Oracle Universal Online Archive

An Oracle White Paper
May 2008

ORACLE

EXECUTIVE OVERVIEW

In the past few years, major changes have been taking place in the content management marketplace. Customers who previously considered content management to be a niche application focused on workgroup or departmental deployments for meeting specialized publishing-oriented requirements are increasingly requesting true enterprise deployments designed to get all of their unstructured information under better control. One consequence of this new, expanded view of content management is the convergence of archiving technology with Enterprise Content Management. Another, more significant consequence is a dramatic shift in the very nature of archiving functionality in this expanded content management view.

The American Heritage® Dictionary of the English Language defines *archive* in the context of electronic technology as "A long-term storage area, often on magnetic tape, for backup copies of files or for files that are no longer in active use." However, issues of regulatory compliance, corporate governance, electronic discovery, and intellectual property protection have significantly expanded the requirements to manage content—whether in "active use" or not. Further, the notion that certain kinds of content should be moved to less accessible media such as magnetic tape is also being called into question as online media prices plummet and the risks and costs associated with tape archiving and backup become more fully understood. Finally, as the focus of archiving shifts from a few especially important documents to the majority of all enterprise information, the complexities and expense of deploying separate, large-scale archives for email, business documents, images, rich media, ERP transactions, etc. are driving customers to look at more general, type-agnostic archiving solutions.

These changes have created an enormous opportunity for Oracle to provide a new class of content management solution. Oracle calls this new class of solution *Universal Online Archiving*. Like other traditional electronic archives, this solution focuses on managing enormous quantities of *historical* information that has reached the point in its lifecycle where content is no longer changing. However, unlike today's archives, much of the managed information needs to be maintained *online* so it is readily accessible to business, imaging, compliance, e-discovery, and business continuity applications. Also, while the information content does not change over time, the *lifecycle* of the information actively continues under a centralized policy manager that controls retention, security, storage management, and even movement

of content into and out of the archive. Finally, the solution is universal in that it manages all different types of static information in a single environment.

This white paper shows how Oracle is combining its content management, database, and middleware technologies with best-in-class Archive Services from Oracle and partners to deliver Universal Online Archive and satisfy the requirements of this changing electronic archiving market.

EVOLUTION OF THE ELECTRONIC ARCHIVE

As the dictionary definition implies, electronic archiving has historically been very much tape-centric. As such, distinctions between archiving and backup were blurred. Full and incremental tape backups were the de facto archiving mechanisms for most content in most enterprises, with the primary driver being disaster recovery. Because of the limitations of magnetic tape in terms of ease of access as well as general concerns about the non-reputability of electronic media, a small number of *really* important documents—as measured by the high value of intellectual property within their content, or significance of the event or action that the document memorialized--continued to be archived in paper form.

This bifurcated, somewhat "low-key" mode of archiving continued for most organizations well into the 1990s. However, just before the start of the new millennium, a "perfect storm" occurred that began changing organizational thinking about archiving. First, having grown relatively steadily for the past 20 years, the volume of electronic content has now exploded. Faced with enormous increases in emails, electronic documents, images, and rich media, organizations were forced to rethink the old tape-centric methodologies. Second, a series of new regulations relating to electronic content began to formalize requirements for content retention and, more onerously, prescribe requirements on how quickly retained content had to be recovered from archival storage. Finally, the risks associated with this deluge of content were dramatically underscored by high-profile corporate scandals linked to content discovery, illicit shredding, or non-compliance with the new government regulations.

While these forces combined to focus the attention of corporate legal and IT organizations on archiving, the spark that really galvanized organizational action on the topic was the formalization of electronic evidence discovery through an update to the Federal Rules of Civil Procedure (FRCP) unveiled in December of 2006. Suddenly, there were now specific rules stating that all electronic content could become evidence in litigation situations and that organizations could now face stiff penalties if legally requested electronic evidence could not be produced in a timely manner. In one fell swoop, archiving requirements could no longer be met via dusty piles of magnetic tapes or a few filing cabinets full of paper documents. Now, essentially all organizational content needed to be electronically archived, retained and destroyed under well-defined policies, and be kept intelligently accessible and deliverable on short notice.

Andersen's massive [electronic] shredding was the corporate equivalent of "wiping down the crime scene before the police get there."

Deputy Solicitor General Michael Dreeben during Enron Hearings

EMAIL ARCHIVING: THE DRIVER OF CHANGE

Overview

"Email is now the biggest content security concern for businesses. Yet for most, email is poorly managed—if at all."

