White Paper

Reclaiming Primary Storage with Managed Server HSM

November, 2013

RECLAIMING PRIMARY STORAGE

According to Forrester Research Inc., the total amount of data warehoused by enterprises is doubling every three years. It is also estimated that 80% of an organization's data is unstructured (Gartner 2010). Controlling the growth of unstructured data is a major challenge. This white paper describes how deployment of file system archiving with deduplication helps take control of data growth and reclaims the primary storage used by duplicate and rarely accessed files.

EXECUTIVE SUMMARY

It's happening with cell phones, email, text messages – even at home on our DVRs. At some level, we're all experiencing the storage explosion. Nowhere is it a bigger problem than in our corporate data centers. Expensive server storage is all too often expanded and replaced with bigger, faster and more costly technology – just to keep pace with the onslaught of new, incoming data. It's common knowledge that disk drives have been getting bigger, faster and seemingly less expensive (per GB). But the comfort in this thought is illusory. You can't just consider the cost of adding more hard drives each time your storage starts getting full. You have to factor in the costs of the resources it takes to deploy, provide power, manage and protect the added storage needed to accommodate the ever-increasing amounts of data. These are the TRUE COSTS of storage.

There are many approaches to solving this problem. For example, some storage vendors offer new, improved technologies such as "scale out" storage which promises to make it simpler to purchase and add more of their hardware. Often, these solutions come with integrated compression and deduplication which helps to squeeze the same amounts of data into less physical space. But how seamlessly do these work with backup, replication and frequent accesses to these files? Is throwing more storage (and dollars) at the problem always the best solution?

This white paper describes a software solution for managing the storage explosion that doesn't require the purchase of new primary storage hardware. Instead, it will describe how companies can continue utilizing existing primary storage and postpone expansion or replacement by archiving inactive and duplicate files – making room for new, active content. Because these inactive and duplicate files are largely removed from primary storage, the benefits go end-to-end; prolonging the useful life of existing storage resources, improving backup, recovery and replication performance and reducing media usage. It will be shown how CaminoSoft's Managed Server HSM file system archiving software provides the foundation for sustainable, efficient management of future data growth.

THE GROWTH OF STORAGE AND ITS IMPACT

It has been said that the world is drowning in data. It's estimated that, each day, 15 petabytes (that's the equivalent of 15,000 one-terabyte drives) of new information is generated. There's little wonder then that corporations are looking to store less data, place it where it's served best and make the best use of their existing assets. The research firm IDC said in 2010 that the world's data already exceeds available storage space. And, according to Forrester Research Inc., the total amount of data warehoused by enterprises is doubling every three years.

It's true that storage innovators are doing their best to keep up; providing solutions to store more and more data. But, the data keeps on coming – outpacing their efforts to provide any sustainable solution. Technical teams around the globe strive to come up with the right mix of solutions that will solve their companies' data problem and – in fact – there may not be any one solution that works for everyone. But, one thing is for sure – lack of action is not an option for any CIO. The problem is not going away and data growth will only continue to increase.

CHALLENGES WITH TRADITIONAL INFRASTRUCTURES

With traditional IT infrastructures, meeting the needs of rapid expansion of storage capacity can be challenging and expensive. End-user-focused applications such as email, productivity tools like Microsoft Office, graphics, multimedia, engineering drawings and countless other data-generating applications are becoming more numerous and complex. As a result, storage demand continues to explode. Sharing many of these files with team members and friends along with saving "copies" here and there only compound the problem by creating duplicate content. These "unstructured" files are now estimated to represent about 80% of all storage used – and they're growing in numbers and in size each and every day.

The simple answer has always been to add more disk space. But, adding more disk space can often be anything but "simple".

THE CHALLENGE OF STORAGE SCALABILITY

Even the more sophisticated storage infrastructures don't always scale up efficiently. Drive densities continue to increase (more data in less space), racks and enclosures still fill up eventually; which creates the need for more racks, more enclosures and more drives. Even in the best of scenarios, storage expansion often means rip-and-replace of hardware and migration of data from old to new.

Storage hardware vendors are convinced they have the answer to the storage explosion. "Let's make it easier for our customers to add more storage to their environments." After all, they are in the business of selling storage.

