

WHITEPAPER

Data Lifecycle Management

Introducing a tiered approach
to data management



Introduction

The information explosion continues unabated. Despite dramatic improvements in data storage, businesses of all types and sizes struggle to keep up with exponential data growth.

The primary culprit, however, isn't just the amount of new data. Nor is it the increasing file sizes that result from richer content. The biggest driver is accumulation of data due to:

- Increase in the number of communication channels
- Evolving regulatory recordkeeping requirements
- Inability to determine when and how to perform disposition of older data

Enterprises need new approaches to data management to cope effectively with a rising tide of data and to avoid the risks that result from uncontrolled accumulation of dark data (data that is collected but not used to derive insights or for decision making).

Existing approaches to data management

Technologists often differentiate between hot and cold data. Hot data is accessed frequently and requires high performance, while cold data is accessed only rarely. Applications often require hot data, which needs capacity tiers to accommodate different data access patterns, with administrators performing data tiering manually.

In the past 20 years, many organizations have deployed solutions to address the problems that come with rampant data growth in a more automated fashion. This includes wide deployment of various data tiering solutions. Among the earliest solutions were email and file system archiving applications that were typically deployed in hierarchical storage management formats. These legacy systems replace content in the source location with a 'stub' or shortcut to the content stored on a less costly storage medium.

More recently, data storage systems have incorporated data tiering technologies to move data among different storage devices within a pool of networked storage. However, these storage systems are typically designed to move dynamic, frequently accessed blocks of data rather than files. Self-tiering storage platforms are usually geared towards caching type operations to enhance performance of frequently updated or accessed tables or indexes within structured data applications, rather than low-cost, long-term storage.

In contrast, purpose-built archiving systems incorporate technology designed to manage individual, non-changing files, and to support large amounts of infrequently accessed data. These attributes are well suited to information governance strategies. However, traditional archive platforms only provide a single class of service, with the assumption that all data within an archive is of equal value and at the final stage of its lifecycle.

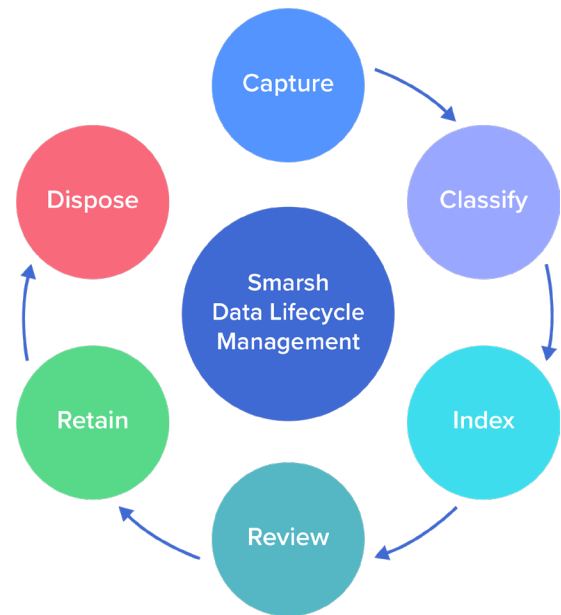
To locate specific files or messages, traditional archiving platforms utilize a full-text index. But because these applications are primarily designed to reduce costs, most vendors have used inexpensive indexing technology. Such index technology has proven inadequate at enterprise scale, and large organizations often suffer with:

- Poor search performance
- Failed and/or inconsistent search results
- Brittle indexes requiring constant maintenance

Smarsh Enterprise Platform

The Smarsh Enterprise Platform addresses the scalability and performance challenges associated with traditional archive solutions.

The search interfaces in the Enterprise Archive and the Enterprise Discovery application modules are both sophisticated and fast, returning millions of results within seconds. This performance allows e-discovery and compliance reviewers to be significantly more efficient as they can iteratively search, filter, and refine results in real time.



Enterprise Archive uses an extensible (XML-based) object model, which maintains a rich set of metadata with every object and supports virtually any content type. This distributed metadata strategy eliminates the need for SQL databases, which become a bottleneck in most archiving applications as the amount of data grows.

