commvault
solving forward®

Volume 4 Issue 1

Turning Dark Data into Smart Data

How Email and File Level Analytics Can Lead to Greater Business Value in the Age of Information

Welcome

Many organizations are discovering that they lack both the policy and technology needed to efficiently manage and understand data for business insight, intensifying the demand for better retention, access, discovery and recovery of business critical information. The growth of data is causing the forward thinking enterprise to finally look to address the issue of dark data, not only to curb mounting storage costs but to gain a true understanding of unstructured data.

The Challenge: Balance Liability with Potential

Dark data is not hype. In fact, understanding data continues to be the most pressing issue when it comes to aging data.[1] The old adage "What you don't know can't hurt you" no longer reigns true in today's business climate and it's what makes the concept of dark data so vexing. According to Gartner, dark data is "the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes..." Dark data can include legacy file shares, backup tapes, archives and former employee email stores that are predominately unclassified and not visible or accessible. The challenge of dark data then becomes to balance its liability with the potential profit gains from using information more strategically. A recent survey found that 70% of organizations reported that dark data will have a negative impact on storage and 53% said it would increase risk[2]. How organizations handle the dark data challenge will become a key foundation for corporate competiveness, business intelligence and innovation, impacting every sector and every business function.

[1] Gartner, Does File Analysis Have a Role in Your Data Management Strategy, March 2014
[2] IDG Dark Data Survey, August 2014
[3] Compliance, Governance and Risk Oversight Council, 2012

Featuring research from
Gartner

Consequences of Data Sprawl-You Don't Know What You Don't Know

With data located in multiple repositories, both inside and outside the data center, organizations today struggle to find information when they need it. Information managers often struggle with having little to no insight into what data is being created, limited control over how it is being stored, and almost no understanding of its business value. ... A common example of this is the increasing amount of data being created on laptops and endpoint devices which is often not well understood. When it comes to information lifecycle governance, organizations often choose to lean on cold storage tape vaults to keep every scrap of data out of the paralyzing fear that they may throw away something of value. Recent studies suggest that 69 percent of a company's stored data has absolutely no value to the organization[3]. Not only is this content taking up valuable and expensive space, but it can also become a liability if not properly managed. The rapidly increasing volume, complexity and disparity of data creates a challenge for organizations as they try to efficiently discover information needed to address corporate litigation, internal investigations, public information and audit requests.

Combine data growth, increasing data diversity and evolving data retention requirements, and you get a trifecta of problems that scream for an integrated solution to unlock the value of all that data. This represents a clear opportunity to address these evolving dynamics with a future-proof strategy that's supported by adaptable technology instead of shortsighted reactions to solving immediate storage needs. In doing so, IT departments can be viewed more strategically while lowering costs and governance risks.

Getting an Information Governance Construct in Place

Central to the idea of managing and understanding dark data, is identifying the ways in which multiple stakeholders across the organization need to use information. By 2018, 25% of progressive organizations will manage all their unstructured data using information governance and storage management policies, up from less 1% today[4]. Yet, this is a big undertaking and often organizations don't know where to start. Different groups have distinctly different needs, and speak different languages. Not all data is created equal and the key consideration is what data is needed to solve a particular business challenge. The industry standard, Information Governance Reference Model (IGRM), offers solid guidance on establishing a governance construct (see Figure 1). As a first step, the IGRM aims to align Legal, IT, Records Management, line-of-business leaders and other business stakeholders within organizations and seeks to facilitate dialogue among these stakeholders by providing a common language and reference for discussion and decision-making based on the needs of the organization.

According to the model, good corporate governance means collaboration to make smart use of the information your business creates every day to get closer you to your clients, understand your opportunities better and differentiate from the competition in order to grow revenues. IT organizations that fail to communicate cross-functionally and centralize policy and retention standards will suffer continuously increasing storage costs and increase their governance risk.

