



1 Introduction

MadWolf Technologies (MWT), through our partner agreements, offers Archive Series Software that enables an Administrator to define file management policies that allow:

- Groups of files to be stored on defined sets of tape cartridges (termed 'volume sets');
- Groups of files to be retained on RAID for predefined periods;
- The option to choose not to version control some file groups;
- The option to choose not to write some file groups to tape; and
- Automatic generation of tape replicas for off-site retention.

This functionality is set up by the Administrator using a Management Console (XMC) to define file groups and volume sets. The XMC is a Microsoft Management Console snap-in and provides a familiar type of interface to an Administrator.

Understanding the concepts of file groups and volume sets is important for successfully implementing the Archive Series data storage solution and these are described in sections 2 and 3 below.

When files are archived to tape, all file versions are retained including deleted files. The concept of file versions and the system's treatment of deleted files are described in section 5.

The implementation of file fragmentation is important for efficient use of storage capacity and fast retrieval of data from large files. This concept is explained in section 6.

The Archive Series Software implements standard Microsoft file security. When files have been deleted or have multiple versions, security concepts are extended and these are described in section 7.

A server running Archive Series software can be used for both archiving of fixed content data and backup of more transient data, as described in section 8.

2 File Groups

A file group is a collection of files that have the same file management policy and consequently are all treated in the same way by the system. Every file that is handled by the system belongs to exactly one file group. Files are assigned to a file group on the basis of their names. This assignment can be based on the name of the directory that contains a file, the name of the file or both. For example, files that have names of the

form *.tmp could be assigned to one file group; files in the directory \data\2001\January\ could be in a different group and files that match the pattern \data\2000\...\aaa???.dat could form a third group.

File group characteristics are defined in the XMC. The following parameters must be defined for each file group:

- File name or path pattern for the file group
- The position of the file group relative to other file groups, which defines the order in which the file name or path pattern selection is performed relative to selection by other file groups
- One of the following storage options:
 - Save all versions to tape and RAID;
 - Save to all versions to tape and latest version to RAID;
 - Save all versions on RAID but do not save to tape; or
 - Do not save on tape and do not save old versions on RAID
- Selection of a defined set of tapes (termed a 'volume set'), if the file group is archived to tape
- Retention periods on RAID, if the file group is archived to tape

The policies defined by the Administrator determine how the system moves files between RAID and tape. The policies do not affect how files appear in the directory listing. For example, files cannot be deleted from the directory listing by using file group rules; file deletion is performed by using standard utilities like Windows Explorer.

3 Volume Sets and Automatic Replication

A volume set consists of a set of tapes with a defined number of replicas that store data from one or more file groups. If replication is not enabled for a volume set, each file is archived to only one tape. Alternatively, if replication is enabled, one or more additional copies of each tape are automatically generated.

The data on replicated tapes in a volume set are kept synchronized whenever the tapes are available to the system. If one or more tapes in a volume set are removed from the library, the system maintains a record of which files need to be written to those tapes. When tapes are reintroduced into the library, the data on them is automatically brought up to date.

With the exception of the blank media set, all media within a volume set must be either WORM or re-writable. Tape replicas must all be of the same capacity.

One special volume set, termed the blank media set, contains all the tapes that are known to the system but are not formatted for storing data. These may be new (unused) tapes, tapes that are unrecognized by the system or re-writable tapes that have been

reformatted by the Administrator. Unlike any other volume set, the blank media set can contain media of different type, i.e. WORM and re-writable.

As more data is written to a volume set, the tapes will eventually become full. At a preset threshold, defined by the Administrator as a percentage of full, the system will automatically take the appropriate number of media from the set of blank tapes and will extend the volume set. If insufficient tapes of the right type are available in the blank media set to create all required replicas, the system will report an error and when the tapes become full, no more data will be accepted from users. The system will then report "disk full" when a user tries to write to the volume set.

Following configuration of the file groups, volume sets and any associated replication requirements, the system operates completely automatically. Files written to the volume (logical drive letter) under the Archive Series control are automatically allocated to file groups. Files allocated to file groups with 'save to tape' enabled have an assigned volume set and are automatically written to both RAID and tape. If replication is enabled for the assigned volume set, this occurs automatically.