Enterprise Archive utilizes a truly enterprise-class indexing technology with self-replication, self-sharding, and self-healing capabilities to create indexes for both object metadata and full text. It supports extremely sophisticated search queries with high performance.

There is a trade-off to this approach: the benefits of significantly better performance, scalability, and resiliency can make Enterprise Archive's upfront costs higher. These costs are mitigated to an extent by enabling organizations to:

- Greatly reduce the amount of data that must be duplicated across the data silos of multiple compliance and e-discovery applications
- Significantly reduce the amount of data—and legal costs—for downstream review
- Increase efficiency while lowering management requirements

Data Lifecycle Management

At Smarsh, our goal is to deliver the best possible total cost of ownership (TCO) for Enterprise Archive customers. One of the key elements that differentiates the Smarsh approach to data lifecycle management from the solutions described earlier is the recognition that not all data is created equal.

Email newsletters and client communications have vastly different value from the standpoint of compliance and discovery. We are developing the Smarsh Data Lifecycle Management (DLM) capability within Enterprise Archive to:

- Greatly reduce the costs of storing and managing data
- Automate the process of managing data throughout its lifecycle
- Continue to deliver the high performance and other benefits that customers expect from Enterprise Archive

This white paper describes the capabilities and implementation of Smarsh DLM.

Introducing Smarsh DLM

Smarsh DLM reduces the costs of operating Smarsh Enterprise while continuing to provide highly differentiated, iterative search capabilities by leveraging:

- Classification
- Multiple storage types
- Granular retention parameters
- Different levels of indexing
- Legally defensible automated disposition processes

These technologies optimize data access speed and total storage costs for different types of content across their entire lifecycle.

Greater control across the data lifecycle

As data ages, it is generally less likely to be needed for investigation, discovery, or surveillance. In addition, there are some types of data (e.g., mass marketing emails with pre-approved content) that are infrequently accessed and less relevant in most searches. Reducing the level of indexing for older and less critical data significantly reduces the size of the indexes needed, which reduces the cost of providing archive services. Smarsh DLM provides a range of tunable parameters to give customers more flexibility to manage data to address:

- Compliance mandates
- Governance mandates
- Cost metrics

Consider a company that has a compliance mandate of seven years and a governance mandate of 10 years. With a legacy archive, that company will probably keep everything on WORM for the entire 10-year period, with no ability to delete anything until the retention period expires. Smarsh DLM provides greater flexibility to span different types of data and satisfy different compliance (WORM) and governance mandates, as illustrated in the following figure.

