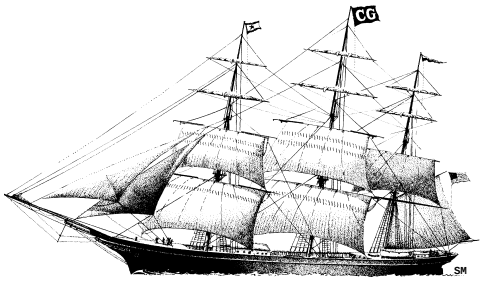
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2006093 October 28, 2006

Index Engines Eases eDiscovery from Backups

Analyst: Dianne McAdam

Management Summary

Running backups at night is a necessary, but not glamorous, task. The importance of a backup becomes very clear when a file, which was accidentally deleted yesterday, must be restored. Then a successful backup can mean the difference from being able to restore that file, or having to recreate the file manually.

However, what if your company has just been sued by a disgruntled ex-employee and your legal department wants a copy of every email correspondence between the ex-employee and the former manager? The person filing the lawsuit was employed for five years. Now you have to search the last five years of backup tapes to find those emails. Is this easy? Actually, the task takes a long time, takes a lot of human intervention, and can cost a lot of money. Moreover, in the end, you cannot be sure that you have recovered all of the emails that were requested, putting your company at a great disadvantage if the employee has copies of email that you did not produce.

There is a company in New England that faced a similar challenge. In order to find all of the emails concerning a pending lawsuit, they hired outside consultants to find the backup tapes, restore each backup tape, and search through each backup to find the needed emails. It took six months and it cost one million dollars! In the end, they *hoped* that all of the emails had been discovered.

Enterprises have been writing backups to tape for many years. These tapes may reside in different locations – such as in the local tape libraries, on racks on the computer room, or in an offsite vault. They may have been created with several different versions of the backup application. In fact, they may have been created by backup applications from several different vendors. Nevertheless, all of these tapes have one thing in common – they contain valuable information that an enterprise needs when faced with a request to retrieve information. The problem is that there is no easy way to determine which tapes contain the valuable information. You have to retrieve each tape, recreate the original backup environment, restore each backup, and search through every backup, to discover the information.

Imagine if you had a way to know which tapes contained the information you required! The retrieval time now becomes minutes or hours and not months. **Index Engines has a solution that can index old backup tapes without having to restore the backups to disk**. This product can eliminate the need to hire outside consultants, save months of time, and save lots of money. Index Engines can also index backups as they are created. Every enterprise that has ever run a backup application should evaluate these solutions from Index Engines. Read on to learn more about how Index Engines can index both current and old backups.

IN THIS ISSUE

Tackling Old Backups

Many surveys conclude that over 80% of businesses (both large and small) within the United States are in the process of being sued on any given day. During the litigation process, many documents that were created years before will need to be retrieved to prove guilt or innocence. In many cases, this information may only exist on backup tapes. And, that is where the problem begins. The backups may have been written by a backup application that is no longer supported. IT administrators must now find a version of this old software, install it, and then restore the backup tapes. Then they must search through the backup to determine if it contains any data pertaining to the current legal case. Then they must find the next tape, mount it, restore the backup, search through the backup, and continue the process until all of the tapes have been examined.

Index Engines Inc., of Holmdel, New Jersey, has developed a product that can search through those older backup tapes without requiring the application be installed first. Index Engine's *Offline Tape Engine* can scan through those tapes and index the data without having to restore the backup to disk, saving a lot of hassle and time.

How It Works

Their Tape Indexing solution is a compact (1U) appliance that contains the indexing software and one terabyte of storage. Each appliance can connect up to eight SCSI or Fibre Channel tape drives.

Index Engine understands the tape formats of the various backup applications and that knowledge of the formats allows Index Engines to scan the tapes without requiring that the backup be restored to disk. In fact, the backup application does not have to be installed – a real benefit for those enterprises that have backup tapes but no longer have the application that created them. The scanning operation begins when a backup tape is mounted into a tape drive connected to the Tape Engine. The indexing software then reads through the contents of the tape cartridge, and then creates full content and metadata indexes for files, emails, and other electronic documents.

Each Engine can store the indexes for about 64 million files. Additional Engines can be clustered to support indexing for up to four billion files. Currently, Offline Tape Indexing supports Symantec/Veritas' *NetBackup and Backup Exec*, EMC/Legato's *NetWorker*, and IBM's *Tivoli Storage Manager* backup applications and can read backups created up to five years ago. Support for other backup applications, and older versions of the currently-supported backups are being developed.

How Long Does It Take?

How long does it take to index one tape cartridge, one hundred cartridges, thousands of tape cartridges? The answer is … *it depends*. The software is limited by the speed of the tape drive – the faster the drive, the faster the indexing process completes. Once indexing is complete, the contents of that tape are immediately available to be searched.

