WHITE PAPER

# Information Governance is a Strategy, Not a Tool

**An Osterman Research White Paper**

*Published September 2015*



**Osterman Research, Inc.**
P.O. Box 1058 • Black Diamond, Washington • 98010-1058 • USA
Tel: +1 253 630 5839 • Fax: +1 253 458 0934 • info@ostermanresearch.com
www.ostermanresearch.com • twitter.com/mosterman

# EXECUTIVE SUMMARY

On a typical day, 325 megabytes of data are created for every man, woman and child on earth; more than 100 billion emails are sent and received; and roughly 500 million tweets are sent via Twitter. Moreover, an enterprise of just 2,500 email users will generate nearly 69 million emails every year and will store nearly 10 terabytes of archivable business records every six years. Add to this the tens or hundreds of millions of files that they store, the multiple gigabytes or terabytes of data in mission-critical databases, the tens of millions of tweets and Facebook posts that are generated every day, the millions of files in file sync and share tools, and the millions of files on corporate- and employee-owned mobile platforms. Now add in things like sensor data, log data, Web pages, .PST files, security videos, and a host of other data types.

The result is that even relatively small enterprises store hundreds of millions or even billions of data records in a large and growing number of databases, email servers, collaboration servers, Web servers, laptops, desktop computers, smartphones, tablets, cloud storage systems, employees' home computers, and other, largely disconnected, silos of content.

Now, consider that this wide range of data types stored in a variety of disparate and disconnected platforms (only some of which are accessible to IT) across a range of venues on-premises and in the cloud is subject to search and production for a variety of purposes: eDiscovery, early case assessments, regulatory audits and other purposes. And, a failure to manage all of this information properly can lead to data breaches that can cost millions of dollars to remediate, greater risk from an inability to satisfy legal or regulatory requirements, legal sanctions, fines, loss of corporate reputation, and a number of other serious consequences.

## WHAT ENTERPRISES NEED

What every enterprise needs, therefore, is a way to manage its information more effectively. It needs to consolidate its content, eliminate duplicate data, implement centralized and coordinated retention policies, and search across all of its data quickly and efficiently. In short, it needs a unified data management and governance strategy and a platform that will enable this strategy to be carried out.

## ABOUT THIS WHITE PAPER

This white paper discusses the growing problem of information governance and what organizations can do to address the problem. The paper also provides a brief overview of ZL Technologies, the sponsor of this white paper, and its relevant offerings.

# THE STATE OF INFORMATION MANAGEMENT TODAY

## THE NUMBER OF DATA SOURCES IS INCREASING RAPIDLY

There are a large number of sources generating business data that must be preserved for long periods. For example:

- **Email**
  In most organizations, email is the single most important application in use. The typical user sends or receives approximately 110 emails on a typical workday, or roughly 27,000 emails per year. If we assume that an organization will retain only 25% of these emails in an archiving system, that emails are retained for six years, and that the typical email is 100 kilobytes, an organization of 2,500 email users will end up storing 9.8 terabytes of archived data at the end of six years. This includes email that is managed on-premises, but increasingly managed in the cloud in systems like Microsoft Office 365, Google Apps and the like.

- **Files and file servers**
  Osterman Research surveys have found that file servers constitute the single largest repository of corporate content, slightly ahead of email in terms of the total volume of content under management. Moreover, a large and growing proportion of organizations' employees store content in cloud-based file sync and share systems – many of them consumer-focused tools – that contain content duplicated on corporate files servers or, in some cases, not backed up to corporate repositories at all.

- **SharePoint and other collaboration and ECM systems**
  Microsoft SharePoint is widely deployed in Exchange-enabled organizations and is used to varying degrees. However, even for organizations that do not rely on SharePoint for mission-critical data storage or applications, these repositories typically house significant quantities of data.

