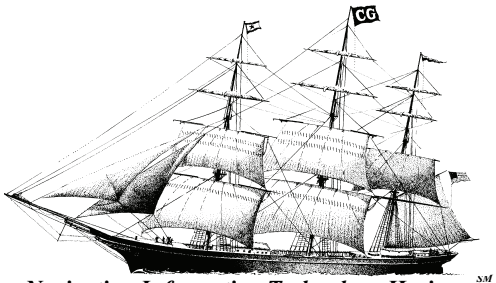
Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER



Navigating Information Technology Horizons℠

New in 2007 Report #TCG2007019 Updated March 31, 2007

Archiving — Choosing the Right Architecture

Second of a Two-Part Series

Analyst: Dianne McAdam

Management Summary

Backup and archiving are two different processes with two different end results. In *Part 1*[1] of this paper, we discussed the differences between backups and archives. Yet, many still use the terms interchangeably. Some enterprises do continue to backup data daily that should be archived only once. That costs time and money.

Consider email systems. We all need them; we all rely on them to get our work done; we all, at some point, hate them. Why? For too many of us, we discover – at the worst possible time – that we have exceeded our stingy mail quota and can no longer receive any incoming mail. At this point, all of our other productive work stops until we delete emails or move them to personal folders. Several months later, the same quota problem comes back to haunt us and again we spend hours cleaning up our inbox. Why do we continue to act like this? Did many of us fail the "Martha Stewart" class in cleaning our inbox? The answer is simple – email has become our personal filing system of choice. We now communicate through email in many different ways - from a laptop, a *Blackberry*, a home PC, or someone else's laptop. So, if we keep all email on the email server – and delete nothing – then we can always retrieve a message from any device, from any location at any time. This practice of saving all emails on the server costs money – just ask any email administrator who keeps adding more and more storage to the email server and struggles to tune the system to deliver good performance.

So what's the answer? Impose stricter email quotas? Hardly – that just makes people "hide' emails in personal folders that may never get backed up. Many enterprises report that over 70% of business transactions are initiated and concluded via email. These emails must be protected as a permanent record of those agreements. The answer is simple – we need an archiving solution that can safely store emails while allowing us to retrieve them anywhere, at any time.

We might solve the email problem by implementing an email archiving solution that solves today's problem. However, we also need to think about future problems that might be introduced by installing a *today-focused* solution. For example, the solution implemented today that can handle today's email traffic may perform poorly as emails continue to grow in size and number.

While we have just focused on the email management problem, there are numerous solutions in the marketplace today to archive different types of data. They are needed to store data such as human resources personnel data, intellectual property in the form of engineering diagrams or pharmaceutical studies, and instant messages. **Understanding the underlying architecture of an archiving solution can ensure that you pick the right solution for solving today's problems while continuing to support future archiving needs.**

IN THIS ISSUE

[1] See the issue of Clipper Notes dated February 1, 2007, entitled *Archiving — Do You Need It? (First of a Two-Part Series)*, and available at http://clipper.com/research/TCG2007018.pdf.

Traditional Archiving Architecture

Let's break down an archiving solution into its common components. The structure of a classic or traditional archiving solution consists of many different servers, storage, and software. This solution may include the following.

