eDISCOVERY BEST PRACTICES

WHEN DEALING WITH EMAIL ARCHIVES

Email archives can play a critical role in eDiscovery, investigation, compliance, and business continuity. They are pervasive throughout the corporate world and have stockpiled huge quantities of data over many years. But when most archives were designed more than a decade ago, no-one anticipated there would be trillions of emails saved forever. As a result, archives struggle to search and extract email quickly and efficiently.

Historically, CTOs and CFOs saw archives primarily as a compliance tool. Once this data was in the archive, it was safely stored away in case a regulator or attorney needed it down the road. Problem solved, or so it seemed.

Email archives ingest data slowly over time but it only takes a few years for a manageable repository to become decidedly unmanageable. Within about five or six years, most individual email users end up with about 10 GB in their archive, around 200,000 emails. If your enterprise has 50,000 employees, you can do the math.

To combat this growth, email archives were designed to expire data automatically once the customer had set up a retention schedule suitable for its industry or geography. The problem is, almost nobody actually set up these schedules or, if they did, they didn't press the "go" button. That's why these archives have grown out of hand.

Email archives were also designed to provide rudimentary indexing of the ingested email. But when they were designed 12–15 years ago no one really thought there would be trillions of emails saved forever. As a result, these archives struggle to search billions of emails efficiently or quickly.

The unfortunate result of this history is that relying on the archive's native searching tools may actually prevent an organization from completing its investigations or meeting its discovery obligations in a timely or cost effective manner. In fact, you may not be able to complete them at all. And the more an organization relies on email archives for day-to-day storage, the more likely it becomes that a regulatory agency or court will rule that it must be searched.

If your organization regularly faces regulatory or eDiscovery requests, you likely have in place internal capabilities and relationships with third party vendors who can assist in the collection process. So you may well ask, "Why would I need more than the capabilities my archive or external vendor provides?"

The answer requires you to understand how email archives actually work. Yes, most email archives have search capabilities. However, those searches can be highly unreliable.

WHY YOU CAN'T TRUST ARCHIVE SEARCHES

Email archives are not designed for day-to-day searching. Legacy archives frequently experience index corruption, which sometimes simply can't be repaired. Archives may return inconsistent search results. For instance, an index may be unavailable if a search operation attempts to query it while an automated process is also accessing it, but may be available the next time you run the same search. The database underlying the archive can also experience corruption.

Many times, index or database corruption is not immediately apparent. Your search result could be flawed without you even knowing it.

If the corruption is widespread enough, the email archive may alert you to it by automatically reindexing the data. Depending on the size of the archive, this could take several days to complete. And this new index is susceptible to the same corruption as the last one. Often, new indexes are corrupted relatively quickly. It can happen literally within the space of a few hours if the archive is being rewritten to enough.

These factors lead to incredibly slow, incomplete, inconsistent, and incorrect search results. They make it virtually impossible to meet deadlines or represent that a search result is reliable. In these scenarios, search requests stack up, holding up investigators and legal teams from doing their jobs.

WHEN "SOLUTIONS" DON'T SOLVE PROBLEMS

A common approach to alleviating these problems is to migrate the archive data to a newer, even more expensive email archive, which boasts native eDiscovery capabilities. While it feels like a logical fix, the end result is often not pretty.

Many of these built-in tools do a poor job of scaling to the size of the archives they are being asked to search. Moreover, the built-in eDiscovery tools are very often rudimentary. They can't deliver results driven by proximity or fuzzy searches, context, clustering, predictive coding, or other complex operations that are required today.

Even if you can identify relevant data, exporting it for production or review is often limited to PST file format, which takes significant time and can cause downstream processing issues, causing a new bottleneck in the discovery process.

Another alternative is to use a third-party tool to search data in place within an email archive. The first issue is that most eDiscovery technologies can't do this. The few that can are forced to leverage the archive vendor's published application programming interface (API)—a sort of "front door" into the data in the archive.

This is problematic for two reasons: One, the API is a single pipeline to the data. As archives become larger, this single path quickly becomes a bottleneck that causes searches to run slowly. Two, the tools rely on the archive's internal index, which, as previously discussed, is often corrupted and unreliable.

Many times, index or database corruption is not immediately apparent—your search result could be flawed without you even knowing it

REAL-WORLD EXAMPLES

Despite these serious issues, data is not necessarily lost when placed in an archive. Accessing it for eDiscovery and other purposes need not be an unsettling and expensive exercise—as three companies we recently worked with found out.

A Fertile Formula for Deleted Databases

A European agricultural company had a real disaster on its hands. Its archive wasn't particularly large and the archive indexes weren't especially corrupt. Even so, the legal team could perform no eDiscovery whatsoever.

