WHITE PAPER:
CLEARWELL TARGETED COLLECTION

Clearwell Targeted Collection

Next Generation Data Collection

Who should read this paper

The time, cost, and risk associated with collecting critical electronically stored information during discovery has grown as data volumes have continued to expand worldwide. However, traditional data collection approaches have largely failed to meet the needs of the legal community in today’s market. Next generation Targeted Collection perfectly compliments Symantec Enterprise Vault™ by providing another way to defensibly reduce data volumes at the beginning of each matter, so less time and money is spent on downstream data processing and attorney review.

Clearwell Now a part of Symantec

Confidence in a connected world. Symantec.

WHITE PAPER: CLEARWELL TARGETED COLLECTION

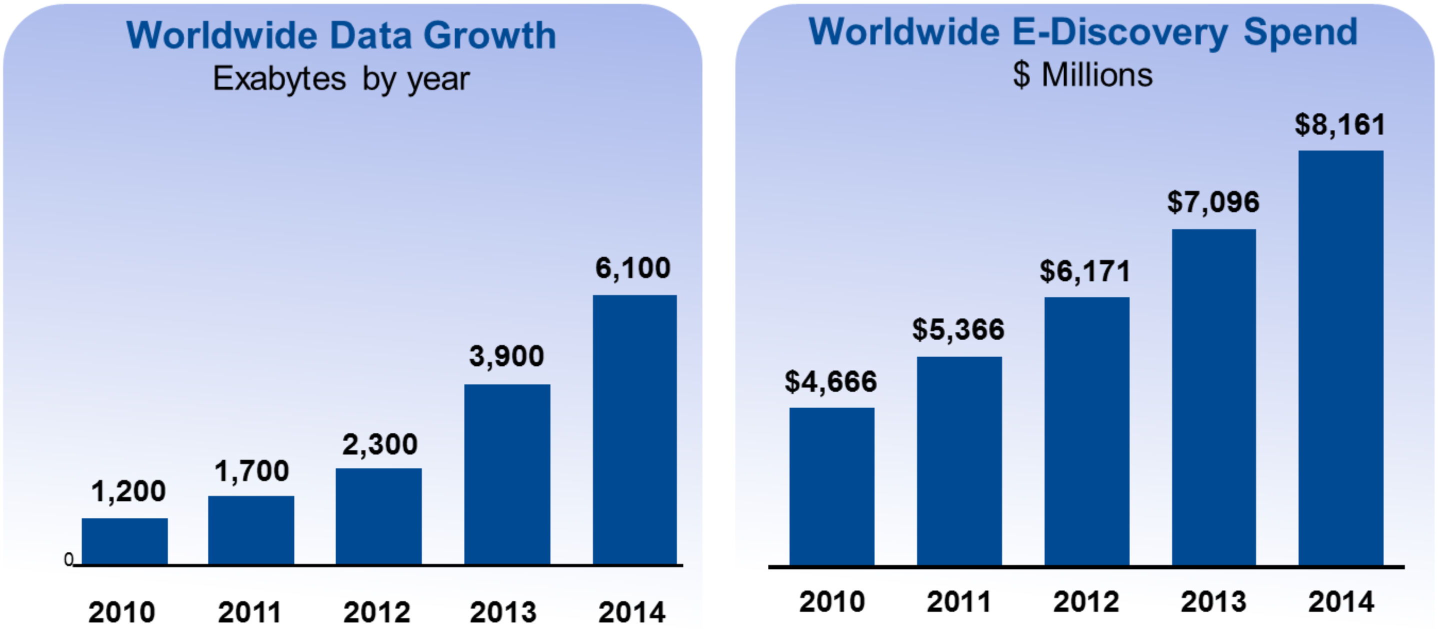
Next Generation Data Collection

Clearwell Targeted Collection

Content

Introduction

A recent article published in *The Economist* states that the amount of worldwide digital information increases tenfold every five years.[1] Given the fact that finding only the relevant electronically stored information (ESI) for a particular matter in the midst of all this digital data is at the heart of the electronic discovery process, it is not surprising that worldwide electronic discovery expenditures are also expected to steadily increase from $4.6 billion to over $8 billion by 2014. (See Figure 1). Combine massive data growth and evolving technology trends, like social media and cloud computing, with stringent legal obligations to produce data, and it is easy to understand why collecting ESI quickly and accurately is increasingly important to thousands of organizations.



