THE CLIPPER GROUP
Navigator™
Navigating Information Technology Horizons℠
Published Since 1993 Report #TCG2004059 July 12, 2004

Capture, Index, Federate = Ready, Set, Go — HP's RISS Solution for an Active Business Archive

Analyst: Anne MacFarland

Management Summary

The value of business information often extends beyond its immediate use. This is easiest to see in E-mail, which has become the poster child for the need to archive enterprise data. As the enterprise's communications mode of choice, it is an obvious source of angst and focus of regulation. **E-mail has brought us prodigious gains in productivity, but in the process, the ability to find out what has gone on across an enterprise of any size has been seriously impaired.** For many of us, our mailbox quota defines our event horizon, and any deeper history usually is selected to our biases. Without an archiving program, e-mail does not leave a reliable and readily usable form of business process documentation. The copy of e-mail that is captured as part of periodic back-ups does not assure that the documentation is complete or accurate, and branch office environments and mobile workers are often inadequately served. Backups are made for restoration, not preservation, with files often in an unsearchable, proprietary format. This does not satisfy government regulations – and it does not satisfy the needs of the enterprise.

The value of data automatically captured by an archiving process is great – but it is further enhanced by the generation of *finding aids,* like indexes. The sender, recipient, date, and subject fields of e-mail are fine as filters, but the meat of the message is its content. If you are looking for malfeasance, it probably won't be listed in the subject heading. If you are looking for expertise, it often must be teased out of a confluence of key words – words that collectively are known as an *IT taxonomy*. Preemptive indexing by these taxonomies, as the e-mail is transferred to the archive, can accelerate the process of *finding*. This finding process must be fast enough to be useful to the enterprise. A scalable, parallelizable architecture is required.

Compliance regulations have driven many enterprises to long-term archiving, but they want the process to be integrated, automated, and as painless as possible. HP's *StorageWorks Reference Information Storage System (RISS)* provides a comprehensive, integrated, federated solution. It includes the hardware, software, services, and support to get the job done. RISS is optimized for e-mail, Microsoft *Office* and Adobe *PDF* documents, and partners for IM archiving. HP has developed connectors to allow other kinds of data to be added to RISS, as each enterprise desires. In time, HP will release a SDK with open APIs to allow integration for applications and data types such as rich media, voice mail, and document management. This archiving will give the enterprise a granularity of knowledge of internal operations never before possible. If this sounds like what you have been waiting for, read on for details.

IN THIS ISSUE

Managing Reference Data

To be demonstrably compliant and to use business information for enterprise-self knowledge in various ways, you need not only to keep the data, but also to be able to find the right information in a reasonable amount of time. This managing data for reference is not just about managing the devices on which it is stored. It is about insuring the integrity and authenticity of the data by assuring it is non-erasable and non-rewritable. It is about preserving the context of the data by culling relevant metadata elements from the application that created or captured it[1]. It is about the need to find the right stuff in a vast sea of information (using search), and it is about parallelizing the *find* process to deliver the results promptly. Finally, enterprises of all sizes need to be able to do this simply and affordably, without wasting capacity or search effort on duplicate data.

How HP's StorageWorks RISS Does Archiving

Aspects of Capture

Compliance Mode vs. Selective Archiving

HP's Reference Information Storage System (RISS) supports both *compliance mode* (archive everything) and *selective archiving.* In both modes, duplicate attachments are discovered and only one copy stored (with pointers to each relevant e-mail). In *compliance mode,* e-mail is captured immediately after it clears the anti-virus and SPAM filters at the network edge – before it hits an individual's mailbox. This removes the opportunity for a recipient to tamper with documentation of his or her actions. RISS supports Microsoft *Exchange* fully. At launch, Lotus *Domino* and *Sendmail* are supported in compliance mode, as their client architectures are not supported[2].

Selective archiving can be used for unregulated business records. Usually policies governing record retention are set within the application generating or capturing them. What unregulated business records an enterprise keeps depends on the nature of the business, its need to use *historic* data, and its tolerance for the liability risk that old records can present.

RISS is connected to end-user applications through connectors or plug-ins. Plug-ins are available for switched and IP telephony, office productivity, AutoCAD, medical imaging, and PeopleSoft, Siebel, Oracle and SAP applications. An SDK is being developed that will allow automated data capture from more data sources, using APIs and/or Web Services. (See box, above.)

Organization: Repositories and Domains

When RISS is set up, the customer sets up multiple *repositories* to reflect the enterprise structure, usually in a tree architecture like an LDAP directory, with similar rules of inheritance. Needs for corporate confidentiality are addressed by the limits of these repositories.

Each repository has a retention period, and if the time stamp of an item has expired, the document is automatically destroyed. The time stamp of an object may be extended or *frozen* (extended indefinitely), but not curtailed.

By the end of the year, RISS will allow content-based retention actions, such as extending the retention of any document containing the name of a nefarious partner, for example, for, say, seven years. Alas, this will often be useful.

Domains are another way RISS partitions data. Domains are used to insure data segregation, which is required by some regulations. Domains segregate classified and unclassified

Document Types Supported

- Microsoft *Word, Excel, PowerPoint, Access, Outlook*
- Adobe *Acrobat* (.pdf)
- Rich Text Format (.rtf)
- HTML
- Text Files (.txt)
- ASCII files
- TNEF (Transport Neutral Encapsulation Font)

Other file types cannot be automatically indexed, but can be retrieved by name.

[1] For more about repurposing IT data for reference, see **The Clipper Group Explorer** entitled *Reference Data Concepts - What IT Folks Should Learn from Libraries and Archives*, dated May 18, 2004, available at http://www.clipper.com/research/TCG4047.pdf.

