

WHITE PAPER

Email Archiving and Continuity

Roaring Penguin Software Inc.

June 2011

Email Archiving and Continuity

Introduction

Email archiving is becoming important for many organizations for several reasons:

- Every business has key employees or a vital business division such as finance which must keep critical records including email. Some industries such as health care are required to archive email in compliance with government regulations, e.g., HIPPA, SOX, FOIA, etc.
- Many companies archive email for purposes related to intellectual property or to facilitate discovery in the event of litigation. A searchable archive can greatly speed up the process and reduce the cost of discovery.
- A searchable archive serves as an email backup and disaster-recovery mechanism in case the primary mail server suffers failure. In this event, the archive's usefulness is dependent on its being up-to-date and how easy it is to search.
- A searchable archive allows administrators to monitor inbound and outbound messages to verify compliance with company policies.

A solution such as Roaring Penguin's CanIt Archiver can be configured to archive email for an individual or department or entire company and can satisfy all of the above requirements.

In addition, CanIt Archiver offers email continuity: If a company's primary mail server is down, employees can still read incoming messages by retrieving them from the archive. This allows timely access to time-sensitive information and ensures continued productivity even if the mail server is down.

Archiving Requirements

At a minimum, any archiving solution must preserve the following information:

- The original unmodified message. (It should not rewrite the MIME structure of the message; rather, it should preserve the message exactly as it came in over the wire.)
- All envelope information (envelope sender, envelope recipients, sending relay IP address, remote system HELO argument).
- The date and time at which the message was archived.

The archiver should permit searches based on:

- Full-text searches of the message subject and body.
- Searches based on envelope information.
- Searches based on metadata such as attachment filenames, message-IDs, relay host, etc.

The archiver should permit the enduser to view and download messages as well as resend them out of the archive. It should clearly mark resent messages with standard Resent-From and Resent-Date headers.

The archiver should allow the user to search related messages so he/she can follow the threads of back-and-forth conversations.

The archiving solution should be integrated with an anti-spam solution so it does not bother archiving spam or viruses. While it is acceptable to put the archiver after an anti-spam filter, this arrangement is less desirable because it is more difficult to accurately preserve envelope information if the archiver is not the perimeter mail relay.

Email Continuity Requirements

For email continuity, the archiver must archive and index messages quickly. Archiving should be done in real-time and the delay between archiving and indexing should be, at most, a few minutes. In order to reduce backscatter, the archiver must be capable of learning valid recipient addresses so it can continue to accept email for them even if the mail server is down, yet reject (or temporarily reject) messages for unknown recipients.

With Roaring Penguin's own CanIt Archiver, email continuity is assured as email is accessible in the event of a mail server outage and email is queued for delivery when mail service is restored.

Email Archiving Configurations

CanIt Archiver is a highly-configurable email archiving system. Email archiving may be configured for a specific end-user or group of users.

On-Premises Archiving

The figure below shows one arrangement for on-premises email archiving:



Figure 1: On-Premises Archiving

In Figure 1, mail arrives from the Internet and flows through the anti-spam filter and archiver to the back-end corporate mail server. Outbound mail flows the other way: from the corporate mail server through the archiver and filter and out to the Internet. Having mail flow both ways allows both inbound and outbound mail to be archived.

Archiving Internal Email

One of the problems with archiving email on a separate machine from the company mail server is that the archiver normally does not see purely internal email (because such email does not leave the corporate mail server). However, many mail servers including Sendmail and Microsoft Exchange have

an option to copy or “journal” email to an external address. By using this Bcc or journalling feature, administrators can ensure the archiver captures all mail (inbound, outbound and purely internal) and has a complete record of the designated streams of user email.

CanIt Archiver supports journalling from Microsoft Exchange 2007 and 2010. It also has a filter for Sendmail and Postfix that mimics Exchange journalling so that Sendmail and Postfix corporate servers can archive their internal mail.

Archiving as a Cloud Service

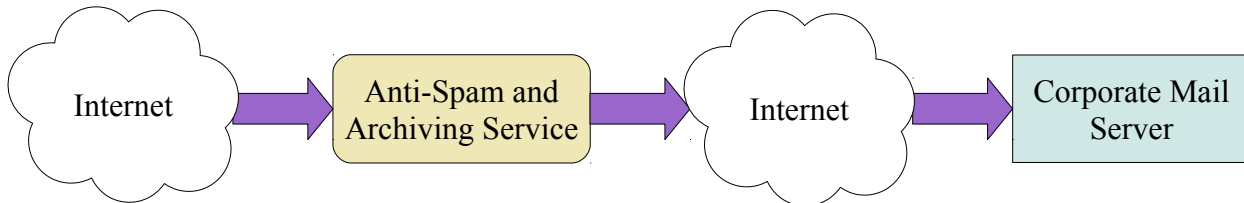


Figure 2: Archiving as a Cloud Service

Figure 2 shows the topology for an outsourced archiving service. Mail arrives from the Internet at the archiving service provider's machine where it is filtered and archived. It is then relayed over the Internet to the customer's back-end mail server.

Note that typically, mail does *not* flow out of the mail server back through the filtering and archiving service. For this reason, journalling or BCC'ing as described above is used to feed outbound and internal mail to the archiver. (Naturally, mail arriving *from* the cloud service need not be journalled back *to* it for archiving.)

Archiving Technical Issues

While archiving seems simple (“Store all the email and index it for searching”), there are a number of technical issues that complicate it.

Disk Space

Archiving messages for several years can consume a substantial amount of disk space. Consider an average-sized organization of 50 people, each of whom receives 1MB of email per day. To archive mail for three years, that organization would need about 53.5GB of disk space (plus additional overhead for the indexes).

To archive the email for 50 people for 10 years would require approximately 178 GB.

A larger organization of 5000 people would need almost 5.5TB of disk space over three years.

Backups

While 600GB or even 6TB of disk space isn't outrageous, the archive *must* be backed up and *should* be stored redundantly. An archive that can't be trusted is worse than no archive at all. Securely backing up a large archive can be time-consuming.

Encryption

The mail archive will contain a lot of sensitive information. For this reason, it should be stored on an encrypted file system and all backups should be encrypted. While an encrypted file system does not protect against an on-line attack, it does protect the archive should the physical server or a backup tape be stolen.

In addition to encrypting the archive, internal mail should be encrypted en route to the archiving machine, especially if it must traverse the Internet. This can be accomplished by using the STARTTLS extension to SMTP.

Access Control and Auditing

The archive system must ensure that users can access only mail that they would normally see (that is, only mail they have sent or received). It should also keep a complete audit trail of searches and message accesses so users and administrators can see exactly who has been searching and accessing the archive.

Archive Integrity

The system must ensure the integrity of the archive. Time stamps must be accurate, original message details must be preserved faithfully, and deletion of archived messages should be prohibited.

Reports

The archiving system should be capable of producing reports such as:

- Messages and bytes archived per day.
- Messages and bytes archived per day per customer.
- Total archive size by customer.
- Number of customer email addresses that have archived messages.

Summary

When selecting an email archiver, the system administrator is wise to consider all of the issues discussed in this white paper. A simple tool to dump masses of email into a file system may not meet the organization's needs. Choosing a dependable, easily-accessible and searchable archiving system requires more care.