

Forensically Sound Preservation and Processing of Exchange Databases

Microsoft Exchange server is the communication hub for most organizations. Crucial email flows through this database continually, day and night; business agreements, negotiations, and conversations. As the system of record for email, the Exchange server environment is at the core of most eDiscovery activities. As such, the core of electronically stored information (ESI) collected to support litigation is email and mailboxes extracted from Exchange. This paper will provide an overview of the current process of ESI preservation and collection from Exchange servers, and propose a more forensically sound approach.

ESI Discovery Challenges

The challenges in performing ESI collection from Exchange are well known. Exchange server repositories, including Exchange databases (EDB's), STM's and log journals, are large, complex and in constant use. Obtaining access to an organization's live mail server is next to impossible, as most IT organizations protect it vigilantly. Shutting down this extremely active database or interrupting it in any way is typically out of the question.

Another challenge in dealing with Exchange repositories is the sheer size of the database. Most organizations have traditionally let user mailboxes grow quite large. An unconstrained user mailbox can easily grow to 10-15 GB. Each gigabyte contains approximately 10,000 emails. Multiply this volume by the number of users in the organization you will quickly find that an EDB can contain millions or billions of objects. Finding specific emails and files in an Exchange repository requires the sorting and culling of these objects, a task that has traditionally been cost prohibitive.

Additionally, Exchange repositories are not designed for streamlined ESI collection. They are a built to manage and protect email, not for easy access and discovery. User data is contained in mailboxes but can also be scattered across the database, according to typical user communication patterns. The user mailbox is simply a starting point. Other repositories, such as other user's mailboxes, deleted items folders, the Dumpster/Tombstone as well as the logs (or journals) also contain valuable information. No single location, mailbox or folder, will produce everything required as email is typically spread throughout the Exchange repository hidden away in various nooks and crannies.

The Typical MAPI Collection Process

Due to the complexity, size and inaccessibility of the live Exchange repository, the collection process is usually narrowed in scope to make it more manageable. A typical request for custodian data focuses on only the email residing in their mailbox folder. If a request is made for a dozen user mailboxes, MAPI tools will extract these mailboxes from the server, and isolate all the content so that ESI discovery can begin. Using this approach you will no longer be dealing with thousands of users' mailboxes, just those that were requested. The scope of collection has been significantly narrowed.

The traditional Exchange discovery tools, like ExMerge or OnTrack PowerControls, use the API library provided by Microsoft (MAPI), which uses a top-down approach. These tools will extract the contents of requested user mailboxes from either the live Exchange server or directly from the EDB (ripping it from the database). Using the MAPI approach for discovery, when searching for access to an email message, the mailbox, the folder, the folder-message link and the message itself all have to be present for the operation to work. In the case of file corruption or truncation, any of the above components can be damaged beyond repair, and MAPI tools will not find the email in question.

ExMerge, a very common MAPI tool, can only create a PST file compatible with Outlook 2000, which limits mailbox size to 2 GB. If a mailbox tagged for collection exceeds this limit ExMerge will truncate the collection without any obvious indication that the collection was partial. Another concern with the MAPI collection approach is associated

with virus protection. Anti-virus software often removes message links within email flagged as virus suspects. MAPI based software won't find these messages because the anti-virus software has eliminated the relationships that the MAPI top-down approach relies on.

The MAPI approach reduces the volume of data, simplifying the sorting and culling process significantly. But collection speed is still a concern. In real world scenarios, MAPI tools collect approximately 1 GB of email per hour. If there are 100 GBs of email to collect, this one step alone can take days. To keep the process more manageable and auditable, each mailbox is typically collected separately. If the collection is time sensitive, this could mean the collection effort must be monitored 24 hours a day to get the job done promptly.

A New Forensic Scanning Process

An alternative approach to using MAPI tools is to forensically scan the Exchange database, to find everything related to the search parameters, rather than just what is contained in specific mailboxes. The forensic scanning approach is a bottom-up process, which indexes all content within the database independent of relationships. So if any part of the mailbox, folder or link is missing, but if the message is there (or partially there), it is still indexed and included in eDiscovery searches. Index Engines technology is an example of commercially available tools that employ the forensic scanning approach.

Forensic scanning is not hindered by the time constraints faced by MAPI tools. Forensic scanning works by indexing recent backups or forensic images of the Exchange server, potentially at speeds of over 100 MB/sec. It is possible with a Microsoft Windows 2003 or 2008 server to capture an exact, bit for bit, point-in-time snapshot of the content of the full Exchange Repository including the EDB, STM and all logs/journals. This coherent copy can be created with negligible downtime to the organization's users. The copy can then be forensically scanned, to find everything related to the discovery project, without interrupting business operations.

Snapshot Images of EDB's

Obtaining a forensically sound point-in-time snapshot of an Exchange server repository is easier than you think. Microsoft has provided IT departments using a Windows server 2003 or later with tools for backing up Exchange servers. The backup process makes a bit for bit copy of the entire Exchange repository including all user mailboxes as well as the dumpster. Asking IT for a backup or a snapshot of an Exchange repository is often much less of a burden than requesting permission to get access to the live server for the extraction of user mailboxes. Generating a snapshot takes minutes, and once it is created, the Exchange server can be returned to normal operation. Using the snapshot the collection process can occur in the background, without any further impact on the live Exchange environment. After the collection from the snapshot is complete, the snapshot can be deleted.

In fact IT may already have thousands of point in time snapshots available - contained in a company's backup tapes. IT may have thousands of tapes stored offsite that contain bit for bit images of the Exchange Repository on the date of the backup. If a specific timeframe is in question you can most likely find a backup tape with an Exchange image within this date range.