Source: "The Digital Universe," *IDC*, 2010

Source: Forrester, Gartner, IDC, and Socha-Gelbmann

Figure 1. Worldwide data growth and eDiscovery spend

Although data volumes continue to grow and evolving technologies change the way data is created and managed, the duty to collect and exchange ESI in federal and most state court systems remains the same. Federal Rule of Civil Procedure (FRCP) 34 defines ESI broadly to include, "writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations—stored in any medium...."[2] This broad definition generally requires parties in litigation (and sometimes non-parties) to honor requests for ESI that are reasonably calculated to lead to the discovery of admissible evidence unless the ESI is privileged, not within the responding party's possession, custody, or control, or is not reasonably accessible due to undue burden or cost.[3] The rules do not excuse a party's duty to collect and produce ESI simply because the organization possesses large quantities of data, has ESI spread across many locations such as the cloud, or because employees use social media technology like Facebook and Twitter™.

What this means is that electronic discovery solutions must evolve to address drastically increasing data volumes and constant technological changes. The good news is that archiving technology from Symantec Enterprise Vault™ enables organizations to proactively address data

1-Data, Data Everywhere, The Economist, Feb. 25, 2010 at http://www.economist.com/node/15557443
2-Fed. R. Civ. P. 34(a)(1)(A)

growth challenges through deduplication, single instance file storage, file compression, and automated records retention for years. Archiving technology is important because it enables organizations to systematically reduce the amount of ESI they retain. Ultimately, storing less data not only decreases storage costs and reduces data security related risks, but it decreases the amount of information that must be collected and evaluated as part of the normal eDiscovery process. Since paying attorneys to review documents is one of the most expensive steps in the eDiscovery process, proactively reducing data volumes saves organizations a significant amount of time and money.

Enterprise Vault has also been enhanced so organizations can now archive social media communications. That means information created using Facebook, Twitter, and other types of social media can be managed, preserved, collected, and produced as part of the normal eDiscovery process, just like email and other files commonly stored in the archive. Symantec Enterprise Vault.cloud™ goes even further by enabling organizations to manage social media communications and other files the same way, but in the cloud. This gives organizations the flexibility to address storage management, regulatory compliance, and eDiscovery either on-premise with Enterprise Vault, in the cloud with Enterprise Vault.cloud, or by using both archiving solutions as part of a hybrid approach.

Another critical challenge facing many organizations today is figuring out how to effectively collect data when potentially relevant ESI resides inside archival systems like Enterprise Vault, as well as outside the archival system and within unstructured data sources, like Microsoft SharePoint®, email servers, file servers, desktops, or laptops. Collecting ESI from the company archive and neglecting to collect ESI from unstructured data sources as part of a legal matter could lead to sanctions. Unfortunately, traditional eDiscovery solutions don't enable users to collect data across multiple systems cost effectively and efficiently, even though data collection is one of the earliest and most critical steps in the eDiscovery process. Instead, traditional eDiscovery technology solutions have largely required organizations to choose between collecting too much ESI in an effort to minimize the risk of sanctions that may result if data is overlooked, or too little ESI in an effort to minimize downstream processing and review costs. Attempts to solve this dilemma using current technology solutions have largely failed, but the same legal obligations still exist even though data volumes continue to grow.

Symantec and Clearwell solve the dilemma with the introduction of Clearwell's next generation Targeted Collection, available in the Clearwell™ eDiscovery Platform. Targeted Collection enables organizations to target and collect only the most relevant ESI for every matter, with pin-point accuracy, regardless of whether the ESI resides inside an archive or outside the archive. This approach enables organizations to reduce the risk of overlooking important ESI during eDiscovery while significantly reducing the time and cost associated with data collection.

