows host agent, sent to the FLR archiver agent, and then populated into the StorageScope repository during the ETL.

The StorageScope repository File and folder tables used for Network File systems are the same tables as host based file level scans. However, one difference is that the network file system records do not contain a host key (see figure 10). Consider the lack of a host key when generating reports or selecting file systems from a specific CIFS server. The "parent directory" field can be used as a filter in cases where a specific CIFS server is required for a report .

Parent Directory	Owner	File Name	Host Key
\\cifs01\drive01\Storage\	#12345	allocation.xls	

Figure 10: example of SRM file record

A detailed list of tables and field descriptions can be found in "EMC ControlCenter 6.0 StorageScope: API and Repository Reference Guide."

StorageScope built in reporting examples

StorageScope provides built-in reports that provide high-level summaries of Network file systems.

The built in Enterprise summary report provides graphs of top users, top file types, and file age distribution. Figures 11 and 12 show some samples from the enterprise summary report.

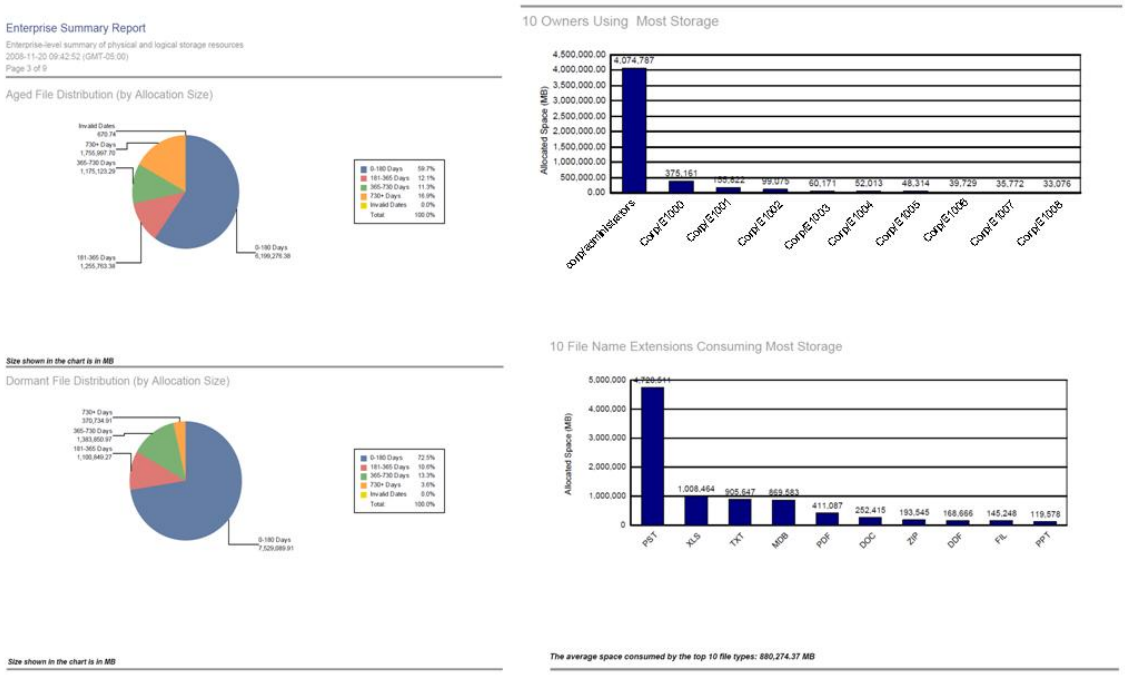


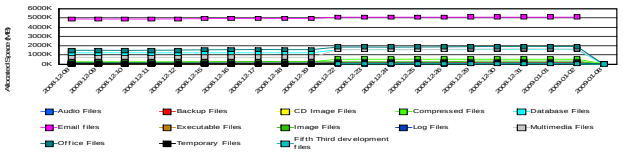
Figure 11: Example of built in reports

In addition to the enterprise summary, there are three trending reports:

- Owner
- File System
- File Type Trending

These reports provide historical trending data on storage utilized by owner, file system growth, and storage utilized by file types.