Simply throwing more hardware at the problem (whether in the form of traditional storage upgrades or more sophisticated scale-out environments) fails to address some of the core issues and costs:

- High cost of new storage
- Rack space for more servers and storage
- Power consumption
- Backup and (more importantly) recovery time along with media consumed
- Replication time and replica storage consumed
- Management of added storage resources
- Sustainability – added storage will eventually fill up again

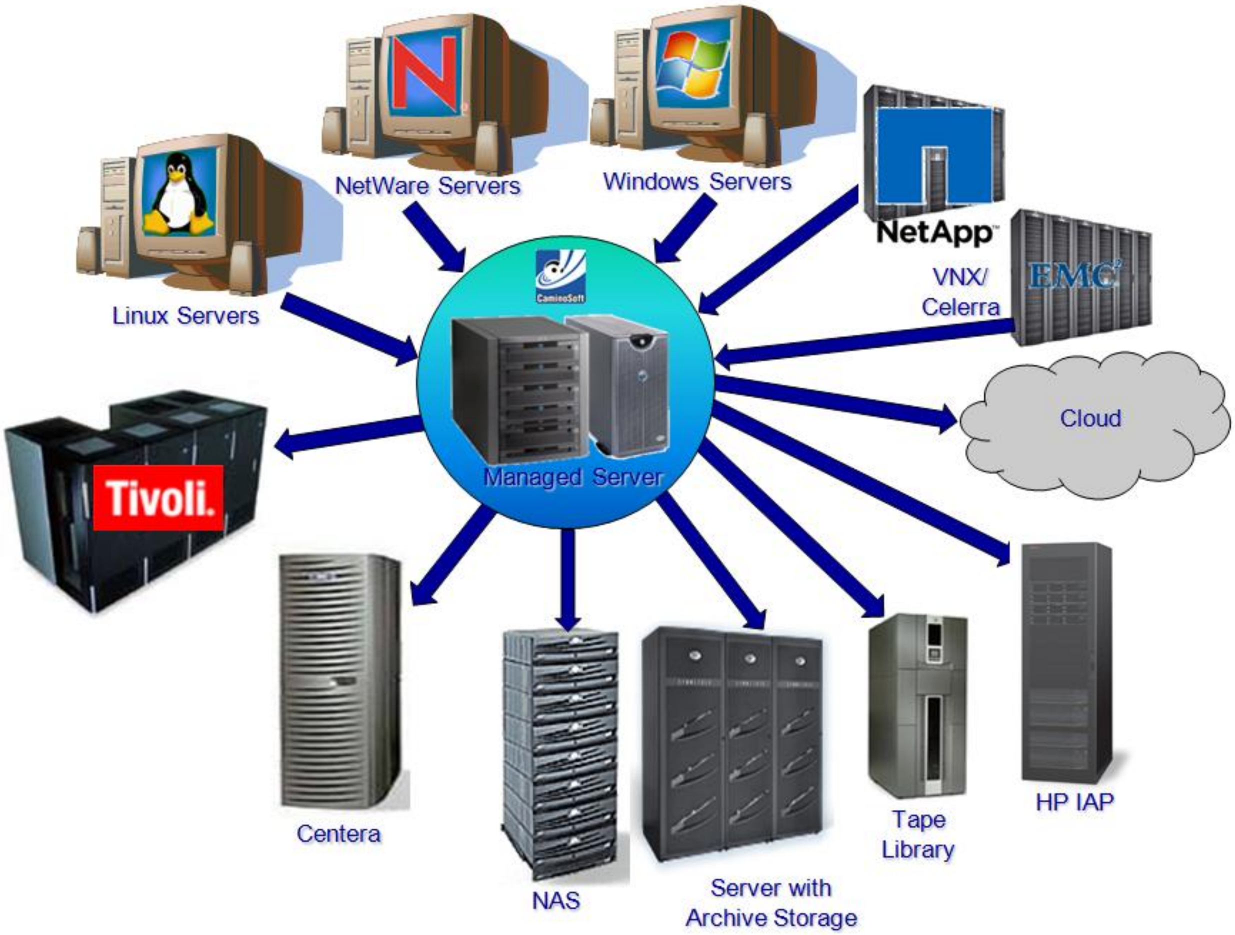
So, wouldn't a better answer to the problem be, "Let's slow down the growth of primary storage and get more use out of the resources we already have?" By laying the groundwork for a sustainable solution that actively manages storage consumption, costly hardware upgrades, replacements and platform migrations can be postponed and planned for; not as a knee-jerk reaction to "running out of disk space" but as actions carefully considered and justified just as any other capital investment would be.

