Portfolio Media. Inc. | 860 Broadway, 6th Floor | New York, NY 10003 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

Document Storage Tips: Inside the Email Vault

Law360, New York (September 04, 2014, 6:28 PM ET)

Jon Kessler

Information — including email — is the lifeblood of many modern businesses, and enterprises depend upon reliable information services such as email servers and document storage systems. However, every piece of information takes up space on an enterprise's storage devices or servers, and to put a new twist on an old cliché — increasingly, "enterprise database storage space (rather than time) is money." Often, older data is removed from production systems so that high volumes of stored information do not slow down day-to-day operations. This also helps to reduce overall costs. Thus, most companies run a combination of email archiving systems, document management systems, and in some cases, cloud-based backup systems.

The combination of live and archival systems allow for the optimal storage balance between production servers and hard-to-access backup tapes. For legal discovery, this balance means that well-run archiving solutions can greatly speed up electronic discovery, and lower the risk of spoliation during litigation or investigation. Risks and costs are lowered by indexing and de-duplicating messages and information, providing audit trails for actions taken, and allowing information to more easily be preserved in place as part of the litigation hold process.

This article will discuss:

- common email archiving and document management systems in use today;
- how they impact electronic discovery;
- what to look for when collecting data for legal discovery from each system; and
- certain aspects of each application's limitations

Armed with this information, counsel can make better-informed decisions when approaching the preservation and collections phases of any e-discovery project involving such systems.

What Are Email Archive Systems and How Do They Impact E-Discovery?

Email archive systems are used so that data can be stored for potential retrieval in the future and/or to reduce IT loads on real-production systems. They are also used to facilitate the preservation of information contained therein for litigation, regulatory, investigatory, audit or compliance obligations. It is important for legal professionals to have a basic understanding of these systems, as archived data can be subject to discovery even though the original implementation of the email archive was not related to a particular litigation.

It is important to note that not all email archiving solutions offer the same features and functionality, as there are many email archiving systems on the market from many providers. Commercial products available often include features such as data de-duplication across content types, retention management, content indexing and even some basic tools for e-discovery, such as legal hold functionality, advanced searching, and data export. Many email archiving systems also offer common policy management for migration, retention and discovery across multiple repositories.

Beyond these core capabilities, and depending on an organization's particular needs, there are a number of additional features to consider that may be added as modules to a baseline system. Potential additional features include:

- Workflow management
- Enhanced legal hold/preservation features
- Extended data storage types
- Advanced search capabilities

As noted above, email archives also reduce the load on active email systems. For this reason, many companies will choose to periodically move old email to less expensive archival storage. By centralizing email in an archive, organizations can reduce or eliminate the costs, risk, and burden associated with de-centralized email management. In particular, organizations utilizing an email archive typically can avoid the cost, risk and burdens associated with identifying, preserving, and collecting locally stored message files, .PST files[1], and other scattered email from across networks and devices.

What exactly is an email archive system, or email vault, as they are commonly known? Simply put, an email archive system preserves email to and from individuals, and makes it more easily searchable for various purposes. The email can be captured by the archive either directly from the email application, or during the transmission of the email from one computer to another. The messages are then typically stored on disk, and indexed to simplify future searches.

Email archives can typically be can be configured in any of the following ways:

- Time-based (archived after a certain time period has elapsed, or a certain date has past)
- Size-based (archived after an email or mailbox exceeds a certain size)
- Attachment-based (all, or certain types of attachments are archived)

- User-defined (custom per organization, and may include one or more of the above boundaries in addition to unique organizational parameters)

Other key email archive system concepts include:

- Email journaling: All emails that are sent and received by any user that is subject to journaling are stored in an email journal database. Since this is tracked at the server level and not the local level, folder structuring created after a message is sent or delivered is not maintained in the journal.

- Original setup: This refers to the process of identifying the original setup of the email archive system. For instance, it is important to know whether old emails from the mail system and archived mail such as PST or NSF[2] files were ingested into the system (thereby populating the archive with legacy email) when it was brought online, or whether the archive was turned on as an empty archive on day one when it was implemented.

- Litigation hold changes/tracking: A litigation or legal hold is a direction to preserve ESI that is usually implemented by counsel for a company when that company is experiencing, or reasonably expects, a litigation to occur, and at any other time a preservation obligation exists. Once a "triggering event" has occurred and a preservation obligation attaches, it may mean that certain email which may be related to a matter or potential matter cannot be deleted and must be archived in order to effect preservation.