Figure 12 shows a portion of the sample File Type Trending report; the historical data rate of change is dependent on the frequency of data collection. For example, if data is collected weekly, as is the case in the figure 12, there will be several days of 0% delta since no new data is being collected during the week. Therefore, you must consider the data collection schedule when interpreting trending data.



File Type: Backup Files

Date	# Files	Allocated Space (MB)	Delta %
2008-12-08	5,972	50,258.55	
2008-12-09	5,972	50,258.55	0.00
2008-12-10	5,972	50,258.55	0.00
2008-12-11	5,972	50,258.55	0.00
2008-12-12	5,972	50,258.55	0.00
2008-12-15	6,004	50,266.33	0.02
2008-12-16	6,004	50,266.33	0.00
2008-12-17	6,004	50,266.33	0.00
2008-12-18	6,004	50,266.33	0.00
2008-12-19	6,004	50,266.33	0.00
2008-12-22	8,051	186,469.09	270.96
2008-12-23	8,051	186,469.09	0.00
2008-12-24	8,051	186,469.09	0.00
2008-12-25	8,051	186,469.09	0.00
2008-12-26	8,051	186,469.09	0.00
2008-12-29	8,054	186,220.46	(0.13)
2008-12-30	8,054	186,220.46	0.00
2008-12-31	8,054	186,220.46	0.00
2009-01-01	8,054	186,220.46	0.00
2009-01-02	8,054	186,220.46	0.00
2009-01-08	6	0.25	(100.00)

Space consumed by the file type shrank from 50,258.55 MB to 0.25 MB in this period, for a total of -50,258.30 MB, or (100.00) %

There were 32 days in this period. Average growth was less than 1 MB per day.

Figure 12: File Type Trending

The “duplicate files” built in report does not report any data on the Network File systems. The stored procedures used for the report depend on value to exist within the host field; the host field is not populated for Network File systems.

StorageScope Queries

StorageScope queries can provide very specific data listing detailed information on files and folders that the built-in reports do not provide. The queries can show file type summaries, top users, and files of a specific age along with specific file names, owners and file systems. The queries are nice for ad hoc data gathering that does not need a formal report output.

In the following example, we used the StorageScope query builder to create a report of the top user folders. The folders are on File system “drive02” and in this particular case the user home folders are at a depth of “2”. The Folder table fields pulled are the Owner, Folder name, Folder Depth, and the Folder Actual Size. Figures 13 through 15 show the key steps and sample output from the query.