AIIM Seminar Introduction, April, 2008

Email has become the "point of the spear" in driving the evolution of archiving. This has occurred for three reasons, all relating to the "perfect storm" described earlier. First, while the volume of all electronic content has soared in the past 5-10 years, the volume of email has exploded. Analysts estimate that almost 7 exabytes (that's 7 trillion megabytes) of business emails are generated every year—with many businesses (including Oracle) routinely having to deal with over 5 million emails per day. Second, a number of government regulations focused on broker/dealers (such as SEC 17a-4 and NASD 3010) specifically called out requirements for email retention. Finally, many of the high-visibility industry scandals of the late 1990's and early 2000's—including Credit Suisse First Boston, Enron, HealthSouth, Tyco, and WorldCom—involved discovery or digital shredding of email.

First Response: Email Silos to Keep Everything Forever

As the pain associated with email archiving became more acute, organizations rushed to find solutions. In many ways, the haste with which solutions were pursued resulted in a number of short-sighted strategies. The most common first generation strategy could be characterized as the "keep everything forever" strategy. The vendors that could deliver fastest to meet the immediate need were typically providing email disaster recovery solutions designed for retaining short-term snapshots of email server content. These were turned into archive products by essentially deploying farms of servers capable of saving snapshots over a longer period of time. Alternative solutions came from ECM vendors who simply journalled emails into their existing file system-based content repositories. Both of these approaches addressed the immediate need, but set the industry off in a direction that ultimately would hit the wall in terms of functionality, scalability, and manageability.

While very few organizations really thought that indiscriminate long-term retention of content arriving at multiple-million object per day rates was sustainable, the combination of rapidly declining cost of spinning storage and the fact that there were no substantially better alternatives made this strategy look interimly viable. In many ways, it was about "what to do until the doctor comes" but at least it kept the C-level folks out of jail while they figured out what to do next.

On Second Thought: Addressing Storage Pain

It did not take long for a different pain to set in. Many executives were surprised to confront the paradox that, while it seemed storage costs were asymptotically going to zero, petabytes of storage were actually still fairly expensive, not to mention the administrative and compliance costs associated with managing an ever larger and unmanageable volume of content and email. Also, it was quickly realized that, while keeping all emails forever satisfied some specific government regulations and eliminated potential accusations of electronic shredding, it also meant that enormous quantities of emails containing potentially damaging evidence was just

sitting in their archives, waiting for some smart lawyer to find them. This gave rise to a second generation strategy in which the focus was essentially on attempting to reduce the number of emails archived.

Actions to reduce the number of emails archived relied on two separate approaches: de-duplication and classification. As the name implies, de-duplication focuses on eliminating the large number of duplicate emails and attachments that result from replying, forwarding, and copying emails to multiple recipients. The ultimate goal of these kinds of capabilities is to achieve true single instancing—where each unique email and each unique attachment is stored in the archive only once. This is a worthy goal. However, actually implementing this capability in an environment in which scalability issues force the archive to be split across multiple servers turned out to be a major challenge. As a result, most second generation products failed to produce significant savings—and some even managed to archive more copies than existed in the live email systems.

Email classification has numerous important uses, but it's most basic value is in understanding which emails actually need to be archived long-term and which can be retained for a cursory period and then thrown away. While there are varying degrees of sophistication to classification schemes, the majority of archive vendors provided only rudimentary capabilities, primarily relying on the identities and roles of the sender or recipient(s) to make decisions about retention management. When more sophisticated classification was called for, the burden was put on the user—a very risky strategy because many users have no idea how to perform this classification—and even fewer will bother to take the time to do so.

Added Focus on Compliance and eDiscovery

The aforementioned retention policies also received increased focus in second generation email archiving products. Starting with simple, built-in retention policy tools, a number of vendors attempted to add on or integrate more sophisticated policy managers and true records management engines capable of providing additional security and defining more comprehensive, multi-phase lifecycles. With this came the need to add legal hold capabilities to prevent normal content dispositions (destructions) from taking place on emails that might need to be produced as litigation evidence. And with the update to Federal Rules of Civil Procedure in late 2006 came a flurry of activity to better discover emails with potential litigation impact—both for actual response to attorney requests and for pre-emptive risk assessment. Much of this product development took place rather hastily and resulted in half-baked solutions and strange lash-ups of repositories, records management products, and eDiscovery tools that were never designed to work together.