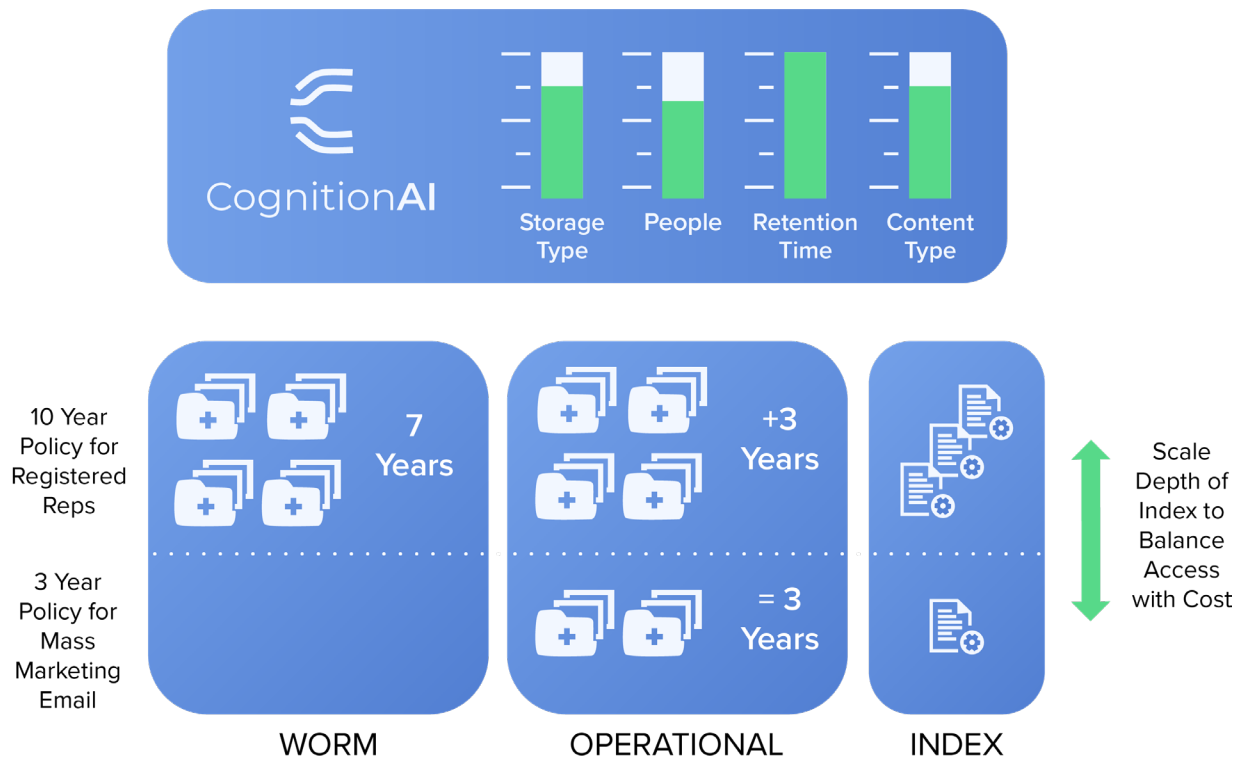


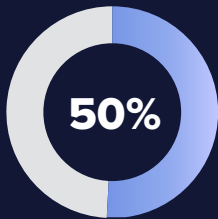
Figure 1. Smarsh DLM provides greater flexibility to address compliance and operational needs.

Smarsh DLM provides greater flexibility to address compliance and operational needs. For the scenario described above, when the 7-year compliance period ends, Smarsh enables the data to change from WORM to operational status. In certain instances (such as meeting GDPR requirements), the records manager can perform privileged delete.

Smarsh DLM benefits

Smarsh DLM provides a set of policies and features that enable customers to capture, access, manage and dispose of data more easily while aligning costs with business priorities. It accomplishes this by partitioning data based on its importance and customer business needs.

This provides customers with:



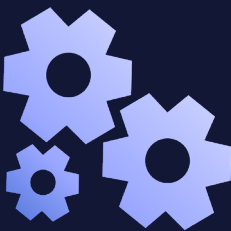
Up to 50% reduced
storage costs

Smarsh DLM eliminates the “one size fits all” policies of legacy archiving solutions. Users can easily define where data lives and for how long.



Improved
recordkeeping

Smarsh enables enterprises to meet growing record keeping demands more effectively by addressing compliance (WORM) and governance (non-WORM) mandates and implementing legally defensible data disposition at scale.



Automated data
classification &
organization

Legacy archiving approaches are reliant on keyword search with limited retention/disposition capabilities. The legacy approach actually increases ‘dark data’ risk. Smarsh DLM will enable AI-based classification to provide more flexible policy-based automation.

How Smarsh DLM works

Smarsh DLM uses a two-tiered approach to indexing data:

Tier 1 data includes full text indexing of both message and attachments content, providing the sophisticated, iterative search capabilities that Enterprise Archive is known for.

Tier 2 data is readily accessible, but leverages a much lighter index, which significantly reduces cost. Searches of Tier 2 data are based on the metadata of the communications, such as custodian and date range. Smarsh has found that data retrieval teams use custodian and date range searches more than 80% of the time.

Search Field	Tier 1	Tier 2
Participant Details Name, Employee ID(s) Communication Aliases	✓	✓
Date Ranges Multiple Ranges, Down to the Minute	✓	✓
Group Attributes Security Groups, Distribution Lists, Custom Attributes	✓	✓
Organizational Details Division, Department & Building	✓	✓
Geographic Details Region, Country, City & State	✓	✓
Content Metadata Communications Type, Subject Line, Attachment Name & Count, Directionality	✓	✓
Content (Body, Attachments) Body, Attachments	✓	✗

Table 1. Search attributes available for the tiers of Smarsh DLM.

