G00296550

Emerging Technology Analysis: Data Classification Sophistication Will Be a Useful Tool in Data Protection

Published: 18 April 2016

Analyst(s): Deborah Kish

Data classification will play an integral part in how organizations handle sensitive data. Data classification product and marketing managers can shape messages and increase visibility of their technology by exploring areas such as big data.

Key Findings

- Organizations need data classification tools that are simple to understand and yield consistent classification, but most solutions address only unstructured data.
- In the past three years, data classification tools have become more robust to include persistent classification, have added flexibility and are migrating toward including automation and machine learning capabilities.
- Business intelligence or big data implementations may provide opportunities for data classification technology providers.
- Organizations often apply data classification in a static "set it and forget it" approach that does not reflect the data life cycle; nor will it be useful with keeping up with regulatory compliance.

Recommendations

- Product managers should focus on ease of deployment and ease of use regardless of skill set in any organization.
- Product developers of data classification tools should build in machine learning and automation capabilities that will alleviate the static set-it-and-forget-it approach.
- Marketing leaders at data classification vendors need to emphasize how the role of data classification is changing, and buyers need to embrace new methods, such as implementing solutions that relate to the life cycle of data.

- Sales leaders at data classification providers need to drive the conversation beyond compliance, and help buyers understand the business risks of failing to correctly classify data.

Table of Contents

Strategic Planning Assumption

By 2017, data classification tools will include machine learning and automation capabilities and assist with the data life cycle through big data initiatives.

Analysis

Data classification tools have been around for over a decade, but in the past three years, they have become more sophisticated. When used as part of an organization's "best practices," classification can identify and prioritize sensitive personal, health or financial data that needs to comply with regulations, or protect intellectual property.

The life cycle of data has changed in the era of digital business through the creation, collection, storage, manipulation, analysis and destruction of data. Organizations have quickly become more mobile in that they have many endpoints that serve as data creation and storage points and therefore additional targets for data exfiltration. Organizations struggle with where their data is located and how to apply automated classification techniques and build policies to ensure that sensitive data is protected throughout its life cycle. Many factors about data and its protection exist, and data classification can act as a barrier to those factors, such as, but not limited to:

- Insider threats — Accidental leakage from employees by being a victim of a phishing attack, attaching the wrong file to an email or innocently sending files or information to personal email

for convenience, as well as malicious or disgruntled employees intentionally wanting to harm a business or enticed by a payoff from outside hackers.

- External threats (traditional and advanced) — Individuals or groups of people who are not authorized for access to an organization's assets and pose a risk to the organization through theft or modification of data via phishing, active or silent malware, or sabotage.
- Shadow data — Data that is created using file sharing applications, such as Box, Dropbox and LogMeIn, where data is stored or accessed.
- Data classification is on the Technology Trigger, heading toward the Peak of Inflated Expectations in the Hype Cycle for Data Security. Businesses need to understand the impact of data classification and how to effectively extract the benefits of it; therefore, it is important to create effective messaging around what data classification tools are and how to use them.

In addition, there are trends toward machine learning and automation where data classification providers should consider either "building in" as part of their product or "partnering with" to extend their capabilities. Trends in enterprise content management (ECM) and big data analytics can help with tracking the life cycle of the data. Also, decision makers within organizations are expanding to include departments outside of information technology (IT), chief information security officers (CISOs) and chief data officers (CDOs), such as human resources and finance. Therefore, it is important to know who the buyers are for data-classification-related technologies.

Technology Description

Data classification is the process of organizing data through an agreed categorization glossary that enables effective and efficient prioritization for data security governance policies spanning quality, security and protection, access, privacy, storage, and retention. The most common use cases are for business value, analytics, compliance, discovery, integration, risk, security and privacy. Data classification involves applying metadata information to the data to facilitate the utilization and governance of the data.

