Alcatraz
The archive for the NEXT 10 Years™
Contents

EXECUTIVE SUMMARY ...........................................3
  About this Paper .............................................3
  The Audience ................................................4

INTRODUCTION ..................................................5
  A Brief History ............................................... 5
  The Future is in the Cloud .................................. 6
  Today’s Archive Challenge .................................. 6
  Distributed vs Traditional Computing ..................... 7
  Alcatraz – the Archive for the NEXT 10 Years™ ....... 8

INTELLIGENT, SECURE CAPTURE .............................9
  10X Ingestion Rates .......................................... 9

LIGHTNING FAST SEARCH ....................................11
  10X Search Speeds .............................................11

DEFENSIBLE, FULL FIDELITY EXPORTS ......................13
  10X Export Velocity ..........................................13

CONTEXT AWARE™ ARCHIVING AND EDISCOVERY ..14
  The Traditional Approach to Social Archiving ..........14
  The Actiance Approach to Social Archiving .............15

CONCLUSION .....................................................17
Executive Summary

About this Paper

Cloud archiving has been touted as the next generation of archiving for the modern enterprise. In truth, not all cloud archives are the same. Many existing archiving solutions are built on traditional computing architectures and have simply been relocated “to the cloud”. These archives do not adequately exploit the advantages of cloud infrastructure and as such may not be able to scale to handle the exponential increase in data volumes expected over the next decade.

Moreover, the variety of communication sources used within the modern enterprise has dramatically changed over the last few years. Business communications are now sent via many more channels than just email. These include instant messaging, unified communications, enterprise social networks and social media. Traditional archives retain these real-time communications as simple uncorrelated emails. In doing so, they fail to capture the context of those communications: their relevance is lost and the cost of review for eDiscovery and regulatory audits is increased.

This whitepaper outlines the technology underlying Alcatraz. Alcatraz, the archive for the NEXT 10 years, is the only archive built using the same elastic technologies underlying today’s most popular online consumer applications. Alcatraz has a distributed architecture incorporating asynchronous processing and highly scalable

“Between 60% and 70% of new instances of compliance archiving for messaging data are using a SaaS deployment model.”

– (Gartner, Critical Capabilities for Enterprise Information Archiving. 12/2014)
queueing, indexing, processing and storage technologies to allow for many tasks to be executed in parallel. This architecture is unique for an archive and allows organizations to ingest, search and export information 10 times faster than with traditional archives, reducing eDiscovery costs and increasing productivity. Specifically, Alcatraz can ingest 200 messages/second (with an average attachment size of 100 KB), and so allow users to view search results in less than a second, while exporting at a rate of 250 messages/second.

Alcatraz is also the first and only Context Aware™ archive. Because it’s an object-based content store, organizations canarchiveany communication channel. Moreover, it uses patent-pending snapshot technologies so they can save time by reviewing social communications in their original format, including edits and deletions.

**The Audience**

This white paper is intended for IT directors and IT professionals who are responsible for selecting and implementing a cloud-based communications archive, and for those responsible for replacing a traditional email archive. Compliance officers and legal technologists that must comply with electronic communication mandates may also find value in reviewing this paper.
Introduction

A Brief History

Archiving has been a strategic initiative within many organizations, large and small, for almost two decades. During that time, the “what, why, where, and how” to archive has continued to evolve. Back in the day, email was the what; the most mission critical business app, the primary form of communication, and the vehicle for storing and managing business relevant data. Storage management was the primary “why.” IT needed to ease the storage bloat on the email servers, eliminate the performance hit taken by mail servers due to excessive data, insure production system backups happen within the backup window, etc.

To a lesser extent, highly litigious and strictly regulated corporations archived email to comply with government regulations and for basic eDiscovery. These first generation archive solutions were located and managed on premises, along with all the licensing, management infrastructure and ongoing maintenance costs (and headaches) that accompanied the archive.

All of these legacy enterprise archiving applications were architected to execute their tasks synchronously, in a linear fashion, rather than perform many tasks in parallel. This methodology worked well for some time and application development continued to be built on this aging technology.

“The limitations of today’s archives are a direct result of the underlying technology that the system is constructed upon, from the platform and the index engine to the archive database.”