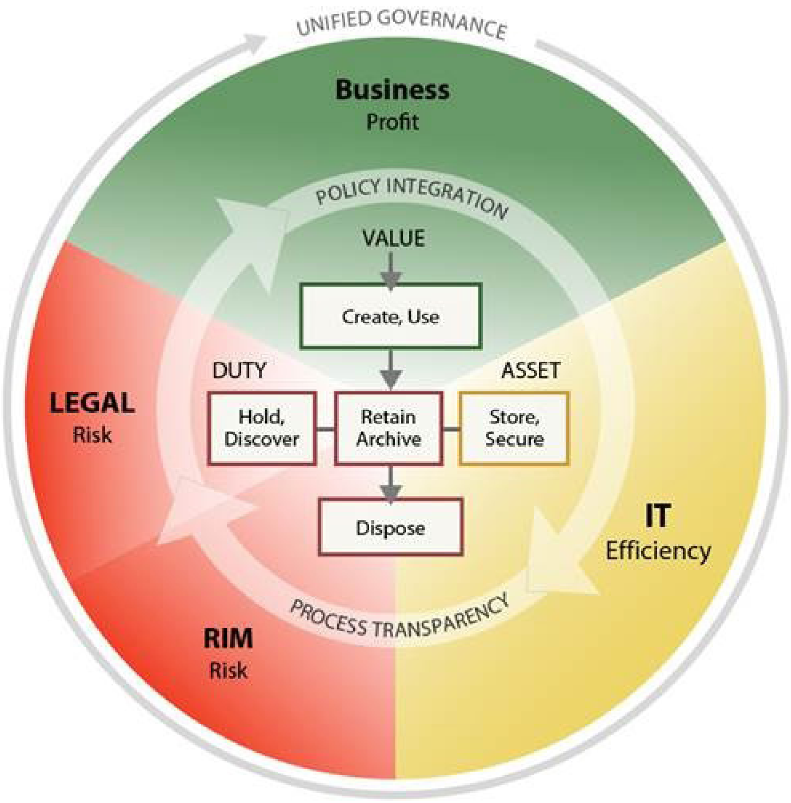
Enter CommVault. CommVault is a software company that is 100% focused on helping customers derive value from their data, across the enterprise. Our flagship product, Simpana™ software, is a single-platform to discover, protect, manage and access information across its entire lifecycle. As a "Leader" in Gartner's 2014 "Magic Quadrant for Enterprise Backup Software and Integrated Appliances" report and a "Leader" in Gartner's 2013 "Magic Quadrant for Enterprise Information Archiving" report, we believe CommVault provides industry-leading technology and superior feature-to-feature performance to address all data and information management needs, providing benefits for every person who interacts with information in the organization.

Simpana for Email and File Level Analytics: No Longer a "Nice to Have"

After establishing the right set of stakeholders across the organization, file and email discovery or analytics is generally the first step in implementing a strategy to address the issue of dark data. Today we see this being utilized in many areas outside of archive, helping organizations find data and the value of it for retention, compliance and or removal/deletion. Most vendors do not

[4] Gartner Inc. Innovation Insight: File Analysis Innovation Delivers an Understanding of Unstructured Dark Data, Alan Dayley, March 2013

FIGURE 1 Information Governance Reference Model, www.edrm.net



Source: CommVault

have robust tools to provide discovery before or after implementation or have reporting tools that utilize a non-integrated product that requires additional management. Imagine having the detail needed to report, find, view and collect information about your environment as you continue to grow or add users and applications. Storage analytics is a key component and one of the main areas companies should be focused on – choosing a solution that can provide detailed analytics to help you make the correct business decisions should be a requirement, not a luxury.

CommVault has perfected methods to intelligently leverage stored content while efficiently managing it and maintaining accessibility throughout its lifecycle. With Simpana, all managed data is kept in the ContentStore which provides a scalable, hardware-agnostic, virtual repository combined with an intelligent index that simultaneously supports data protection, archive, and storage infrastructure reporting operations. This includes often siloed information such as data from endpoints. The notion of a single data store that eliminates redundancies and separate silos is compelling on many levels, including the opportunity to reduce the strain on congested IT networks, restricted hardware/software budgets and overburdened administrative teams. A holistic approach that captures data once and then repurposes the copy for data protection and preservation is the key to getting the right data into the hands of the right people so they can turn it into something more meaningful and actionable for the business.

