INFORMATION MANAGEMENT BRIEF

MicroForensics Targets ESI Collection Inefficiencies

Date: January, 2009

Author: Brian Babineau, Senior Analyst

Abstract: In an effort to help customers improve PC and file share ESI (electronically stored information) collection, MicroForensics announced its Titan Collector solution, which analyzes and extracts relevant information from data sources rather than creating a full PC or share image. Targeted collection helps customers capture data more quickly and cut ESI review times—both of which allow attorneys and litigators to make more informed decisions much earlier in the overall electronic discovery process.

Introduction

Increasingly, organizations that are consistently responding to electronic discovery requests have been investing in data processing, analysis, and review technologies and services in order to deal with the massive amounts of information collected and preserved after an inquiry is received. While these investments have helped, they treat the symptoms (growing volumes of potentially responsive corporate data), not the cause (over-collection of corporate data).

Current collection solutions image PCs and network shares, grabbing every byte on these devices and contributing to the over-collection problem. This method is great for criminal investigations and was adequate for civil procedures and regulatory investigations when PC drives were measured in the single gigabytes; however, PC drives—along with the file shares many PCs connect to—are getting bigger, increasing the amount of time it takes to complete an image-based collection. Image-based collection also gathers a significant amount of irrelevant data, the culling of which creates more work during the ESI review process.

Despite the issues and trends impacting ESI review, there has been very little innovation around data collection tools in the marketplace—until the release of MicroForensics Titan Collector. Titan Collector automatically analyzes data on PCs and network file shares; provides a detailed report of relevant content based on criteria specified by corporate counsel, regulator, or other investigating party; and, if required, extracts the selected content to a read-only file for collection and preservation purposes. Along the way, Titan Collector can remove duplicate content and common NIST files to further reduce the amount of data that is ultimately collected.

One of Titan's unique capabilities is its ability to survey a PC and file shares building a file inventory as it processes each file. The inventory is the equivalent of an index, but it is generated must faster due to Titan Collector's analysis capabilities. Customers can see how much information exists and sample the data previewing the quantity of files that will meet a specific set of selection criteria. As a result, corporate counsels can determine if a particular PC or file share has relevant information and whether or not to run a collection operation against these data sources.

Titan Collector enables customers to conduct a survey of potentially relevant data from several custodians' data sources in a short period of time. This, in turn, allows investigators to determine whether or not to proceed with a case. Corporate counsels also can prepare for Meet & Confer sessions earlier and conduct early case assessments. Customers can then execute a targeted collection from PCs and network shares which reduces the amount of information included in the electronic discovery process, cutting litigation expenses and mitigating risks.

The Unstructured Data Collection Problem

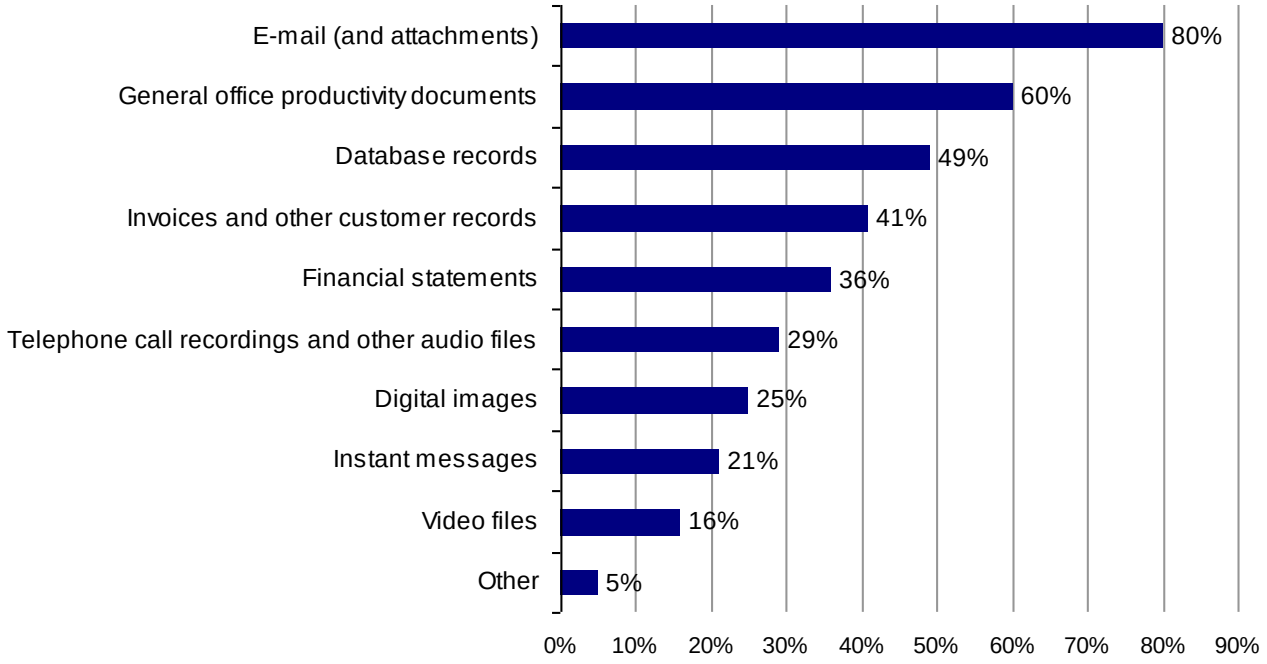
E-mail and files, often referred to as 'unstructured data,' make up 74% of corporate information[1] and are the most common sources of ESI requested during discovery (see Figure 1). These information sources share two