- **Databases**
  There are a number of databases maintained by most organizations, including customer relationship management (CRM) databases for salespeople, as well as more specialized databases for HR, product management, inventory control, e-commerce, and a wide range of other applications. The general rule of thumb is that a single database will contain one gigabyte of storage for every 1,000 users[i], resulting in enormous quantities of data storage even for mid-sized organizations.

- **Social media content**
  A significant and growing proportion of corporate users employ various social media tools in the workplace. While a growing proportion of organizations have deployed enterprise-grade social media tools like IBM Connections, Jive and Yammer; some of the more popular non-enterprise social media tools – including Facebook, Twitter and LinkedIn – are used widely for work-related tasks. These systems generate significant amounts of data, much of it image-based, along with various other types of business records that must be retained for long periods.

- **Cloud-based applications**
  There are literally thousands of cloud-based applications in use in the workplace, including at least 100 different file sync and share applications, tools for Internet telephony, online conferencing, instant messaging, cloud storage and other specialized functions. These solutions contain enormous amounts of corporate data – some of it constituting business records – that must be preserved for long periods of time.

- **Sensor data**
  As the Internet of Things (IoT) develops, the number of sensors of various types – and the data they generate – will increase exponentially. For example, various forecasts call for the number of active wireless connected devices to increase from 16+ billion in 2014 to nearly 41 billion by 2020[ii], the number of Internet-connected automobiles to increase from 23 million in 2013 to 152 million in 2020[iii], and the number of smart utility meters to grow from 313 million in 2013 to 1.1 billion in 2022[iv]. Another forecast calls for the number of IoT units to increase eight-fold between 2013 and 2020, from 3.03 billion to 25.01 billion[v].

- **Mobile applications**
  Osterman Research surveys have found that roughly five percent of all corporate data is stored on smartphones and tablets, and this proportion is growing as the workforce becomes more mobile and as formal telework programs become the norm.

- **Log data**
  Organizations store enormous amounts of log data from various corporate applications, servers, security systems and other sources. While some log data is

kept for only short periods of time, there is a trend to maintain data for longer periods because of its value in Big Data applications.

- **Other content**
  There is a variety of other content that organizations store and manage, including scanned paper documents, images, data from various legacy systems, archived Web pages, purchasing records, data on thumb drives, content on employees' home computers, .PST files, geolocation data, stills and videos from security cameras, and the like.

## DATA VOLUMES ARE MEASURED IN THE BILLIONS…OR MORE

These data stores house vast and growing quantities of data. So much so, that data quantities stored in the typical company no longer number in the millions, but increasingly in the billions and tens of billions. Underscoring just how enormous data quantities have become and are increasing, IBM estimates that every day 2.5 quintillion bytes of data are created – the equivalent of 325 megabytes of data for each of the 7.33 billion people on earth. Given that data creation is not equally distributed, the quantity of data generated and stored in information-focused enterprises and in more affluent nations is significantly greater than the average. It is not uncommon, therefore, for the typical information-focused organization to have multiple billions of data elements under management.

## LAWS AND REGULATIONS ARE CHANGING

Most nations impose some form of regulatory obligations for maintaining records that direct what information must be retained and for how long. Information subject to these retention obligations should be treated with great care, much like information subject to eDiscovery, due to the potential for penalties and fines for not following various laws adequately. Data subject to compliance requirements that is not managed and retained in accordance with regulations can trigger government information requests, formal audits and the like. These can quickly transform into expensive legal proceedings, fines and possible jail time.

Many government regulations include obligations for handling and retention of certain types of information under the organization's control. There are at least two types of sensitive data that organizations should take pains to control and secure:

- Employee and customer data, typically called Personally Identifiable Information, Personal Financial Information and the like, and

- Intellectual property

Accidental release of a customer's or employee's social security number, bank account number, health-related information or tax data can trigger lawsuits, massive costs, and penalties, as well as negative publicity for the organization.