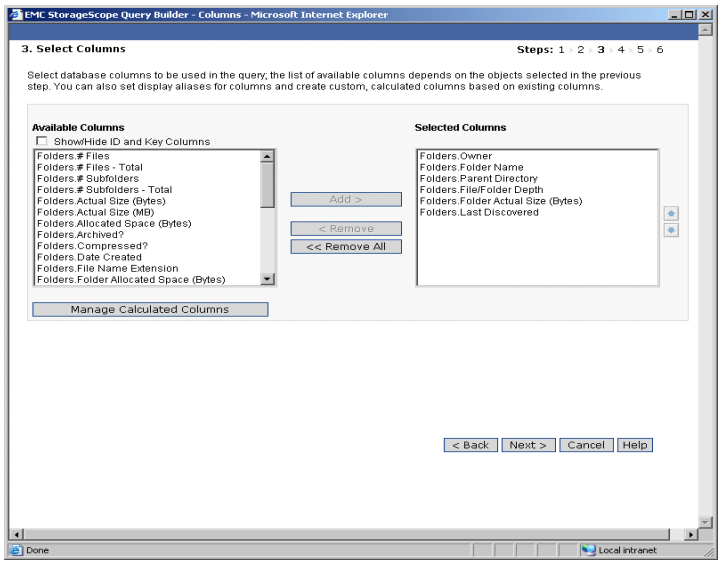


Figure 13: Folder columns used for report

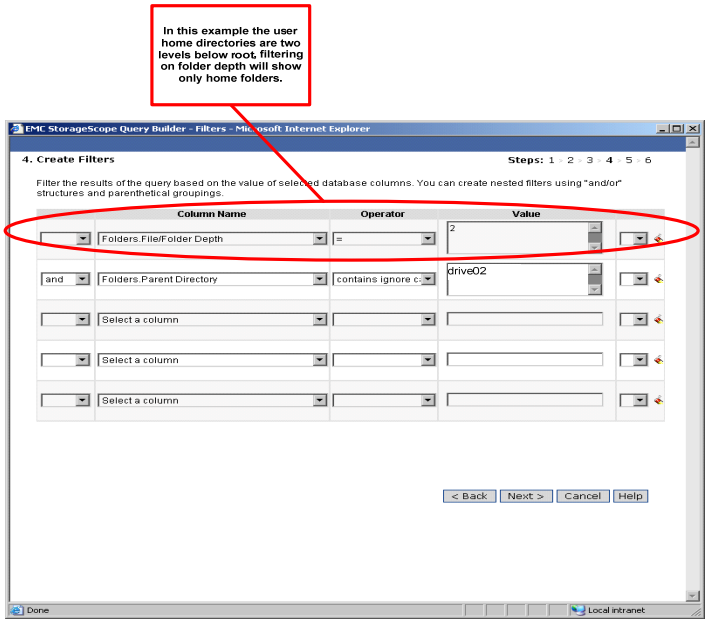


Figure 14: filtering for home directories

OWNER	PARENTDIR	FOLDERNAME	Folder TOTALACTUALSIZE(Bytes)	Folder TOTALACTUALSIZE(GB)
E1000	\\cifs01\drive02\Users\	E1000	15200280741	14
E1001	\\cifs01\drive02\Users\	E1001	12029984368	11
E1002	\\cifs01\drive02\Users\	E1002	11420982836	11
E1003	\\cifs01\drive02\Users\	E1003	11053530661	10
E1004	\\cifs01\drive02\Users\	E1004	10917038997	10
E1005	\\cifs01\drive02\Users\	E1005	10804079714	10
E1006	\\cifs01\drive02\Users\	E1006	9622873584	9
E1007	\\cifs01\drive02\Users\	E1007	9497003592	9
E1008	\\cifs01\drive02\Users\	E1008	8987231945	8
E1008	\\cifs01\drive02\Users\	E1008	8672733443	8
E1009	\\cifs01\drive02\Users\	E1009	8418624497	8
E1010	\\cifs01\drive02\Users\	E1010	8400378321	8
E1011	\\cifs01\drive02\Users\	E1011	8329826195	8

Figure 15: Sample output from top folder query

The StorageScope Query filter can be used to pull a quick list of files within a specific age range. In the case of figure16, the filter will pull files last modified within the dates specified for a specific file system. The limitation is that the dates are hard coded and are not useful for a periodic report.