There are two ways to treat data: by applying tagging and classification labels. The fundamental difference between the two is simple. Data tagging is essentially providing basic information about the file. It is identifying the "type" of file, such as a Word or Excel document. Classification of data is taking the tag to the next level, where it is informing users of the sensitivity of the context of the data within the file or document. These classification tools have a major flaw in that they help only with unstructured data. Very few products can help with structured data, as they are part of either data loss prevention (DLP) or data-centric audit and protection (DCAP) solutions.

In "How to Overcome Pitfalls in Data Classification Initiatives," Gartner identified three main classification categories:

- "Public" — This is data that is published on your publicly facing website or other official external communications, such as social media feeds and various product collaterals.

- "Internal" — This data is for internal use only, but consists of your routine business communications and documents created as part of your normal, day-to-day activities. This includes the majority of internal email.
- "Confidential" — This data is your sensitive data that typically requires special handling procedures. This can be data subject to regulations, intellectual property, or information that is not publicly known or available internally, such as merger and acquisition documents, corporate financials and human resources data.

Technology Adoption

Organizations are evaluating classification tools by a combination of methods, such as user-driven classification and soon by using automation and by performing data scans to find where their data resides. We also noted in "How to Overcome Pitfalls in Data Classification Initiatives" how organizations build complex and ambiguous data classification schemes that are difficult to understand. Therefore, they also need tools that are easy to implement and use regardless of users' skill sets. In addition, data classification can provide a value-add to organizations that are implementing big data analysis on targeted datasets, therefore exacerbating the need for tools that operate over structured and unstructured data.

Methods of Classification

There are many specific use cases and methods of classification:

- User-driven classification — Where people who create and manipulate data files can apply classifications based on their knowledge of the content. Some solutions allow for right-clicking a file and choosing the appropriate classification (for example, confidential, internal use or restricted) and/or building "one-click" classification into Microsoft Office. Some solutions also help manage the life cycle of classification by interacting with the user when an email is being sent or a file being saved to ensure the content is classified before the action is completed. User-driven classification can be difficult, because it requires training of both data owners and those with access and permission to use, view and manipulate files and data.
- Automated classification — Where the tools are configured to automatically classify documents and files based on content or when manipulated to include content that would require the classification to change. There are two methods of automated classifications:
  - Real-time classification typically requires some type of agent.
  - Scheduled/batch classification can be performed with a local agent or remotely.
- Data-at-rest scans — Where unstructured data (such as Microsoft Office documents and PDFs) and semistructured data are stored and shared either on common servers or in cloud environments, such as Box, Dropbox or OneDrive, where classification can be applied once a scan has been done to find documents and files that have been tagged and require classification based on specific content.

Why Classification

Organizations are evaluating data classification to assist in traditional use cases for reducing risk and increasing accuracy and retrieval for discovery and compliance with privacy mandates and company policies:

- Regulatory compliance — Organizations, such as those in the financial and healthcare sectors, that need to adhere to regulatory compliance, such as the PCI and the U.S. Health Insurance Portability and Accountability Act. All organizations need to appropriately handle personally identifiable information (PII) for their own employees versus customer information. Therefore, the appropriate data classification rules and policies need to be in place based on specific business units, such as sales and human resources.
- Intellectual property protection — Organization-specific sensitive data, including contracts or settlements; product/marketing requirements; service manuals; formulary data for pharmaceutical, chemical and manufacturing markets; engineering content (computer-aided design [CAD]/computer-aided manufacturing [CAM]) for the manufacturing industry; source code; financial modeling applications and calculations; and actuary tables. Not all data classification tools are capable of classifying CAD/CAM diagram files.

Factors That Will Drive Adoption

Differentiated Use Cases

Organizations are targeting specific use cases for data classification today as data growth, endpoint diversity (mobile devices) and adoption of cloud continue to increase, and finding and utilizing information become more and more difficult. In addition, data classification can assist with life cycle management of data through big data analysis on targeted datasets instead of reviewing the entire organization's volume of information. However, it will be useful only with unstructured data. Again, this compounds the need for tools that include both structured and unstructured data. Partnerships for specialist classification tool providers should extend to DLP and DCAP providers, as well as those that can make for complementary solutions to monitor the data life cycle, such as in data access governance.

