Preservation Archiving Storage Technologies

Keith Rajecki; Oracle; Redwood Shores, California/United States of America

Abstract

This poster presentation is intended to provide a recommendation for standards implementation and best practices for a viable, cost effective, and reliable long term archiving storage system. This solution is based on a combination of storage technologies to provide long-term viability.

We will examine operating system platforms and filesystems for archive storage systems. We will examine the performance and archive capabilities of the filesystems including ZFS, Lustre, and SAM/QFS. We will also discuss the Hierarchical Storage System, or HSM, as a key software element of the archive. The HSM provides one of the key components that contributes to reliability by through data integrity checks and automated file migration. The HSM provides the ability to automate making multiples copies of files, auditing files for errors based on checksum, rejecting bad copies of files and making new copies based on the results of those audits. The HSM also provides the ability to read in an older file format and write-out a new file format thus migrating the format and application information required to ensure archival integrity of the stored content. The automation of these functions provides for improved performance and reduced operating costs.

Oracle's Sun StorageTek Storage Archive Manager (SAM) software provides the core functionality of the recommended Archive Storage Architecture. SAM provides policy based data classification and placement across a multitude of storage devices from high speed disk, low cost disk, or tape. SAM also simplifies data management by providing centralized meta-data. SAM is a self-protecting file system with continuous file integrity checks.

The digital content archive provides the content repository (or digital vault) within Sun's award winning Digital Asset Management Reference Architecture (DAM RA). DAM RA enables digital workflow and the content archive provides permanent access to digital content files. With SAM software, the files are stored, tracked, and retrieved based on the archival requirements. Files are seamlessly and transparently available to other services. SAM software creates virtually limitless capacity. Its scalability allows for continual growth throughout the archive with support for all data types. The policy based SAM software stores and manages data for compliance and noncompliance archives using a tiered storage approach with integrated disk and tape into a seamless storage solution, SAM software simplifies the archive storage. Allows you to automate data management policies based on file attributes. You can manage data according to the storage and access requirements of each user on the system and decide how data is grouped, copied, and accessed based on the needs of the application and the users. Helps you maximize return on investments by storing data on the media type appropriate for the life cycle of the data and simplifying system administration.

Introduction

Digital Repositories and Preservation Archiving have become increasingly important as institutions move away from the data management practices of deleting files and disposing of older books, magazines, and newspapers to a data retention mind set of preserving materials. Repositories help museums, national libraries, and educational institutions manage and capture institutional assets. This transformation in thinking has created a new set of challenges for these institutions financially, organizationally, and operationally. Past practices of deleting older files was an easy way to avoid the additional costs associated of increasing storage capacity. Removing older books and other materials from circulation in libraries made those shelves available for new materials.

Museums and national libraries are relying on digital repositories in order to preserve history . The ingest or digitization of historic photographs, films, maps, painting, and other items prone to deterioration allows for their long term preservation. The digitization and archival of these important historic artifacts is often times made available through the internet, enabling scholarly research, exhibitions, and other research and educational activities. Museums and national libraries have begun to leverage the digitization of these objects and technology to provide never before seen detail of manuscripts, paintings, and other artwork. Museums and libraries have even begun to develop three dimensional renderings of objects such as historic places, buildings, and people.

Higher Educational Institutions have turned to digital repositories for both academic and administrative purposes. A digital repository can hold a wide range of materials for a variety of purposes and users. It can support learning, research and administrative processes. Digital repositories provide a means of centralizing institutional data while providing greater levels of access and controls. Administrative digital repositories are primarily used for managing digital objects related to the administration of the institution such as student records, electronic invoices, and institutional policies. Academic repositories such as those in digital libraries manage books, papers, theses, and other works which can be digitized or were 'born digital'. 'Born Digital' objects are those that remain digital and are stored within the repository as they are created. Many organizations have begun digitizing their collections in an effort to provide open access.

The increase in the number of students taking online courses through the world wide web has also had an impact on the traditional service delivery of university libraries. These remote students are deprived of the vast majority of the resources traditional library have to offer. This has forced many university libraries to make their reference materials available online. This digitization effort has required a new business model for the university library. Libraries are no longer solely relying on the IT organization for support of their Integrated Library Systems (ILS). Librarians, IT Architects, and Systems Administrators now work collaboratively as part of communities to develop and implement institutional digital repositories and preservation archives.