[a party must produce] any designated documents or electronically stored information — including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations — stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form

Federal Rules of Civil Procedure, Rule 34. Producing…Electronically Stored Information

"…people that I talk to are asking the right questions now. It's a great opportunity for a vendor to come out with a fresh look at the problem. Not just about message archiving. People have had real bad experience with email archiving. ..They want more."

Craig Le Clair
Senior Analyst
Forrester Research

GENERATION NEXT: A NEED TO RETHINK TIRED TECHNOLOGIES

As is often the case with technology products that evolve in parallel with the problems they are trying to solve, most existing email archive solutions today have become overly complex and under-deliver on current customer requirements. In addition, over the 10-15 year timeframe since many archive solutions were first conceived, the volume of content that needs to be managed has grown so dramatically, virtually none of the leading solutions on the market can efficiently scale to meet the requirements. This lack of scalability (and consequential email-specific workarounds that have been developed to deal with it) make most solutions shipping today technologically unable to survive the next required step of transforming from siloed email-only archives to generalized and type-agnostic content solutions. Finally, the leading vendors in the space have been slow to effectively integrate their archive solutions with enterprise-class facilities for developing and securely implementing comprehensive lifecycle policies for content migration, records, retention and storage management, and risk mitigation.

Enter Oracle: The Universal Online Archive

The evolution of archiving requirements has created a major opening for an entirely new class of product. Oracle is addressing that opening with the release of Oracle Universal Online Archive (UOA) 10g and Email Archiving Services (EAS). This product raises the bar on archiving capabilities in a number of important ways—summarized as follows:

True Enterprise Scalability and Performance

The transformation of electronic archiving from a collection of siloed environments, each specializing in a particular content type to a single universal archive facility fundamentally improves the cost, implementation, and administration requirements and greatly broadens the utilization models. This is the reason that every major archiving vendor is beginning to tout their ability to handle different content types. However, in order to provide true universal archiving, a solution needs to not only be able to functionally process the varieties of content that need long-term management, but also must be able to deal with the enormous total volume of content that an enterprise generates. Existing archive solutions can barely keep up with enterprise email volumes. When one adds instant messages, files from shared drives, Microsoft SharePoint, desktops, legacy ECM systems, digital audio and video, and business applications content in the form of invoices, receipts, bills of lading, checks, faxes, etc., no solution available can keep up with the ingestion rates and manage the aggregate content volumes—until today. Universal Online Archive, based on the industry-proven Oracle Database 11g, provides a level of scalability and performance previously unheard of in the archiving space.

Through its standards-based High-Volume Import (HVI) technology, UOA 10g can ingest content at a rate of over 1 million objects per day—on each processor.

That roughly translates to the average email traffic associated with 20,000 mailboxes. So an eight-processor system running UOA with EAS, for example, could ingest the emails for an enterprise with over 100,000 mailboxes—and still have bandwidth to take in another 3 million files per day from other content sources.

UOA leverages the capabilities of the Oracle Database 11g to provide unmatched repository scalability. Secure, robust management of many billions of objects is available in a single UOA instance—and Database Partitioning will extend that number even further in the future.

Finally, through a combination of Oracle and partner technologies, UOA provides the fastest search and retrieval performance in the industry. At least one industry competitor glibly states, "archiving is easy, retrieval is hard." In reality, true enterprise archiving is hard, and retrieval of enterprise content is even harder. Search and retrieval has been the Achilles heel of the current archiving industry leader—who uses antiquated search technology resulting in searches taking hours or days—that on UOA will take seconds. Only Oracle UOA provides the performance and scalability to meet the enterprise requirements for both archiving and retrieval.

Efficiency and Cost Control

Being able to effectively manage and search for billions of objects in a scalable, true single instance system is a market-changing innovation. However, with UOA, these improvements don't stop with scalability and manageability. A major additional benefit is controlling the spiraling costs associated with the explosion in unstructured content volumes. UOA helps to control these costs in two important dimensions: dramatically reducing storage requirements and preventing out-of-control growth in the number and size of live servers.

While the unit cost of storage is unquestionably declining, the concurrent growth in the amount of storage required means that the overall cost of storage in most enterprises is significantly increasing over time. For example, an enterprise with 50,000 email users can expect to see annual storage growth in the 30 Terabyte range. Such growth is not cheap, particularly if it's being maintained on an ongoing basis in the high-performance enterprise-class storage typically required for live servers. Add the other content types that need to be stored and the costs pretty quickly get out of control.