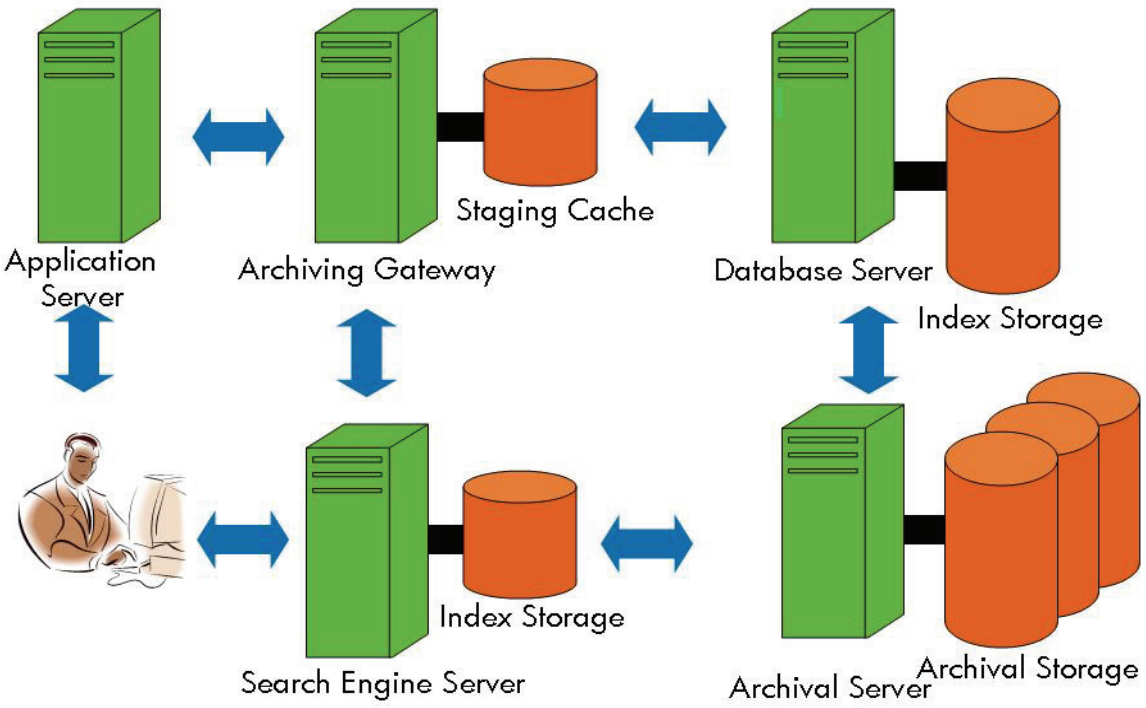
- **Applications server** (such as an email server);
- **Archiving middleware software** (running on its own server), which receives the messages;
- **Database and search engine servers, software, and storage**; and
- **Content Addressable Storage (CAS) software, server, and storage**.

The applications, archiving, database, search engine, and CAS software can run on different platforms with different operating systems connected to different storage systems. They may support different protocols and use different interfaces to communicate with each other.

Let's follow an email message through one of these typical archiving structures.

1. The email messages is sent to the archive gateway to begin the process of categorizing the message.
2. The archive gateway must have sufficient disk storage to hold the messages until the database server processes it.
3. The database server receives the messages and indexes it under several specific categories. For example, the database server will look at the properties of the messages and determine when it was sent, who sent it, who it was sent to, who was carbon copied, who was blind copied, and the subject of the email. This information is put into an index that supports future searches on these various fields.
4. The messages may also be sent to a search engine server. (This process may not be available in all systems.) Here, the actual content of the message is examined. The search engine then updates the index file, which later can be searched to retrieve emails based on text strings within the message or the attachment.
5. The database server (and the search engine server, if available), send the messages to the archival server to be stored.
6. The message is finally written to the archival storage.

Exhibit 1 — Typical Archiving Architecture



Source: HP. Used with permission.

This process sounds relatively straightforward. However, each one of the components can be a potential bottleneck when faced with a large volume of email. Consider a typical large enterprise, Company XYZ, with about 40,000 employees.

- The average email user at Company XYZ sends and receives about 60 to 65 emails a day.
- The average email is about 650 KB in size. Some emails have large attachments (6 MBs or greater in size), while other emails are very small, with short replies.
- On any given day, Company XYZ will have to manage about 2,500,000 messages or about 29 messages a second.

The archiving gateway must be able to process, on average, about 29 emails every second. But email message traffic has peaks and valleys as workers start and end their workday. So, the peak message traffic for this company can be two or three times greater (say, 58 to 87 messages per second) during high-processing periods. Gateway servers are the first components that must be able to handle large volumes and must be clustered and load balanced to provide high availability during periods of high-demand.

Next, the database server needs to be properly sized to index the incoming messages. Each message field, such as *From*, *To*, *CC*, *BCC*, *Date*, and *Subject*, requires its own index. If the database server only indexes on the four most common fields (*To*, *From*, *Data*, and *Subject*), the database will be updated by 10,000,000 table updates every day. Indexing on more fields will result in larger indices and larger number of table updates. A high number of table updates can cause table splits and requires that database administrators monitor the system - for performance and tuning.

While the database server is indexing on standard fields, the search engine server has an even more processor-intensive job. Each email must be scanned for common words and phrases that must be indexed. Attachments in different formats, such as *Word*, *Excel*, or *PowerPoint*, must also be scanned for common words.