To allow users of the system to perform global searches, Enterprise Archive provides a unified search interface that allows users to search both Tier 1 and Tier 2 at the same time. Search results from the unified search interface yield a single result set that includes data from both tiers and allows users to view and/or export content.

In addition, Smarsh provides an Index on Demand capability, allowing Tier 2 data to be searched using the same extensive search parameters available for Tier 1 data when additional review of Tier 2 data is needed.

E-discovery use case

The e-discovery use case illustrates how Smarsh DLM can be applied in practice, yielding cost savings with minimal or no impact to users. Consider the steps in the electronic discovery reference model (EDRM) workflow. Enterprise Archive automatically carries out the necessary information governance and identification steps, ensuring that all data is proactively retained and identified.

The Smarsh Enterprise Discovery application utilizes the same unified search interface as the Archive Management application, eliminating the need to learn and access different tools for different functions.

After performing an initial search using metadata, data from the results set can then be collected into a case. The collection process includes an automated Index on Demand function that provides sophisticated search of all data using the Review interface of the Enterprise Discovery application, as illustrated in the following figure.



Figure 2. The Enterprise Discovery application enables a single query to collect data from both the primary and secondary tiers, and then automatically perform Index on Demand when the search results are collected into a case.

The Enterprise Discovery application enables a single query to collect data from both the primary and secondary tiers, and then automatically perform Index on Demand when the search results are collected into a case.

After an Enterprise Discovery user performs a collection action, all relevant data on the Tier 2 is elevated to Tier 1 where the message body and attachments are full text indexed and become available for full review. The following table shows various attributes of small, medium, and large cases, including time to activate.

Note that even for large cases, time to activate is typically less than two hours.

Search Field	Small Case	Medium Case	Large Case
of Custodians	10	50	100
Date Range	6 months	12 months	2 years
of Documents	180,000	1,800,000	7,500,000
Total Data	25GB	250GB	1TB
Attributes Search (any time)	Seconds	Seconds	Seconds
Time to Activate	< 5 minutes	< 30 minutes	< 2 hours
Content Search (after one-time activation)	Seconds	Seconds	Seconds

Table 2. Attributes of small, medium, and large cases showing time to activate and time to search for Smarsh DLM.

Smarsh DLM versus legacy archives

The following figure compares time to results for a legacy archive solution versus Enterprise Archive with Smarsh DLM. With a legacy archive, every search takes approximately the same amount of time to return results.

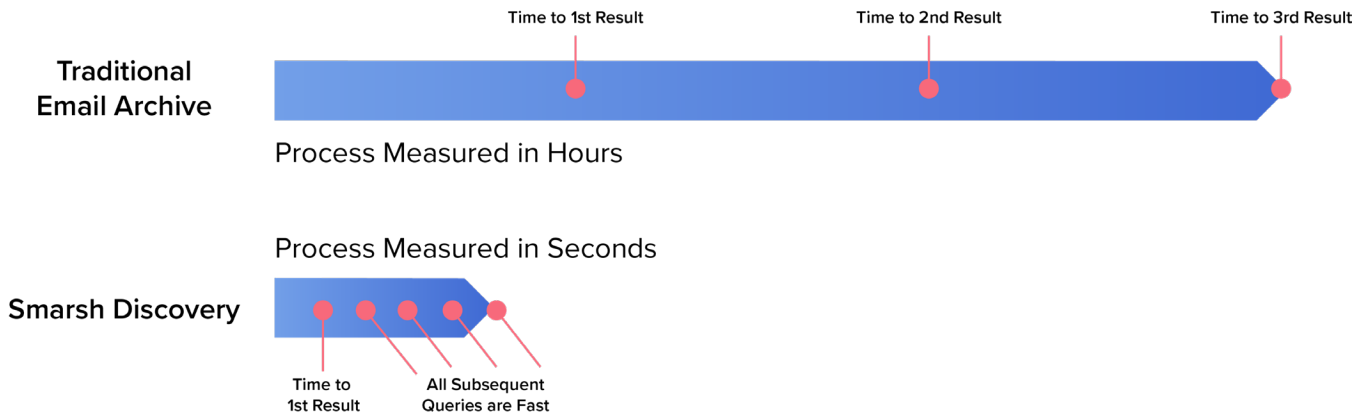


Figure 3. Introducing a tiered approach to data management provides lightning-fast search results with more sophisticated syntax and filtering.