The limitations of traditional data collection approaches

Three of the most common technological approaches for collecting data during eDiscovery are described below.

1. Traditional manual forensic collection tools

Many organizations continue to use off-the-shelf forensic collection tools to make copies of files for each new matter. This approach often results in more files being collected than necessary, or what is referred to as the "over-collection" of ESI. The benefit of collecting more files than necessary is that the risk of files being deleted, altered, or lost is minimized and so is the risk of data spoliation. On the other hand, relying on internal Information Technology (IT) employees or paying external collection specialists to use off-the-shelf forensic collection tools can be very time consuming and expensive.

For example, potentially relevant ESI often exists across many different systems and in multiple offices for many organizations. Collecting data using manual forensic tools normally requires a specialist to physically connect a collection device to each relevant desktop, laptop, and server to collect or "copy" files for every new matter.[3] The process is not only time consuming and disruptive for employees whose data is collected, but many organizations feel compelled to collect more data than necessary to both minimize the risk of files being lost or deleted and to avoid being forced to revisit and recollect data from the same sources repeatedly if the issues in the case change or expand. Typically, resources are instructed to copy each relevant employee's (custodian's) entire hard drive, home directory, and email folder at a minimum every time there is a new matter. The end result is that far more data is collected than necessary which leads to an expensive domino effect, beginning with the high cost of downstream data processing and filtering, and ending with even costlier attorney document review.

To put the cost of "over-collecting" data into perspective, consider that a few years ago Gartner, an independent industry analyst, indicated that reviewing one gigabyte of data could result in approximately $18,750 in legal review costs.[4] This figure doesn't even account for common fees charged by third-party vendors to process, filter, and load data into review tools which also impact overall costs. Given the financial impact of legal document review, a key objective for any astute corporate legal department faced with high eDiscovery costs is to limit the amount of data sent to outside counsel for processing and review. Unfortunately, although traditional off-the-shelf forensic tools can be used to decrease organizational risk through "over collection" of data, the reduction in risk is often overshadowed by substantial financial costs since more data is typically collected and passed downstream for further processing and attorney review than necessary.

2. Traditional network-based forensic collection tools

In response to the demand for more automated collection functionality, providers of manual forensic collection tools were among the first to offer new tools to enable network-based data collection. Unfortunately, these tools continue to suffer from both technical and proprietary limitations that have impeded their adoption. For example, some of these tools require the installation of software agents on laptops and desktops before data can be collected. Installing, upgrading, and maintaining these software agents is often a tremendous burden for IT departments—especially in large organizations with hundreds or thousands of employees. However, the risk of failing to properly maintain these agents is high because it could result in insufficient data collection and ultimately sanctions.

Other serious limitations with these traditional tools include a lack of user friendliness and search flexibility. For example, non-technical users incapable of writing simple programming language, known as scripts, will be required to hire or train additional staff before they can even use these complex solutions. Similarly, these tools do not have the ability to conduct federated searches by leveraging indices that have already been created by native applications. Instead, the technology must scan every file contained in every application in order to conduct each new search—a process that is painstakingly slow. Depending on the size of the environment being searched, each search could take days or even weeks. The amount of time required to use this type of technology is often not practical in the real world of eDiscovery, where deadlines tend to be pressing.

The other key challenge with network-based forensic collection tools is that once files have been collected, they are often rendered unusable by common eDiscovery processing, analysis, and review tools. In an overabundance of caution, the providers of these tools decided to develop their systems so that collected files are wrapped in a proprietary "evidence" file to preserve the integrity of the files for evidentiary purposes. Maintaining the integrity of the evidence is unquestionably important. However, using a proprietary format to properly secure evidence is unnecessary to maintain integrity. Some would argue that the decision to use these proprietary wrappers is motivated more by financial gain than evidentiary integrity. Supporters of this argument point to the fact that "opening" the evidentiary files to review substantive ESI files requires the use of proprietary software that costs money.