CAMINOSOFT MANAGED SERVER HSM

CaminoSoft's Managed Server HSM (Managed Server) software provides a solution that helps companies take control of their storage growth. Managed Server enables storage administrators to immediately recover lost storage space, slow down the growth of precious primary storage and provide the foundation for sustainable, efficient management of future data growth.

HOW IT WORKS

Managed Server enables administrators to configure and utilize policies to archive seldom accessed and duplicate files. By constantly monitoring the file system, Managed Server persistently identifies duplicate and inactive files and transparently archives them from expensive, primary server storage to less expensive or special-purpose storage (low cost NAS, archive appliances, tape libraries, the cloud, etc.). The archived files are replaced by small stub files. So, users continue to locate and access their files from the same folders in which the files were originally saved. If and when the stub files are opened by users or applications, Managed Server automatically and seamlessly recalls the archived file content. The recall process is completely transparent to users.



INSTALLATION

The Managed Server installation is broken into two segments: Engine and Management Console Graphical User Interface (GUI).

The Managed Server engine is installed by running the appropriate setup.exe program from the installation media. Customers choose between 32- bit and 64-bit implementations. Both are provided on the installation media.

The GUI (used for system administration) can be installed on the same server as the engine and on any supported workstation (XP, Vista, Windows 7, etc.). To install the GUI, navigate to the GUI folder on the installation media and run the setup.exe from that folder. Simply follow the steps in the wizard to install the program.

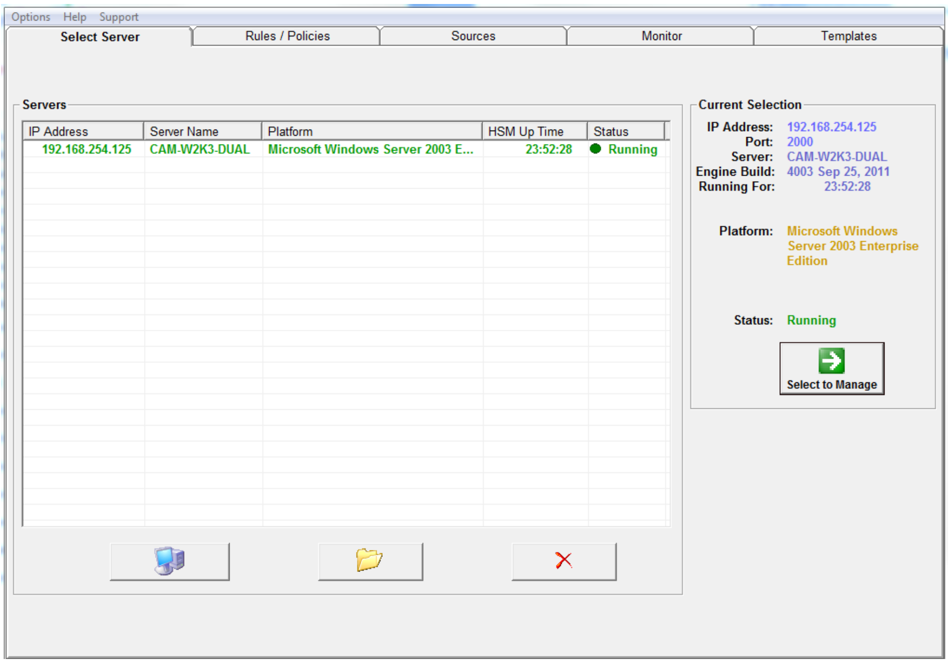
When you install Managed Server, the functionality of the complete Managed Server product (policies for file archiving, deduplication and deletion) are enabled for trial purposes.