Digital repositories are also a key component to preservation archiving. The digital repository provides a mechanism by which the digital objects can be stored and thereby preserved. While the

objective of the digital repository is typically related to collecting and disseminating digital objects, the objective of the preservation archive is focused on long term preservation of the digital objects. These preservation activities have expanded in scope to include audiovisual objects, datasets, presentations, text-based materials and research works. The preservation of these objects contributes to the re-use of digital content.

Digital Repository and Content Management Solutions

For the purposes of this solution architecture, digital repository refers to the system by which objects are stored for preservation archiving. There are a number of viable repository solutions available that provide the capability to store, manage, re-use and curate digital materials. Repository solutions range from the traditional commercially available content management systems to open-source alternatives. Repository solutions support a multitude of functions and can be internally developed or extended. The repositories must be sustainable and supportable in order for the underlying storage system to operate. The following repository solution was highlighted for its ability to integrate into a tiered storage architecture and support for interoperability.

Oracle Universal Content Management

Oracle Universal Content Management (UCM) is a comprehensive repository solution providing the most unified enterprise content management platform for a broad range of purposes including document management, Web content management, digital asset management, and records retention. The extensible functionality of the Oracle UCM allows organizations to build and complement business applications. The Oracle Universal Content Management solution enables organizations to build strategic enterprise content management infrastructure for content and applications that helps reduce costs, easily share content across the enterprise, minimize risk, automate and streamline manual processes, and consolidate multiple repositories onto a single platform for centralized management. Through user-friendly interfaces, roles-based authentication and security models, Oracle Universal Content Management empowers collaboratio and preservation in a highly secure environment.

Requirements and Design Criteria

Whether you are building a digital repository and preservation archive for historical preservation, to store data for business compliance, or meet the evolving business needs in higher education, a tiered storage architecture provides you with the most reliable and cost effective solution. Access and performance requirements are also important factors to consider when architecting your solution. Regulations often require that information be located and retrieved very quickly. If architected incorrectly, data searching and retrieval can be time consuming and costly. Traditional tape only archival methods simply can not meet the access requirements of many of today's repositories and long term archives. Likewise, storing all the data on disk requires greater administration and is more costly. The proposed preservation archiving solution provides a proven solution with a balance between disk and tape storage hardware to support long term archiving.

Compliance

Higher Educational Institutions are facing complex compliance regulations for student records, financial, human resources, and donor information. While there are a variety of archiving products available, many products address specific applications, such as e-mail, while others enable broad archival for unstructured data, such as documents and audio/video files. This solution provides a tiered storage architecture that can be leveraged across a multitude of software solutions with a centralized policy based active archive capability . This preservation archiving solution addresses compliance requirements for non-rewritable, non-erasable format, the ability to verify automatically the quality and accuracy of the recording process, serialize the original and duplicate units of storage media, store separately a duplicate copy, and providing audit trails

Manageability

As with any multi-component solution, management complexities becomes a concern. The Oracle preservation archiving solution design allows you to improve manageability of the technical infrastructure. The storage and archival functionality lead to improved manageability of the overall operational capabilities for the entire digital preservation organization. The software and hardware components take advantage of open standards allowing fewer operators to manage greater storage capacity.

The ability to automate data management policies based on file attributes, enable you to manage data according to the storage and access requirements of each user on the system and decide how data is grouped, copied, and accessed based on the needs of the application and the users. This hands off approach to managing digital assets allows you to better utilize your valuable IT resources. The strategic placement of digital assets according to policy allows you to improve storage utilization and maximize return on investments by storing data on the most appropriate media type for the life cycle of the data while simplifying system administration.

Scalability

The ability for the entire technology infrastructure to expand and contract as services and storage requirements increase and decrease is increasingly important. The ability to dynamically reassign compute resources and storage as an organization moves from ingest to access drastically improves utilization. The Oracle preservation archiving solution provides a highly scalable solution that takes advantage of server and storage virtualization to allocate system resources as needed. The components of the preservation archiving solution also enable additional resources to be added while the systems are on-line, resulting in live upgrades and transparent migrations.

Security

Digital archiving security requirements reflect concerns for long term access and preservation. Digital repositories and preservation archiving share the same security requirements as most enterprise applications with the added complexity of distributed object level policy based access. When long-term preservation spans several decades, generations, or centuries, the security of digital objects becomes critical. Open, standards-based

access control, single sign-on and federation services are required in order to enable long term preservation while helping to control costs and minimize the risks of security obsolescence. The security solution must provide integrated user provisioning and identity synchronization services for securely managing identity profiles and permissions throughout the entire identity lifecycle.

