minereye
See Beyond Data

White Paper

Data Leakage Prevention, Evolution or Revolution?

By Malcolm Harkins, Chief Security & Trust Officer, Cymatic
Yaniv Avidan, CEO & Co-founder, MinerEye

October 2020

Contents

An Obsolete Concept: Data Leakage

Stop using the term "data leakage". The concept that data can leak out from the perimeters of an organization's network has become obsolete. That horse has left the barn some time ago, since the unstructured data pandemic[1] and the use of cloud infrastructure, platforms and services.

What is Unstructured Data Anyway?

A good place to start is by first understanding what is considered ***structured data***. Some structured data examples include airline reservation systems, inventory management, ERP systems, CRM platforms – systems in which data fits neatly into a relational or hierarchical database. This rigid structure allows tools to easily query and analyze the data to make business decisions.

In contrast, ***unstructured data*** is usually not well organized. It's stored in easily accessible and in shareable formats. Examples include Word documents, PDFs, spreadsheets, text messages, and emails. These formats make communication simple and quick. Unfortunately, that ease-of-use and access also make unstructured data more vulnerable to unauthorized access. Official records are often in the form of unstructured data, including documents like business plans, product designs, contracts, commercial information, and often enough, customer data. Despite its volume and frequency in organizations' networks, it's estimated that around 90 percent of unstructured data is never analyzed[2]. According to projections from Gartner, 80 percent of worldwide data will be unstructured by 2020.

Gartner coined the term "dark data" describing it as data that is collected but not used for anything more than its intended purpose. Dark data can be interpreted as a subset of big data which constitutes the biggest volume. Dark analytics focuses mainly on raw text-based unstructured data that has not been tapped or analyzed before for example, text messages, emails, audios, videos, images, customer information, log files, previous employee information, raw survey data, financial statements, and account information.

Dark data can be found in the following formats:

- **Traditional unstructured data**: Includes untapped data that remains available within organizations but is not explored such as emails, documents, and messages in text-based form that remain untouched.

[1] https://www.information-age.com/the-unstructured-data-pandemic-123481132/, Kevin Widdop, InformationAge, March 2019

[2] https://www.analyticsinsight.net/understanding-dark-analytics-potential-advantages-risks/ , IDC

- **Non-traditional unstructured data:** This dark analytics dimension is comprised of different categories that cannot be mined and analyzed using traditional analytics techniques. This includes audio and video files, still images that could not be explored until now. This data can be analyzed for insight on customers, employees, markets, and operations.

That folder where everything is "dumped" as a "resource"

How many times have you looked through a shared drive hierarchy that others claimed to be well organized but proved to be the dumping ground of partially completed files? Did that folder on the file share have an innocuous name? Does anyone really know what's in that folder and who has access to it? Who created it and for what purpose? Typically, it's filled with items for someone else's use but others looking through decided to copy the files to be organized in their own way. But what's in there? Is it that unfinished spreadsheet from HR with employee phone numbers and addresses, and other sensitive information? Most likely, those employees going through the folder nor even the creator may realize the content's sensitivity or that saving it in draft format could present any kind of security risk.

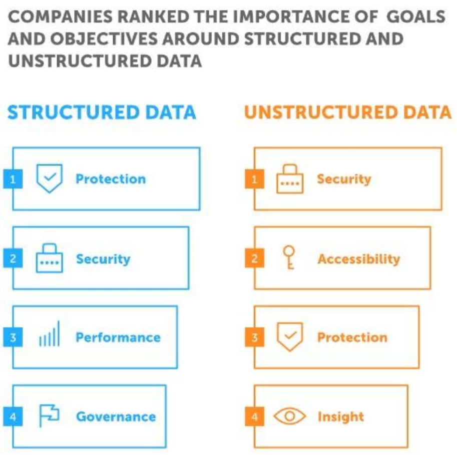
The hard facts about unstructured data[3]:

- The typical organization reports its unstructured data grows 23 percent annually, which means it will double every 40 months. Roughly one fourth (24 percent) cite growth rates of over 40 percent, where total unstructured data doubles every 24 months.
- 82% of organizations manage more than one billion files and objects, while 59% manage more than 10 billion files.