MANAGED SERVER ENGINE

The Managed Server Engine is the engine component that performs the archival and recall of files. For Windows servers, the engine is installed on each server where files to be managed reside. For OES Linux environments, Managed Server is installed on a Windows application server and CaminoSoft OES Linux agents are installed on the Linux servers. For NetApp and IBM N series environments, Managed Server is installed on a Windows application server and the NetApp/IBM N series filer is configured to communicate with the application server.

MANAGEMENT CONSOLE (GUI)

The Managed Server Administrator GUI allows you to define, manage, and monitor your archiving storage environment (file system sources and target archive storage).



Servers

IP Address	Server Name	Platform	HSM Up Time	Status
192.168.254.125	CAM-W2K3-DUAL	Microsoft Windows Server 2003 E...	23:52:28	● Running

Current Selection

IP Address: 192.168.254.125
Port: 2000
Server: CAM-W2K3-DUAL
Engine Build: 4003 Sep 25, 2011
Running For: 23:52:28

Platform: Microsoft Windows Server 2003 Enterprise Edition

Status: Running

The interface is comprised of the following tabs:

Select Server
Use the Select Server tab to select and manage servers.

Rules/Policies
Use the Rules/Policies tab to manage archiving schedules, volume thresholds, and other global policies.

Sources
Use the Sources tab to manage archive and deduplication profiles. In the profiles, administrators dictate the rules to be followed in identifying candidate files for archiving and/or deduplication. This includes the rules for the location of files to be managed, how old they have to be to be considered “inactive” and the target storage where the inactive and duplicate file content will be stored when archived.

Monitor
Use the Monitor tab to view activity logs and statistics including files archived , files recalled and exceptions/errors.

Templates
Use the Templates tab to view and manage templates for volume/share threshold rules, schedule and global policies, and archive and deletion profiles. Templates can be useful in larger environments where administrators wish to clone rules and policies from one server to other similarly configured servers.

HOW ARCHIVING/DEDUPLICATION PROFILES WORK

Rules determine which files are candidates for deletion, deduplication, archiving or re-archiving. These rules can be summarized as follows:

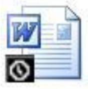
- By default, files that have not been backed up are not archived. If necessary, you can change this setting on the Rules/Policies tab.
- For archiving rules, files in managed folders that reach a certain age are candidates for archiving. Administrators can specify this age in days, and can also specify the basis for determining the age of a file (by creation date, last access date, or last modified date).
- For deduplication rules, files in managed folders that are duplicates of other files are candidates for archiving without regard for their ages.
- File archiving can be restricted to files that meet defined minimum KB and maximum MB size limits. The default minimum file size is 10 KB and the default maximum file size is 100 MB.
- Routine file archiving can be restricted to specific times of day so that archiving does not interfere with production activity, backups, or other critical activities. Emergency file archiving can take place any time. In addition, you can specify a blackout period as a window of no activity, so that no archiving of any kind occurs during that time.

HOW THE ARCHIVING PROCESS WORKS

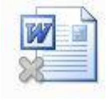
Managed Server monitors the status of managed volumes on production servers. Based on rules that administrators define and maintain using the Administrator GUI, seldom-used data files (and/or duplicate files identified in a deduplication profile) are archived from the production servers to the designated target archive storage when they meet the rule set criteria.

The following points describe how the archiving process works:

- The Engine runs in the background on each installed production server, and constantly monitors assigned storage resources.
- When critical volume thresholds are reached, Managed Server quickly scans the managed volumes and determines candidate files—files that meet defined criteria such as access date and size—and archives the identified files to target archive storage where they remain available for recall at any time. For deduplication profiles, scans looking for duplicate files are not dependent on volume thresholds and, instead, are launched according to schedules set by the administrator.
- Managed Server updates the directory of each relocated file with a archive key (stub) that identifies the file as archived and deletes the data portion of the file. This is how more room is made available for the working set of data files.
- When a file is archived, its content is stored in the specified target archive storage.
- The user will see the icon for the archived file in its original location on the managed server. The icon of an archived (stub) file includes a small black clock face (or X in the case of Windows 7) overlaying the icon to indicate that the file was archived, as shown in the following sample icons of an archived file:



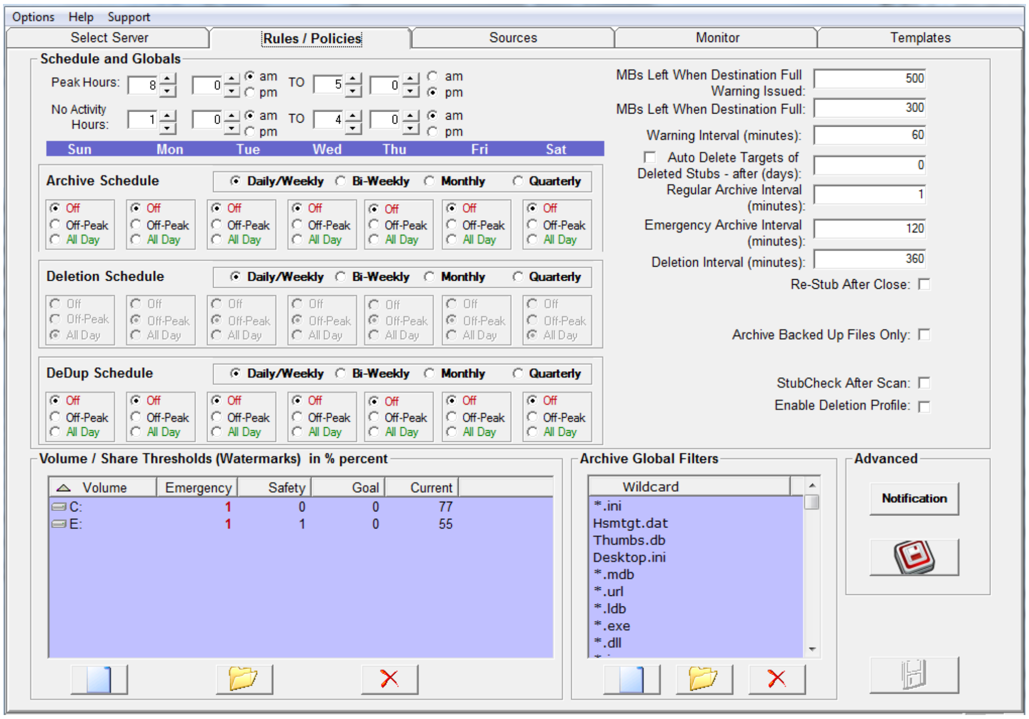
or



- When a request is made to open an archived file, Managed Server intercepts the open call and uses the file's archive key to quickly and transparently retrieve the file from the target archive storage. The file is placed back in its original location and is released to the application or user that requested it. Other than a possible slight pause while the file is retrieved, the user is totally unaware of any of the processes taking place. As far as the user is concerned, the file was never moved.
- One or more production servers can share the same target archive storage.

SETTING GLOBAL RULES AND POLICIES

The administrator determines the when, what and how of file management. First, the schedules are established to determine when Managed Server will scan the file system. The scheduling allows the flexibility to run scans in off-hours when servers are least busy. Next, the administrator specifies what volumes are to be managed (local volumes or volumes on configured Linux or NetApp servers) and their thresholds (levels of “fullness” where action needs to be taken). Administrators can specify global filters which is a list of file types that are to be avoided by Managed Server. Lastly, several options are provided for warnings (target storage low or out of space), notifications, deletion management (housekeeping for deleted files), scan intervals and others.



Schedule and Globals

Peak Hours: 8 : 0 ● am ○ pm TO 5 : 0 ○ am ● pm

No Activity Hours: 1 : 0 ● am ○ pm TO 4 : 0 ● am ○ pm

Archive Schedule: ● Daily/Weekly ○ Bi-Weekly ○ Monthly ○ Quarterly

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Off	●	●	●	●	●	●	●
Off-Peak	○	○	○	○	○	○	○
All Day	○	○	○	○	○	○	○

Deletion Schedule: ● Daily/Weekly ○ Bi-Weekly ○ Monthly ○ Quarterly

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Off	○	○	○	○	○	○	○
Off-Peak	○	●	●	●	●	●	○
All Day	●	○	○	○	○	○	●