Because searching in a traditional archive is so slow, many organizations eschew trying to perform any review within the archive. This often results in users performing only a single, basic search (e.g., custodian and date range) which leads to over collection and preservation, and unnecessarily large exports. Exporting from traditional archives is already notoriously slow and costly. So the net effect is extended time and excessive (per GB) fees for the export and import into downstream review platforms.

With Smarsh DLM, all data collected into a case is fully indexed and available for review in about the same amount of time it takes to complete an initial search in a traditional archive. Once activated, all subsequent searches return results in seconds.

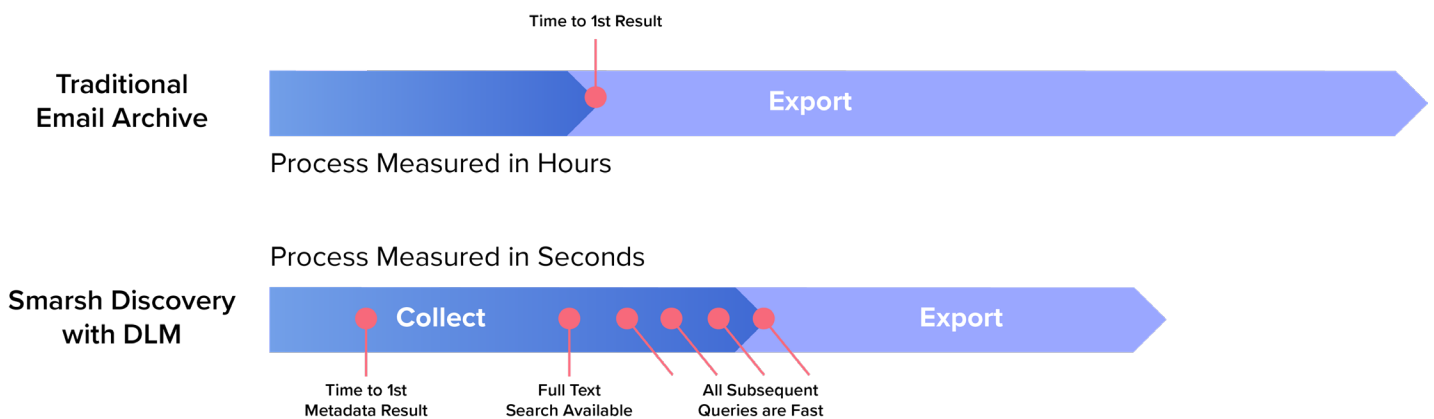


Figure 4. Smarsh DLM with Index on Demand lowers costs, but still provides premium search when needed.

Policy-based tiering

Smarsh DLM will also support policy-based tiering. With policy-based tiering, the Smarsh Policy Engine automatically promotes or demotes data between the two tiers according to the policies that organizations define.

For instance, a company may wish to move data for certain participants to Tier 2 after three years. Or, they may wish to import data from a migration directly to Tier 2, but leverage the Policy Engine to ensure that the most recent data is indexed in Tier 1.