Finally, the message is transmitted to the archive server, which stores the message for its specified retention period. This message may be stored in a form that cannot be modified, if dictated by policies. If the archiving server does not detect and reduce multiple occurrences of the same inbound email (which can occur when the same email is sent to numerous people) then the amount of archive storage required might grow by 1.6 GBs per day (2,500,000 messages x 650 KB = 1.6 GBs). This is good news for a vendor selling storage but can be very expensive for the enterprise.

Security

Performance (i.e., minimizing bottlenecks) is one important aspect of an archiving solution, but security is also important. Companies would be liable if a financial trader was bilking clients, for example, and was able to edit archived transactions to cover up unscrupulous behavior. Many archival storage products support policies that prevent messages from being modified or deleted before their expiration date. However, these messages are vulnerable before they reach the final archive store.

If we review the path of the object through our previous architecture, there are numerous points where security can be breeched. In fact, the first vulnerability is when the message still resides on the email server, where it can be tampered with or deleted.

The next point is when the message is temporarily stored on the staging cache for the archival gateway. Here the content can be tampered with before it is sent to the database server.

Next, the tables within the database server can be deleted, which removes pointers to messages, making the messages impossible to find. And indices within the search engine can be corrupted with the same effect – making messages impossible to locate.

The missing element in these traditional archive solutions is a unified platform that provides security throughout the process while scaling to meet future demands. Thus, you need an integrated architecture.

Consider an Integrated Architecture

While the traditional archiving solution described above may meet the performance needs for today's workload, it may not meet tomorrow's needs and can introduce security risks. More integrated solutions now are available that include all of the required components under one secure umbrella. These solutions are based on a grid architecture and add more processing power when additional storage is added. This grid approach allows solutions to scale as

the number of items that are archived continues to grow. And since the components reside under one "cover", data is secure from the moment it enters the system until it is stored on disk.

Enterprises that are experiencing large growth in the amount of data that needs to be archived - or expect that internal policies or external regulations will require them to archive large amounts of data - should evaluate integrated solutions over point solutions. Integrated solutions will not only scale to meet future performance needs but save money as well. Point solutions may initially cost less. As additional servers, storage, and software licenses are added, the cost for a point solution may greatly exceed that of an integrated system.

Questions to Ask

It can be difficult to choose the right archiving solution for your environment. The following questions may help to narrow down the choices.