DeDup Schedule: ● Daily/Weekly ○ Bi-Weekly ○ Monthly ○ Quarterly

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Off	●	●	●	●	●	●	●
Off-Peak	○	○	○	○	○	○	○
All Day	○	○	○	○	○	○	○

MBs Left When Destination Full Warning Issued: 500

MBs Left When Destination Full: 300

Warning Interval (minutes): 60

□ Auto Delete Targets of Deleted Stubs - after (days): 0

Regular Archive Interval (minutes): 1

Emergency Archive Interval (minutes): 120

Deletion Interval (minutes): 360

Re-Stub After Close: □

Archive Backed Up Files Only: □

StubCheck After Scan: □

Enable Deletion Profile: □

Volume / Share Thresholds (Watermarks) in % percent

Volume	Emergency	Safety	Goal	Current
C:	1	0	0	77
E:	1	1	0	55

Archive Global Filters

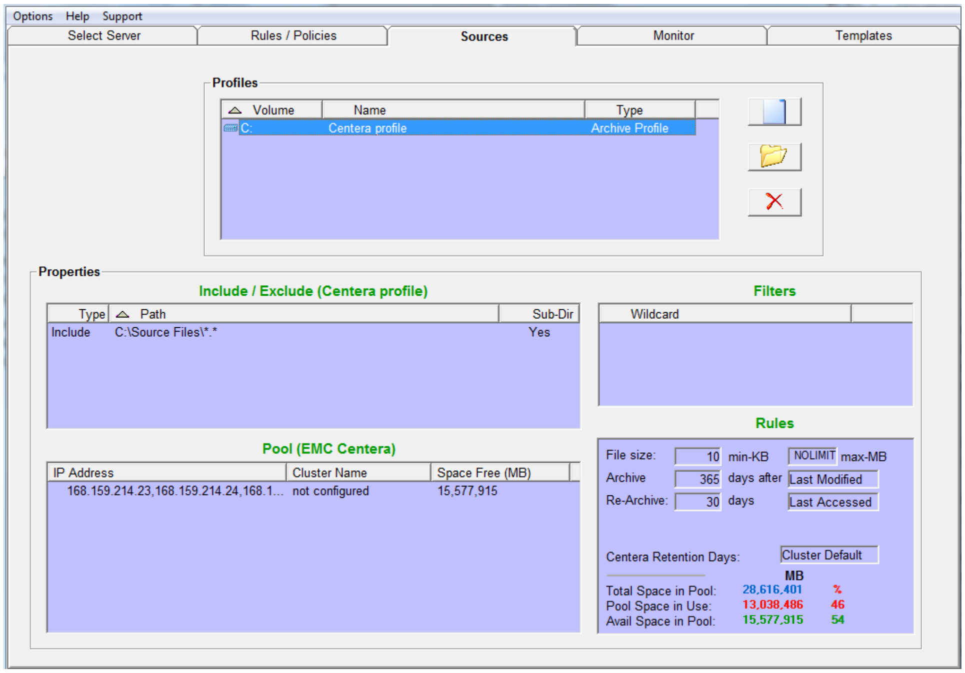
Wildcard
*.ini
Hsmtgt.dat
Thumbs.db
Desktop.ini
*.mdb
*.url
*.ldb
*.exe
*.dll

Advanced

DEFINING THE SOURCES, TARGETS AND RULES

Administrators use the SOURCES tab of the GUI to create profiles which determine which files are to be managed, the rules to use for selecting candidate files for archiving and the target archive storage to use when files meet the rules criteria. There are 3 types of profiles available.

- **Archive Profile** – This is the most common and includes the locations of the source files, rules and target archive storage definitions for files that are archived on the basis of file age.
- **Deduplication Profile** – This includes the locations of the source files, rules and target archive storage definitions for files that are analyzed for duplicates. Files need not have the same names or be in the same volumes or even on the same servers to be identified as duplicates.
- **Deletion Profile** – This is available but not commonly used. It includes the locations of the source files and rules to use in determining files that are to be automatically deleted by Managed Server. These deletions can be configured for certain types of files (video, music, etc.) or files that have simply become too old to retain.