We already know that archives hide your data in complex proprietary container files and that archive indexes can become too large or too corrupt for everyday use. But there is another component to this already complex architecture: Back-end databases. These databases, which typically run on Microsoft SQL Server, contain a ton of information about the files including who has access to which files, which files are on legal hold, where the archive has stored parts of the message, and many other vital things you need to know about the data imprisoned in the archive.

This company accidentally deleted those databases. Uh-oh.

The archive tools were now completely broken. eDiscovery searches simply wouldn't work. Users couldn't even retrieve their own archived emails. Everything was lost, or so the company thought.

It turned out that even though the database was broken, there was no problem with the data itself. Nuix targets this data directly on the archive's file system and at very high speed returns it to its usable format: Email. When Nuix explained how we could absolutely give the company its data back in just a few weeks, you have to imagine the relief. Even the contractor that was paid to maintain the databases and archive infrastructure was pardoned from certain dismissal because using Nuix made it possible to return all the email and attachments to 1,200 users. Phew!

Prayers to the Technology Gods Didn't Help—but Nuix Did

A well-known athletic equipment manufacturer had to conduct hundreds of small internal investigations each year to deal with human resources matters. Due to its unique intellectual property profile, the company was also involved in multiple patent suits. So when it came time to use its email archive's search tools, this company had to hit the ground running.

What the legal team experienced was maddening. The paralegals responsible for creating custodian-level legal holds and running keyword searches on their aging archive were pulling their hair out.

Searches would start and never finish. The archive tool had no monitor to show how far searches had progressed. Sometimes searches would run for days and appear to be going OK, cross your fingers, and then crash. Each crash left the archive's indexes more corrupt. These people literally said special prayers and incantations in an effort to coax the searches and productions to complete.

The IT services company that managed the archive couldn't help and eventually the archive manufacturer conceded that its tools simply weren't designed for the rigor of a heavily litigated customer.

Nuix processed the archived data at about 1.5 TB per day. With a new and reliable index in place, the company trained its paralegals on the Nuix user interface, recreated all its eDiscovery cases in our tool, and began reliably executing hundreds of custodian and keywords searches with ease. With this problem solved, the company decided to migrate the data out of the legacy archive and into a more manageable environment. Nuix could help again, using the same software and hardware we used to solve the original urgent eDiscovery problem.

Hospital Finds the Right Treatment for its Out-of-control Archive

One hospital chain in the mid-western United States received thousands of third-party subpoenas a year and had to set up a records custodian to specifically handle these requests. This significant cost left the hospital's legal team with little patience for any additional investigations and eDiscovery events. Add to the story a legacy email archive containing no less than 1.5 billion emails in it. Like most archives, it had no retention schedule, it just kept all the emails forever. And its built-in search just didn't work.

So the hospital brought in a bolt-on discovery product to tap into the archive. The legal team believed this tool could produce data from the archive at a reasonable speed with a reasonable level of accuracy. They were wrong.

This archive had so many individual indexes—part of the growth issue again—that searches would take literally months to complete, if they finished at all. Even simple custodian and keyword searches produced dubious results.

Nuix helped this hospital group by deploying several Nuix servers to index the data. With this complete index in place, search response times were much, much faster. Additionally the legal team could perform complex searches as well as fast deduplication, clustering, predictive coding, and other operations they never dreamed possible. While the legal team could not eliminate the necessary evil of responding to eDiscovery search requests, they made the process much less drawn out and considerably less painful.

QUESTIONS TO ASK

To ensure that an email archiving tool, internal eDiscovery tool, or third-party eDiscovery vendor can help you meet investigative and legal requirements, here are four key questions you should ask:

1. Can your tool search data in place inside an archive or does the data have to be exported first? If you need to export data before responding to regulatory and legal requests, this makes it exponentially more expensive and time-consuming than searching in place. It also creates potential information governance concerns down the road when you need to account for all copies of data.
2. What search and production speeds will you achieve with the volume of data in our archive? Ask for a demonstration and insist on hard throughput numbers.
3. Can we run complex searches requiring multiple operators and search logic? If the answer is "yes," ask for a live demonstration on something that approximates your needs and environment.
4. Will the searching reliably scale across the volumes of data in the archive? Again, this claim should be able to be easily demonstrated and backed up with client references.

TO FIND OUT MORE ABOUT eDISCOVERY BEST PRACTICES VISIT
nuix.com/ediscovery

ABOUT NUIX

Nuix protects, informs, and empowers society in the knowledge age. Leading organizations around the world turn to Nuix when they need fast, accurate answers for investigation, cybersecurity incident response, insider threats, litigation, regulation, privacy, risk management, and other essential challenges.

North America
USA: +1 877 470 6849
» Email: sales@nuix.com

EMEA
UK: +44 203 786 3160
» Web: nuix.com

APAC
Australia: +61 2 9280 0699
» Twitter: @nuix