Enterprise Features

Monday, December 31st, 2012

Search

Effective Records Management – When Going From Millions To Billions Of Files

by Paul Rudo on 10/03/11 at 5:39 pm

ZL provides cutting-edge enterprise software solutions for e-mail & file archiving for compliance, litigation support, corporate governance & storage.

The Company started in 1999, and set out to solve very large scale messaging problems. The initial target was the carrier space, where green field opportunities presented in areas like ASP, ISP, and wireless carriers, where the need for messaging services for millions of subscribers far exceeded anything available at the time. They further differentiated their massively scalable email system by adding secure delivery, tracking, audit trails, etc.

As an analogy, they were now offering “FedEx” vs. snail mail delivery.

Today, I'll be interviewing Kon Leong, CEO of ZL Technologies to talk about records management.

Can you please explain Records Management for our readers? How is it different from other forms of archiving?

Records management is the proactive management of business records, including capture, classification, indexing, and auditing, in anticipation of requests from regulatory, compliance, or e-discovery queries sometime during the life of the record, up to and including disposition.

Archiving, depending on which definition is used can be as simple as the basic storage of records. In itself this definition lacks sufficient depth if questions regarding retention, searchability, and classification are raised.

Unfortunately, most enterprises still believe that archiving and in some cases "records management" is simply the act of storing data, satisfied by something as simple as back-up. This is not the case.

What are some of the biggest records management trends that we can expect to see within the next few years? What are some of the driving factors behind these trends?

The biggest trend in records management is the realization of the number of challenges that are brought about by the sheer volume of data records managers and records management systems are now being asked to manage.

Efficient storage and de-duplication of data to minimize cost and impact is an obvious one, but relatively easily addressed. Ingestion, search efficiency, accuracy, and 100% data accountability, auto-classification, and effective disposition are all many times more difficult to perform effectively under scale.

So whereas traditional records management solutions have been accustomed to a few million business records to manage, the growth of retention periods from years to decades, and most importantly, the inclusion of email as a viable business record, suddenly pushes a records management solution from a few million to a few billion, and none of the traditional solutions have been designed, engineered or tested to deal with those kinds of volumes, and are failing, and the costs are lost legal cases, regulatory violations, fines and sanctions, as well as 10x+ costs to redress and fix the situation, which can only be solved by migrating to new, more scalable solutions.

From a records management perspective, what are some of the biggest mistakes that companies make when storing their data?

They fail to test and evaluate solutions for volumes based on long term scales, instead of basic functionality tests against minimal volumes. What functions a system can easily perform under the weight of a few GBs of data can completely fail under the weight of hundreds of TBs.

Speed, performance, manageability, and administration of systems going from GBs to PBs are night and day, and companies are failing to do the kind of testing that even remotely approximates these requirements. The unfortunate thing is that without proper testing, companies won't find out their mistake, until one or two years into the project, at which point, sufficient volumes of data will have been ingested into the failing system, enough that the decision to migrating out will be weighed against spending more money on promises that the solution will fix itself.

Unfortunately, we see this all too often, and by the time millions of dollars have been spent, the company no longer has enough money nor the determination to move from the failed system and so it sits, and so all the benefits such a system promised will be unrealized.

One quick way to determine how effective a solution is in storing data is to determine how many different databases the system requires to operate across multiple data types and to scale as the volume increases. If it requires more than one database, the customer should look closely at why the system requires more than one database to perform its function.

What are some of the biggest mistakes that companies make when processing discovery requests?

Insufficiently determining the time it will take for the solution to ingestion, index, reconcile, and export its discoverable information. This can lead to delayed meet and confers, discovery meetings, and very upset judges depending on the amount of time delayed.

Any quality solution will be able to provide a linearly scaling architecture that will enable it to ingest, index, reconcile, categorize, and ultimately export its discoverable data and do so by the addition of virtual machines to increase throughput.

If a solution is only able to go as fast as a single machine can go, or can only scale across multiple machines by dividing the workload across separate deployments, then reconciliation and discoverability of that data will be highly problematic since users will have to perform each search, review, and discovery work across every separate deployment instead of one.

