

File Archiving Solutions

A Guide To Managing The Growth Of Unstructured Data

Executive Summary

Organisations of all sizes are facing accelerated growth in the amount of unstructured, file-based information they have to store. IT research houses all agree that the burden this places on primary disk storage arrays is:-

- Becoming more and more excessive leading to greater costs,
- An increasing IT administration burden,
- Inefficient storage utilisation,
- Difficulties in obeying legal and regulatory frameworks regarding information retention and deletion,
- Disk space wasted through file duplication.
- Information becoming harder to find.

The answer to this set of problems is to recognise that most unstructured information is not needed for immediate access. It can safely be moved to less expensive, capacity-centric storage arrays and amalgamated in a single logical file archive that provides:-

- Cost-savings:
 - Using very much more cost-effective storage for older, less active files,
 - Faster backup of primary data with lower space needs,
 - Ability to have several archival storage tiers to match storage cost to information value
 - Flexibility to use any storage supply source you wish
 - Removing duplicated files,
 - Simplified file management
- Security and compliance
 - Automated file retention and deletion
 - Reporting facilities to demonstrate compliance
- Improved management
 - IT department set policies for file deletion and migration
 - IT department set policies for file retention
- User-transparency
 - File-access paths preserved for applications and users
 - Better file revision management facilities for users

A file archive system can enable you to manage the seemingly uncontrollable onrush of unstructured information that needs to be stored and impose sensible and practical and automated management processes on it to save you money and your users' time.

Introduction

Organisations store more and more files. This near ceaseless growth in file storage is putting growing pressure on Windows servers as their hard drives fill up. It takes longer and longer to backup file server contents and the administration of tens if not hundreds of thousands of files, even millions in larger organisations, is an increasing burden on IT departments.

An executive or manager's office can be characterised as having two storage resources for paper-based information: the desk with its working surface and In, Out, and Active trays on the one hand, and the filing cabinets for less often needed but still desired reference information. This is a two-tier scheme and it makes good sense so as not to overload the desktop and its file trays with old information that gets in the way of dealing with the newer and more active information.

This is not the case with digitally-stored information though. Expensive online storage is being used to store both current files which are often accessed, and also less active files which may only be accessed once a month or less. This growing file estate requires more and more disk capacity to be purchased.

By archiving less active files to secondary storage, less expensive than frontline drive arrays, capacity purchase can be deferred and file server backup completes faster. The file servers themselves may well perform their work faster as well.

Yet not all file archival products are equal for business use. This Waterford Technologies white paper will help you understand what to look for in a file archival solution and make a good choice for your business.

Storage Market/File Archiving market today

File data is often described as unstructured or semi-structured data in contrast to highly structured data base records. It can be viewed as all user-generated and non-database information stored on servers and their disk drive arrays, comprising Word documents, spreadsheets, presentations with slides and graphics, PDF documents, scanned images and documents, and also e-mails.

E-mail archiving and file archiving are two distinct sides of the same coin and e-mail archiving has its own unique requirements¹

¹Waterford Technology's MailMeter addresses e-mail archiving requirements and several white papers discuss different aspects of the subject. See <http://www.mailmeter.com/WhitePapers/index.asp>

There are no shortage of research reports describing the inexorable rise in the amount of unstructured data enterprises have to deal with. For example, the Taneja Group consulting firm surveyed file storage activities in business. In its report, "Next Generation File Management and Controls Market Overview,"² the Taneja Group found:-

- 73 percent of the users in the survey indicated more than half their data, 60 percent or more, was unstructured,
- Just over half, 53 percent, had 11TB or more of unstructured data in their systems,
- Unstructured data growth rates between 16 and 75 percent were reported by 62 percent of the users.

The major software drivers for this growth were Microsoft Office (78 percent), e-mail attachments (66 percent), and backup and archive (81 percent).

Another factor in the growth of unstructured data is the generation of multiple and distributed copies of file data from content creation and collaboration. Users pointed to Windows as their standard file storage platform at the server and storage level, housing more than 26 percent or more of their unstructured content.

Finally, the majority of respondents expected their file management and control budgets would grow by up to 20 percent in the next 12 months in verticals such as government, professional services, financial services, retail and telecom.