Let's say that you need to index one thousand LTO-2 tape cartridges. For this example, we chose LTO drives since they have gained wide market acceptance. LTO-3 drives, which are about twice as fast and store twice as much data per cartridge, are available today, but remember – these are older backups! Each LTO-2 drive can read data at about 40 MB per second and each LTO-2 cartridge can store about 200 GB of data in an uncompressed format. Let's assume that each tape contains about 100 GB of data that will be indexed. Note – not all data stored on the tape needs to be indexed.

One single tape engine appliance can support eight tape drives at one time. So, one the Offline Tape Engine can index about 320 GBs of data per second (i.e., 40 MB per second times 8 drives) or about 1 TB per hour.

That means that one TB of data (or 10 tape cartridges that contain a total of one TB of data), can be indexed in an hour. Similarly, one thousand tapes would take about 100 hours.

The elapsed time can be shortened by clustering the engines. So, if you clustered 10 Offline Tape Indexing engines, you would only need 10 hours to complete the job. That shortened elapsed time makes the task of indexing older backups easily achievable.

Searching for Documents

Now that the backups have been indexed, the contents can be searched to find the data required. Searches can be very granular. One can search on several different criteria, such as:

- Key words across all documents
- Contents of the properties entries, such as author, title or subject
- File name or file type
- Size or age of the data

- Date of last modification or access
- For email messages, include the following:
  - Who sent the email
  - Who received it
  - Who was copied (or blind copied)
  - Subject
  - Date

Say the legal department requires all emails sent from one individual to another within the last five years. These search arguments can be entered into the management screen and a list of emails that match that criteria (and the tape volume serial numbers that contain those emails) are displayed on the screen. A request can then be submitted to the backup administrator to schedule that the tapes be mounted and the emails retrieved.

Tackling New Backups

Once older backups are indexed, it is important to keep up with current backups. Index Engines has a product to help here as well. The eDiscovery Engine sits in front of existing tape and disk drives on a storage area network (they also have other versions that index information as it is replicated, snapshoted, archived, or vaulted. It examines the backup data as it is being sent to the target device and immediately indexes that data. Just like the Offline Tape Engine, no data is copied. Only an index and pointers are created. Since no data is copied, the storage requirements for the indexes are minimal. Enterprises should estimate that the indexes would require about 5% to 8% of the original size of the data. Each Engine can index backups at about 2 Gigabits per second.

Enterprises with multiple backup servers need multiple eDiscovery Engines. These Engines can be clustered. Up to 64 Engines can be clustered together to support one unified search over 4 billion documents.

Currently, the Enterprise eDiscovery Engine supports Symantec/Veritas NetBackup and Back_up Exec, EMC Legato NetWorker, and IBM Tivoli Storage Manager. Support for other back-up applications is planned for the future. All common unstructured file types are currently supported, such as documents, spreadsheets, text, HTML, and PDF files. Microsoft Exchange and PST files are also supported.

Conclusion

Backups can contain a wealth of information that may be needed to satisfy audits or respond to the demands of legal discovery processes. However, backups are not designed to be searched for particular information. They are designed to restore data that has been accidentally deleted or corrupted.

There are several products are the market today that archive data and allow enterprises to search the archives for relevant information. However, these products cannot tackle the problem of searching data that exists on old backup tapes.[1] Index Engines can!

Index Engines provides two solutions - the online eDiscovery Indexing Engine and the Offline Tape Indexing Engine - that make backups easy to be searched. Both solutions can index data at a fast rate. Both require minimal amount of storage and both are affordable. For example, the entry level Offline Tape Engine can index up to 2 million emails or files and costs $29,500. That is a lot less expensive than hiring lots of contractors to restore lots of backup tapes to find the data, hopefully. Moreover, unlike hiring contractors after the request to find the data is received, the Index Engines solution allows you to index backups ahead of time. Now, when the request comes in to find the data, the search can be completed in seconds or minutes. Only the backup tapes that contain the required information need to be mounted. **Simply put, every enterprise that runs backups should evaluate Index Engines' solutions.**

[1] See *The Clipper Group Navigator* dated October 8, 2006, entitled *"Archiving – Do You Need It?"* and available at http://www.clipper.com/research/TCG2006089.pdf.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Dianne McAdam **is Director of Enterprise Information Assurance for the Clipper Group**. She brings over three decades of experience as a data center director, educator, technical programmer, systems engineer, and manager for industry-leading vendors. Dianne has held the position of senior analyst at Data Mobility Group and at Illuminata. Before that, she was a technical presentation specialist at EMC's Executive Briefing Center. At Hitachi Data Systems, she served as performance and capacity planning systems engineer and as a systems engineering manager. She also worked at StorageTek as a virtual tape and disk specialist; at Sun Microsystems, as an enterprise storage specialist; and at several large corporations as technical services directors. Dianne earned a Bachelor's and Master's degree in mathematics from Hofstra University in New York.

- ***Reach Dianne McAdam via e-mail at dianne.mcadam@clipper.com or at 781-235-0085 Ext. 212. (Please dial "212" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **The Clipper Group Voyager**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, "*Navigating Information Technology Horizons*", and "*teraproductivity*" are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.