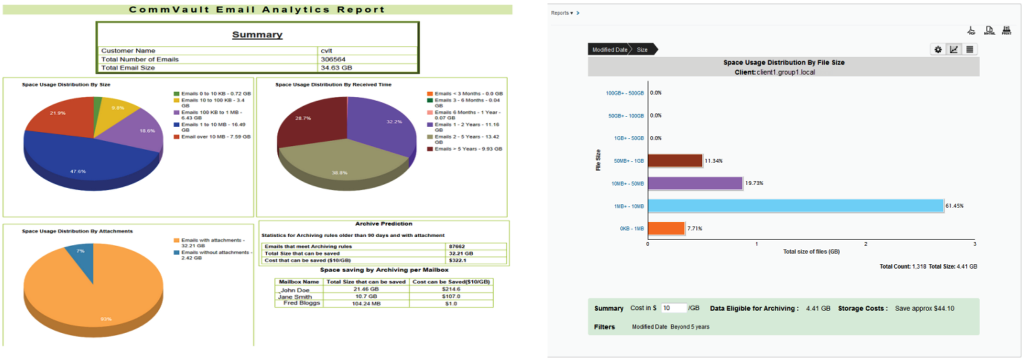
Using the ContentStore, CommVault has laid a foundation for improved email and file analytics, providing a holistic view into content, how many copies exist, who owns it, why it's kept it and where it exists. Simpana analytics tools differ from traditional storage reporting tools because the technology doesn't just report on simple file attributes, but can also provide critical contextual information; with the ability to analyze, index, search, track, report on metadata, and even content. Aligned data sets can be used for trending, patterns or other means of analytics. Centralized reporting enables business and IT leaders to make more informed decisions with their data while bolstering analytical skills. Easy export and view options are available and storage costs can be assigned to provide customized, real-world predictions related to your environment. Putting a cost to the report gives you an immediate look into the savings applied by the archive rules or disposition.

Organizations also can extend their view into the business with embedded intelligence and analytical tools that provide granular insights into the ever-evolving role data can, and should play, in driving business direction. With robust reporting and predictive tools, it's much easier to forecast, analyze and budget properly for the ongoing onslaught of data without compromising data integrity, security, accessibility or accountability.

CommVault® Simpana® Software Delivers

In the world of Dark Data, any opportunity to simplify and centralize the management of information to gain a more granular understanding of data is a step in the right direction. Organizations that are still using legacy retention methods and "keeping everything forever" will encounter insurmountable challenges in accessing massive amounts of information for discovery and compliance purposes. Forward-thinking companies will be able to take full advantage of a future-proof solution that elevates overall information management while providing appropriate access to business-critical information as it ages. By embracing the Simpana Email and File Analytics features and CommVault, you will be able to:

FIGURE 2 Simpana Email and File Level Analytics Reports



CommVault Email Analytics Report

Summary

Customer Name	cvlt
Total Number of Emails	306564
Total Email Size	34.63 GB

Space Usage Distribution By Size
21.9%
9.8%
18.6%
47.6%
Emails 0 to 10 KB - 0.72 GB
Emails 10 to 100 KB - 3.4 GB
Emails 100 KB to 1 MB - 6.43 GB
Emails 1 to 10 MB - 16.49 GB
Email over 10 MB - 7.59 GB
Space Usage Distribution By Received Time
28.7%
32.2%
38.8%
Emails < 3 Months - 0.0 GB
Emails 3 - 6 Months - 0.04 GB
Emails 6 Months - 1 Year - 0.07 GB
Emails 1 - 2 Years - 11.16 GB
Emails 2 - 5 Years - 13.42 GB
Emails > 5 Years - 9.93 GB
Space Usage Distribution By Attachments
7%
93%
Emails with attachments - 32.21 GB
Emails without attachments - 2.42 GB

Archive Prediction

Statistics for Archiving rules older than 90 days and with attachment

Emails that meet Archiving rules	87662
Total Size that can be saved	32.21 GB
Cost that can be saved ($10/GB)	$322.1

Space saving by Archiving per Mailbox

Mailbox Name	Total Size that can be saved	Cost can be Saved($10/GB)
John Doe	21.46 GB	$214.6
Jane Smith	10.7 GB	$107.0
Fred Bloggs	104.24 MB	$1.0

Total Count: 1,318 Total Size: 4.41 GB

Summary Cost in $ 10 /GB Data Eligible for Archiving : 4.41 GB Storage Costs : Save approx $44.10

Filters Modified Date Beyond 5 years

Source: CommVault

Space usage reports for both files and email. CommVault analytics for file, email and PST's can be executed without a CommVault installed environment and made available via the CommVault cloud. Once a CommVault environment is deployed, ongoing analytics can be enabled to provide the same detail on premise to continue investigation or monitoring.