3-If data is stored in the cloud, this collection method may not even be available since third party cloud providers are understandably hesitant to allow an outside collection specialist onsite to collect data due to data security risks.
4-Debra Logan, John Bace, Gartner, E-Discovery: Project Planning and Budgeting 2008-2011, Feb. 2008.

3. Traditional "index-everything" network-based collection tools

Different kinds of network-based collection tools evolved to address some of the challenges with the forensic collection tools, but the risk, time, and cost of managing these tools has largely diminished their value—leaving many organizations feeling that the network-based collection tools they purchased to automate their collection process have under-delivered. These tools can generally be described as "index-everything" solutions.

The "index-everything" approach seeks to address common dilemmas faced by organizations utilizing manual forensic collection tools by creating a searchable index of all files within an organization's environment. The idea is that a resource in the IT department can leverage a centralized technology solution to simultaneously search across all data sources, custodians, and file types from a single application to identify ESI every time there is a new matter. Once identified, potentially relevant ESI can be collected or copied to a secure repository through the company network automatically with a few simple mouse clicks. Theoretically, the "index-everything" approach minimizes the risk, time, and expense of the traditional forensic collection approaches because leveraging an "index" makes searching faster. In reality however, these solutions tend to present some of the serious challenges listed below that often lead to increased organizational risk and cost.

IT challenges: Implementation and maintenance

Implementing and maintaining "index-everything" technology solutions across an unstructured environment is different than indexing data stored in a centralized archive and often becomes an unexpected and monumental burden for under-staffed IT departments. The reasons for these challenges stem largely from the fact that even midsized organizations tend to have massive amounts of data spread across multiple systems and different offices that are dynamic. In order to create this searchable index, "index-everything" solutions must often crawl or scan files contained in each dynamic system in the network. The dynamic nature of the different data systems makes indexing across multiple systems and offices much more difficult and risky than indexing data that has been centralized in an archiving tool because more data sources, software applications, file types, and storage systems must be maneuvered and more network connectivity and bandwidth issues are likely to exist.

To complicate matters further, creating a searchable index across the environment could take months or years and may never be complete. Users continually create, modify, and delete files so all systems must continually be indexed to ensure the accuracy of searches. In reality, creating a comprehensive index of an entire organization's data systems is unrealistic, even for relatively small companies because so much information is constantly created and modified. Some common IT challenges are listed below:

- Decreased network speeds (bandwidth) impact end users
- Decreased network speeds make completing routine data backups difficult
- Increased storage costs as a result of maintaining the index
- Installing and updating agents on local computers

Taken together, these issues add complexity to the deployment and management of "index-everything" solutions and often result in organizations not being able to fully deploy the solution, even after months or years of hard work.

Risk: Incomplete searches could result in sanctions

Maintaining complete and accurate data indices of dynamic content can also lead to incomplete eDiscovery searches and even sanctions. Most "index-everything" solutions originated as enterprise search tools to help employees find information more easily and were not originally designed to address more stringent legal requirements related to electronic discovery. Unfortunately, that means these solutions are often unable to search and collect important files during discovery, such as files within files (container files), encrypted files, or uncommon file types. The risk is that organizations with millions of documents may overlook thousands of potentially relevant files because all the files were never properly indexed. This can create obvious challenges for lawyers who believe the company's environment has been comprehensively searched for ESI, only to find out that the search was not complete. In the worst case scenario, lawyers could end up unwittingly making false representations to the court and opposing parties about the comprehensiveness of their data collection and production which could lead to sanctions and penalties.

Cost: Deployment and maintenance costs are usually much higher than expected

Organizations who purchase technology solutions with the belief that they will be able to create a searchable index for the entire organization almost always find that the deployment and maintenance costs far exceed what they expected. In addition to high initial purchase and deployment costs, challenges related to using and monitoring the solution often require increases in headcount. Other hidden costs might include additional expenses to increase bandwidth, storage costs to maintain the index, and ongoing training and support for managing a very complex system. Since data volumes for most organizations continue to grow and many lack the budget and resources necessary to support a complex indexing system, attempting to deploy an "index everything" solution is often cost prohibitive for organizations even if the internal IT burden and risk of sanctions could be minimized.