– Scott Whitney, VP of Product Management, Actiance.
The Future is in the Cloud
Organizations that have spent years managing and maintaining these first generation email archives likely find it easy to see why there is an ongoing shift from on premises archiving to archiving in the cloud. According to Gartner, between 60% and 70% of new instances of compliance archiving for messaging data use a SaaS deployment model.[1] With cloud based archiving, organizations aren’t faced with the enormous set up costs, or the on-going costs of managing the archive. Cloud archives have minimal staffing requirements, as management and maintenance is handled securely by the service provider. And, as a company grows, the cloud archive can easily provision users to the system without concern for infrastructure modifications.

The fact is, the technology that satisfied basic archiving requirements in the past simply doesn’t meet archiving expectations or requirements of today’s workplace, or tomorrows. On-premise archives were designed and built 10-15 years ago while many cloud archive solutions are over 5 years old. Vendors moving their on-premise archives to the cloud have not made the required investments in a distributed architecture. As such, many older cloud archive providers suffer from the same legacy technological limitations as on-premise archives. With the huge amount of data being generated today and the expected ongoing growth rate, this legacy technology reliance is a questionable strategy. Organizations that continue to utilization legacy technology will struggle to meet their capture, search and export SLAs.

Today’s Archive Challenge
The challenge facing organizations today is to assess whether their current (legacy) archive will be able to meet the volume, velocity and variety of data produced now and in the future. Archiving applications haven’t kept up with the communication modalities that employees are utilizing today. Nor are they designed with the flexibility to handle future communication applications, many of which are still to be invented!

Unified communications, instant messaging networks, enterprise social networks and social media are just some of the communication channels used in addition to email in today’s the modern enterprise. In order to remain compliant with corporate governance initiatives and regulatory requirements, data from these additional sources must be captured, archived, managed and made available. Traditional archives are not built to handle these emerging social communications. A traditional archive processes these real-
time communication channels as multiple email messages (if they process them at all), losing metadata, context, as well as point in time edits and deletions along the way. While capturing social communications as an email is a stop-gap archival solution, this methodology makes the process of eDiscovery and supervision difficult, if not impossible!

Social communications are enveloped in metadata that tell the full story of a communication thread, at any point in time, and thus should be captured when archiving these new forms for communication. Without the metadata, review and governance of these communications are next to impossible as there is no correlation of one post, thread or tweet to another.

Along with the ever growing data sources that should be archived, governed and made discoverable, today’s archive also has to keep up with the exponential growth of data overall. Enterprise users generate millions of documents every day in various forms including, emails, instant messages and persistent chats, blogs, wiki pages and social media posts. The enterprise cloud archive must be designed to handle ingestion of both real-time data and historical data from an existing messaging source or an archive. This could require dynamic scaling to handle the increase in data volume and variety.

Understanding how the archive is built and performs couldn’t be more important, and the key lies in identifying the technologies that lend themselves to supporting and presenting the vast amount of data that will occupy the archive over time.

**Distributed vs Traditional Computing**

Changes have swept through online consumer technology that allow everyone to gain instant access to the content they want. For example, when a consumer grabs a glass of wine and sits back to watch a movie on Netflix, they expect it to start almost instantly. This expectation is now being demanded in the workplace, and is fast becoming a basic requirement.

Enterprise applications in the past were written to execute tasks synchronously in a linear fashion rather than performing many of them in parallel. For traditional archiving technology, that has meant that they can’t keep up with the constant ingestion, delivery and export of the massive amounts and various types of data demanded of them.

One of the primary advantages of the cloud is the ability to provide a virtually limitless pool of compute power as a utility as opposed to a finite amount of compute power from a defined number of enterprise grade servers. This elasticity allows today’s archive cloud provider to provide that consumer-like experience demanded by the workforce.
In economics, elasticity is the measurement of how responsive one variable is to a change in another. Elastic computing is much the same - resources vary based on the demand for that resource. With on-demand availability of compute, storage and networking resources, the notion of distributed computing has become a reality, allowing asynchronous execution of several tasks to increase overall throughput. An archiving service built on an elastic computing model can allocate resources on-demand to perform tasks, such as initial ingestion of data, or to produce the results of an eDiscovery request in a timely fashion.

**Alcatraz – the Archive for the NEXT 10 Years™**

The decision to purchase an archive is a decision that organizations will have to live with for many years as the purpose of having an archive is to be able to keep information for longer periods of time, for example, 10 years. Thus, when organizations consider purchasing an archive, they should give considerable thought as to whether it is built using technology that is likely to handle their archiving needs well into the future.