[2] In time, they will be.

data, or data from different business initiatives, on separate cells. Such segregation can be useful when a part of the business is spun off, taking their archives with them.

Security

RISS has many layers of security to protect the integrity of the data. There is a firewall within the RISS system, and NAT devices protect against packet sniffing and packet snooping. Each smart storage cell (about which more later) runs a secure, locked-down operating system. And, finally, there are digital signatures and fingerprints on each data object. Audit trails can be kept of access to each data object.

Replication gives another classic form of data security. There is constant mirroring at the data object level between cells in a RISS environment. Additional mirroring for multiple redundancies is possible, if required. By the end of the year, remote mirroring will provide the protection of distance.

Capture Throughput

RISS captures and indexes documents at a rate of 40 documents per second. RISS can process up to 2 million e-mail messages a day. Large enterprises may need to use multiple RISS repositories for different lines of business, which will probably have different taxonomy requirements.

Indexing and Search

The need for quick access demands indexing to taxonomies of key words as information is added to RISS. This is full-content indexing, including attachments, which can be cracked, broken into separate pieces if they are in a proprietary format, and transformed so that all information is accessible. The indexing is done to the enterprise's specific taxonomic requirements. HP RISS allows users to evolve taxonomies and can re-index - in the background - should that become necessary.

As well as indexing, HP's RISS technology has the ability to go into the applications of the RISS partners (listed in the box on page 4) and harvest metadata about information, which it then keeps in a database. In this way, RISS preserves the context of the information that defines where the information is relevant.[3] This information augments the file system and facilitates more efficient queries.

RISS supports a number of extended search capabilities in addition to the standard application-based search on headings, full-text search, and Boolean fine-tuning. RISS can search for words in a specified degree of adjacency (such as *within five words*), and can search for a *sounds-like* term, generating spelling alternatives.

Federation

RISS is HP's first implementation of storage grid architecture. The federation of grid architecture changes the traditional paradigm of cost and scale and is the basis for ensuring data availability on a large scale. The RISS combination of grid architecture and smart storage cells is key to assuring a sub-three second response time, no matter how many objects are searched, and one archive can scale to 67 TB[4]. Up to 900 simultaneous retrievals are supported by the 4 TB base system.

The RISS Smart Storage Cell

The cell is the basic functional unit of the RISS system. At present, this cell consists of a 2.4 GHz Pentium 4 processor with 240 GB of storage. Each cell acts as its own storage controller. The hardware, both processors and storage, will evolve over time. RISS supports heterogeneity so the customer can use multiple generations of cells in the same enclosure. In the future, HP is considering smart cells using other media – magneto-optical or tape.

Metaservers

There is a sub-set of cells, called *metaservers,* which act as an out-of-band controller for the whole federation of cells. They delegate queries and aggregate the results for presentation to the user. If one of these metaservers fails, the software can select another cell that will migrate its data elsewhere and take on the metaserver role. Because of the transactional nature of grid architecture, this automated fail-over is transparent to the end user. So the smart

---

[3] See **The Clipper Group Explorer,** ***Reference Data Concepts - What IT Folks Should Learn from Libraries and Archives*** dated May 18, 2004, at http://www.clipper.com/research/TCG2004047.pdf.

[4] A very large corporation usually consists of autonomous business units, which could each have a separate archive.

RISS/ILM Partners

E-Mail Archiving	Database Archiving	HSM
CommVault	Princeton Softech	Camino Soft
IXOS		CommVault
KVS		GRAU Data Storage
Legato	**Rich Media**	Pegasus
Mirapoint	ADIC	Qstar
E-Mail Policy Management	**Disk-Based Backup**	**IM Archiving**
Orchestria	Avamar	Akonix
		IMLogic

cells are self-healing, ands more capacity can be added as the need arises.

Management

The RISS management software is known as *Reference Information Manager,* or *RIM.* As well as system monitoring and management, RIM includes the following:

- Access management
- Security management
- Content indexing and storage
- Finger printing
- Digital signatures
- Data authentication
- Space management
- Data retention management
- Load balancing
- Routing
- Failover

How to Buy

The base system, 4 TB of usable storage (8 TB raw), occupies two full racks, and costs $425,000 (list price). More smart cells can be added in drawers of 1.2 TB (usable) at a cost of $98,000. Because of the grid architecture, its capability increases as it grows – and the response time for searches does not. One archive can scale to 67 TB. The power and environmental requirements mandate its placement in a data center.

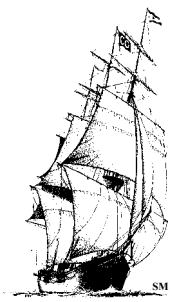
RISS Services

HP has developed services to accelerate and ease the RISS deployment process.

- *Design and Integration Services* - This includes the design of the RISS solution, the implementation of the solution (install and setup) and the integration into the application.
- *Legacy Data Services* - Many early RISS customers, realizing the value of a searchable archive, want to use legacy back-up data for forensics, and therefore need to repopulate the archive with data going back several years – often only found on back-up tapes. Since multiple versions, and perhaps even multiple backup applications are involved, extracting the information from these tapes into a usable, searchable format is not a simple process. HP has a service practice in Atlanta, focused specifically on this problem, which can accomplish the migration expeditiously.
- *IT Electronic Vaulting Services* - HP Hosted Services provides electronic vaulting to a remote location to those who request it. This is a clear indication of the high value given to this appliance by customers who use it.

Conclusion

HP has looked carefully at what is needed to create an active business archive, optimized for e-mail, which is complete, incremental to buy, easy to use, and extensible. If the need for a secure, comprehensive and searchable e-mail archive is what keeps you up at night, consider the HP RISS system.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Enterprise Architectures and Infrastructure Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 28. (Please dial "1-28" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.