



White Paper

The Undue Burden Undone

Author – Jim McGann, Index Engines, VP of Information Discovery

Liability of Offline Data

Over the past 20 years corporations have realized the importance of electronic data. This data is the knowledge that drives their business. As a result IT departments have implemented environments that protect this data, backing it up on tape and storing these tapes offsite for safekeeping. In the event of a catastrophic disaster, IT can confidentially restore this data and minimize loss of knowledge.

Corporations have amassed piles of these offline tapes; all but the most current tapes being useless for data restore purposes.

Why doesn't IT get rid of these tapes? If they are no longer being used for disaster recovery, why don't they destroy them? Not knowing what is contained on these tapes, the specific files and email, there is really no choice but to keep them around. Therefore, many organizations have tens or hundreds of thousands of tape archives. These archives have served their purpose, and have now turned into a liability. Regulations such as the California eDiscovery Act (Chapter 5) and the Federal Rules of Civil Procedure (FRCP) require that the information locked away on backup tapes be made accessible for electronic discovery.

Tapes Explained

Let's take a closer look at the contents of these backup tapes. The content exists at the base level: working documents (presentations, spreadsheets etc...), email, databases, and system files (exe's, dll's, etc...). Responsive data is primarily made up of documents and email; databases contain transaction data not typically of interest and system files are of no interest at all. Email is typically encapsulated in a proprietary semi-structured database such as Microsoft Exchange. At the file and email level, data is not very difficult to access. However when it is backed up to tape it is placed in a backup format, or container of documents and email. This container is proprietary, based on the backup software and the specific version of the backup software that is utilized. Finally this data is placed on tape. Many tape formats exist, and these tapes require the appropriate tape drive or library in order to read the contents. The most common formats are LTO-2 and DLT. As a result the issue when attempting to perform discovery on tape contents is a complex software and hardware endeavor. Most companies do not have the appropriate hardware and software available to quickly access the data contained on tape in order to respond to legal requests. Therefore, the data remains locked away and tape discovery becomes an expensive proposition.

Traditional Tape Discovery

Let's examine how these issues impact the traditional eDiscovery process when tape is involved. Before discovery can begin the data must be restored from tape. Before this can happen an inventory must occur to profile the tape contents: what software was used to back it up, what version of this software, what email software was used, how much storage is required, etc... Tapes are cataloged and analyzed in order to gather this information. If the company has gone through a number of mergers and acquisitions, then many flavors of backup software might be associated with the tapes that were acquired, adding a new dimension to the challenge.

Armed with the knowledge of the amount of storage required to bring the data back online, as well as all the backup software necessary, including the specific versions, the data restoration process can begin. The process of restoring the data back online is the most time consuming step of the traditional tape discovery process due to the sheer volume of data and the technical resources required to manage the project.

Once the data is back online the indexing can begin. All keywords and metadata are indexed and made searchable. Many tools exist to perform indexing – speed is typically an issue since in this scenario the data has not been culled and the volume is significant. Processing 1TB of data at 5MB/second would take about 56 hours, where processing at 50MB/second would get you to a much more reasonable 6 hours. Indexing occurs on everything, including spam email, system files, and other irrelevant content, adding a lot of unnecessary time.

With the indexing complete the discovery process can begin. First system files and duplicates must be filtered out. Then a query can occur in order to find responsive data. The query typically consists of a combination that includes custodian, date, and specific text content. As responsive data is found it is then delivered to the legal team.

The problem with this scenario is all the processing that occurs prior to the query being performed is very time consuming and expensive. If you are dealing with a few hundred tapes the cost to catalog the tapes, determine the contents, and then restore it online could be in the millions of dollars and take many months.

These lengthy timelines and huge cost used to be an accepted burden argument. However, under the amended FRCP requirements and Chapter 5, courts are requiring organizations to shoulder these burdens. Legal teams must find a way to discover tape content fast enough to please the courts, and affordable enough to not cripple their clients.

A Better Way

The main obstacle to overcome within the traditional tape discovery model is to compress the time it takes before a user can begin to query the data. The biggest time and cost associated with offline tape processing is gaining access to the data so that it can be indexed. The traditional method of moving all the data off tape before processing does not need to occur. Indexing the data directly from tape, without copying it or moving it off tape, is the key to accelerating the tape discovery process.

Indexing data on tape is a tricky proposition. You first need to understand all the backup formats, including historical versions, in order to get inside the tape. Once you accomplish this task you need to get inside the email databases and compressed file containers that may exist. In other words, you need to have a deep understanding of backup and data formats and have the ability to process them on the fly, as the tape spins. Sounds tough? Yes, but new technology has broken through these tape access barriers and is eliminating the need to restore tape contents.

Eliminating the requirement to restore content from tape in order to begin discovery can save 50 to 70% of the time and cost versus traditional methods. Indexing data directly from tape allows for the discovery

to occur quickly. In fact, the tape discovery process is completely flipped. Rather than restore all data to begin discovery you can discover first and then restore what you need. Given the fact that on average less than 1% of data on tape is responsive, now only 1% of tape data needs to be extracted from tape, rather than the 99% of irrelevant data.

Burden Undone

The burden of producing archived electronic files for litigation support has been lifted. Gone are the weeks and months spent restoring taped archives to then filter through reams of irrelevant data. Index Engines new approach to eDiscovery of backup tape content allows archived tapes to be indexed without bulk restoration and without recreating legacy backup and application environments. This new methodology saves time, money and ensures compliance to electronic data handling regulations. Courts are aware of this new approach to tape discovery, before you approach the bench with a burden argument in hand, be sure you are too.