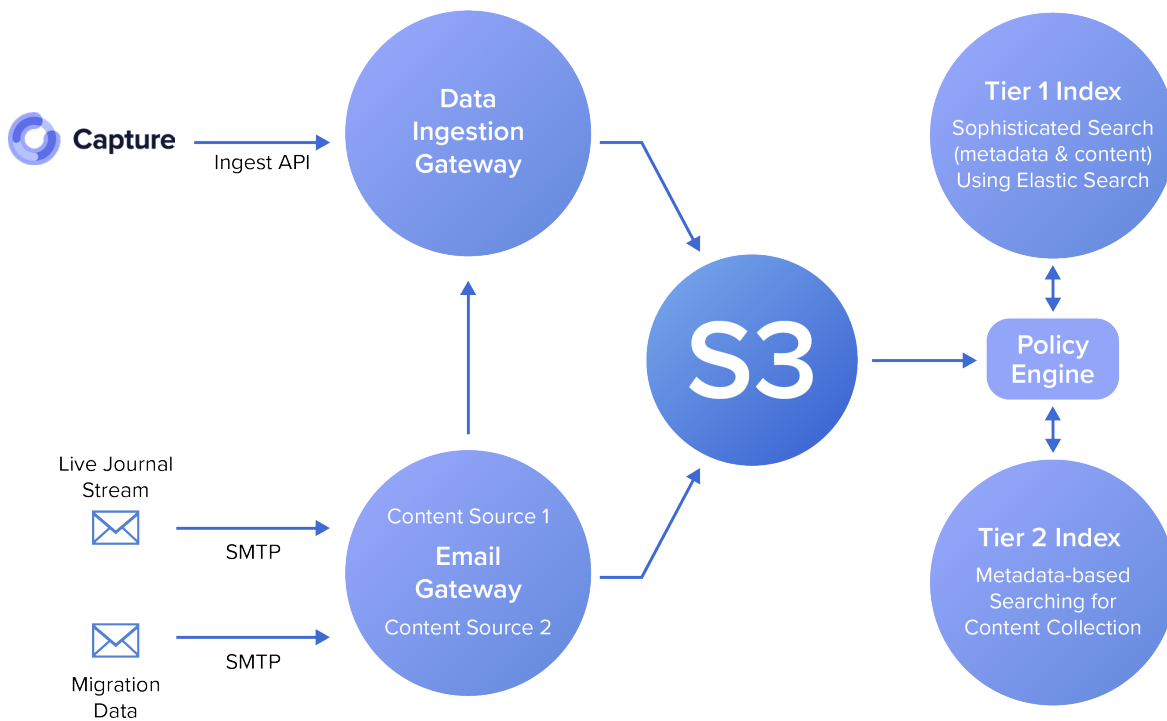


Figure 5. Smarsh DLM with Index on Demand lowers costs, but still provides premium search when needed.

There are a variety of policy-based strategies that customers can use to achieve their data management and compliance objectives by:



**Data
Age**

In most cases, data is accessed less and less as it ages. Aging data out of Tier 1 to the Tier 2 can reduce cost significantly without affecting operations. For example, for a supervision use case, an organization may want data fully indexed for six months. At that point, access rates have fallen, any potential issues have been identified and resolved, and metadata indexing is likely sufficient.

**Participant
Group**

Customers can also create policies based on individuals or groups. For instance, an organization may want to specify that communications from a particular group of users, based on criteria such as Active Directory membership, location, etc., is always indexed in Tier 1.

**Content
Type**

An organization may want to create policies for specific types of content. For example, if an organization conducts important business using Microsoft Teams or Zoom, that data may need to be retained in Tier 1 for six months, a year, or longer. There may be other types of content such as spam email, mass marketing, and newsletters that may not need to be retained in the primary tier at all.

Smarsh DLM will provide ‘pre-flight’ metrics when data is being collected, but prior to Index on Demand. Customers can use a dashboard to see how much data is in each tier and adjust policies to achieve the desired cost and performance objectives.

Classification: Powered by CognitionAI

Soon, the Smarsh Policy Engine will incorporate the classification capabilities of CognitionAI to provide intelligent upfront routing of data to the proper tier. CognitionAI is a mature, machine learning analytical engine currently implemented in Enterprise Conduct that analyzes communications and surfaces alerts for targeted signals. Rather than manually defining what data belongs in which tier, CognitionAI will:

- Inspect the data
- Classify the data
- Determine where the data belongs
- Send data directly to the proper tier

This eliminates the need for full indexing of ingested data ultimately destined for Tier 2. CognitionAI automatically classifies data as it enters the system, assigning tags such as “spam,” “financial transaction,” etc. This classification data is then used to identify and organize each data object and apply the right retention policy, providing additional intelligence and additional cost savings across the entire data lifecycle.