Next generation Targeted Collection

Next generation Targeted Collection introduces a more flexible and simple method for streamlining data collection that solves the time, cost, and inefficiency issues that have plagued traditional collection approaches for years. Instead of forcing organizations to create massive data indices before searching and collecting ESI, next generation Targeted Collection provides the flexibility to conduct federated searches of key data sources by leveraging native indices created by large systems or by connecting directly to smaller data sources. Combining this flexibility with a simple interface enables organizations to collect data faster, more cost effectively, and more efficiently, regardless of whether data resides inside or outside of archival systems like Enterprise Vault.

Most traditional tools are inflexible and they can only search data sources directly or they must create a new index of the targeted data source before search and collection can even begin. The main challenge with both of these approaches is that they are painfully slow and/or complex. For example, searching data sources like archives, email servers, and SharePoint directly is slow and cumbersome due to the large amount of files contained in these systems that must be scanned. Many traditional tools have attempted to solve this problem by creating a searchable index of these systems to make search and collection from these sources faster. Unfortunately, the time, risk, and complexity associated with trying to create and maintain these indices have largely made this approach untenable for most organizations, as described earlier.

To circumvent the problem, next generation Targeted Collection introduces the ability to search the native source index that these large applications have already created so new indices need not be created. Other smaller data sources, such as desktops and file servers, can be searched directly using Clearwell. That means Clearwell can be used as the central hub to conduct federated searches directly into systems like desktops and file servers, while simultaneously searching the source index of systems like archives, email servers, and SharePoint. (See Figure 2).

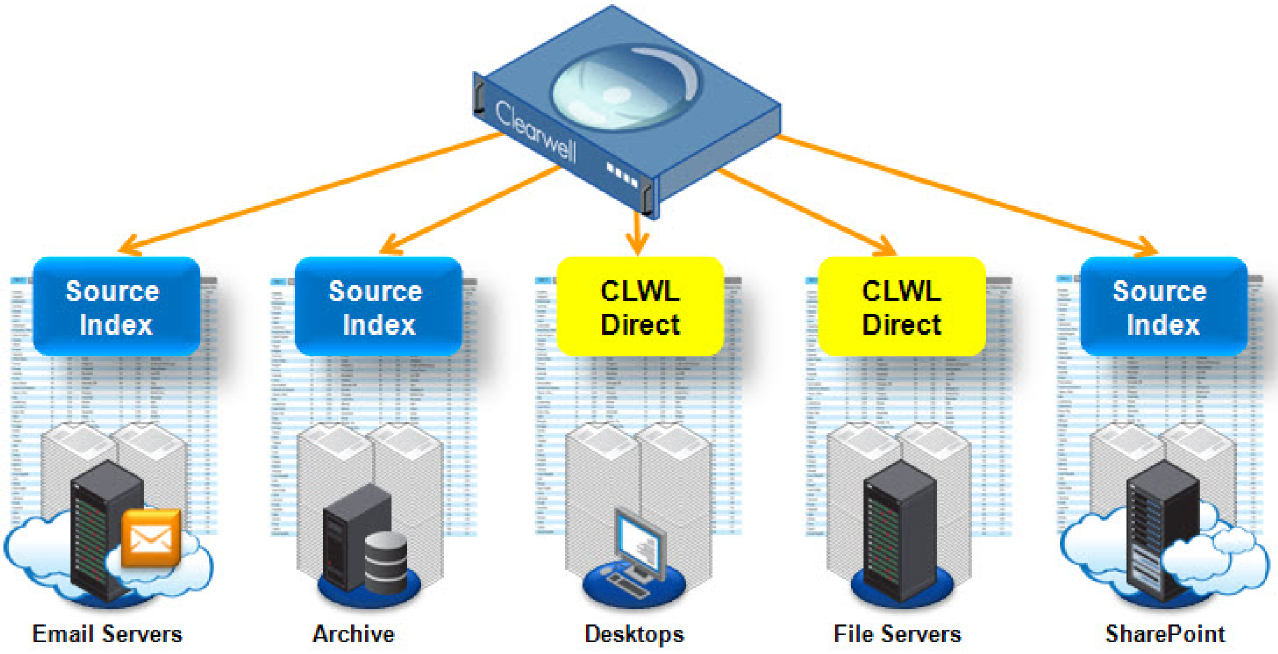


Figure 2: Targeted Collection provides user's with the flexibility to leverage indices created by native applications like email servers, archives, and SharePoint to search and collect data or to search other data sources, like desktops and file servers, directly without the use of agents.