According to research firm The Info Pro, a typical Fortune 1000 enterprise will see its data storage needs grow by up to 230 percent by 2010

IT research firm IDC published its report 'The Expanding Digital Universe, 2007' and stated that the amount of digital data world-wide in 2006 was 161 exabytes³. It would grow sixfold to 988 exabytes in 2010 with enterprises responsible for storing 85 percent of it. The vast majority of all the new data created would be unstructured.

In an update to this report published in 2008, 'The Diverse and Exploding Digital Universe,' IDC said that the 2007 digital universe total, was actually 281 exabytes, 10 percent more than it estimated in the 2007 report, and would grow to ten times that total in 2011, 2.81 zettabytes. Enterprises would still be responsible for storing 85 percent of it.

In everyday office terms the storage of all this information of fast, front line drive arrays is akin to having no filing cabinets and storing everything on the desk. It is impractical. Storing all file information on a single set of fast drive arrays will become unaffordable and will choke the array's performance, make backups take more timer and space, as well as making array management inefficient, time-consuming and very costly.

Cited in InfoStor, December 2007.

³1,000MB is 1 gigabyte. 1,000GB is 1 terabyte. 1,000TB is 1 petabyte. 1,000 PB is 1 exabyte. 1,000EB is 1 zettabyte or one billion billion megabytes.

Customer requirements

What is needed is, in digital terms, the re-instatement of filing cabinets for reference data. The front-line, high-performance storage arrays need less active and unwanted files removed and placed on more appropriate storage. It means that backup of primary data is substantially faster, as less information needs to be backed up, saving both time and backup disk space.

The reference information in these files is still valuable. Viewed as a whole it represents part of the life blood of an enterprise and cannot be lost, discarded or stored ineffectively.

The fundamental requirement is for a digital file archive and software to manage it. There needs to be a way of detecting less active and unwanted files, capturing them, and moving them to a second tier of storage which can archive the files and keep them secure and safe for the time when they are needed. If the time comes when the files are no longer needed or must be deleted because a legally mandated retention period has expired and data privacy rules specify that files containing individual's details must then be destroyed, then a digital file archive must ensure prompt and effective deletion.

Additional requirements include:-

- The first requirement in implementing a file archiving solution is that the organisation needs to know the size of the issue that needs to be addressed. This means that a comprehensive file reporting module is needed to show age of files, created date, file ownership and duplication of files that reside on file systems, file servers and across multiple file storage devices. It is important that the reporting offers a consolidated view of all of the files in an organisation as well as providing granular reports so that organisations can focus in on specific problem areas.
- Businesses differ in the criteria they use for deciding whether information in a file is no-longer active, what the file retention period should be, and what the end-of-life action for a file should be. These are all policies that should be settable by customers.
- Once set the policies need implementing in an automated way. This dramatically reduces the file administration burden and also much reduces the potential for error; the more people involved in file management there are then the larger the possibility of human error leading to mis-kept files or mis-deleted files and inefficient archive management. The archive should be able to recover from accidental deletions. It should also be able to restore files to primary storage if they become more active.

³1,000MB is 1 gigabyte. 1,000GB is 1 terabyte. 1,000TB is 1 petabyte. 1,000 PB is 1 exabyte. 1,000EB is 1 zettabyte or one billion billion megabytes.