Because intellectual property represents a potentially enormous investment by an organization, leaks that result from theft or inadvertent disclosure can cost an organization millions or even billions of dollars, loss of market share, loss of shareholder equity, and ongoing negative publicity.

In short:

- The number of data management requirements is increasing

- Laws and regulations are becoming more complex

- Retention and management requirements are increasingly difficult to satisfy

- Courts and regulators are expanding the data types that are subject to eDiscovery, regulatory retention, etc.

## DATA IS STORED PRIMARILY IN SILOES

Today, most data is stored in individual and separately managed siloes, both on-premises and in the cloud using a variety of data formats:

- Most organizations have deployed Microsoft Exchange, and so email is stored in on-premises Exchange Server databases, but also in Personal Storage Table (.pst) files and other formats. IBM Domino, on the other hand, uses a NoSQL document database format (.nsf), while Zimbra uses a MySQL database as its content data store.

- The worldwide CRM market is 47% SaaS and 53% on-premises, distributed across several vendors (led by Salesforce, SAP, Oracle, Microsoft and IBM) that use a variety of data formats[vi].

- There is a wide range of document file formats in use, including Office Open XML (.docx), OpenDocument (.odt), HTML (.htm, .html), PDF, Postscript (.ps), Rich Text Format (.rtf), Uniform Office Format, DocBook, ASCII, UTF-8, Open XML Paper Specification (OXPS) and many others.

- Gartner has identified 20 vendors in its Magic Quadrant for eDiscovery software, 20 for enterprise information archiving, 19 for enterprise file sync and share tools, 17 for on-premises application integration suites, and 15 for security information and event management solutions. This does not include the much larger base of vendors whose products are used across thousands of enterprises.

From a functional standpoint, the current data management model makes sense, since systems and their data repositories are typically provided by and/or managed by a variety of vendors, using multiple data types and in multiple data formats. However, these siloes make data management very difficult for a number of reasons:

- **Data access is difficult**
  Accessing data is difficult and time-consuming because a large number of different repositories must be accessed, often in a sequential fashion. Moreover, because a growing proportion of data repositories are under the control of individual employees (e.g., consumer file sync and share tools or social media accounts), accessing all of the data repositories may prove to be impossible in many cases. When an organization must gain access to all of its data in a timely way, such as during an eDiscovery exercise or a regulatory audit, an inability to access all data – and to do so quickly – makes it more likely that an organization will not be able to fulfill its data production obligations.

- **Duplication of data**
  Most organizations have large quantities of duplicate data, such as emails that are stored both in Exchange databases and individual .pst files, or data on corporate file servers and in employee-managed file sync and share tools. Because it is virtually impossible to single-instance across the various siloes in use in the typical organization, storage costs are substantially higher than they would be if single-instance storage was implemented. More importantly, however, when data is produced for eDiscovery, for example, multiple copies of the same files are generated, driving up legal and related costs and slowing the production of needed information.

- **Multiple interfaces**
  Because data is so widely distributed across multiple systems in the typical organization, content must be accessed from a large number of interfaces. This provides disparate views of the same data, resulting in no single or complete

version of the truth. For example, if a critical email message exists in both an Exchange database and in a senior manager's .pst file, which is the true and verifiable "original" of that message?

- **Disjointed retention policies**
  Another problem of the typical, siloed content management "system" in place in most organizations is that each silo has its own retention policies and retention management capabilities, resulting in a lack of consistency in retention across the various siloes. For example, an organization that needs to retain email, files, and instant messaging content will employ at least three different platforms to do so, which is contained in at least three different data repositories, using three different retention management capabilities. This results in the very real risk of over or under retention of various types of content.

- **Disparate search problems**
  The use of separate siloes, multiple interfaces and disjointed retention policies results in searches across siloes that will return different and non-integrated results. For example, searching for all relevant email and instant messaging conversations in support of an eDiscovery exercise will produce different and often incomplete results. Moreover, the data will not be integrated so that an instant messaging response to an email, for example, will not be presented in chronological order and in the context of the email thread. This not only increases the cost of finding the right data, but also decreases the certainty of finding the right information.