The flexibility of Clearwell's federated search capability is significant because it allows organizations to avoid the tremendous amount of time and energy required to create a searchable index from scratch. Similarly, this flexibility allows organizations to leverage the major investments they have already made in business critical solutions, like Microsoft® Exchange, SharePoint, and Enterprise Vault, by searching the indices already created by these systems, instead of investing in new solutions to create new indices. The ability to search and collect data quickly without spending months creating and maintaining indices is critical since deadlines for producing ESI are often short. The end result is that organizations are able to use Clearwell's simple interface as the primary data collection tool for the organization's most critical data sources.

The simplicity of Targeted Collection also enables organizations to decrease costs at the beginning of the collection process because data can easily be searched and filtered by file type, date range, custodian, and other criteria at the point of collection. This means the number of irrelevant files that are ordinarily collected with less sophisticated tools can be drastically reduced, which reduces downstream eDiscovery processing and attorney review costs. The cost savings are even enhanced further by enabling users to search and collect ESI from multiple data sources using keyword searches. The ability to search and collect ESI with pin-point accuracy instead of relying on less sophisticated traditional approaches often results in savings of thousands or even millions of dollars for a single case because less data is sent to outside vendors and counsel for processing and review.

In addition to curbing the historical challenges associated with searching large data sources, Targeted Collection also simplifies the data collection process for both IT resources and end users. For example, unlike traditional solutions, Clearwell does not rely on "agents" to search laptops and desktops. Eliminating the need to install and maintain hundreds or thousands of laptop and desktop agents removes a tremendous burden for the organization's IT department. Similarly, instead of requiring users to craft complex scripts to search for potentially relevant ESI, Clearwell's easy-to-use interface enables users to simultaneously search and collect data across multiple data sources with a few mouse clicks. Not only does simplifying the end user's ability to craft search terms reduces the burden on often strained IT resources, but it also empowers end users to find important data on their own quickly.

Finally, next generation Targeted Collection technology is part of a complete end-to-end eDiscovery solution that maintains file integrity through every step of the eDiscovery process without requiring the purchase of additional software to open and review proprietary "evidence" files. Instead, Targeted Collection technology relies on the creation of hash values and chain of custody reports to prove authenticity. This approach streamlines the ability to quickly cull, analyze, and review within the same Clearwell system without causing unnecessary delay or expense and without sacrificing the integrity of the evidence.

Conclusion

The time, cost, and risk associated with collecting critical ESI during discovery has grown as data volumes have continued to expand worldwide. However, traditional data collection approaches have largely failed to meet the needs of the legal community in today's market. Next generation Targeted Collection perfectly compliments Enterprise Vault by providing another way to defensibly reduce data volumes at the beginning of each matter so less time and money is spent on downstream data processing and attorney review.

About Symantec

Symantec is a global leader in providing security, storage, and systems management solutions to help consumers and organizations secure and manage their information-driven world. Our software and services protect against more risks at more points, more completely and efficiently, enabling confidence wherever information is used or stored. Headquartered in Mountain View, Calif., Symantec has operations in 40 countries. More information is available at www.symantec.com.

For specific country offices and contact numbers, please visit our website.

Symantec World Headquarters
350 Ellis St.
Mountain View, CA 94043 USA
+1 (650) 527 8000
1 (800) 721 3934
www.symantec.com

Symantec helps organizations secure and manage their information-driven world with **IT Compliance**, **discovery and retention management**, **data loss prevention**, and **messaging security** solutions.