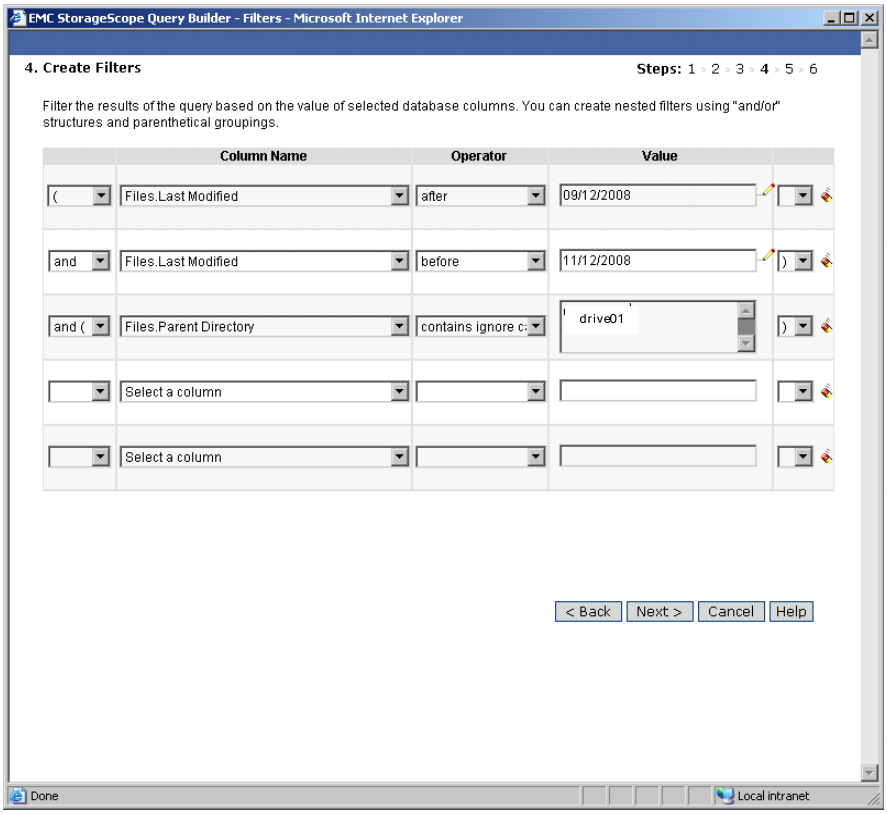


Figure 16: filter for file age selection

Crystal Reports

Crystal reports can provide very complex and customized reports in a structured format that StorageScope queries cannot provide. Crystal Reports provide some very useful date functions e.g. “aged31to60days” that can be used to categorize files by age without having to provide a specific date range as in StorageScope queries. Crystal Reports can also be imported into StorageScope and run automatically on a schedule.

> This article is not intended to be an overview of StorageScope. There is an informative whitepaper called “EMC ControlCenter 6.0 StorageScope Best Practices Planning” that provides more details on StorageScope.

Crystal reports running total field option combined with its ability to evaluate totals based on a specific formula makes it possible to create a file age distribution report spanning multiple file systems. Figure 17 shows an example of a Running Total Field in which the total number of files last accessed 31 to 60 days of the date the report is generated using the ACCESSTIME field from the SRMFILE table.

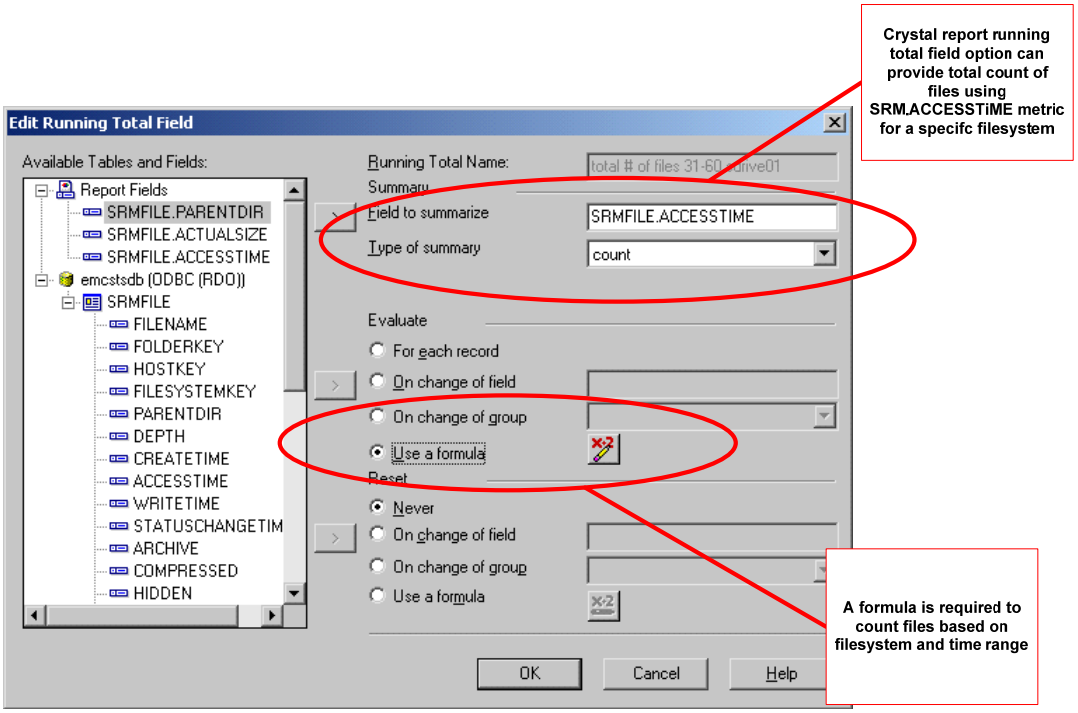


Figure17: Running Total Field

Figure 18 shows the evaluation formula for the running total. The expression **{SRMFILE.ACCESSTIME} = Aged31to60Days** will evaluate to true if the access time is within 31 to 60 days of the date the report is ran, and to count only files for a specific file system; a string expression: **"Drive01" in {SRMFILE.PARENTDIR}** is used.