Interoperability

Interoperability is the ability of software and hardware components to be functionally and logically interchangeable by virtue of their having been implemented in accordance with open standards. Interoperability of the digital repository and preservation archiving solution for institutional repositories is a complex problem. Interoperability is typically addressed between a specific software or hardware component. In this preservation archiving solution, interoperability is achieved among multiple software and hardware components through the use of open standards. In this model, different services and components can communicate with each other through open interfaces, and clients can interact with them in an equivalent manner. When repositories and digital objects are created in this manner, the overall effect can be a federation of repositories that aggregate content with very different attributes, but that can be treated in the same manner due to their shared interface definitions.

Logical Architecture

The primary functions of a digital repository and preservation archive is to support the acquisition, organization, preservation, and access to digital objects regardless of their format. The creation of functional requirements and identification of key policy issues for the digital repository are essential to building the appropriate architecture. The functional requirements for the Oracle preservation archiving solution are based on the Open Archival Information System (OAIS) Reference Model[1] and include ingestion of digitized and born-digital materials, metadata generation, data management, archival storage, access, preservation planning, and administration.

An OAIS-compliant repository is an organization of people and systems, which has accepted the responsibility to preserve information and make it available for a designated community. OAIS provides a conceptual framework to define the core requirements of the digital repository.

Preservation Archiving Solution Components

Open Storage

Open Storage refers to the systems built with an open architecture using industry-standard hardware. An open architecture allows the most flexible selection of the hardware and software components to best meet preservation requirements. A closed storage environment restricts the available technical componenets such as disk drives, controllers, and proprietary software resulting in higher costs and limitations to extending functionality. Long term preservation is directly dependant on the long term viability of the architecture and associated software components. Open standard solutions offer the most viable long term option with open access and community based adoption and support.

Hierarchical Storage System – Tiered Storage Solution

The Hierarchical Storage System, or HSM, is a key software element of the archive. The HSM provides one of the key components that contributes to reliability through data integrity checks and automated file migration. The HSM provides the ability to automate making multiples copies of files, auditing files for errors based on checksum, rejecting bad copies of files and making new copies based on the results of those audits. The HSM also provides the ability to read in an older file format and write-out a new file format thus migrating the format and application information required to ensure archival integrity of the stored content. The automation of these functions provides for improved performance and reduced operating costs.

The Oracle's Sun StorageTek Storage Archive Manager (SAM) software provides the core functionality of the recommended archive solution architecture. SAM provides policy based data classification and placement across a multitude of tiered storage devices from high speed disk, low cost disk, or tape. SAM also simplifies data management by providing centralized metadata. SAM is a self-protecting file system with continuous file integrity checks.

Oracle's Sun Storage Archive Manager addresses compliance by applying policies to files, copying and moving files based on those policies and maintaining audit information on files. SAM indexes files for searchability and writes multiple copies to specific media based on the compliance retention policies.

Designed to help address the most stringent requirements for electronic storage media retention and protection, Oracle's Sun Storage Archive Manager Software provides compliance-enabling features for authenticity, integrity, ready access, and security.

Key Benefits of Storage Archive Manager Software

- Enforces retention policies at the storage level
- Software-controlled disks implement non-rewritable and non-erasable files
- Enables data integrity checking
- Provides flexible storage configurations

Storage Archive Manager software supports write-once read-many (WORM) files that are nonrewritable and nonerasable. Robust security features such as audit logs, user authentication, and access controls, combine to help safeguard the integrity of the digital information. In addition, the critical metadata attributes cannot be changed.

Infinite Archive System

The Infinite Archive System provides a pre-installed and configured hierarchical storage solution for digital repository and preservation archiving. The Infinite Archive solution scales easily providing petabyte scalability. The Infinite Archive System provides a three tier storage system consisting of the following components.

- Working Data Set, Online, on fast Fibre Channel (FC) Storage (Oracle's Sun StorageTek Storage Array)
- First Level Archive, Midline, high capacity SATA storage (Oracle's Sun StorageTek Storage Array)
- Second Level Archive, Nearline, high-performance tape storage (Oracle's Sun StorageTek Modular Library System)

- Remote Archive provides a further level of archiving, with remote off-site storage of archived tapes

The Inifinite Archive System takes advantage of Oracle's Sun SAM/QFS software to manage the placement and retention of the data to ensure the most cost effective use of your storage resources.