Therefore many collection efforts are performed on data from backup tapes. Once you have access to this backup tape you still have the challenge of finding the relevant email and content. A MAPI based approach would traditionally require the database to be restored back online before mailbox extraction can begin. This process reintroduces the issue of scale and complexity. However, when collection is mandated by the courts this approach has been the widely accepted, yet painful practice. The forensic scanning approach, offered by Index Engines, eliminates the need to restore backup data, allowing discovery to begin immediately.

Automated Collection from EDB Images

The MAPI approach implements extraction tools that get you access to pre-defined chunks of data or mailboxes. Forensic scanning of EDB images on tape or disc provides a complete view into the Exchange database. The scan produces a searchable index of all the content including the dumpster. This technology sees even email that has been purged from the dumpster and not written over. These emails can then be reclaimed allowing discovery to go well beyond the dumpster. This detailed forensic approach delivers a far more comprehensive collection of ESI versus the limited approach provided by mailbox extraction. One may argue that those emails are inaccessible for production purposes, but the fact that technology exists to recover these emails means that the duty to preserve is in effect regardless of inaccessibility arguments.

Another alternate source of valuable email information is the Exchange Server transaction logs. These can be quite convoluted because they often have internal references to the specific EDB for which the transactions are logged. By carefully parsing a full EDB and the subsequent log files it is possible to recreate all the emails that came in or out of the Exchange Server. Simple collection and preservation of mailboxes will always miss this important secondary source of emails. Most importantly, the user has no ability to influence the content of the Exchange server logs. Thus, in an environment where the user is somehow bypassing the dumpster, the logs will still contain many of the emails. A forensic scan of the Exchange server will also index these logs and make them discoverable.

A fully indexed Exchange image also allows for deeper discovery and a more cost effective approach. Not only are full user mailboxes searchable, but also complete conversations that may reside in other user's mailboxes and not with the custodian in question can be uncovered.

Case Study – Norcross Group

Norcross Group, a service provider and forensics specialist, has implemented Index Engines for ESI collection and discovery. Norcross has proven experience with this new forensic scanning approach for Exchange discovery. Prior to using Index Engines Norcross went onsite to collect custodian mailboxes. This was a very invasive process, and typically took about one hour per GB of mailbox data collected. For example, a collection of six user mailboxes may require 12 hours or more of onsite time depending on the mailbox size.

Using Index Engines forensic scanning approach Norcross Group is now able to reduce the time spent at their customer sites, and also benefits from a more comprehensive and legally defensible collection process. Norcross can now request a snapshot of an entire EDB. This can either be provided by the client by creating a standard full backup tape, or with Norcross Group's assistance in the snapshot creation and preservation process. Using this snapshot Norcross can typically index the entire database in a few hours and then begin the discovery and culling process. This approach allows Norcross to process more data, deliver more forensically sound results, and bypass all the MAPI limitations and the complexities of Exchange.

Working with the full Exchange repository allows Norcross Group to not only perform more comprehensive collection, but also speed up processing times and thus reduce costs. What makes this possible is technology that has cracked the backup formats. Since backup software is used to generate the snapshot of the Exchange server on tape, getting inside this backup format or container is the key to making this scenario possible. Index Engines technology has achieved this task and can perform a forensic scan of an Exchange image on tape or disc and make it searchable without restoring the data. Once the search is refined and finalized, individual email can then be extracted from mailboxes and dumpsters. This approach eliminates the constraint of the mailbox and performs a significantly more sound collection. In fact, using this new approach tends to find more relevant email versus traditional methods. Included in the results is content that is typically passed over including email that is truncated,

restructured by virus software, corrupt or inconsistent. Since Index Engines performs a bit for bit scan of the EDB this corrupt email is processed and is discoverable. Using this technology nothing is ignored or left behind.

Conclusion

The challenges of collecting and searching email for litigation are many; disruption of live Exchange servers, tremendous volumes of email data, locating pertinent email through the communication stream, and collection tool limitations are just a few. Forensic experts working with new technology have implemented a process to overcome these obstacles. By using backup tape or mail server snapshots as a forensically sound image of the Exchange server, the Norcross Group is able to employ Index Engines Tape Engine to quickly, efficiently and defensibly find all the data requested. The enterprise email server is not disrupted, and even the email residing in deleted folders, journals and dumpsters is indexed and searched. By moving away from tools never meant for eDiscovery in the first place, these forensic experts have been able to provide their clients more effective, thorough and cost conscious support.

About Index Engines

The patent-pending Index Engines discovery platform is the only solution on the market to offer a complete view of electronic data assets. Online data is indexed in-stream at wire speed in native enterprise storage protocols, enabling high-speed, efficient indexing of proprietary backup and transfer formats. Index Engines' unique approach to offline records scans backup tapes, indexes the contents and extracts relevant data, eliminating the time-consuming restoration process. Index Engines provides the only comprehensive discovery platform across both online and offline data, saving time and money when managing enterprise information. For more information on Index Engines, please visit <http://www.indexengines.com>.

About Norcross Group

Norcross Group provides a full range of electronic discovery services for litigation support and subpoena compliance, including complex digital forensics, all in accord with the recent Electronically Stored Information (ESI) amendment to the Federal Rules of Civil Procedure. The firm's deep knowledge of digital investigation and discovery helps organizations simplify and streamline the retrieval and retention of critical information including misplaced, erased, or damaged data. For more information on the Norcross Group, please visit: <http://www.norcrossgroup.com>.

© Copyright 2009, The Norcross Group, Inc. and Index Engines. All Rights Reserved Worldwide.

Norcross Group, The Norcross Group and the Norcross Group logo are trademarks of The Norcross Group, Inc. Index Engines is a trademark or registered trademark of Index Engines, Inc. All other marks are the property of their respective owners, and are used here in an editorial context, with no intent of infringement.