Control Growth

According to Gartner, the top storage concern for organizations in 2014 is to manage data growth to support business demands[5]. When it comes to getting to grips with the mammoth task of dark data, Simpana email and file analytics tools deliver enterprises with the understanding required to clean up legacy and current data, by identifying what can moved to lower-cost storage, and that which can be deleted.

- Gain actionable understanding of large data volumes
- Simplify management of data across on-premises, cloud and hybrid environments
- Keep only what has business, compliance or evidentiary value, cutting storage costs by 70%

Support Compliance & eDiscovery

Ultimately, analytics tools allow for faster location of data. When it comes to making a critical business move or managing governance processes, it can be extremely labor intensive to sift through the masses of irrelevant information contained within dark data. This 'run around' either consumes IT management's time and budget, leaving less bandwidth for immediate business needs, or requires a costly outsourced response or settlement. In a corporate landscape, where the risks associated with locating necessary data in the event of a breach or in response to legal action are significant, Gartner expects eDiscovery benefits to drive the adoption of analytics tools as companies begin to address their dark data[6].

- Automate and enforce retention and defensible deletion to reduce risk
- Provide control over access rights and security processes
- Enable cost-effective long-term retention to meet information governance standards

Extract Value

How to get the most business value out of data throughout its lifecycle is a top concern in dealing with aging data. The classification process enabled by Simpana can support a well-defined data strategy and be used to enforce information governance policies. Once data has been evaluated and indexed properly, organizations can better determine how and where to store that data – whether it's locally, in the cloud, or using some combination of these.

- See a complete view of data across all repositories
- Apply, audit and leverage data classifications for better insight, control and security of unstructured data across the enterprise.
- Eliminate the manual processes often associated with the retention and search of corporate information

Conclusion

The main challenge organizations face in adopting email and file level analytics tools is that they are reluctant to face the abyss that dark data represents, which has been ignored for so long. But organizations must bear in mind that at its core, dark data represents untapped opportunities to transform business, and this can only be realized through a better understanding of your data assets. Simpana software's granular view of your data provides the clarity and transparency that's often lacking in today's chaotic data management state, enabling you to lower costs, reduce risk and operational complexity. Maximum value can be extracted from all business critical information in ways that produce tangible business benefits. CommVault's holistic Solving Forward® philosophy offers unrivaled advantages over competitive options and ensures that your organization is well poised to address data challenges today and into the future.

About the Author:

Emily Wojcik is a senior product marketing manager for the Information Management business at CommVault. Emily has specialized in Information Management technologies for approximately 15 years and has developed significant experience in the areas of archiving, compliance, eDiscovery, enterprise search and retention lifecycle management. Emily graduated with honors from Michigan State University with a Bachelor of Science degree in Communications. Emily is a member of ARMA, EDRM and ACEDS.

Source: CommVault

[5] Gartner Data Center, December 2013

[6] Gartner Inc. Innovation Insight: File Analysis Innovation Delivers an Understanding of Unstructured Dark Data, Alan Dayley, March 2013

Research from Gartner

Innovation Insight: File Analysis Innovation Delivers an Understanding of Unstructured Dark Data

Explosive, unstructured data growth is forcing IT leaders to rethink data management. IT, data and storage managers use file analysis to deliver insight into information about the data, enabling better management and governance to improve business value, reduce risk and lower management cost.

Key Findings

- Unstructured data growth is rapidly outpacing structured data and is poorly controlled, stored and managed on file shares, on personal devices and in the cloud.
- Organizations have little awareness of the volume, composition, risk and business value of their unstructured data.
- Instead of addressing the holistic picture of unstructured data, including content, data access and data storage, IT leaders tend to view unstructured data only from the perspective of age, and do little if anything to support information governance.

Recommendations

- Organizations should review the scope of their unstructured data problems by using file analysis (FA) tools to understand where dark unstructured data resides and who has access to it.
- Identify the value and risks of unstructured data, and prioritize unstructured data management needs for classification and information governance, file and identity governance, storage management and content migration.
- Delete redundant or unneeded data once unstructured data is classified and mapped, then move legal, regulatory and stale data for compliance or low-touch retention reasons to lower-cost storage, and assign policies for retention and access.