4 More on ILM and Records Management

In many industries, files must be retained for defined periods. Often at the end of the retention period, files will be destroyed. The Archive Series Software produces a controlled number of file copies, unlike backup software, and consequently file destruction can occur in a well-controlled fashion.

In most cases, the ability to create multiple sets of tapes (volume sets) allows file management to be aligned with the planned destruction of data. This is illustrated by an example from the US banking industry where check images must be retained for 7 years. For this case the administrator might set up policies whereby:

- Check images created in 2002 are written to a file group that saves the files to a volume set that is due for destruction at the end of 2009
- Check images created in 2003 are written to another file group that saves to a different volume set that is due for destruction at the end of 2010
- Etc.

In addition to destruction of data on tape, the administrator can ensure that the files are removed from RAID by deleting them over the network, having first set a policy parameter that ensures non-current files are flushed from RAID.

At the time of scheduled destruction of a volume set, there may be some files that the organization must keep beyond the minimum retention period. Perhaps these 'exception files' are the subject of a subpoena or ongoing litigation. In this case, the exception files can simply be copied to another file group that writes to another volume set that is not scheduled for destruction until a later date.

5 Version Numbers

The History Explorer provides access to current file versions, old file versions and deleted files.

If a file is updated with a newer version by overwriting or appending, the software assigns a new version number. A file's version number increases by one every time a file has data written to it. Note that the version number does not increase for every individual write operation, just for every file open that is followed by a write. Version 0 of a file never contains any data; the first time an application writes to the file, the version number is incremented to 1.

If a file is deleted and then a new file of the same name is created, the system starts again with version 1 of the new file. The deleted file and its prior versions may be accessed by enabling "Show deleted files" within the History Explorer.

6 File Fragmentation

The term "fragmentation" refers to the way in which computer systems break large files into smaller, more manageable units for transfer to or from storage devices. With hard disks, fragmentation is required because gaps are created when files are deleted. For a hard disk, fragmentation leads to performance degradation that can be corrected using de-fragmentation utilities. Tape drives do not inherently suffer from file fragmentation problems because they use a medium that is recorded linearly from one end to the other, with individual files recorded as complete entities. However with large files on tape, there are occasions when controlled file fragmentation can lead to significant performance improvements. The Archive Series software allows the Administrator to enable file fragmentation for any file group.

Disabling file fragmentation is preferred for file groups that predominantly consist of small files, less than a few tens of megabytes in size. If file fragmentation is not enabled, the system has the following characteristics:

- When a file is modified, the new version of the file will be completely written to tape; and
- When a file or portion of a file is read from tape, the complete file will be read.

Enabling file fragmentation typically provides benefits for file groups that predominantly consist of large files, greater than a few tens of megabytes in size. If file fragmentation is enabled, the system has the following characteristics:

- When an application modifies a large file by appending, the appended data will be written to tape as one or more additional fragments. Unchanged fragments will not be rewritten, saving time and space on the tape cartridge.
- If an application modifies a small part of a large file, for example by updating an index at the beginning of the file, then only the fragments containing modified data will be written to tape; and
- When a portion of a file is read from tape, only the applicable fragments will be read, saving both transfer time and space on the RAID cache.

If the Administrator enables file fragmentation for a file group, the fragment size must be defined. Recommended fragment sizes depend on the application but will typically be around 1MB or 1% of the average file size, whichever is larger.

7 File Security

Archive Series Software integrates fully with the Microsoft Windows security model, based on Active Directory. Files and directories have user-definable security attributes just as they do with standard Microsoft file systems and access control checks are performed in the same way. The security model is extended to deleted files and old versions of files made available to users via the History Explorer, or another application written using the our Archive Series API. In these cases, the security applied to deleted files or directories or old versions of files is the same as that applied to the most recent version, regardless of the security applied when the old version was originally in use. This feature allows the Administrator to update access controls for old files based on changing business requirements.

The Archive Series API, combined with MadWolf Technologies knowledge of use of the Domino.Doc API set, offers a viable, secured document management solution, based on sound best practices, file retention and archive policies.

8 Combining Archive and Backup

The Archive Series is designed for archiving applications but it also supports leading backup software, allowing the tape library to be used for back up in addition to archiving. This makes for a very cost effective combined solution.

Backup software supported includes Microsoft Windows Backup, Brightstor ARCserve and Veritas Backup Exec. Typically WORM tapes are used for data archive and re-writable cartridges are employed for backup.