[3] State of Unstructured Data Management, Igneous report

Comparing Unstructured vs. Structured Data

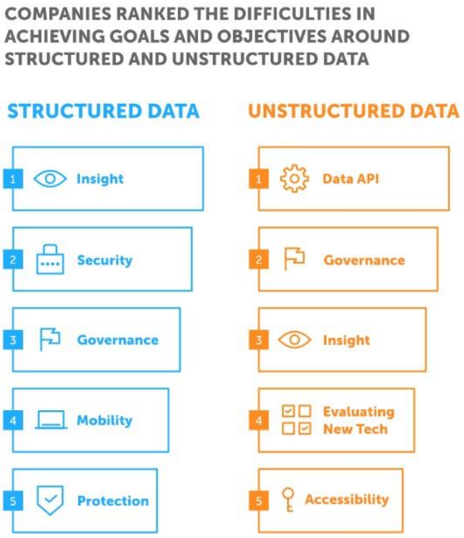
In comparing unstructured vs. structured data management goals, organizations place a high value on data access, regulatory compliance and insight for unstructured data, something they don't rank as highly for structured data. For structured data, values are more for the traditional aspects such as data protection, security and capacity management. This reflects a fundamental shift in what is important when managing structured versus unstructured data. It is much more difficult to achieve access, governance and insight over tens of billions of files and objects than it is with thousands of databases and VMs.



COMPANIES RANKED THE IMPORTANCE OF GOALS AND OBJECTIVES AROUND STRUCTURED AND UNSTRUCTURED DATA

	STRUCTURED DATA	UNSTRUCTURED DATA
1	Protection	Security
2	Security	Accessibility
3	Performance	Protection
4	Governance	Insight

When it comes to unstructured data, organizations face greater challenges with automation, governance and insights than they do with structured data. It makes sense that automating data workflows through API's is more difficult for unstructured data due to the randomness of the data set in size and number of files. The specific APIs one uses to access files varies depending on where the data is stored, and these APIs are not particularly rich. Data accessibility also makes sense for similar reasons.



COMPANIES RANKED THE DIFFICULTIES IN ACHIEVING GOALS AND OBJECTIVES AROUND STRUCTURED AND UNSTRUCTURED DATA

	STRUCTURED DATA	UNSTRUCTURED DATA
1	Insight	Data API
2	Security	Governance
3	Governance	Insight
4	Mobility	Evaluating New Tech
5	Protection	Accessibility

The too frequent "Reply to all" as a multiplier of unstructured data

There are many sources of unstructured data in an organization, which accounts for much of the data growth that we've seen in the enterprise. Estimates dating back to 2003 note that structured data only represented around 15% of all the data we use on a daily basis, and everything else is considered unstructured. Right now, the most significant sources of unstructured data are email and file services; both generating high volumes of data. File services do not just include spreadsheets and Word documents, but also video files, audio files and image files rich data that are exceedingly difficult to control. With email, consider how we "Reply to All" and forward messages that duplicate and proliferate an email many times very often with attachments. Ultimately, the unstructured data is significantly piling up and cluttering our file servers and other infrastructure.

How to manage "unstructured data" in the cloud era

A 2018 survey by IDG, a technology media company, stated that 73% of companies had applications or infrastructure in the cloud, with another 17% expected to make the move in the coming year. In the cloud era, unstructured data growth is hardly news anymore. In fact, the challenge is no longer exponential growth, but rather about keeping control over the data. How can an enterprise protect data while giving access to users, applications, and devices distributed globally?

This scenario has been evolving very quickly. In the last ten years we have gone from storing data locally on premises, to storing it throughout several repositories diverse in their nature and in the way they are accessed. This trend is not going to change any time soon particularly since the "work-from-home" pandemic era. Multi-cloud strategies combined with a myriad of SaaS software creates and distributes data at an alarming rate with little control or protection.