- **Increases in cost and administration time**
  Yet another problem associated with the use of siloed information across a variety of systems is that the cost of administration is higher, as is the cost of training. Since administrators must be trained on multiple interfaces and must access content from one silo after another, the overall cost of data retention and retrieval increases.

- **Greater chances of not finding required content**
  Finally, and perhaps most seriously, the use of multiple siloes increases the chance of not finding all required content. This can lead to spoliation of data during eDiscovery searches or regulatory audits, which can carry with it a variety of serious consequences, including legal sanctions, regulatory sanctions or – in a worst-case scenario – an adverse inference instruction from a court. Such an instruction might permit a jury to infer that one party's inability to produce relevant information during eDiscovery, for example, is evidence of its culpability in a legal matter.

## THE SOLUTION IS UNIFIED DATA MANAGEMENT AND GOVERNANCE

The solution to the problems described above is to build a strategy to manage all enterprise data under a single, unified system. Increasingly, the concept of the "data lake" has emerged as a potential strategy for pooling and accessing all data sources within one singular repository, but the individual silos mentioned above have functionally impeded progress in most organizations. Furthermore, many existing "data lake" environments do not cleanse or govern content: they simply pool and aggregate. It is important to note that not all data lakes are created equal – some become data "dumps", as noted below.

Implementing a unified "data lake" strategy requires two key steps:

- **Consolidate content**
  The first step is to consolidate all enterprise into a single repository, a concept that some refer to as a "data lake." By managing all data in a single repository, four important problems with the status quo are eliminated:

- Duplicates of data are eliminated because superfluous copies of the same data are deleted as part of the consolidation process. This prevents multiple copies of data from being produced during searches, resulting in the production of streamlined data sets that are less expensive to review by legal staff and others.

- Search is more consistent because a single repository is being accessed instead of independently managed siloes of data. This allows a single set of search criteria to be applied to a single set of data, eliminating the somewhat haphazard amalgamation of results that are often returned when accessing multiple siloes.

- Disjointed retention policies and practices are eliminated because now a single retention policy and set of retention practices can be applied to enterprise data.

- Enterprises are better able to delete data that is no longer required. Because defensible deletion policies are important to reduce data storage costs and to mitigate the risks associated with preserving data for longer than necessary, having a single data repository means that data can be deleted in a more coordinated and consistent way across the enterprise.

- **Implement good data management and governance**
  Data management and governance of these data lakes has largely been overlooked in most organizations, with many so-called data lakes instead becoming de facto dumping grounds for unmanaged and uncleansed data. However, a managed data lake allows all data to be managed holistically for a variety of purposes, including records management, eDiscovery, regulatory requirements, storage and analytics.

  In reality, the data lake needs to be treated as a living ecosystem rather than a simple pool of unmanaged content. Ongoing governance and filtering is required; de-duplication of content, addition of emergent data sources, and consistent management of lifecycle policies all ensure that the data lake will provide a representative population of enterprise data from which accurate analysis can be derived. While pooling massive volumes of electronic data is admirable in theory, such an approach also requires active management in order to minimize risk while maximizing the potential for analysis.

  In the past, enterprises could get by with ignoring the vast majority of electronic data as transitory and focus on only hardcopy records. For those small proportion of electronic documents there were considered a record, organizations instructed individuals to ”print them out” and file them as they would hardcopy records – destroying any metadata attached to them. Now that roughly nine out of ten information workers spend their day on their desktop computers, laptops, smartphones and tablets creating work-related content, communicating with email and instant messages, storing content in file sync and sharing tools, and conducting marketing activities on social media, records management, data management and information governance have taken on new roles and meanings.