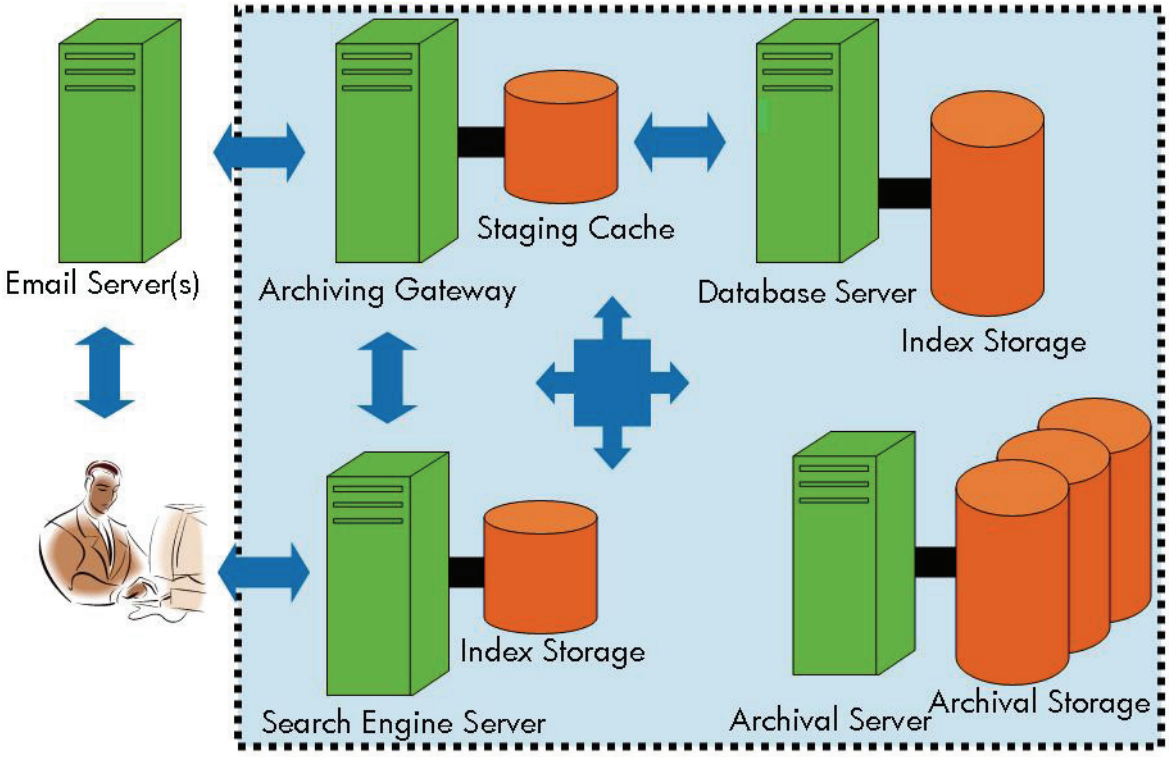
Hardware

1. *Does the archive solution use industry-standard hardware? How many and what types of servers are required? Has the solution been built to vendor specifications? What is the cost of this hardware? If the hardware fails, can it be easily and quickly replaced? What is the vendor's escalation procedure when hardware fails?*
2. *What storage devices are supported? Are WORM devices supported? WORM support is critical for data that must be stored in an unmodified format.*
3. *Can additional storage be added nondisruptively?*

Operations and Support

4. *Does the product integrate with existing backup/restore applications? Archiving solutions should integrate with and not replace existing backup processes. Are specific APIs available to make backups easy to implement?*
5. *How long will it take to implement the solution? Is it customer-installable? If not, are there additional charges for installation? Are professional services required for implementtation? What is the cost of this service?*
6. *How often is the software updated? Can the updates be installed nondisruptively?*

Exhibit 2 — Typical Archiving Architecture



Source: HP. Used with permission.

7. *Is training required or recommended? What is the cost of that training?*
8. *What support is available from the vendor? Are support costs included in the purchase price? If this is a multi-vendor solution, will one vendor take ownership for problem determination? Who gets the first call when problems occur?*
9. *What is the warranty period for the hardware and software? What is the cost for maintenance in the future?*

Supported Environments

10. *Which email servers (for example, Microsoft Exchange, Lotus Domino) are supported? Which email clients (for example, Outlook 2000, Outlook 2003)? Are these solutions certified by the email vendors?*
11. *What other archive applications are supported (e.g., Princeton Softech database archiving, IM Logic instant messaging)?*
12. *Are interfaces (APIs) available that allow file and print services to direct their output to the solution? Are APIs available for database applications?*
13. *Does the solution support end users who need to access data remotely through various means, such as Web-based access?*
14. *Is support available for end users who regularly use home computers to read their email? Is support available for remote users who normally access email "on the road"?*
15. *What applications will be supported in the future?*

Architecture

16. *Are agents required on the application servers, such as the email server? Is an agent required on the client? How are agents installed? Can they be installed remotely?*
17. *How is the solution implemented? Will the solution consist of various point solutions? If so, have the vendors certified each other's solutions?*

Performance, Scalability, and Availability

18. *What is the performance impact of the archive solution on the application?*
19. *How many end users can the archiving solution support?*
20. *How many messages/files can be stored per hour? What are the maximum theoretical rates to process data for each component – archival gateway, database engine, search engine, archival store?*
21. *What has been the experience of other data centers that have installed the system? How much data are they ingesting per hour and per day? If the amount of data increases, can the solution adequately support the additional growth? Do you need to install more servers?*
22. *What is the upgrade plan? What is the cost of these additional servers (and software licenses)? Given your current growth rates, what will the solution cost the first year, second year… and fifth year?*
23. *How many files/messages can it store? Thousands? A hundred thousand? Millions? Billions?*
24. *Indices are created when data is stored to enable quick access. Where are the indices stored? How much storage is required to store them? If the index is corrupted, can it be rebuilt?*
25. *How quickly can the software search through, say, 100,000 messages to find all of the messages sent by "Bob Smith"? When the number of stored messages increases from 100,000 to ten million or 100 million, is search performance degraded significantly? How quickly can the solution find all of the email messages, instant messages, Excel spreadsheets, and other files that were used to create the last financial quarterly statement?*
26. *How many application servers can be supported by the archiving system? How much storage can be supported?*
27. *Is the archive engine on a clustered server? If the server fails and it is not clustered, what is the impact on the email application? Can the application continue processing? How do you recover operations when the failed archive engine is replaced or repaired?*
28. *Does the archiving solution compress data to reduce storage requirements? If so, what compression rates can be expected? What compression rates have other customers experienced? Are other data reduction techniques, commonly called "data deduplication", used?*
29. *Does the solution detect and remove multiple copies of the same file or message?* This is usually called "single instance store". *Does the solution detect and remove multiple copies of the same attachments?* Storing only one instance of the same message or attachment can significantly reduce storage requirements. *Are other data reduction techniques, commonly called data deduplication, used?*