> ***Now that unstructured data within cloud repositories results in easier access and data sprawl, is DLP still relevant?***
>
> ***Given the fact that the cloud was designed to increase data availability and agility, is DLP the right mindset?***
>
> ***Are we fighting yesterday's war?***

GDPR: Its own set of rules around data management

Moreover, in this multi-cloud scenario, new regulations like GDPR demand a totally different approach to data management. Without the right tools, it is practically impossible to comply with changing laws. Another problem is that most cloud solutions create a separate silo of unstructured data storage that has unique application capabilities to enhance data availability.

IT teams need to manage this separate silo independently using an entirely different storage software stack than what the organization might be using on-premises. As a result, the organization ends up running one file system in the cloud and another file system on-premises. One of the most critical aspects of cloud security is to ensure that only authorized personnel have access to the documents and files stored in the cloud. Ultimately, the responsibility of securing cloud data lies between the CIO/CISO and the cloud storage vendor. Liability is a new factor in the already complex equation of managing unstructured data in the cloud.

Many organizations fail to understand that they are still fully liable for managing and protecting their data despite data being remotely backed up, maintained and managed by cloud storage vendors. Companies need to take a few steps to ensure the cloud service they choose can guarantee the security of their data. When it comes to the cloud, there are security and safety concerns especially now that trust in tech giants is broken.

The more companies move their data to various clouds – public, private and hybrid clouds, cloud storage environments, software-as-a-service applications, and so on, the more complicated protecting and securing all their data across multiple environments can be.

Here’s a quick look at the serious implications of data sprawled throughout various cloud environments:

a) Organizations no longer know where all their applications and data lie.

b) With most of their applications and data housed on multiple third-party infrastructure, organizations do not have unified visibility into who is accessing and using their applications and data, which devices are being used for access, or how their data is potentially being used or shared.

c) There are too many data policies to manage including emerging and evolving ones.

d) Organizations do not have insight into how cloud providers are storing and securing their data.

Clearly, the most impactful parameter on security of data across multiple environments is the fact that different cloud providers have varying capabilities, which can result in **inconsistent cloud data visibility, protection, and security.**

How to protect "unstructured data" in the cloud era

The largest data breach in history, courtesy of Yahoo, compromised three billion accounts. Yahoo is also responsible for the second-largest breach, thought to have impacted 500 million users. From the likes of Yahoo and others such as Uber and Bupa that suffered significant breaches last year, we have witnessed an interesting trend. Hackers were found to breach user accounts, not necessarily with a goal of infiltrating corporate applications and databases, but to gain access to extremely sensitive data residing in email and other unstructured file stores.

Think about all the sensitive files that could be associated with just one breached account: Tax or financial statements, personal healthcare data, or banking and credit card information. Not surprisingly, this is the type of information that hackers are after today: sensitive data that is ripe for the picking. Analysts estimate that unstructured data (emails, PDFs and other files that exist outside application or database boundaries) comprises 80% of all enterprise data today. This is a significant challenge for companies, particularly for those who lack adequate visibility into their stored data. Not only do companies struggle to understand what data exists in these unstructured data stores but because hackers often steal copies, it's impossible to know what specific data was taken.

How to reduce the attack surface in the cloud era

Ever since organizations have shifted their business to remote operations due to the COVID-19 pandemic, there has been a dramatic rise in the number of data breaches. In the first half itself, cases of data breaches have been reported in 81 global companies from 81 countries![4] With every hack, consumers become increasingly wary, and boardrooms across the globe face increasing pressure to protect their organizations' data. Yet, when we examine our data vulnerabilities, what are we really looking at? If your organization is only considering the neat rows and columns of a database, then a big part of the picture is missing.

That narrow view leaves out a tremendous amount of data, referred to as unstructured data, and that can leave potential liability unaccounted for. With the growth of unstructured data comes the unfortunate truth that it's much more difficult to control and secure than structured data. For example, if an employee is taking information in the form of unstructured data and moving it elsewhere, they may store the original document or picture on a local file share or send it in an email as an attachment. Within one organization, the process for handling documents could vary across employees and teams, and it's very likely that management has no idea this is happening.

Following the unstructured data explosion and due to the cloud evolution, it is of paramount importance that organizations can automate the analysis of their file / document information. They must be able to respond to changing data and access patterns and requirements, and to satisfy customers, stakeholders, employees, data protection controls, auditors, regulators etc.