Strategic Planning Assumption

By 2018, 25% of progressive organizations will manage all their unstructured data using information governance and storage management policies, up from less 1% today.

Analysis

Innovation Description/Definition

FA differs from traditional storage reporting tools not only by reporting on simple file attributes, but also by providing detailed metadata and contextual information to enable better information governance and storage management actions. These tools analyze, index, search, track and report on file metadata and, in some cases, file content, to assist in taking action on files according to what was collected.

FA tools offer a variety of options, for example:

- Storage management FA tools focus on the frequency of unstructured data use, identifying data associated with different applications and taking action on that data, such as migration to an archive or a tiered storage layer, or to be deleted.
- File and identity governance tools focus on who has access to which files and can identify and correct anomalies directly through the tools or through integration with Active Directory.
- Another class of FA tools provides a full content index, and is used for classification and information governance. These tools focus on what actions to take on unstructured data for information governance, e-discovery (such as legal hold), archiving, defensible deletion and storage management.

Business Impact

FA provides business value in the following ways:

- Reducing risk by identifying which files reside where and who has access to them, allowing remediation on areas such as eliminating personally identifiable information, corralling and controlling intellectual property, and finding and eliminating redundant and outdated data that may lead to business difficulties, such as multiple copies of a contract

- Reducing cost by reducing the amount of data stored
- Classifying valuable business data so that it can be more easily found and leveraged
- Supporting e-discovery efforts for legal and regulatory investigations

Unstructured data is growing at a much faster pace than traditional relational database management system (RDBMS) data, now accounting for well over half of all data storage by organizations and presenting a major challenge to manage. This data resides in file shares, email, SharePoint, file sync and share (FSS) applications, and individuals' laptops and desktops. Organizations are better equipped to take action when data access, usage, associations, redundancy and content are fully understood. Identifying the users and groups with access to data, matching them to who should or shouldn't have access and recognizing anomalies reduces security risks and increases organizational effectiveness. Understanding data's use and its associations with applications can identify the data to be moved to lower-cost storage or to be deleted. Going beyond the metadata and understanding the content of dark data (data gathered by companies that is not part of their day-to-day operations; see Note 1) can provide even more value as organizations initiate information governance strategies.

FA improves e-discovery readiness through searching, indexing and categorizing unstructured data that can be fed into archiving, enterprise content management (ECM) and e-discovery tools.

FA tools enable IT to create a visualization of unstructured data that can be presented to others in the organization so that they can make decisions based on the data. Key to this process is the creation of an effective cross-functional team of IT, lines of business (content experts), legal and compliance stakeholders that work together to use the data generated from the FA to make better information governance decisions.

FA tools enable views into an organization's unstructured data, much as master data management (MDM) does for structured data. As organizations visualize the content of their unstructured data, the use cases will move beyond storage management and governance into business support.

Use cases for FA include:

- Classifications for information security purposes
- Enforcement of information governance and retention policies
- Support for archiving/e-discovery and business reporting
- Storage management
- Data center or server consolidation
- Cloud migration
- Support for management of data as a result of mergers or acquisitions
- Data deletion/legacy data cleanup
- Copy data management

Examples of specific use case scenarios:

- Organization: Manufacturing Company:
  - Use Case: Storage management
  - Objective: Cleanse a file share environment that contained 30TB of file system data.
  - Implementation: Initially delayed because of the newness of the FA approach internally. Once permissions were received, the file discovery and analysis project took less than three months to complete.
  - Outcome: A total of 50% more content was identified beyond the original 30TB. After analysis, almost 60% of the data was identified for removal. As a result, the CIO authorized policies for the deletion of the data (currently being implemented). The ROI (payback) is two years, not including the resultant cost avoidance deferral of the storage hardware purchase.
  - Buyer: CIO and storage team

