WHITE PAPER

The Critical Importance of File Analysis

An Osterman Research White Paper

Published March 2017

Osterman Research, Inc.
P.O. Box 1058 • Black Diamond, Washington • 98010-1058 • USA
Tel: +1 206 683 5683 • info@ostermanresearch.com
www.ostermanresearch.com • @mosterman

EXECUTIVE SUMMARY

Organizations create and store massive amounts of information, much of it in file shares. However, most data owners and those in IT do not know much about the information in their files shares, nor have they established the policies, practices or systems required to analyze, classify or remediate this information. The result is a greater risk of non-compliance with legal or regulatory obligations, a greater chance for data breaches, and higher costs.

KEY TAKEAWAYS

Here are the key takeaways in this white paper:

- Organizations create and store enormous quantities of digital information and these content stores are growing rapidly.
- Nearly one-half of this content is stored on corporate file shares, yet most decision makers know little about the content that is stored on these systems.
- The result of poor file management is an inability to properly classify information, poor data governance, excessive storage of content, and a negative impact on user productivity.
- To address these issues, organizations should implement a robust file analysis solution that is supported by policies and processes focused on good data governance.
- The file analysis process cannot be a one-time or occasional exercise, but an ongoing process with appropriate buy-in from all relevant stakeholders in an organization.

ABOUT THIS WHITE PAPER

ZL Technologies sponsored this white paper. Information about the company and its offerings is provided at the end of the paper.

THE PROBLEM OF DATA GROWTH

At a macro level, the volume of digital data is exploding: on a typical day, 2.5 quintillion bytes of data are created; Internet traffic will grow from 100 gigabytes per day in 1992 to 50,000 gigabytes per second by 2018; 12.2 billion emails are sent each hour, and 12 hours of YouTube footage is uploaded to Google's servers each minute. Add to this the fact that 90 percent of data is unstructured[1].

Paralleling these macro trends, data growth within the enterprise is an enormous problem that continues to get worse year by year. For example, if we conservatively assume that the typical enterprise stores 100 gigabytes of content per user in 2016, and that storage is increasing at an average of 40 percent per year[2], an organization of 10,000 users will see its current storage requirement of 980 terabytes increase to 3,750 terabytes by 2020. Moreover, if we assume that the typical enterprise user generates 30 emails and five files per workday, an organization of 10,000 users will generate 87.5 million user-generated files per year. Assuming that an organization retains its files for seven years, a total of 613 million of these files will be retained at any given time.

However, we must also consider that enterprises are generating and storing vast and growing amounts of other data types, including video and still camera feeds, social media posts, instant messaging conversations, data in SharePoint or similar types of systems, voicemails, web pages, source code, CRM data, ERP data, server logs,

[1] http://www.vcloudnews.com/wp-content/uploads/2015/04/big-data-infographic1.png
[2] Source: Enterprise Strategy Group

databases, HR management data, purchasing records, customer transaction records, text messages, customer service call logs, geolocation systems, and a wide variety of other data types. The net result is that an enterprise may end up storing well in excess of several billion files at any given time.

Most decision makers are familiar with the problems associated with escalating data volume. IT has learned how to store large amounts of data by regularly throwing additional storage resources at it, and they have some level of knowledge about how to work with the hundreds of different formats of unstructured data – the "variety" of data – with app-to-app connectors. The most significant – yet least addressed – issue with data is the speed at which this data is now being created, captured, stored, copied, and shared – the so-called "velocity" of data.

WHERE ARE A COMPANY'S FILES?

The content that users and applications generate can be stored in a wide variety of locations: file shares, email systems, users' desktop and laptop machines, collaboration systems like SharePoint, FTP servers, smartphones, tablets, file sync-and-share systems like Dropbox, users' home computers, cloud storage systems and other locations. However, Osterman Research surveys have found that nearly one-half of all corporate content is stored on corporate file shares. Yet, many data owners in IT cannot accurately determine what is stored in those venues. Their failure to identify, classify and manage this large and growing volume of corporate information results in a significant and unnecessary increase in their organization's legal and regulatory risk, reduces their ability to protect their intellectual property or their employees' and customers' privacy, and increases their data-related costs.

FILE GROWTH RESULTS IN A NUMBER OF PROBLEMS

The sheer volume of files stored by the typical enterprise, not to mention the rapid pace of their growth, creates a number of vexing problems for data managers:

- **Poor data classification**
  TechTarget defines data classification as, "the process of organizing data into categories for its most effective and efficient use." Data classification is an essential component of any information governance strategy because it helps an organization to manage its data in important ways:

  - Functionally, by allowing users to more easily find and access critical data objects.
  - Strategically, by restricting and logging access to sensitive and confidential data.
  - In accordance with legal and regulatory compliance obligations.