[1] Source: ESG Research Report, *Medium-Size Business Server & Storage Priorities*, June 2008.

characteristics that make collection for electronic discovery purposes extremely challenging: First, the sheer volume of messages and files created by employees continues to increase at a rapid pace. Second, this data is dispersed across an organization, making it very hard to locate when a discovery request does arrive.

FIGURE 1. MOST FREQUENTLY REQUESTED RECORD TYPES DURING ELECTRONIC DISCOVERY

To the best of your knowledge, which of the following record types has your organization been asked to produce in a legal proceeding or regulatory inquiry? (Percent of respondents, N = 107, multiple responses accepted)



Source: ESG Research Report, Electronic Discovery Requirements Escalate, November 2007

With data being stored in many different locations, most electronic discovery requests center on custodians and the data sources accessed by individual employees. Translated, most collections begin by gathering information from PCs and any associated data sources (network file shares and removable drives) as well messaging systems. That being said, a recent survey conducted by ESG found that one-third of respondent organizations improved e-mail collections by implementing e-mail archive solutions that centralize and index content, allowing parties to conduct searches for information that falls within the scope of the discovery request.[2] This form of archiving replaced cumbersome restoration operations involving the manual recovery of backup tapes.

PC collections have yet to go through the same transition; they still involve the creation of a full image copy of the entire device. With disk drive capacities consistently increasing (desktops can have over 500 GB drives and laptop drives are often over 100 GB), the imaging process can take several hours. As a result, either the ESI review process is delayed or the investigating party must summon additional staff to complete the collection on time.

The argument for full image PC collection is that it ensures that all data was captured in a forensically sound manner (meaning the information was gathered without modifying the associated data or metadata). Additionally, an image was the only means to capture 'deleted files' and 'slack' as well all active files. These requirements were in large part a result of criminal evidence requirements where preservation of every fragment of data during collection was paramount to achieve conviction. In civil procedures, an estimated 2% of cases actually go to and

[2] Source: ESG Research Report, *E-mail Archiving Survey*, November 2007.

are resolved by trial—the rest are settled or dismissed.[3] In addition, the forensically defensible collection of active files is all that is required for the vast majority of civil litigation. As such, a majority of ESI collected may never be produced in court; rather, it is used to help determine case facts, such as timelines that uncover 'who knew what and when they knew it.' Further, much of a PC image includes applications and operating systems, which are not useful in most matters. In all, collecting the entire image can be extremely wasteful in terms of time and the sheer amount of data corporate counsel is forced to review.

MicroForensics Delivers Targeted Collection

MicroForensics Titan Collector facilitates 'targeted collection'—where an entire information source is automatically analyzed and only the data meeting specific criteria is captured. This is ideal for PCs and any accompanying network file shares or removable drives because only relevant information is captured, allowing the process to complete faster.