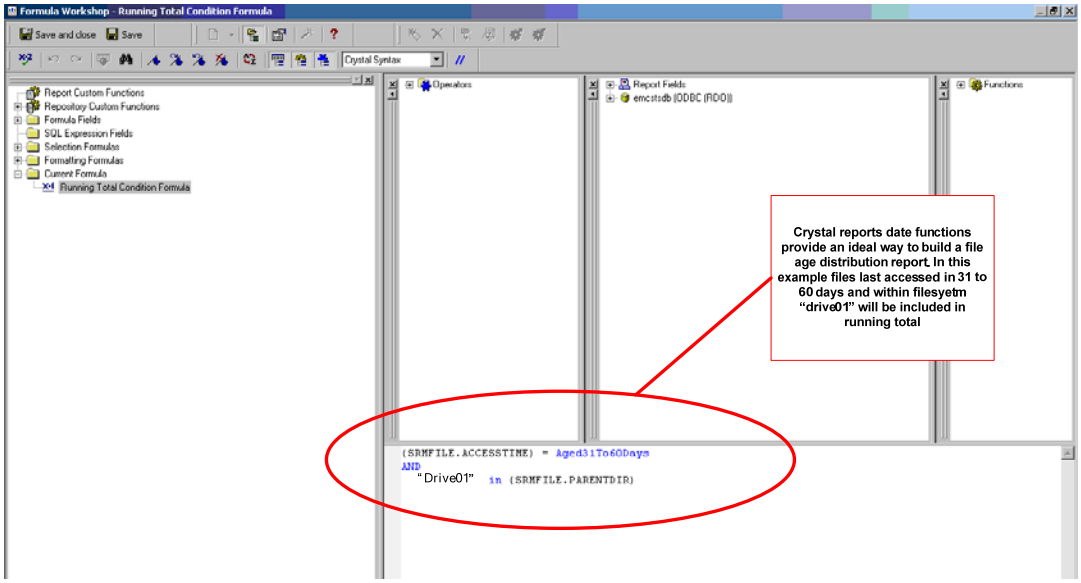
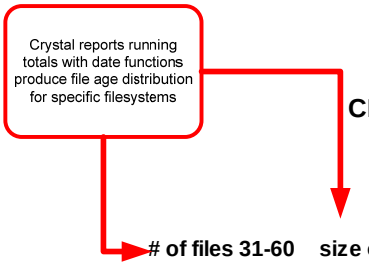


Figure 18: Evaluation Formula

The complete report requires the creation of running total fields for each file system, within each desired date range. The values in the sample report in Figure 19 are running total fields, one for counting each file within the specified date range, and one to calculate a sum of the actual file size within the specified date range.



CIFS File Age Summary

	# of files 31-60	size of files 31-60	# of files 61-90	size of files 61-90
drive02	9,132	35,534	9,540	22,007
drive03	15,678	64,450.10	19,408	54,464.03
drive04	13,809	39,512.31	16,209	38,075.06
drive05	3,093	30,796.39	3,999	15,770.13
drive06	585	4,994.06	816	16,685.31
drive08	3,259	21,911.88	2,546	9,811.56
drive09	1,373	4,863.96	1,269	7,175.26
drive10	1,778	5,852.10	2,471	20,713.30

Figure 19: sample report output

Conclusion

EMC ControlCenter's ability to collect file level data from a UNC can be a powerful tool for Celerra CIFS share data collection when compared to Windows native tools or scripting. That collection ability, when applied with best practices and combined with StorageScope's reporting features, can save significant time in providing CIFS utilization reports.

Biography