- Organization: Oil and Gas Company:
  - Use Case: Migration to SharePoint
  - Objective: Clean up many sites worldwide that had unknown tens of terabytes of unstructured data prior to migration to SharePoint. Remove sensitive data, get a document countdown and provide good-quality data.
  - Implementation: The implementations and initial cleansing at all sites worldwide were completed in one year. Massive savings were realized, as the tool identified more than 30% of data that could be removed prior to the migration. The FA product generated metadata about the files to be tagged, and to assist in reorganizing the storage. This information was passed on to a migration tool. The project is still running and is being funded with new success factors focused on improving business user engagement.
  - Outcome: The FA tool reduced the time and costs for migration, and provided metadata tags that could not have been practically generated by manual processes.
  - Buyer: CTO
- Organization: Financial Services:
  - Use Case: Reporting
  - Objective: The organization had previously deployed another tool and had achieved limited success without delivering to full expectations. While storage savings was a factor, the overall drivers weren't 100% clear. However, one long-term objective was to add metadata prior to migration to SharePoint.
  - Implementation: During the six-week project, the tool identified 100TB of data — of which only 35TB were unique. Of the 35TB, the FA tool identified 15TB for removal.
  - Outcome: At the onset, there was data everywhere that was being poorly managed. The storage team was able to identify the potential to go from 100TB to 20TB of necessary data.
  - Buyer: CIO and storage team

As organizations implement FA tools to assist in general information governance activities, more use cases will become apparent. The impact of understanding and taking action on unstructured data will be greatest on organizations that generate millions or billions of files from many applications. The potential for a high payback will help drive the adoption of these tools (see Figure 1).

IT Impact

The impact of FA tools on IT can be dramatic. Storage administrators now have a tool that shows detailed information about the data being stored to take to business owners so that more-informed decisions can be made on data retention and optimized data protection policies. File shares can be dramatically cleaned up by deleting old, orphaned and irrelevant data, greatly reducing the burden on IT when an e-discovery, regulatory or compliance request is presented. Data generated by FA tools can be integrated with data loss prevention (DLP) tools to provide proactive management of intellectual property.

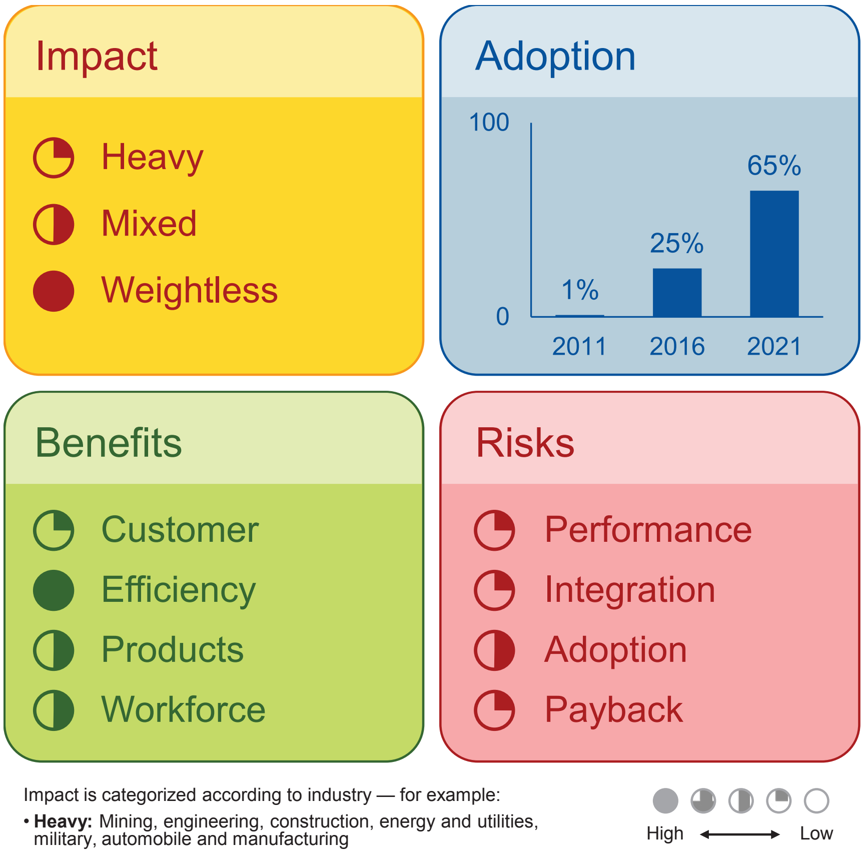
Adoption Rate

Gartner considers FA to be a high-impact technology, and estimates that it will take two to five years before it reaches mainstream adoption. Adoption rates will differ according to use cases. FA for storage management purposes, namely for migrations or technology refreshes, may evolve more quickly as organizations view massive amounts of data stored on file shares as cumbersome to move in totality.