UOA can help control these costs in three important ways. First, through a combination of Oracle Archive Service technologies and Database 11g SecureFiles, all content stored in UOA, regardless of source, is de-duplicated. IDC (International Data Corporation) says that each item of content in an enterprise is duplicated an average of eight times, so UOA de-duplication immediately reduces the amount of content to be stored by a factor of eight. Second, the content that is stored in UOA can be compressed via SecureFiles Advanced Compression—typically resulting in storage savings of between 25 and 50%. Finally, content in

UOA can often be stored on less costly, lower speed storage devices—particularly as the content ages and is accessed less frequently. The cost differential between such devices and the storage typically required in a live server can be a factor of four or five. So let's again consider the incremental 30 Terabytes per year of email that was discussed earlier. Adding this capacity of high-performance storage (in the $6,000 per Terabyte range) would cost $180,000. Using de-duplication, we would expect the 30 Terabyte email volume to reduce to 3.75 Terabytes. Let's assume the email and attachments (which typically compress well) would then be reduced to 1.875 Terabytes. The content could then be stored on a lower-performance device at about $1,500 per Terabyte—for an annual cost of $2,812—or a saving of over 98%!

Live servers running applications such as Microsoft Exchange, Lotus Notes, and Microsoft SharePoint have been designed to handle active content. The content is being accessed and/or changed frequently and access and update times need to be very quick. As time passes, new content is added and older content becomes more static and is accessed less frequently. In most enterprises, this historical content typically accounts for more than 90% of the information under management. The accumulation of such content in a live server pushes the application beyond its comfortable operating envelope and performance begins to suffer. The typical organizational response as this occurs is to add servers, driving up hardware, license, and administrative costs and increasing complexity of use. A better approach is to migrate this historical content to UOA. In addition to getting this content under better control for security and risk mitigation reasons, and dramatically reducing the cost of storing the historical information, such migration permits the live servers to perform better and greatly reduces the need to constantly expand the server count. Further, in many instances, UOA can serve as an online backup for live servers, eliminating the need for arcane, specialized backup facilities and greatly reducing the time to restore live server information.

Security and Risk Mitigation

Getting content under "adult supervision" with UOA in order to effectively deal with compliance, eDiscovery, and lifecycle management requirements is also a prime driver for customer adoption. A universal, single instance system will not only simplify content administration but also reduce costs associated with legal and regulatory requirements. As a natural extension of the Oracle technology stack, there are several inherent security and risk mitigation advantages when implementing UOA.

"…you could be using a secure, compressed, de-duplicated, encrypted, archive without ever noticing. Throw in a Records Management Agent, and you'll also invisibly comply with dozens of regulation and laws... **no matter where you store your information.**"

Bex Huff
Bexhuff.com

First, UOA provides "security in depth." UOA utilizes the Oracle Database 11g as the content repository, which inherently provides significant security, disaster-recovery, and referential integrity advantages over hybrid, file-base content stores. With optional capabilities such as Database Vault and SecureFiles encryption, database-resident content remains safe even from a DBA or in cases where the physical media is compromised. The UOA application itself adds powerful security features to ensure all content is protected from unauthorized access and

inappropriate actions. In addition, an important new content locking mechanism known as a Lifecycle Lock virtually eliminates the possibility of content under lifecycle management being tampered with by users. Finally, UOA audits essentially every event that occurs within the system so that any attempts to breech the other layers of security can be quickly identified and dealt with appropriately.

Second, UOA works in conjunction with Oracle Universal Records Management (URM) to provide a centralized enterprise environment for managing retention and records policies. Unlike a number of competing products, URM utilizes an adapter model that was specifically designed to work with content in a wide range of repositories. As a result, URM's integration with UOA is seamless and efficient, intelligently dividing the workload between the two environments to meet realistic enterprise demands. In addition, URM's adapter model also allows it to work with a wide range of third-party repositories—providing centralized policy management for both consolidated content in UOA and federated content in its native repositories. Policy centralization means dramatic reductions in policy management overhead—making it much more likely that policies will be correctly defined, kept up to date, and actually used. Moreover, having well-defined and accessible policies has time and again been proven to be the best way to maintain compliance with often vaguely written regulations. If one can show that there is a reasonable policy in place and can prove it was enforced, the risk of being out of compliance—or being accused of shredding or other serious wrongdoing—can be greatly reduced. Finally, being able to get content under some level of centralized policy control regardless of where it resides—as opposed to waiting until a consolidation program can be fully executed—provides improved security and risk mitigation in a way that much more realistically reflects the situation in most enterprises.