Titan Collector's architecture is unlike any other collection solution in that it does not need to be installed on the source system. It can execute over the network or from a removable storage device (USB drive, etc.) connected to the PC it will scan, analyze, and eventually collect data from. It can also be uploaded to the PC, run, and then removed when the collection operation is complete. A small software footprint and agent-less implementation makes it very easy for investigating parties to quickly deploy across several PCs.

How it Works
Within Titan Collector's CLI or GUI, customers establish configuration parameters that determine how the analysis and collection process will run, including:

- **The selection criteria, which will identify documents to be collected.** The selection criteria can be based keywords, the date a file was created or modified, the date a file was last accessed, who owns the file, and the type of file.
- **What, if any, file-types will be automatically collected** regardless of whether or not they meet any of the keyword criteria. Very often, customers will collect personal mail archives (PSTs) and PDFs as content and documents can be nested within these file types. The content can be disaggregated during final processing and prepared for review at that point.
- **How duplicate files and NIST files are to be handled.** Titan Collector can collect duplicates or a single copy of a duplicate file. In either scenario, it can generate a report on how many duplicate files are located on a specific data source. The same collection alternatives and reporting capabilities are available for NIST-listed files.
- **Where the collected information will be stored** once it is extracted from the data source. In most Titan Collector implementations, customers store collected data on a removable drive that is temporarily connected to the data source. The removable drive can be sent to a central location where it would be joined with ESI gathered from other sources for additional processing and review.
- **The log files that will be created by Titan Collector** as it executes the analysis and collection. The logs track all of actions taken in Titan Collector, as well as the results of these actions. The duplicate and NIST file reports are generated based on particular logs. In addition, any files that could not be analyzed or extracted can be tracked and reported on.

After customizing Titan Collector to meet the specifications of a particular discovery request, the customer runs the software on the targeted data sources. Titan Collector analyzes information on the data source at a rate of roughly 30-40 GB per hour, allowing an investigating party to complete PC and file share collections up to 80% faster than a full image copy.

When a file meets any of the preconfigured selection criteria, it is copied into a FileSafe—a logical evidence file which is stored in the location established during initial set up. Before it is copied, Titan Collector creates a digital fingerprint based on a hash algorithm. The fingerprint is also generated after the data is stored in the FileSafe, ensuring that the file collected is the same as the original identified on the PC. If there is misalignment between the original data and what is collected, Titan Collector creates a log which can be reported on. Fingerprinting via hashing any time the data is moved within Titan helps prove the authenticity of the information.

[3] Source: http://www.ojp.usdoj.gov/bjs/glance/tortperctrial.htm

At the end of the collection, a customer is left with a FileSafe for the data sources it captured information from. Depending on the initial set up of Titan Collector, the FileSafe file contains deduplicated and de-NISTed information that meets specific criteria. The FileSafes can be used for early case assessments and sent for final processing and review with other sources of ESI. MicroForensics has also developed integration with Symantec's Enterprise Vault e-mail and file archiving solution. The integration allows a FileSafe file to be ingested and indexed by Enterprise Vault, enabling customers to search and review all collected unstructured data from a single interface.

When the collection is complete, the customer also has an inventory of what data was not collected. The investigating party can sample the uncollected data to ensure that no data that the met initial selection criteria was missed during the collection. More importantly, this permits the investigating party to sample the uncollected information in the event the discovery scope is modified to see if additional collection operations need to be run.
Titan Collector reduces processing and review costs because it collects a much smaller amount of data when compared to solutions that only gather information via a full PC image. Because the FileSafe file is created and stored in read-only format, the collected information cannot be modified or deleted, satisfying ESI legal hold and preservation requirements.

Addressing the Requirement for Forensically Sound Collection
One of the reasons why customers continuously choose to create full PC images and over-collect information is because the image also captures the data without altering any metadata. In an image-based collection, all of the data remains as it was on the PC, including the directory structures.

Even though it does not create a full PC image, Titan Collector maintains the directory structures of any captured files. The fingerprinting and read-only FileSafe logical evidence file ensures that what is collected is the same as it was on the PC and that no one can alter the data. In addition, Titan Collector true-types all files as they are analyzed, which prevents employees from spoofing file extensions so that those files will miss collection. Unlike other collection tools, Titan Collection supports long file paths, ensuring that it does not miss any data even though it may be buried several subdirectories deep.