FA for classification, governance and e-discovery will increase in frequency and importance as organizations ascertain the legal, compliance and intellectual property loss potential around their unknown unstructured data and associated costs to manage.

Figure 2 shows the responses of organizations at the December 2012 Data Center Conference in Las Vegas to the question, "Do you have management tools in place to help better understand your unstructured data?"

FIGURE 1 Innovation Window for File Analysis



Impact is categorized according to industry — for example:

- **Heavy:** Mining, engineering, construction, energy and utilities, military, automobile and manufacturing
- **Mixed:** Consumer packaged goods (CPG), logistics, retail, pharmaceuticals, local government, education and healthcare
- **Weightless:** Insurance, media, banking, advertising and intelligence

Source: Gartner (March 2013)

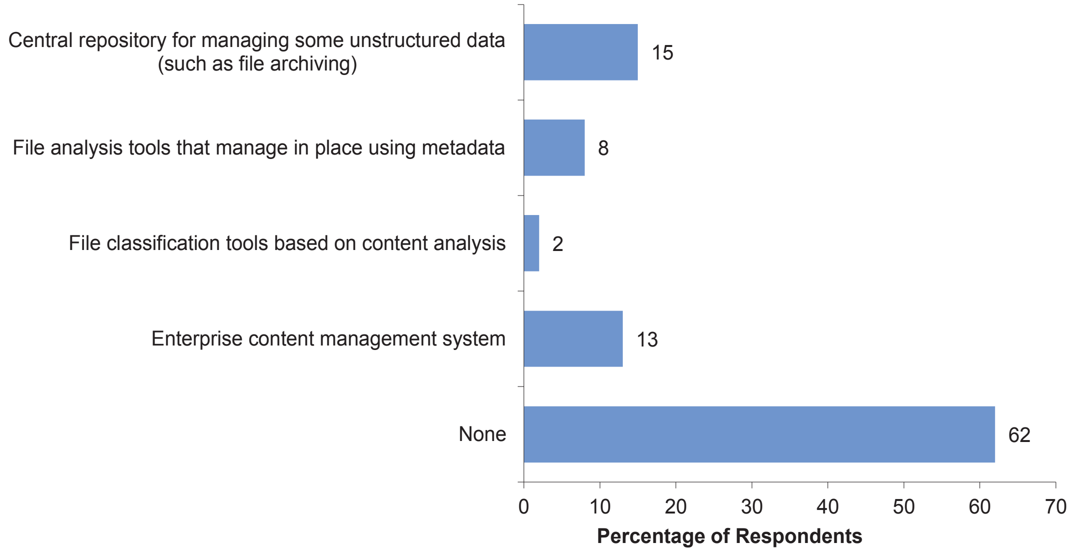
Figure 3 shows the responses of organizations at the December 2012 Data Center Conference in Las Vegas to the question, “What type of aging data represents your biggest challenge?”

Risks

The main challenge organizations face in adopting FA is finally facing the black hole of data they have ignored for too long. Some organizations have literally said they are afraid of what they might find.

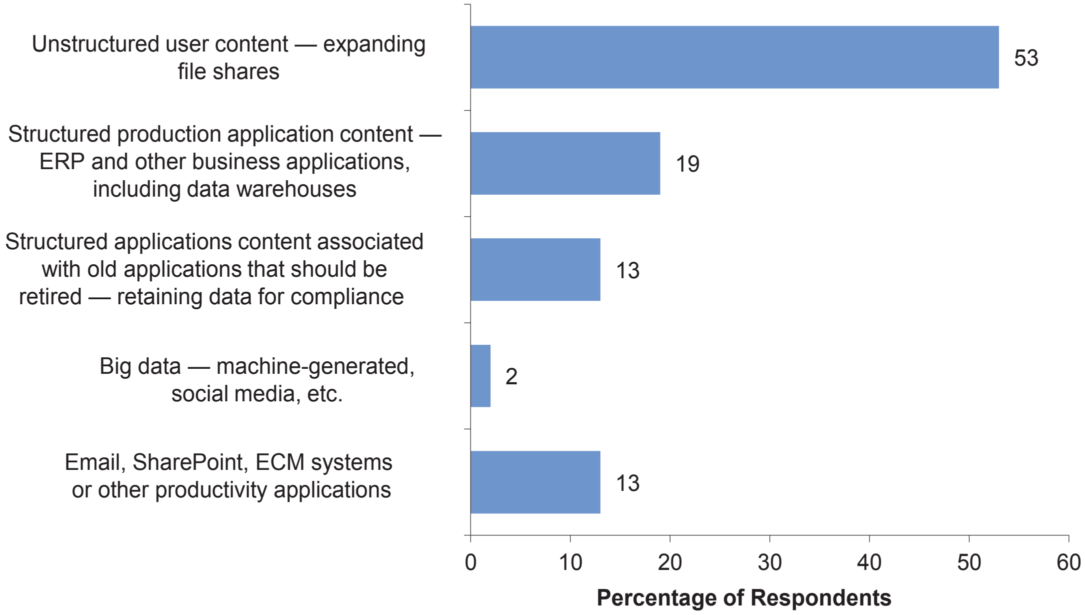
Yet, FA technology is relatively risk-free, as the main outcomes of data scans are data reports and visualizations. Based on these, organizations can take action on segments of the data. Risk arises as a result of poorly defined policies on what to do with the data once it’s classified, and/or the improper movement of the data in response to the classification. For example, if an organization runs a report on access times for documents and then deletes everything that has not been touched in three years, issues can arise if regulations require the organization to keep some of the data longer.