Modular Tape Library System

The Modular Tape Library Systems provide scalable solutions up to 56 petabytes and 70,000 tape slots. This makes them the ideal platform for tape archives for off-line or dark archives. The Virtual Tape Libraries VTL enable tape consolidation with the low cost, cartridge removability, and long-term retention capabilities. This tiered storage solution is managed by policies on the VTL, so the overall solution reduces your labor costs for virtual and physical tape management.

The Modular Library Systems provide greater levels of reliability ensuring access to your data. The robotic mechanism maintains reliability regardless of the number of expansion modules and helps to increase the stability and predictability of backups. Redundant, hot-swappable components, such as power supplies and fans, minimize disruption. An advanced digital vision system automatically calibrates the library to reduce wear and tear on the cartridge, drive, and robot. Dynamic worldwide naming and firmware code uploads eliminate single points of failure.

Conclusion

The Oracle preservation archiving solution is an ideal framework for any institution looking to deploy a digital repository and preservation archive. The software and hardware components of this architecture support the long-term maintenance of digital objects by storing the digital object information along with the object itself. These components are organized into a system that supports not only long-term viability of the repository, but also the digital information for which it has responsibility. The components can be upgraded and changed out independently with overlapping technology life spans to support the long-term requirements of the repository. The policy based distribution of digital objects throughout the repository further supports fiscal responsibility and sustainability. Incorporating industry accepted conventions and standards such as the OAIS model ensures the ongoing management, access, and security of deposited materials.

The Oracle approach of working with leading digital repositories to integrate directly the digital repository software into storage software enables trusted digital repositories to manage the integrity and authenticity of records. This provides streamlined mechanisms to validate assertions about trustworthiness and provide the preservation processes that implement the required control and management capabilities. There are multiple technologies available today that can be used to build a digital repository capable of maintaining the authenticity and integrity of ingested objects.

The components outlined in the preservation archiving solution can be combined into a single system to validate, store, audit, secure, and preserve digital objects. The tiered storage architecture provides the most cost effective solution for object repositories and long-term archives while supporting scalability. The extent at which those storage tiers are deployed is dependant on the access patterns and archival policies. Although this architecture is not intended to cover all business requirements, it can be applied in a modular approach to address specific business requirements where one or more tiers may not be feasible due to business or technical requirements.

Key Benefits

The preservation archiving solution identifies key system components and processes that are required to achieve high service levels and scalability. It provides the following major benefits to educational institutions:

- Higher service levels — The architecture is designed to optimize service levels with redundant components and automated failover using storage virtualization and cluster technologies.
- Reduced cost — Virtualization technologies enable consolidated solutions with higher resource utilization and tiered storage helps customers avoid overprovisioning or underprovisioning their systems. Best practices for management can also reduce the cost of maintaining the solution environment.
- Faster time to delivery — Accelerates deployment by providing proven and tested configurations with simplified installation to be up and running almost immediately.
- Reduced risk — Validated hardware and software configurations greatly reduce the risk of unforeseen problems in a production implementation.

Reference Preparation

Note that for references a tab should be placed between the reference number and information; a hanging tab is set to .15 inches. Samples for references styles are shown below. Reference [1] style should be used for books, Reference [2] style should be used for Journals, and Reference [3] style should be used for Proceedings.

After your references, apply the template tag “Title”. This tag will align your 2 columns to an equal depth.

References

[1] Consultative Committee for Space Data Systems (2002). "Reference Model for an Open Archival Information System (OAIS)". CCSDS 650.0-R-1 – Blue Book. Available at: http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

Author Biography

Keith Rajecki is Senior Director of Solutions for Global Public Sector, Education, and Research at Oracle. Keith Rajecki joined Sun Microsystems as a Solutions Architect with Global Government, Education, and Healthcare in 2007 to lead the development of education industry solutions. In this role Mr. Rajecki worked to promote the development, adoption, and integration of Sun technologies to create strategic technology solutions that meet industry demands.

Mr. Rajecki possesses over 14 years of innovative experience. Mr. Rajecki completed both his Bachelor and Master of Science in Computer Information Systems at Golden Gate University.