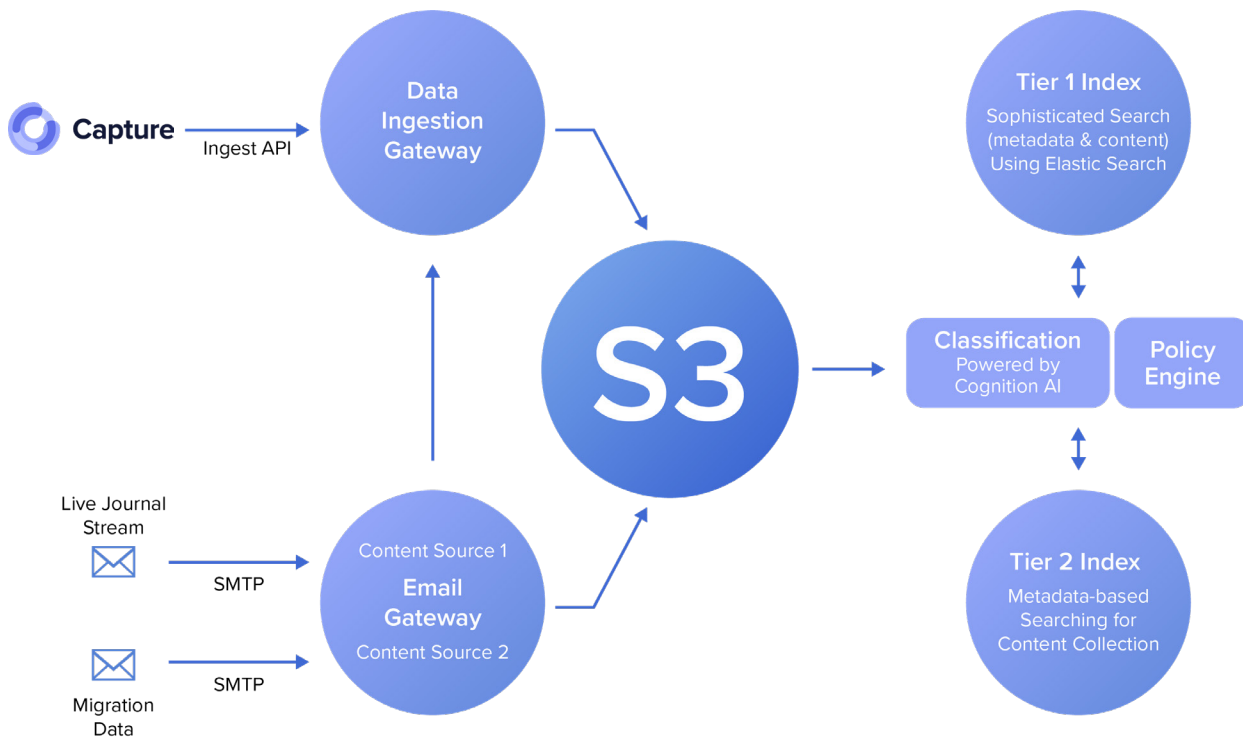


Figure 6. CognitionAI classifies data automatically as they enter the system.

Conclusion

Smarsh DLM provides enterprises with the sophisticated and high-performance search functions they need. It delivers greater reliability and resiliency than competing solutions while reducing the infrastructure costs associated with long-term data retention. With the Smarsh Enterprise Platform, Enterprise Archive, and Smarsh DLM, data archives become an active, intelligent, and valuable resource — not just a poorly managed dumping ground for aging data.

Contact us to learn how Smarsh DLM can help you better satisfy your compliance and data governance needs.



Smarsh enables companies to transform oversight into foresight by surfacing business-critical signals in more than 100 digital communications channels. Regulated organizations of all sizes rely upon the Smarsh portfolio of cloud-native digital communications capture, retention and oversight solutions to help them identify regulatory and reputational risks within their communications data before those risks become fines or headlines.

Smarsh serves a global client base spanning the top banks in North America, Europe and Asia, along with leading brokerage firms, insurers, and registered investment advisors and U.S. state and local government agencies. To discover more about the future of communications capture, archiving and oversight, visit www.smarsh.com.

Smarsh provides marketing materials for informational purposes only. Smarsh does not provide legal advice or opinions. You must consult your attorney regarding your compliance with applicable laws and regulations.

Whitepaper - 10/22



1-866-762-7741



www.smarsh.com



@SmarshInc



SmarshInc



Company/smarsh

© 2022 Smarsh, Inc. All rights reserved