All of the aforementioned capabilities enable Titan Collector to be considered a 'forensically sound' collection solution. While ESG is not an expert in computer forensics, a seasoned computer forensic examiner has compared Titan Collector with other image-based collection tools considered to be forensically sound and has come to the same conclusion.[4]

Flexibility is Key

Many customers have standardized on an image-based collection tool, building electronic discovery processes around it. Changing standard operating procedures for a new solution like Titan Collector can be disruptive, but customers should find a case to test it out. If a customer wanted to see the technology in action without executing a full collection operation, there are two other ways to use Titan Collector.

The first is extremely easy and may be the best way for customers to witness the speed at which Titan completes data analysis and extraction. Any organization using Guidance Software's EnCase can leverage Titan Collector to extract data from EnCase's evidence files (.E01). Oftentimes, this process can be difficult and time consuming, leaving customers wondering why they collected data if they cannot access it. With Titan Collector, customers establish the criteria for information extraction and execute the software against any number of EnCase Evidence files or logical evidence files. At the end, a subset of information gathered by EnCase's full PC image is extracted with Titan Collector. The process runs very similar to a normal Titan Collector implementation, but the customer gets the best of both worlds—a full PC image capture and then a targeted collection based on that image.

Customers can also experience the benefits of Titan Collector without running a complete collection exercise by implementing it as an ESI identification solution. Customers that want to know what information exists on custodians' PCs can configure Titan Collector to execute a survey. This survey option scans a data source and creates a metadata or full content index of all the data on a PC. At the end of the process, an investigating party can see how much data, including duplicates and common NIST files, exists on a PC. With this insight across all

[4] Stenhouse, David P. *Validation Testing of Titan Electronic Discovery Tool*

custodians' PCs, the investigating party can understand the corpus of information that may be included in a collection. The investigating party may use the index to conduct a sampling exercise where queries are executed within Titan Collector to see how much data matches certain criteria. For example, the party could see how much information on all custodian PCs meet the date range criteria spelled out in the discovery request. This assists the investigating party in arguing for reduced scopes in Meet & Confer sessions and appropriately budget staff for the review tasks.

The Bottom Line

A recent study by The American College of Trial Attorneys found that over 75% of the respondents believed that discovery costs, as a share of total litigation costs, have increased disproportionately due to the advent of electronic discovery. [5] One reason for the increase in costs is the over-collection of information at the beginning of the electronic discovery process. And a large portion of this over-collection issue is the direct result of image-based PC collection. With disk drive capacities increasing, the amount of time it takes to complete a PC collection, as well as the amount of data captured, continues to grow.

There are plenty of cases, especially criminal investigations, where an image-based PC collection makes sense. However, forensically sound, targeted collection solutions will more than satisfy the needs of an investigating party. ESG has witnessed targeted collection solutions emerge for other data sources, including e-mail; it only makes sense for similar solutions to be used in PC data gathering.

Titan Collector delivers all of the benefits of targeted collection, including speedy and accurate data analysis and extraction—characteristics that are currently not associated with image-based collection. While it may be hard for customers to switch from a preferred collection tool, Titan Collector is flexible enough to be implemented in several ways. It is very easy to configure and run on several data sources, further expediting what has traditionally been a very cumbersome process.

The December 2006 amendments to the Federal Rules of Civil Procedure cemented ESI as a common source of evidence and brought electronic discovery processes to the forefront of many corporate counsels and regulators minds. Over the past two years, many companies have upgraded the technology used to support electronic discovery, with many of the investments focused on the automation processing, review, and analysis. It is now time for organizations to address other phases of the electronic discovery process and the easiest payback can be achieved by improving an extremely tedious and costly aspect of ESI capture: PC collection. Titan Collector allows for more efficient PC data collection, which frees more time for investigators and attorneys to make critical decisions. It also helps cut the amount of data that has to be sent for final processing and review, making the balance of the electronic discovery process run smoother for less cost. Once these benefits hit home, customers will find that they may never image a PC for ESI collection again (unless they really have to).

[5] American College of Trial Lawyers, *Interim Report on the Joint Project of The American College of Trial Lawyers Task Force on Discovery and the Institute for the Advancement of the American Legal System*, August 1, 2008.