  The current state for most organizations is either improper or incomplete data classification, in which only some data is properly classified; or there is no data classification whatsoever. Either state results in increased risk to an organization because it is more vulnerable to data breaches, data loss, incomplete eDiscovery and/or an inability to satisfy compliance obligations.

- **Poor data governance**
  Most organizations not only have enormous amounts of information, but they are acquiring new content quickly, and they lack an appropriate infrastructure to effectively manage this content: in short, they have a data governance problem. As noted above, the vast majority of this data is of the unstructured variety, usually controlled and "managed" by individual employees, not by IT, compliance, or an information governance function. Much of this unstructured data is considered "dark data" because it is invisible to IT and not easily accessible by individual users. Instead of being stored within an archiving solution or on managed enterprise content management systems, this dark data

is usually stored on employee workstations, removable media, enterprise file shares, or even outside the organization's control on employee-controlled personal clouds. Dark data poses a growing cost and liability to organizations because it is still considered an organizational asset and within its scope of responsibility.

The end result of poor data governance is an inability to fully satisfy the compliance requirements an organization faces, either those imposed upon it by regulatory or legal considerations.

- **Overcollection for eDiscovery**
  The majority of organizations tend to overcollect information when performing eDiscovery to be on the safe side and ensure they have defensibly captured and protected all potentially relevant content. However, when an organization accumulates too much data, the eDiscovery process then incurs additional costs as attorneys are forced to review every data object to determine if the content is relevant to the lawsuit or is privileged and not subject to the eDiscovery request. Moreover, many data objects are duplicates, and therefore add to the overcollection problem that most organizations experience. Overcollection can add millions of dollars to the cost of defending even a single case, while undercollection can cause an organization to lose a case before it even goes to trial because of spoliation and hiding evidence.

- **Excessive storage requirements**
  A critical issue that most decision makers will recognize is the ongoing need to add additional storage resources. Because of the increasing data velocities and volumes that most organizations face, IT departments are regularly forced into purchasing additional storage resources to keep up with demand. Even though the price of storage continues to fall, the volume and velocity of enterprise information continues to outpace the price reductions.

  The cost of storage is calculated using the current average fully loaded cost of the storage tier that an organization is using for file system, SharePoint and email storage and other applications, multiplied by the total amount of storage being used. Many will simply insert the cost of raw storage, such as what they might see at their local electronics store. This is a misguided approach that overlooks the additional costs of performance tiers, floor space, power/cooling, and the cost of backup and disaster recovery, not to mention the labor costs associated with sourcing, deploying and managing storage. The fully loaded cost of tier 1 or 2 storage will be many times the cost of personal computer hard disks. Even if we assume the use of very low cost storage capabilities, such as Amazon Web Services' S3, just the direct cost of an additional 20 terabytes of excess data will be $6,390[3] – a volume that an organization of 10,000 users will generate every 4.1 days!

  However, it is important to note that excessive storage results not only in additional costs, but also in duplicate files, superfluous files, or files that simply should not be stored either because they no longer have any business value or their retention constitutes an unnecessary risk.

- **End user access to content suffers**
  Employees spend a significant amount of time searching for old content for reuse and reference, such as customer contacts, older presentations, contracts, purchase agreements, communications with key customers, and the like. When they cannot easily find the data they need, they either spend an inordinate amount of time searching for this content, or they end up spending time to recreate the data they couldn't find.

[3] Based on Northern California pricing for Standard Storage at $0.026 per gigabyte per month. Source: https://aws.amazon.com/s3/pricing/

Illustrating just how expensive end user search for content can be in a large organization, consider the following example[4]:

- 10,000 employees
- Each employee works 49 weeks per year, five days per week
- Average employee salary is $20 per hour
- Each employee searches for content 15 minutes per day

Using these assumptions, employees spend a total of 612,500 hours per year searching for content, costing $12.25 million in lost productivity. If we very conservatively estimate that the amount of time that employees spend searching for content could be reduced by just 50 percent, this would save an organization $6.13 million annually, or $612.50 per user per year.

An effective data management program will ensure that data can be found quickly, eliminating the need to recreate lost information. While most employees do not search for large quantities of older information on a regular basis, most usually need selected bits of data from older content stores.