- Organisations need any application (such as Excel, Word, etc) access to archived files to be unaffected by their removal from primary storage to the archive. It should not be necessary, in fact it should be unacceptable, for application access to such files to break and require remedial application coding to fix.
- User access to files in the Windows file/folder infrastructure should not be broken either. There needs to be a way instituted for both user and application access to files to continue, after files have been migrated into a file archive.
- It should be possible to retain files in the archive even when deleted from the primary store. This caters for reference information needed for compliance or legal purposes with no current access needs at all.
- Customers need a file archive to be efficient and not waste space. There is no need to store multiple copies of the same file. Any duplicates should be deleted and their access path switched automatically to the single master copy. The archive should be compressed to further save disk capacity.
- The archive's contents should be capable of being encrypted to prevent unauthorised access and data leakage.
- The archive software should be schedulable so that it regularly and automatically inspects the primary store looking for files to migrate to the archive – i.e. no access within a settable time period – or delete – all MP3 files more than 7 days old.
- The archive should have excellent search and retrieval capabilities for administrators. This will reduce the time and cost of searches for information that the business needs to find quickly, such as contract agreements.
- The archive should be agnostic about the storage medium used to give customers flexibility. The supervising software should function regardless of whether the archive is stored on serial ATA (SATA) hard drives, storage area network (SAN), network-attached storage (NAS), optical disks or tape.
- Customers will need good management information about their file archive so as to understand its size, its contents and the history of files within it. They will need management facilities to monitor and control the archive's size, files added to it, files contained within it, files deleted from it, file access histories, and the storage resources it relies upon.

What to look for in a file archiving solution for your business

Here is a checklist of functions and capabilities to look for in a file archive solution:-

Comprehensive file storage reporting – This should provide reports on age of files, last modified date, ownership and file duplication. The reports should offer a consolidated view of all of the organisations file storage devices and also provide granular reports at the file server and file system level.

Transparent access to archived files - A pointer or smart stub should be placed at the original file location to redirect file access to the new location. In effect the archive and primary store are virtualised into a single logical file store.

File archive policy setting – the IT department should be able to set policies for an automated archive to detect, delete or capture less active and unwanted files, place them in an archive and retain them there until they reach end of life.

Schedulable file archiving activity – Primary storage should be regularly and automatically inspected for files that can be taken away and retained in the archive or deleted. The schedules for this should be user settable.

Automated policy implementation – file archiving actions should be automatic to increase efficiency and reduce the likelihood of human error.

File archive reporting – The IT administrators should have clear and full information pertaining to archive contents and activity so that the archive and its users can be readily managed to increase the organisation's efficiency and the optimise the efficiency of the archive

File archive retention and deletion facilities – The archive should be an actively-maintained store for the secure and reliable retention of needed reference information.

File archive security – Encryption of the archive's contents to prevent unauthorised access to data.

File archive space efficiency – Use of single file instancing and compression to avoid storing duplicated data needlessly.

File archive management capabilities – Reporting facilities and archive management functions to optimise its size and efficiency.

Storage medium independence – The ability to use any chosen storage medium in one or more archive tiers of SAN, NAS, direct-attached disk, optical disk and/or tape to store information on the most cost-effective medium for its current value to the organisation.

Fast searching – The ability for authorised users to search the entire archive, or relevant sections of it, quickly and effectively so they can carry out their tasks effectively.

Reduced administration costs – The saving of IT management cost through enabling fewer administrative people to manage more data in a much more effective way.

Reduced storage costs – The saving of disk storage cost through migrating substantial amounts of less actively-needed data to less expensive storage, providing good access to it but not using expensive primary storage to do so.

Conclusion

A good file archive product will remove up to 80% of the under-accessed files on an organisation's primary drive arrays and delete or migrate them, removing duplicated files en-route, to more cost-efficient secondary storage. The migration and deletion will be carried out using IT department set policies. The file archive will implement retention policies also guided by pre-set policies and will deliver comprehensive reporting facilities. It will be storage-agnostic

The net results will be:-

- A potentially huge reduction in the amount of storage needed for unstructured information,
- A reduction in the cost of storing it,
- Lower IT administration costs,
- Faster backup of primary data,
- Simplified overall file management,
- Better secured and more searchable reference data storage,
- Better utilisation of primary disk storage,
- Deferred primary disk capacity upgrades,
- Extended life for current primary storage and possible retirement of a proportion of it.

Overall organisations will be able to place a firm grip on the deluge of unstructured information they are facing and store it in the most efficient way that suits their individual needs and the legal and regulatory frameworks they operate within.

For More Information

For additional information on Email Intelligence, Archiving, Compliance, Discovery or Storage Solutions, please visit the Waterford Technologies website at:

www.waterfordtechnologies.com

email :

USA: +1 (949) 428 9300

UK : +44 121 601 7041

Canada +1 (416) 603 6920

Ireland +353 51 334 967