Of course, when all different types of content can be consolidated in a single environment such as UOA—with true single instance storage, comprehensive content security, and extremely fast search—the cost and risk associated with eDiscovery can be dramatically reduced. In addition, the efficiency of content discovery in the UOA environment means preemptive risk assessment searches can be performed to help make better decisions about how to respond to litigation.

"Where I think they are doing it right is by not trying to force companies into replacing their current applications," he added. "I don't think anyone would be interested in another Exchange or another SharePoint just to implement better compliance software."

David Roe, Lead/architect, ECM Group, Ironworks Consulting

User Transparency

All the capabilities and attributes described thus far become completely meaningless if the solution does not get used. For that reason, one of the most important design criteria for UOA was to ensure that, when organizational policy so specified, migrated content could continue to be seamlessly accessed by users from the live server environments.

While there will certainly be situations where content and metadata are completely removed from live servers during the archiving process, it is expected that the more frequent use case will be the utilization of *stubbing*. This technique allows the applications associated with live servers to continue to manage the information *context*—access control, metadata attributes, folder hierarchies, etc.—but replaces the content in the live server with a link to the archived content in UOA.

Depending on the specific Oracle Archive Service used with UOA, the link can be supplanted with an abbreviated version of the content (e.g. the first few lines of an email or the first few paragraphs of a document) or the stub can apply only to an email attachment, while the email itself remains in the live server. Stubbing allows continued user access to the content but, under policy control, the content in the archive typically remains the "source of truth" and cannot be modified by the user.

The net result of this approach is that users do not have to change the way they work when archiving with UOA is implemented. In addition, they actually gain a very significant benefit in that the apparent capacity of their application can be greatly increased. Email users essentially see an "infinite inbox" and do not need to constantly deal with over-quota situations, while SharePoint users are not driven to create new sites and instances as their databases fill up and performance suffers. And this increase in apparent capacity provides security and risk mitigation benefits in that users are less likely to need to engage in the dangerous processes of archiving emails to desktop PSTs and NSFs or removing content to thumb drives and other unmanageable storage environments.

…And it's Open and Standards-Based

While this may seem like motherhood and apple pie, being open and standards-based is a true differentiator in the electronic archiving space since many of the competing products are anything but. Oracle Universal Online Archive gives the customer the choice between using a hot-pluggable solution that easily fits in their standard IT infrastructure, or a highly specialized, proprietary, siloed solution.

UOA is based on proven content management infrastructure technologies in use by thousands of customers in a wide variety of demanding enterprise applications. This content infrastructure leverages standard Oracle technology stack components and—unlike competing archive products which use complex and proprietary lash-ups of database and file system—UOA relies on an off-the-shelf Oracle Database as the complete content repository. This means infrastructure and repository management skills are either already on-board or readily available. It also means that the massive investments Oracle has made and is continuing to make in advancing its technologies can be easily leveraged by the UOA customer and a wide range of Oracle Database options and third-party add-ons "just work" with UOA.

The fact that UOA is based on the standard Oracle technology stack also means UOA is based on industry standards such as Java, Java EE, JMX, LDAP, SQL, JAX-RPC and WebDAV. UOA also provides an open environment for integrating with new content sources in the form of the SQL-based High Volume Importer. In addition, unlike many of its "closed" competitors, UOA supports a comprehensive set of Web services and Java APIs, a capable scripting mechanism, and a facility for event-based custom processing—all of which can be used by Oracle Consulting or professional services partners to extend UOA functionality and further integrate it into the customer's environment.

Finally, again unlike many archiving competitors that tend to be highly Microsoft-centric, both UOA and the initial Oracle Archive Services are available on a range of hardware platforms and operating systems. In addition, EAS works with Microsoft Exchange and Lotus Notes, as well as other email systems supporting IMAP/SMTP.

CONCLUSION

The changing landscape for electronic archiving has created a major opening for an entirely new class of product. Oracle is addressing that opening with the release of Oracle Universal Online Archive 10g. UOA leverages open and standards-based Oracle technology to provide the first truly universal archiving solution, delivering unmatched capabilities in terms of scalability and performance, security and risk mitigation, efficiency and cost control, and user transparency. With Universal Online Archive 10g, Oracle is truly raising the bar on electronic archiving to an entirely new level.

FOR MORE INFORMATION

For more information about Oracle Universal Online Archive, please visit: www.oracle.com/goto/uoa

Oracle Universal Online Archive
May 2008
Author: Rich Buchheim
Contributing Authors: Raghu R. Kodali

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com