Artificial Intelligence and Machine Learning play critical roles in designing automated, intelligent data security strategies:

1. The combination of AI and ML enable staging and organizing the information about data from both content and context perspectives.

2. It can enable and accelerate the multi-dimensional analysis of information attributes with an output of the state of data in near real-time, especially when there is no other centralizing business unit or tools currently doing that.

3. It can play a significant role in simulating an organization's policies as they emerge or evolve before their invoking a label on data thereby reducing the risk of hampering business objectives.

4. It can orchestrate multiple actions such as moving, archiving, encrypting, redacting, preventing access and sharing quickly, while resolving potential policy conflicts before they happen.

It's time to change how we think about DLP

It's time for CIOs, and CISOs to change how they think about information security and part of this process is to adapt the terminology to fit the current landscape. DLP (Data Leakage Prevention) is one of those rigid concepts in data security. Data exists. It is shared and accessed. It is not leaking anymore. It exists everywhere. The challenge is to reduce unstructured data's footprint to only where it should reside and be accessed by the people and machines that must access it at that specific time.

The quality of our information security becomes more and more directly affected by the quality of the way we manage our data. When you say your security is data driven, it does not mean that your data classification relies on pre-defined regular expressions – this is human driven security. We must integrate data engineers into our information security teams. We must start monitoring data much more than systems and security events. That is a first critical step that puts us ahead of the threat actors instead of being reactive to clues and breadcrumbs.

Innovation in Data Governance

Similar to what data warehouse and BI concepts did for operational information systems in the late 90's, MinerEye is doing for data protection controls of the present. It fills in the evolving gap between the data and the privacy, protection, and management policies. It becomes a precursor to policy design, security architecture implementation, disaster recovery, data management and more.

When looking at the zero-trust framework essentials, perimeters have increased in number, becoming more granular, shifting closer to the logical entities they protect, i.e. the data, identities of users, applications, devices, and workloads. Data is one of the segments that requires technologies that continuously analyze and aggregate it, for the relevant people and devices across workloads and networks. Excessive trust, like excessive risk, represents waste and a latent cost to the organization. Continuously assessing risk leads to a leaner, more effective and efficient trust.

Benefits of multi-dimensional file classification

MinerEye takes a multi-dimensional approach in data governance by analyzing unstructured data based on its content, context, user access, and intended use. This approach delivers multiple benefits to an organization such as:

1. **Data Protection**: Protect your sensitive, classified information in a way that doesn't cripple your business operations
2. **Secure Collaboration**: Enable remote working with confidence that files containing sensitive information will not arrive in the wrong hands
3. **Data Discovery**: Identify and manage your dark data – your sensitive data is now completely visible to you including automated analytics that can track changes to its context and benefit from tight management for use cases in data compliance, file storage optimization and reduction of overall risk to undermanaged data.
4. **Data Mapping**: Multi-dimensional data segmentation for use cases like legal hold, intellectual property, privacy regulations, and separation of responsibilities among business divisions.
5. **Data Privacy Compliance**: Fulfill customer requests according to consumer privacy regulations in minutes, match privacy regulation articles to each individual file to check and monitor compliance.
6. **Smart Cloud Optimization:** Before and after cloud adoption, ensure that only required files are stored, thus decreasing processing costs up to 40%, while enabling more efficient use of information by employees.

Three case studies in automated unstructured data governance

The following are three case studies of MinerEye customers. All three required comprehensive visibility of unstructured data alongside with granular, dynamic and multi-dimensional analysis capabilities.

Customer: A global insurer headquartered in Switzerland

One of the world's largest insurers, operating globally throughout 25 countries, is modernizing its IT by moving to the Azure cloud. The cyber defense department that is accountable for securing the data in the cloud selected MinerEye's data protection solution to optimize the move of unstructured data to the cloud by:

- Discerning redundant and obsolete data to save on costly operations on the way to and while in the cloud,
- Discovering personal information location and usage,
- Adding virtual multiple file labeling on sensitive data using AI before integrating with Microsoft Information Protection (MIP),
- Resolving labeling conflicts created by wrong user manual labeling of sensitive file data.