- Stubbing on the active email server: "Stubbed" emails are partial mail messages where one part of the message exists on the active mail server (and is visible in the user's mailbox) and the rest of the message (such as the body or attachments) are stored and linked to within the email archive system. Where collection of a user's mailbox subject to "stubbing" is contemplated, one should consider how to avoid the collection of stubbed emails (i.e., those messages that have been stripped of attachments and other data). In order to avoid costly re-work and incomplete collections, it is advised that the collection team develop a complete understanding (ahead of the collection process) of the policy that creates the stubbed email.

- Follow-up collections: Time-based research of new data or a refresh collection. For example, if a collection is performed on Aug. 1, 2012, for all mail prior to that date and another collection is required in the future, a date filter can be applied to only collect data created after the original collection.

Although email archives tend to be feature-rich, there are certain things they DO NOT do very well. Examples include:

- Non-searchable text: Text that cannot be "read" cannot be indexed, and therefore cannot be searched by an email archive's search tool(s). It is important to remember that many scanned documents, such as contracts and other documents that have signature execution, may not have text in a form that can be readily searched. Therefore, if a search term is in a document, but the archive system doesn't "OCR"[3] documents, the document/image would be missed if the chosen method of searching for documents was only through keyword searching.

- Encryption: Encrypted data is typically neither searched nor reported on as an exception in email archive systems.

- Embedded objects: Embedded objects such as Excel spreadsheets or other embedded content that are pasted into Word or other similar documents/files are typically not searched by email archive systems.

- Search syntax: Email archives typically employ search engines that are limited in their ability to apply complex search criteria or techniques.

- Extraction speed issues (e.g. — one week or more to extract email from a single custodian)
  - Collection operations may need to split the size of the extraction criteria or run smaller incremental extractions based on dates to alleviate issues caused by large volumes of data
  - Export format (typically limited in file formats and accompanying metadata)
  - Available local or network storage for extraction of the email collected, as well as the log/temporary files required to perform the extraction

- Searching for email attachments: These sometimes cannot be searched as the email archive's indices may not include the content of attachments.

Some enterprise email archiving solutions can present additional challenges, especially when it comes to the actual collection of archived mail. For example:

- The need to define custodians with multiple aliases

- Limitations in search capabilities or indexing

- Inconsistent results or wildcard issues

- Interface issues

- Corrupt indices

- Fragmented backend databases

Not understanding these potential challenges could cause an incomplete, corrupt or incorrect export from the archive system. This, in turn, can cause extreme downstream challenges if not discovered until full-blown discovery is underway. Even worse, without a thorough understanding of the limitations of an email archive, certain omissions may not be readily apparent to the operator of the archive's search and export function. If a thorough understanding of the system and rigorous quality assurance/quality control processes are not in place prior to production, those omissions may be discovered by opposing counsel after production occurs. When the integrity of a party's production is called into question, serious consequences may follow. Whether judicially imposed, imposed by a regulator, or simply economic in nature, it is clear that when dealing with archives, an ounce of prevention — or understanding in this case — is worth a pound of cure.

Conclusion

The ability to make an informed decision when approaching any evidentiary collection or e-discovery project — ideally before your company's IT team or legal counsel is under the gun of an active investigation or litigation — is key to smoothing what can be a complex and expensive process. Taking the time to understand your email archiving system, assessing it for compatibility with your e-discovery processes, and testing the results of a controlled exercise will likely be very informative.

The outcome of the investigative process and the test case results may well lead to a formal assessment or "health-check" for your organization's email archive. This proactive approach could yield an email archive update, repair, or "hotfix" where necessary, and is crucial to maintaining a reliable source of archival information. A holistic review of archiving tools, policies and processes is part of a forward-thinking information governance strategy, and will inevitably lead to a reduction in downstream costs, risks, questions and frustrations.

—By Jon Kessler, Epiq Systems Inc.

Jon Kessler is director of Epiq Systems' forensic consulting group

The opinions expressed are those of the author(s) and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] Microsoft Outlook email container file

[2] IBM Lotus Notes email container file

[3] Optical Character Recognition – Computer recognition of text that has been scanned or otherwise exists in a graphical format. The text in these graphical formats cannot typically be searched by computers unless and until the text is scanned and converted by a computer into processable text/words.