Unique amongst cloud archiving services, Alcatraz has a distributed architecture incorporating asynchronous processing and highly scalable queuing, indexing, processing and storage technologies to allow for many tasks to be executed in parallel. This architecture enables Alcatraz to ingest, search and export content 10 times faster than traditional archives. The key components of this infrastructure include the best technologies available today for queuing storage, indexing and processing:

- A scalable front-end message-based queuing infrastructure to truly distribute processing of critical workloads.
- A massively scalable open source store to process metadata that is used at some of the most well-known consumer web sites including Netflix, Google and Amazon.
- An extremely scalable indexing and searching technology that is a popular and mature open-source search engine library.
- An index and search subsystem that extends the functionality to make storing, indexing and searching more efficient with a multi-instance infrastructure.
- A multi-instance infrastructure that starts with an "instance", or a collection of compute and storage nodes, to which additional nodes can be added within an instance or an additional instance of the compute and storage nodes in order to scale out.
- A processing pipeline topology that consists of a set of compute nodes that create a pipeline of tasks to ingest, search and export in parallel. It is configurable so that additional processing, index and storage nodes can be easily added to handle enterprise storage growth transparently over time.

Alcatraz is also an object-based content store so organizations can archive any communication channel. Moreover, it uses patent-pending snapshot technologies to save time by reviewing social communications in their original format - including all edits and deletions.
Intelligent, Secure Capture

For legal and compliance teams to respond to eDiscovery requests or regulatory audits in a timely manner, email, chat and social communications sent by employees must first be captured and indexed. Alcatraz’s technology stack and its matchless integration with Actiance’s Vantage and Socialite products allow it to handle the unpredictable nature of incoming data types and volume. (Together, Socialite and Vantage let you capture, control, and monitor email and all of your critical business communications before sending them to Alcatraz to be archived.) Once content has been captured, Alcatraz simply distributes the processing pipeline between multiple components that horizontally scale to extract, index and store data in a consistent, secure and defensible manner.

Alcatraz ingests data as XML to enable schema-less archiving. Any type of data can be stored, searched and exported as needed without having to rewrite any part of the application and no changes have to be made to the application when new file types and forms of communication are introduced into the market. Utilizing SHA-2 encryption, Alcatraz ensures full fidelity of content and establishes a defensible chain of custody.

Alcatraz provides a durable and immutable data store to allow organizations to meet the electronic storage requirements of the Securities and Exchange Commission (SEC) Rule 17a-4(f), the Commodity Futures Trading Commission (CFTC) Rule 1.31(b) as well as other regulations. In fact, a 3rd party assessment provided by Cohasset Associates, Inc. has concluded that Alcatraz meets the electronic storage requirements of SEC Rule 17a-4(f) for the recording, storage, retention and management of record objects. Alcatraz is also built with high availability and fault tolerant infrastructure to manage and ensure continuous index health.

10X Ingestion Rates

Alcatraz ingests data 10 times faster than traditional archives. Recent performance tests have proven that Alcatraz can ingest more than 250 messages per second (with attachments averaging about 100KB). That translates to approximately 21 million messages per day - a true cloud scale archive.

The diagram below (Figure 1) shows how Alcatraz utilizes distributed parallel computing and the elasticity of various components to achieve a very large throughput to ingest content from a multitude of sources.
As the figure 1 illustrates, processing and storage subsystems in Alcatraz scale by distributing the workload into multiple processing “bolts,” i.e. parallel computing tasks. To accommodate the ingestion of very large content, individual storage subsystems allow for seamless expansion via additional nodes as needed. As data comes in to the system, the system validates the data and puts the raw XML into the message processing queue. The data is then picked up by the Alcatraz processing pipeline where it is then split and stored into the appropriate storage subsystems asynchronously. This allows other content to be processed in parallel maximizing the overall throughput.

Additional compute and storage nodes can be added dynamically into the system to improve the overall throughput to the desired limit. The underlying index and search engine can distribute the “shards” between available nodes dynamically to improve the overall indexing and search performance. Other systems within Alcatraz support a similar dynamic expansion and distribution technique to allow the entire ecosystem to perform at scale.
Lightning Fast Search

Google has set an expectation for search speed. Users anticipate seeing search results instantly, often in less than a second, irrespective of the volume or variety of data being searched.