Results delivered:

- By sifting through hundreds of terabytes of diverse and dispersed files storages in a few days, MinerEye's solution had prepared the organization to mass migrate file data in a secure and compliant fashion.
- By modeling and fine-tuning dozens of granular and tailor-made data protection policies for files were optimized prior to integration with Microsoft file labeling.
- 46% reduction in unstructured data footprint by archiving and eliminating duplications and near duplications.
- Significant cost reductions in private cloud infrastructure by leveraging Kubernetes and Docker technologies in a hybrid cloud environment.

Customer: U.S. branch of an international commercial bank

Mega ICBC is a Fortune 500 international commercial bank that operates through 107 branches in Taiwan and 21 branches worldwide. The bank is a major international financial institution with approximately $103 billion in assets, including $9 billion at its New York branch. In 2016 Mega International Commercial Bank of Taiwan was fined a large penalty for not securing Non-Public Information (NPI) in distributed unstructured data repositories, as required by the New York State

Department of Financial Services (NYDFS) requirements. Since then, the bank uses MinerEye AI-powered discovery and classification to encrypt files that contain NPI. The deployment included the following objectives:

- Sampling the bank's existing datasets to train the AI algorithms to discover multi-lingual documents that contain NPI,
- Grouping the NPI file data virtually by business categories to feed into groups such as what should be minimized, encrypted, and labeled for the specific user access rights management,
- Automating an accurate and selective encryption of discovered NPI documents including on-premises desktops.

Results delivered:

- Provided a unified view to NPI in unstructured data that never existed, enabling compliance with NYDFS Consent Order as dictated in 2016 by the New York State Department of Financial Services (NYDFS).
- Savings accrued estimated at several hundreds of thousands of dollars a month for a few months in comparison to a manual service that the bank considered.

Customer: A financial auditor headquartered in Canada

Richter Auditing firm, based in Toronto, Canada, has a cyber risk practice that serves primarily financial services. This business unit is managing their clients' cyber risk and compliance activities. The Cyber risk unit was challenged by the huge duration of time spent by a full team to sit and wade through inboxes and OneDrives of compromised accounts following a breach. One company that hired Richter was required to submit a detailed breach disclosure that included personal information (PI) that was probably compromised by the account hijacking and subsequent data exfiltration.

The service included the following objectives:

- Scanning specific account sources such as employee mailboxes, OneDrive folders to extract PI hidden in files and emails,
- Organizing the PI into a comprehensive and accurate report that can be disclosed to the regulator.

Results delivered:

By using MinerEye's Incident Investigation and Breach Notification solution, Richter was able to reduce its resources dedicated to its customer by 40%.

Will Xiang, VP Cyber and Privacy in Richter said, "*Beyond shaving weeks off of a typical manual scan, by using MinerEye we were able to uncover data files that were randomly labeled and contained thousands of sensitive personal information (PI) within minutes. This confirmed that the combination of MinerEye's AI-based data discovery solutions and our professional cyber services is clearly the fastest and most effective approach in addressing incident response.*"

About MinerEye

Since 2015, MinerEye has enabled organizations to overcome the challenges of analyzing and accessing unstructured data for automated AI-based information governance, data privacy and protection. It automatically scans, indexes, categorizes, and virtually labels every file in unstructured data via proprietary Interpretive AI™, machine learning, and computer vision. MinerEye understands that an effective and efficient organization begins with complete visibility and access to its "crown jewels" in its unstructured data. Despite unstructured, dark data comprising more than 80% of a network according to analysts, it remains a black hole that must be overcome for organizations to protect its network.

To solve this difficult issue, MinerEye automates the instant discovery, mapping, indexing and classification of files holding personal information (PI) and business sensitive information. Customers receive insightful analysis that set file-sharing policies, achieve precise classification, comply with privacy regulations, optimize cloud usage and discover compromised data. MinerEye's customers range from financial services, retail, manufacturing, defense and other sectors around the world. MinerEye is a private, VC-funded company, headquartered in Israel.

Visit us: www.minereye.com Write us: info@minereye.com