  Moreover, Big Data analytics, the process of examining enormous amounts of data to uncover beneficial hidden patterns, unknown correlations and other useful information – is adding to the necessity of good information governance. With so much activity around digital information, simply managing business records for regulatory requirements and legal purposes becomes almost secondary. There lies massive potential in human communications and behavior patterns associated with the creation of day-to-day business content, and to focus only on the “final” record is to miss the big picture. Unstructured data analysis is rapidly becoming more sophisticated, and the failure to manage this

human business content today risks missing out on groundbreaking strategic insights in the not-so-distant future.

Unlike current records management solutions, information governance is a superset that now *includes* records management. An information governance solution will ideally provide for the enterprise-wide identification/indexing of all information, the centralized management of all information via coordinate retention/disposition policies, and automated information-sharing based on security levels, all while supporting the enforcement of information governance policies across business functions, locations, and information silos.

## SUMMARY

The number of data sources employed by the typical enterprise is large and growing, and the data they house is increasing at almost exponential rates. Moreover, this data is managed in a variety of disconnected silos, accessed from a variety of interfaces, and managed for retention and disposition by a number of disjointed policies. The result is that data searches for legal, regulatory and other purposes take longer than they should; return incomplete or inconsistent data sets; and expose organizations to significant risk of data spoliation and an inability to respond to information requests in a timely or thorough manner.

What enterprises need to overcome these problems, therefore, is a unified data management and governance platform – a managed "data lake" -- that will eliminate the problems associated with duplicate data, inconsistent searches and production of information, and the risks that current practices create.

## ABOUT ZL TECHNOLOGIES

ZL Technologies makes Unified Archive® software (ZL UA) to enable large enterprises to implement information governance all unstructured content such as email, files, and instant messages to satisfy corporate needs for eDiscovery, records, compliance, and storage management. By providing singular and comprehensive data management architecture, it also enables business content to be leveraged proactively for analytics and competitive advantage, via ZL Enterprise Analytics™. ZL UA's unique differentiator is its unified architecture, which consolidates all applications and billions of documents under one platform, thus eliminating today's fractured data silos which significantly raise operating costs, increase legal risk, and derail effective Big Data analytics initiatives. Demonstrating a proven track record with Global 500 customers and strategic partnerships with major players such as Oracle, Unisys, PwC, and SunGard, ZL has emerged as the technology leader in harnessing unstructured "Big Data" for strategic advantage.

Customers come first for ZL. The company's long-term vision for clients and platform is unique in an industry where product acquisitions and software integration are the norm for building out functionality. ZL Technologies was founded in 1999 and is employee-owned and controlled, free from the short-term focus of investors. The freedom to concentrate on long-range corporate goals gives ZL the flexibility to make prudent, strategic decisions for customers. ZL's long-term outlook has culminated in a clear differentiation in product quality: a characteristic consistently echoed by customers.

With reliable products and services, talented people, and constant collaboration between partners and customers, ZL has created a profitable and sustainable business model that has endured over time. This philosophy, combined with relentless dedication to the needs of forward-leaning enterprise clients, is one that continuously drives ZL's products and services in new and groundbreaking directions.

For more information on analytics, please visit www.analytics.zlti.com. For governance and company info, visit www.zlti.com.



## REFERENCES

[i] http://support.snapcomms.com/customer/portal/articles/1102859-what-is-the-sql-database-size-in-typical-use-

[ii] https://www.abiresearch.com/press/the-internet-of-things-will-drive-wireless-connect/

[iii] http://www.autonews.com/article/20140110/OEM06/301109910/the-race-to-market-the-connected-car

[iv] http://www.navigantresearch.com/newsroom/the-installed-base-of-smart-meters-will-surpass-1-billion-by-2022

[v] http://www.gartner.com/newsroom/id/2905717

[vi] http://www.forbes.com/sites/louiscolumbus/2015/05/22/gartner-crm-market-share-update-47-of-all-crm-systems-are-saas-based-salesforce-accelerates-lead/