However, there are some important differences between a search performed in Google and one in Alcatraz. Unlike a Google search, Alcatraz searches are typically more discovery oriented. In enterprise searches, users review the entire result set rather than just a few pages of results. Alcatraz users also expect consistency and fidelity with the search results every time the same search is run.

10X Search Speeds

Alcatraz employs a similar parallel computing technique, as it does with ingestion and indexing, to perform high speed search and retrieval of content. This technique is illustrated in Figure 2 below.

![Figure 2. Alcatraz 10x Search Process](image)

Dues to its unique architecture, Alcatraz users can start reviewing search results in less than a second while the search application prepares the remaining pages of search results in the background. It does this by using the index and search engine, and the in-memory grid technologies. In fact, in a recent performance test by a third party, the first page of a 119 million message results set were available in under 1 second.

MITIGATING SOCIAL MEDIA RISK AT THE FLORIDA BAR ASSOCIATION

The real gain [of Alcatraz]: the time it takes to do complete searches of the records went from days and weeks at a time, to minutes and hours. It saves the organization substantial time in processing requests, and makes the public customers much happier.”

– Jonathan Israel, IT Operations Manager for the Florida Bar
In parallel, Alcatraz also prepares facets so that users can filter the search results based on several categories including communication channels, networks within each channel, and alphabetically arranged user groups that are involved in the communication, as shown in figure 3 below.

Figure 3. Networks and Communication Channels
Defensible, Full Fidelity Exports

There are typically three main challenges with exporting large amounts of archived data. The first challenge is to retrieve, package and produce the archived content. The second challenge is to preserve the full fidelity of content to ensure defensibility of the documents being produced as part of a eDiscovery response. A final bottleneck in the eDiscovery workflow is the transfer of data from the archive to an eDiscovery platform for processing, analysis, review and production.

Alcatraz is architectured to handle all of these export challenges with extreme efficiency, at scale.

10X Export Velocity

According to performance tests, Alcatraz can export up to 250 documents/second while retrieving, formatting and creating PST/NSF, EDRM XML or HTML files for export; a rate several factors higher than competitive solutions. As a result, Alcatraz can export a million messages in minutes. (If requested, Actiance can then make exported data available using a data shuttle service to further expedite the production process.) With Alcatraz, the days of scheduling search and production jobs overnight are gone forever!

As illustrated in Figure 4 below, Alcatraz scales to accommodate extremely large volumes and numbers of export requests by distributing the workload into multiple processing bolts with individual storage subsystems allowing for a seamless expansion via additional nodes. The export operation is streamlined by splitting the request into multiple parallel requests. Each operation fetches, formats and packages a portion of content into a multitude of formats including PST, NSF, HTML and XML. Additional compute and storage nodes can be seamlessly added to further accelerate the export process.

By 2019, 75% of organizations will treat archived data as an active and “near-line” data source, and not simply as a separate repository to be viewed or searched periodically, up from less than 10% today.

- Gartner Magic Quadrant 2014, Enterprise Information Archiving
Alcatraz uses patent-pending technology to store data as snapshots. Snapshots can then be threaded to allow events of significance to be combined for additional context, enabling a true context aware archive. Snapshots are especially important to connect social conversations. Otherwise, they can be time consuming to construct and review, as for example, with traditional archiving systems that can only present them as a series of disconnected (seemingly unrelated) emails.

The Traditional Approach to Social Archiving

The traditional approach to archiving social content is to create a separate email for every single social interaction (a comment, reply, edit, etc.). For example, within IBM Connections, when one individual posts a topic, it generates one email. If a second individual posts a reply, it generates yet another email in the archive. Perhaps the poster then deletes a comment, which in turn generates yet another email, and so on. The main problem is that there’s no quick way for the reviewer to appreciate that all of these separate emails are in fact related.
With a traditional email archive, reviewers can’t understand the relationship between the various emails unless they spend a great deal of time combing through the metadata or looking through the actual content of each individual message to try to thread them together. With some traditional archives, metadata is not captured and various pieces of the interactions are relegated to attachments, making it nearly impossible to make the connections between messages. In short, the conversation becomes an undiscernible mess, as shown in Figure 5 below.