Volume	Name	Type
C:	Centera profile	Archive Profile

Type	Path	Sub-Dir
Include	C:\Source Files*.*	Yes

Wildcard

IP Address	Cluster Name	Space Free (MB)
168.159.214.23,168.159.214.24,168.1...	not configured	15,577,915

File size: 10 min-KB NOLIMIT max-MB
Archive 365 days after Last Modified
Re-Archive: 30 days Last Accessed

Centera Retention Days: Cluster Default

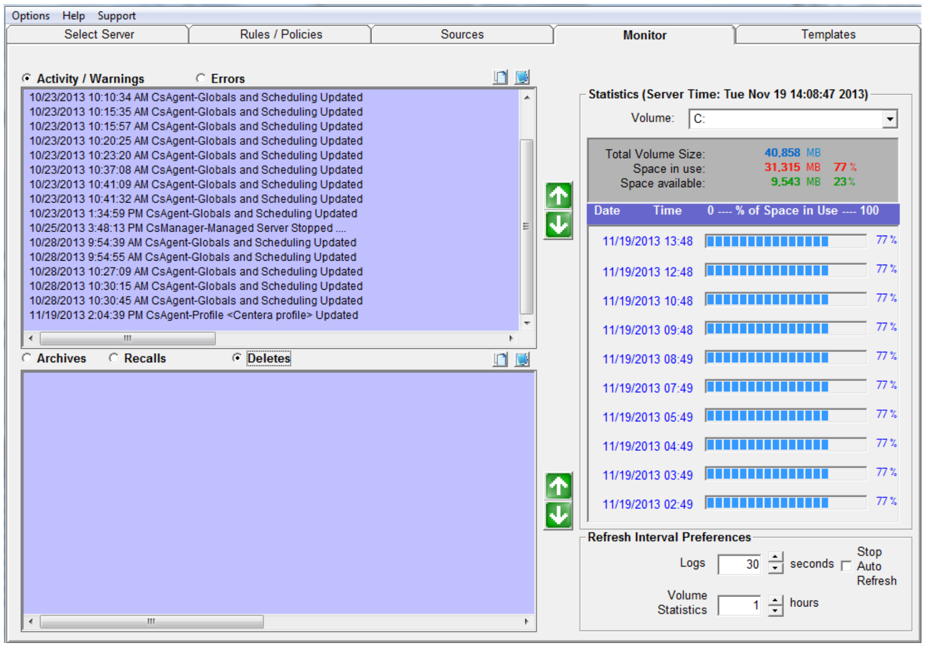
	MB	%
Total Space in Pool:	28,616,401	
Pool Space in Use:	13,038,486	46
Avail Space in Pool:	15,577,915	54

-Sample Archive Profile-

MONITORING ACTIVITY

The monitor feature provides detailed information about the running processes by allowing real-time access to the logs on the Monitor tab. There are also statistics for a selected volume and the ability to set preferences for the displayed logs and statistics.

From the monitor screen, administrators can monitor general activity and critical events, archived files, recalled files, exceptions/errors and track volume usage. There are even tools provided to select sections of the logs to be saved to text files or placed on the clipboard.



● Activity / Warnings ○ Errors

```
10/23/2013 10:10:34 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:15:35 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:15:57 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:20:25 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:23:20 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:37:08 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:41:09 AM CsAgent-Globals and Scheduling Updated
10/23/2013 10:41:32 AM CsAgent-Globals and Scheduling Updated
10/23/2013 1:34:59 PM CsAgent-Globals and Scheduling Updated
10/25/2013 3:48:13 PM CsManager-Managed Server Stopped ....
10/28/2013 9:54:39 AM CsAgent-Globals and Scheduling Updated
10/28/2013 9:54:55 AM CsAgent-Globals and Scheduling Updated
10/28/2013 10:27:09 AM CsAgent-Globals and Scheduling Updated
10/28/2013 10:30:15 AM CsAgent-Globals and Scheduling Updated
10/28/2013 10:30:45 AM CsAgent-Globals and Scheduling Updated
11/19/2013 2:04:39 PM CsAgent-Profile <Centera profile> Updated
```

○ Archives ○ Recalls ● Deletes

Statistics (Server Time: Tue Nov 19 14:08:47 2013)

Volume: C:

Total Volume Size: 40,858 MB
Space in use: 31,315 MB 77 %
Space available: 9,543 MB 23 %

Date Time	0 ---- % of Space in Use ---- 100
11/19/2013 13:48	77 %
11/19/2013 12:48	77 %
11/19/2013 10:48	77 %
11/19/2013 09:48	77 %
11/19/2013 08:49	77 %
11/19/2013 07:49	77 %
11/19/2013 05:49	77 %
11/19/2013 04:49	77 %
11/19/2013 03:49	77 %
11/19/2013 02:49	77 %

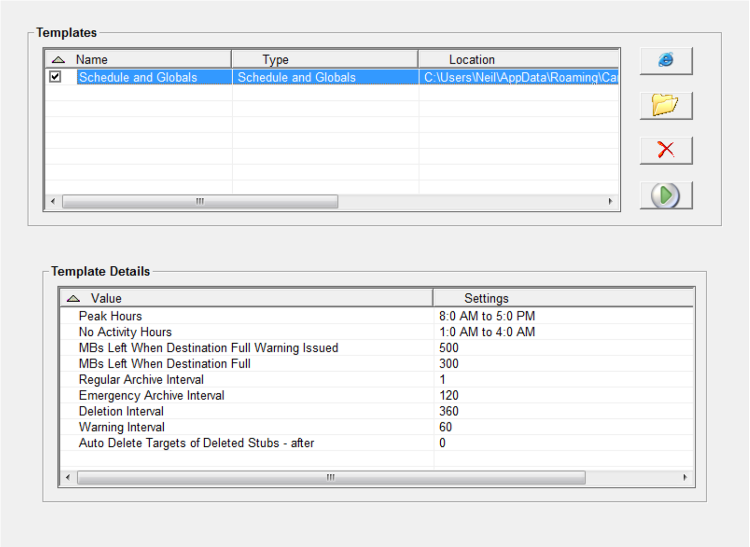
Refresh Interval Preferences

Logs 30 seconds □ Stop Auto Refresh

Volume Statistics 1 hours

TEMPLATES

Templates are settings for managing archiving that can be saved and applied to more than one available managed server. Templates can be generated for volume/share threshold rules, schedule and global policies - and archive, deduplication and deletion profiles.



Templates

	Name	Type	Location
☑	Schedule and Globals	Schedule and Globals	C:\Users\Neil\AppData\Roaming\Ca

Template Details

Value	Settings
Peak Hours	8:0 AM to 5:0 PM
No Activity Hours	1:0 AM to 4:0 AM
MBs Left When Destination Full Warning Issued	500
MBs Left When Destination Full	300
Regular Archive Interval	1
Emergency Archive Interval	120
Deletion Interval	360
Warning Interval	60
Auto Delete Targets of Deleted Stubs - after	0

Once Managed Server has been deployed and the rules and policies configured, there is little need for day-to-day administrator intervention. Archiving scans and file recalls occur according to schedules and automatically as needed, respectively, and notifications can be configured to alert administrators when exceptions occur.

SUMMARY

Enterprise servers are drowning in data – and it's doubling every 3 years. Many believe that continuing to add ever-increasing amounts of primary storage is the only way to keep up with the storage explosion. And, the storage hardware vendors are only too happy to accommodate. The trouble is that throwing more hardware at the problem is simply not sustainable. It's only a temporary fix.

CaminoSoft's Managed Server HSM software offers a solution that:

- Does NOT require the purchase of additional or replacement primary storage hardware.
- Is simple to install, configure and manage.
- Enables storage administrators to immediately recover lost storage space (frequently in excess of 50%) by archiving seldom accessed and duplicate files.
- Helps to slow down the growth of primary storage usage by persistently identifying files and archiving them to target archive storage – assuring that ONLY active files remain in costly primary storage.
- Reduces backup and recovery time and media used.
- Provides the foundation for sustainable, efficient management of future data growth.

REFERENCES

1. Wikipedia: Unstructured Data

http://en.wikipedia.org/wiki/Unstructured_data

2. CIOs Struggling With Data Growth - Doug Balog

http://www.enterprisestorageforum.com/management/features/article.php/3911686/CIOs-Struggling-With-Data-Growth.htm