THE NEED FOR FILE ANALYSIS

The problems with file growth and the resulting issues with data classification, data governance, excessive storage and end user access to content will continue to get much worse if they are not addressed in a serious way. To properly deal with these issues, organizations must focus on file analysis.

REASONS TO CONDUCT THOROUGH FILE ANALYSIS

There are a number of compelling reasons to conduct thorough file analysis to remediate the problems discussed above:

- **To determine what can be defensibly deleted**
  A survey conducted by the Compliance, Governance and Oversight Counsel (CGOC) found that in the typical organization, one percent of data is subject to legal hold, five percent is subject to governmental regulatory retention requirements, and 25 percent has some ongoing business value. The CGOC's analysis determined that the remaining 69 percent of corporate data had no apparent business value and could be disposed of without incurring any legal, regulatory or business consequences. For example, for the majority of employees, the need to search for and review an email message older than two weeks almost never occurs. As a result, the probabilities of overall data reuse drop off rather quickly, approaching one percent after just 15 days.

  There are three important issues to consider when defensibly deleting data:

  - First, it is essential to ensure that the disposal process is part of an up-to-date and well-documented corporate policy.
  - Second, information can be safely deleted only if it is not subject to any current legal holds or government requests.
  - Third, employees must be educated about the need to delete unnecessary data and the use of retention/disposal policies to avoid conflicts that may arise if IT, compliance or some other group deletes "their" data.

  Timely and defensible disposal of information reduces the risk of involvement in a future legal case or government information request, reduces the cost of eDiscovery review and storage, and increases employee productivity. Defensible

[4] http://www.zlti.com/wp-content/in%20the%20news/12.14.15_ZL%20(ACC)_Article_ROI%20for%20IG%20(L.%20Sharp).pdf

disposal is an important variable when calculating the return-on-investment (ROI) of a data governance program.

- **To determine what must be retained**
  Some retention obligations are fairly straightforward, such as those that affect broker-dealers or healthcare providers, for example, although even these regulations are open to interpretation by regulatory authorities and the organizations they oversee. For most organizations, however, determining the content that must be retained is often a difficult task because retention obligations fall into something of a "gray area" for many types of content. For example, content that may be necessary to retain for litigation related to wrongful termination or product liability does not have a legally prescribed retention obligation associated with it. However, legal counsel can establish best practices based on legal opinions, precedent, statutes or regulatory rulings for the retention of various data types.

  A well-implemented and managed file analysis capability can properly classify content and help an organization determine what it must retain and when it is no longer subject to retention obligations. In the absence of such a file analysis capability, decision makers cannot properly determine what should and should not be retained, and so face the risk of retaining too little and being out of compliance with legal and regulatory obligations; or retaining too much, thereby incurring higher costs from excessive storage and satisfying their compliance obligations.

- **To determine what cannot easily be classified**
  While some files can be easily classified because their contents are identifiable as subject to a compliance obligation, other files may not fall neatly into an appropriate classification. A robust file analysis capability can help organizations determine the types of information that cannot be easily classified. Once this information has been identified, a robust file analysis solution should be able to perform additional content-based analysis to help further classify this information.

- **To find and protect sensitive data such as PII, PHI, PCI**
  The ability to find and protect sensitive data, such as personally identifiable information (PII), protected health information (PHI) or data that is subject to PCI DSS requirements is a compelling reason to conduct a thorough analysis of files. If this information is left unprotected and unmanaged, the result can be a breach of this data that can lead to damaging consequences. For example, a breach of PII could result in a violation of state-level data breach notification laws, the Gramm-Leach-Bliley Act, or FINRA requirements, depending on the industry in which an organization is involved. A breach of PHI could result in significant fines from the US Office for Civil Rights and the Attorneys General in the state(s) where the breach occurred. A PCI non-compliance violation can result in fines of up to $100,000 per month.

- **To determine and implement appropriate retention periods and access rights**
  Finally, file analysis can help an organization determine the appropriate retention periods and access rights for various types of content. Based on the content of a file, legal and/or regulatory retention periods can vary from one year to indefinitely. Further complicating the issue of determining retention periods is that litigation holds require open-ended retention of various types of content. Robust file analysis capabilities can help decision makers determine how long to retain various files so that legal and regulatory compliance is satisfied. The ability to take policy-based actions directly from the file analysis solution is an important capability.