This will add 10x the level of work and complexity.

Electronic Records Management is still a relatively new field, but I understand that there are now efforts underway to create industry-wide standards across vendors. Can you tell me about that?

There is a lot of work and interest in establishing certain standard formats for communicating between solutions within records management and e-discovery.

Due to the number of vendors and disparate goals of each, it will take some time to achieve a comfortable level of transparency, but customers are pushing very hard and demanding that vendors become more interoperable because there are so many solutions involved throughout the life cycle of records management.

There is an alternative school of thought by some solutions which have proposed the benefits of a unified solution, one that is able to consolidate all the functionality of each component of the ecosystem, or at least all major parts of it, and to do so in a unified manner, reducing complexity, cost, and increasing performance and efficiencies.

These vendors have found some early success.

Can you please explain “Manage-In-Place” capability? How does this work, and what are the benefits?

Traditional indexing for search, required proactive ingestion of data, effectively creating an additional copy to a dedicated archive. For many customers, particularly ones with a significant amount of stored data, the notion of creating copies of their already burgeoning data stores is

unappetizing to say the least.

As a result, some vendors developed search engines with the ability to perform In-Place Indexing, which effectively scoured through existing data stores and created indices for that data without creating a copy to a dedicated archive. This enabled customers to search for their data and when necessary, point to the existing data store and attempt to retrieve it. It also enabled vendors that had weak scalability to compete in a new market by eliminating the need for them to create a controlled and managed archive.

The problem with this approach was that while the search engine could access its in-place index, it had no control over the existing data stores or the data inside. If the native application elected to delete the data, the In-Place search engine could attempt to search for the data, but finding nothing, would fail to return a result.

These systems would also be unable to perform preservation holds. For this reason, many firms that have attempted In-Place archiving solutions due to their appeal of low storage overhead, have found in real world situations that these solutions are not effective for practical reasons.

Manage-In-Place is a more advanced look at this concept that goes beyond the simple generation of an index for the data, such that in the event of certain actions, such as preservation request or other trigger, the system will go out to the location for the item and collect and copy that data, ingest and archive it so that from that point forward, no action against the native record, deletion or otherwise will affect the archived copy of the mail.

The benefits to MiP are:

1. Full search capability and retrieval benefits
2. Minimal storage overhead
3. Capture and control of data occurs as soon as required
4. Prevent spoliation once records are identified as important

How do the records management needs of very large organizations (such as banks) differ from those of smaller companies?

Very large organizations suffer most from:

Massive volumes of data make the simplest actions incredibly difficult if the solution is not engineered properly.

- Large firms have more types of data.
- More complexity in deployment architecture potentially required due to multiple regions. Complexities around privacy laws between legal jurisdictions.
- Need for sophisticated approval workflows when dealing with searches across legal lines or country borders. Ongoing support and manageability of such a complex solution.
- Increased scrutiny of data accountability because of the moving parts and complexity.

For smaller or medium-sized organizations, what are some tips that you can give when selecting and implementing a records management solution?

Look for a solution that unifies multiple functions to minimize the number of solutions they need to work with, and identify methods to prove scalability, so that when volume of the firm does increase, they are not forced to move to a new platform at a late stage.

Also determine the flexibility of the solution to adapt to new business requirements. This can often be determined by how much of the solution is based on internally developed vs. externally OEM'd pieces. The more 3rd party OEM'ing, the more difficult it will be to adapt the solution and be flexible to changing business needs because the vendor does not have knowledge or access to improve those core exports.

Related posts:

1. OSAM Document Management and Imaging Software, Services and Products
2. 1E NightWatchman – Energy Saving PC Power Management Systems For CO2 Reduction And Green Computing
3. Interview With Solera – Network Forensics and Threat Management
4. Interview With BeyondTrust Priveliged Identity Management

Leave a Reply

Name (required) Email (required) - will not be published Website (optional)

Leave comment