FIGURE 2 Organizations With Management Tools in Place to Better Understand Unstructured Data



N = 52
Source: Gartner (March 2013)

FIGURE 3 Organizations' Aging Data Challenges



N = 52
Source: Gartner (March 2013)

Legal problems also may arise if the data is moved from one repository to another and the chain of custody is not maintained. While FA tools may have a slight impact on system performance, most are configured to run at a low rate of impact on CPUs, or to run during idle periods.

Key Technology and Service Providers

Technology providers with FA capabilities have varying backgrounds, including storage management, e-discovery, indexing and FA, which may be the providers' primary product areas. The providers offer at least one of the following capabilities for either file metadata or content reporting: storage management, file/identity governance, classification/information governance and content migration. Sample vendors include Acaveo, Active Navigation, Aptare, Autonomy (an HP company), AvePoint, Clearswift, Content Analyst, dataglobal, Dell-Quest Software, EMC, Equivio, FileTek, IBM-Stored IQ, Idera, Imperva, Index Engines, Litera, Metalogix, Northern, NTP Software, Nuix, Proofpoint, Recommind, RSD, Symantec, Tarmin, Varonis Systems and ZyLAB.

Note 1. Definition of Dark Data

Gartner defines dark data as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value.

Gartner Research G00250750, Alan Dayley, 28 March 2013

About CommVault

A singular vision—a belief in a better way to address current and future data management needs—guides CommVault in the development of Singular Information Management® solutions for high-performance data protection, universal availability and simplified management of data on complex storage networks. CommVault's exclusive single-platform architecture gives companies unprecedented control over data growth, costs and risk. CommVault Simpana software employs individually-licensable modules designed to work together seamlessly from the ground up, sharing a single code and common function set, to deliver superlative backup and recovery, archive, replication, search and resource management capabilities. More companies every day join those who have discovered the unparalleled efficiency, performance, reliability, and control only CommVault can offer. Information about CommVault is available at www.commvault.com. CommVault's corporate headquarters is located in Oceanport, New Jersey, in the United States.

Turning Dark Data into Smart Data is published by CommVault. Editorial supplied by CommVault. is independent of Gartner analysis. All Gartner research is © 2014 by Gartner, Inc. All rights reserved. All Gartner materials are used with Gartner's permission. The use or publication of Gartner research does not indicate Gartner's endorsement of CommVault's products and/or strategies. Reproduction or distribution of this publication in any form without prior written permission is forbidden. The information contained herein has been obtained from sources believed to be reliable. Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Gartner shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The opinions expressed herein are subject to change without notice. Although Gartner research may include a discussion of related legal issues, Gartner does not provide legal advice or services and its research should not be construed or used as such. Gartner is a public company, and its shareholders may include firms and funds that have financial interests in entities covered in Gartner research. Gartner's Board of Directors may include senior managers of these firms or funds. Gartner research is produced independently by its research organization without input or influence from these firms, funds or their managers. For further information on the independence and integrity of Gartner research, see "Guiding Principles on Independence and Objectivity" on its website, http://www.gartner.com/technology/about/ombudsman/omb_guide2.jsp.