FILE ANALYSIS IS NOT A ONE-TIME EXERCISE

It is essential to note that file analysis is not a one-time *project*, but rather an on-going and continuous *process*. A single file analysis exercise will address the immediate problem of classifying an organization's data at a point in time, but the continual addition and modification of files will almost immediately make any sort of file analysis obsolete – over time, it will become more so as additional content is added, as existing files are modified, and so forth.

WHAT ORGANIZATIONS NEED TO DO

The details of establishing a file analysis process will depend on a variety of factors, including the regulatory environment in which an organization operates, the jurisdictions in which it has operations, the legal obligations it faces, management's tolerance for risk, and other factors. However, we recommend the following with regard to creating a workable file analysis process:

- **Focus on continuous file analysis**
  As noted above, continuous file analysis is an essential best practice. A one-time file analysis project will yield minor, temporary benefits, but generally will not be worth the effort unless it is ongoing. As noted earlier in this white paper, file growth averages 40 percent per year, and so a single file analysis effort will be largely out-of-date and of little value just a few weeks after its completion.

- **Establish long-term policies**
  It is also essential that organizations establish long-term policies focused on creating appropriate data classifications that will allow them to properly classify existing and future data. This is a key element that will enable organizations to identify critical data assets, apply appropriate retention to them, and archive them properly.

- **Integrate file analysis into existing governance programs**
  Finally, organizations must integrate their file analysis process into the existing corporate information governance strategy and processes to ensure that file analysis efforts support the superset of policies and processes focused on governance. File analysis cannot be a separate, disjointed process or else it will not add value to the organization's information governance capability.

RECOMMENDED BEST PRACTICES

There are a number of best practices that Osterman Research recommends in the context of establishing a file analysis process:

- **Identify the key stakeholders**
  An essential first step in the development of a file analysis system is to identify all of the relevant stakeholders in an organization: data owners, data managers, IT and records management, as well as any external stakeholders like business partners. In short, the individuals who own and manage key data assets must first be identified before any file analysis processes or technologies can be implemented.

- **Implement an appropriate mechanism for analyzing file metadata and content**
  Because most of the information required to classify information can be obtained from file metadata, it is important to implement an appropriate mechanism to analyze this metadata. However, we recommend the use of an iterative process whereby metadata can be analyzed as a first step to provide an initial assessment of an organization's files and an appropriate remediation strategy. Metadata analysis is a useful initial step in file analysis, but in order to adequately analyze Protected Health Information (PHI), Personally Identifiable

Information (PII), or data subject to the Payment Card Industry Data Security Standard (PCI DSS), content analysis is necessary.

- **Apply appropriate classification to existing data**
  Next, it is necessary to apply the appropriate classification to existing data, such as data that can identified as candidates for deletion, data for long-term retention, data that is subject to legal hold, or data that requires further analysis to determine exactly how it should be classified. The iterative process noted above can be used to address file identification, classification and remediation in a phased approach that is more manageable and cost-effective.

  As a part of this effort, it is essential to identify and protect critical data types that must be encrypted in-transit, at-rest and in-use. These include PHI, PII, data subject to PCI DSS, and customer-owned data.

- **Implement other changes, as needed**
  Finally, it is important to be able to assign or change access privileges for data based on its sensitivity, confidentiality, coverage by regulations, "need to know" and the like. For example, HR records – particularly medical information – should be accessible only by appropriate HR staff members, benefits administrators and senior management in order to comply with privacy guidelines. Sensitive customer data, such as payment information or Social Security numbers, should be accessible only by those who have a direct need for this information. Intellectual property should be protected with specific access privileges so that the potential for theft of this information is minimized.

SUMMARY

File shares contain close to one-half of the content that organizations possess, yet most have not yet identified, classified or appropriately managed this content. The result is that organizations increase their risks of non-compliance with their legal or regulatory obligations, they increase the risk of data breaches and loss of intellectual property, and their data storage costs are higher than they need to be. To address these issues, organizations should implement an ongoing file analysis process that will properly inventory corporate data, classify it properly, and allow decision makers to establish appropriate retention periods and deletion practices for this data. The file analysis process should be part of an organization's overall information governance strategy designed to minimize corporate risk and maximize its ability to comply with its legal and regulatory obligations.

ABOUT ZL TECHNOLOGIES, INC.

ZL Technologies makes **Unified Archive® software** (ZL UA) and ZL File Analysis and Management to enable large enterprises to manage all unstructured content, such as email, files, and instant messages to satisfy corporate needs for eDiscovery, records management, regulatory compliance, information governance, 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 EA). ZL'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, 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.