Management and Administration

30. *How is the system managed? Can it be managed remotely?*
31. *Can administrators be assigned different levels of access?*
32. *What types of reports, if any, are available?*
33. *Are alerts generated when certain error thresholds are reached?*
34. *Are logs or reports generated that track which employees have retrieved messages from archives?*

Security

35. *Is data encrypted on the archive store?*
36. *How are administrators authenticated into the system?*
37. *Can administrators be granted different levels of access to allow or restrict access to data?*
38. *How are the indices protected? Can they be modified/deleted by disgruntled employees?*
39. *What levels of security are provided? Can employees view and retrieve all data in the archives or are they restricted to viewing only certain messages?*
40. *Who has the authority to delete data in the archives?*

Policies

41. *How granular are the policies? Can different policies be set for different departments or for different groups of workers?*
42. *Can policies be set to determine when data is sent to the archive? By date of last access?*
43. *For regulated industries, can policies be established that require managers to review employees' emails?*
44. *Can retention periods be extended when necessary? This is important when a discovery or audit process is taking place and there is a danger that data required for the discovery or audit process may expire before the process is completed.*

Email End-User Interface

45. *When a message is archived, how does it appear to the end user? Is the message flagged with an icon to signify that it is archived?*
46. *When an end user needs to search the archive for messages, is the search menu easy to use?*
47. *Can end users be restricted to viewing and retrieving only certain messages?*

Replication to a Remote Location

48. *Does the solution have policies to route data both locally and to a remote location? If not, can existing replication products be used to replicate the contents of the archive to a secure, remote location?* Archival storage must be protected at a remote location for disaster-recovery purposes.
49. *What are the distance limitations?*
50. *Will replication impact performance?*

Pricing

51. *How is the product priced? By the capacity of the archive? By the number of application servers supported?*
52. *What does the solution cost today? What will the solution cost in several years when more and more data is archived?*

Market Acceptance

53. *How many solutions have been sold?*
54. *How many are installed in production environments? Which applications are in production?*
55. *Is the product a recent addition to the market or has it been available for months or years?*

Future Concerns

56. *If the email system is upgraded to a newer release level (Exchange 2000 to Exchange 2003, for example), are any changes required to the archive system? When the email system is upgraded to the latest version two years from now, will the archive vendor support these new releases? What are the archive vendor's plans to maintain compatibility with future releases?*
57. *The same questions can be asked about other applications currently supported. What are the vendors plans to maintain compatibility with future release?*
58. *Is the data stored in a "future-proof" manner? If an audit occurs seven years from now, will you be able to read the data archived today?*
59. *What plans are on the vendor's roadmap? What enhancements are planned? What are their plans to deliver higher performance and greater capacity?*

Conclusion

Many archiving solutions are available today. Choosing the right one can be a difficult decision. So, it is critical to evaluate them very carefully. **Archiving data is a long-term requirement. It requires a solution that will meet the needs of your enterprise today and into the future. Choose wisely.**

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

Dianne McAdam **is Director of Enterprise Information Assurance for the Clipper Group**. She brings over three decades of experience as a data center director, educator, technical programmer, systems engineer, and manager for industry-leading vendors. Dianne has held the position of senior analyst at Data Mobility Group and at Illuminata. Before that, she was a technical presentation specialist at EMC's Executive Briefing Center. At Hitachi Data Systems, she served as performance and capacity planning systems engineer and as a systems engineering manager. She also worked at StorageTek as a virtual tape and disk specialist; at Sun Microsystems, as an enterprise storage specialist; and at several large corporations as technical services directors. Dianne earned a Bachelor's and Master's degree in mathematics from Hofstra University in New York.

- ***Reach Dianne McAdam via e-mail at dianne.mcadam@clipper.com or at 781-235-0085 Ext. 212. (Please dial "212" 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.