collabspace

Transform Records Management & Gain Business Insight with the Collabspace Data Lake

Bringing Data Lake Clarity

While the term 'data lake' has been floating around the RIM world the past few years as more enterprises have begun using data lake services, the concept itself is still a bit murky.

We're here to clear up the muddy waters around what it is, how it compares to traditional data warehouses, the benefits and how it can transform and bring insights to your organization's records management. Let's dive in.

What is a Data Lake?

A data lake is a type of data repository which has a flat architecture providing storage capacity for large volumes of information. Whether text-based or image-based, processed or raw, this reserve can hold all types of data that can be streamed in from all different sources. Each data element in the lake is tagged with a set of metadata tags, so users can easily search, access and gain the most insights from whatever content has been stored.

Where Did the Data Lake Come From?

In the past two decades, organizations began to find that traditional data storage solutions were no longer sufficient to hold all of their information. A new solution that could hold large volumes of unstructured data was necessary, so data lakes were created.

The term itself was coined in 2010 by James Dixon, then-CTO of Pentaho. He used it to contrast data marts, which are smaller data repositories with limited storage (a subset of data warehouse, which we'll explore below). Dixon described the term in the following way:

"If you think of a data mart as a store of bottled water—cleansed and packaged and structured for easy consumption—the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."

How Does the Lake Compare to the Warehouse?

The data lake is often compared to an older repository-type, the data warehouse. While we outline their differences below, we will first state: one was not created to replace the other. Both serve their own purpose. Organizations should look at factors of each and understand their RM needs before deciding to incorporate one or both.

Data Warehouse

- Centralized data storage
- Hierarchical structure
- Can retain processed data
- Limited data storage
- Established, standardized

Data Lake

- Decentralized data storage
- Flat architecture
- Can retain all data types
- Holds large volumes of information
- More flexibility

Data Lake Benefits for RM

Data lake solutions are beneficial to records management because they provide organizations with a single source to backup and safeguard all their information. With large storage capacity and ability to stream from various sources, the data lake is a reservoir from which users can manage and access all types of content from across platforms.

This not only ensures that all records have been streamed into one repository to easily apply appropriate document lifecycle workflows, but they are easily searchable from one place for eDiscovery, GDRP, and FOI requests.

Gaining the Most Value from Data Lakes

While the data lake offers incredible potential for records management, organizations need to be mindful of where they are streaming their information, or they can find themselves with a data swamp repository that is not well protected and where valuable information gets lost amongst all other data. *It is critical to select a solution that not only helps organizations securely store and manage their information, but also provides clarity for improving business processes.*

How to ensure this? Storing information in a data lake that is intelligent. A data lake repository, like Collabspace, stands apart because it implements artificial intelligence to automatically stream and categorize content from systems without impacting use or operations. This solution offers standard data lake capabilities, enabling information retention and protection with compliant archive and backup into a ransomware-proof WORM-compliant storage system. In addition, OCR capabilities and transcription allow users advanced, unified search of their text-based documents, images, scanned PDFs, audio, and video content.

But it is the AI compute, automatically streaming and categorizing content, that ensures visibility and avoids staff interruption, so you have more time to unlock analytics, new insights and intelligence for improved records management and overall business process.

Stream Content from Multiple Sources

By connecting all your content sources together in the Collabspace Data Lake, automatically stream for holistic and cross-platform records management control, multi-system search/discovery, analytics, insights and more so your team has access to the data they need while compliance is handled in the background.

The best way to get the most value from this repository type is selecting a solution where the stream flows both ways: organizational data is streamed in to centralize, back up and secure from one central hub, while the application of AI allows for an outflow of visibility and business insights.

Contact Us

For more on how Collabspace data lake capabilities bring clarity, visibility, and insights to your organization, contact us for more information and a personalized demo.

collabware

www.collabware.com

contact@collabware.com

1-855-268-0442