![Figure 5. Typical Delivery of Archived Social Communications](image)

**The Actiance Approach to Social Archiving**

Alcatraz was built from the ground up to archive business communications, whether those communications occur in email, or other real-time communications.

Alcatraz uses snapshots to display the complete history of a conversation, as shown in Figure
6 below. When executing a search to review or preserve conversations, Alcatraz displays the conversations as a series of snapshots. Alcatraz saves the entire thread for an accurate chronological representation of the conversation, even if a portion of a conversation has been edited or deleted. Reviewers can thereby appreciate the relevance of complex interaction events, such as real-time chat, blog entries, discussion board comments, etc. in a single view.

![Figure 6. Alcatraz Snapshots](image)

Alcatraz, the Archive for the NEXT Ten Years™
Conclusion

Alcatraz, is the only archive built using the same elastic technologies underlying today’s most popular online consumer applications. Alcatraz has a distributed architecture incorporating asynchronous processing and highly scalable queuing, indexing, processing and storage technologies to allow for many tasks to be executed in parallel. This architecture is unique for an archive and allows organizations to ingest, search and export information 10 times faster than with traditional archives - reducing their eDiscovery costs dramatically. Specifically, Alcatraz can ingest 250 messages/second (with an average attachment size of 100 KB), allow users to view search results in less than a second, and export 250 messages/second.

Alcatraz is also the first and only Context Aware™ archive. It is an object-based content store so organizations can archive all of their communication channels. Moreover, it uses patent-pending snapshot technologies so they can save time by reviewing social communications in their original format, including edits and deletions.

Alcatraz is in reality, the archive for the NEXT 10 years.
About Actiance

With business communications evolving and proliferating quickly, you need to prepare for the future, today. Actiance, the foremost innovator in business communications management, allows you to unleash your business so you and your colleagues can better engage with customers and collaborate with each other. Our portfolio of cloud services helps you reduce your eDiscovery costs, automate your compliance processes and accelerate your organization’s adoption of existing and new communication channels.

Alcatraz, the archive for the NEXT 10 years™, is the only archive built using the same elastic technologies underlying today’s most popular online consumer applications. Its unique architecture allows you to ingest, search and export information 10 times faster than with traditional archives. Alcatraz is also the first and only Context Aware™ archive. It uses patent-pending snapshot technologies so you can save time by reviewing social communications in their original format, including edits and deletions.

Together, Socialite and Vantage let you capture, control, and monitor email and all of your critical business communications before sending them to Alcatraz to be archived. In addition, Vantage allows you to add disclaimers to outgoing messages, establish ethical walls and federate your communications with other organizations. Socialite also enhances your regulated sellers’ ability to engage their network by allowing them to use social media websites directly, and not be limited by a social portal.

Actiance software is used around the world by over 2.5 million professionals across dozens of industries. In the financial services industry, 10 of the top 10 banks in the US, 8 of the top 10 banks in Europe and all of the 5 biggest banks in Canada use our software. Our sustained success means more resources for you: Actiance spends over 30% of its revenue on R&D. Today, Actiance helps you manage Microsoft Exchange, IBM Lotus Domino, and SMTP email as well as communications in the leading social media, unified communications, collaboration, and IM platforms provided by Facebook (FB), LinkedIn (LNKD), Twitter (TWTR), Google (GOOG), Yahoo! (YHOO), IBM (IBM), Jive (JIVE), Microsoft (MSFT), Cisco (CSCO) and Salesforce.com (CRM).
More information:
actiance.com
sales@actiance.com

Worldwide Headquarters
1400 Seaport Blvd.
Building B, 3rd Floor
Redwood City, CA 94063 USA
(650) 631-6300 phone
info@actiance.com

EMEA Headquarters
Asmec Centre, Merlin House
Brunel Road
Theale, Berkshire RG7 4AB
United Kingdom
Tel: +44 (0) 1189 026 468
eMEA@actiance.com

Follow us
facebook.com/Actiance
linkedin.com/company/actiance-inc
twitter.com/actiance
youtube.com/actiance
slideshare.com/actiance

©2015 Actiance, Inc. All rights reserved. Actiance, the Actiance logo, Socialite, and the Socialite icon are registered trademarks of Actiance, Inc. Vantage is a trademark of Actiance, Inc. Alcatraz is a trademark of Actiance, Inc. All other trademarks are the property of their respective owners.