ECM and Metadata

ECM is used to create, store, distribute, discover, archive and manage unstructured content (such as scanned documents, emails and office documents). It is both a framework and an architecture that supports all types of content throughout the content life cycle.

In "Automatic Classification and Tagging Make Metadata Manageable for ECM and Search," we found that metadata is entering the mainstream and is important to effective ECM and that three broad categories for machine-assisted tagging are: rule-driven, machine-learning-based and hybrid (a combination of both). ECM vendors have made metadata-driven functionality central to their platforms.

Big Data Analytics

The growing interest in and use of cloud have increased vulnerability to most organizations' data. Data discovery used as a tool to find data will enhance the journey toward successful classification strategies. In "Forecast Snapshot: Data Discovery, Worldwide, 2016," we define data discovery as an enabling technology used to perform a range of tasks, from provisioning dashboards to applying predictive models and enabling users to perform data exploration tasks that previously required an expert.

The business intelligence (BI) and analytics markets are expanding and expected to continue a compound annual growth rate through 2020 of 5.3% (see "Forecast: Enterprise Software Markets, Worldwide, 2013-2020, 1Q16 Update").

Purchasing decisions have shifted from IT leaders to business unit or line-of-business executives and users seeking agility and more customizable options. The BI market is driven by several factors, including (but not limited to) new vendors entering the market with innovative product offerings, market awareness and adoption of smart data discovery, organizations seeing support of real-time events, and streaming data capture in support of IoT use cases. For more information on BI and analytics, see "Magic Quadrant for Business Intelligence and Analytics Platforms."

In "Big Data Needs a Data-Centric Security Focus," we call out the impacts that big data initiatives will have, such as that they will require data to move between structured and unstructured data silos, and that DCAP tools that cross the data silos are forcing CISOs to find pragmatic strategies to implement data security governance policies. As mentioned earlier, we see decision makers moving beyond those in the IT department, but in the case of organizations with big data implementations either in plan or in place, the CISO is your target audience.

Data classification tools can be a complementary tool for those organizations that are implementing big data analytics platforms as more organizations are seeking more information about their data and simultaneously need to apply classification, depending on the specific use case.

Factors That Will Inhibit Adoption

In the past, data classification processes and technologies have been used and misused for many different initiatives within organizations at times. The strategy to identify, tag and store more data than is necessary results in an overabundance of false positives. User-driven classification is a classic way of experiencing misclassifications and requires organizations to continually train employees or data users to do it right the first time. Therefore, automation and machine learning with the use of pop-ups to remind users that they need to think about what they are doing before they close, save or send a file or document can be useful.

Data classification tools will not go away, nor will adoption be inhibited because the market direction is in constant motion in terms of consolidation between vendors. There is a merging of technologies, particularly with the DLP and DCAP markets, and we will soon see consolidation within the ECM markets, as well. The overall data security space is crowded with vendors with pointed solutions that cater to data at rest (DAR), data in use (DIU) or data in motion (DIM). This includes the DLP market as a whole, not just the parts where solutions can call for both structured

and unstructured data. More pointedly, these capabilities may be consolidated or merged with cloud access security brokers (CASBs).

As defined in "Market Guide for Data-Centric Audit and Protection," DCAP is a category of products characterized by the ability to centrally manage data security policies and controls across unstructured, semistructured and structured repositories or silos. Based on data security governance (DSG) principles, these products encompass the ability to classify and discover sensitive datasets and control access to the sensitive data by centrally managing and monitoring privileges and activity of users and administrators. The segmentation for DCAP products extends to CASBs, which include data security controls that cross DCAP and DLP (which also includes integrated DLP solutions such as classification).

Technology Impact

Data classification will have a profound impact on the overall data security picture, because classifications of data through the use of tags and class (such as public, internal, confidential) are useful identifiers to aid in protecting sensitive data, as well as assisting with the life cycle of the data. Classification is largely driven through policies and rules, and organizations need data classification schemes that not only are logical, but are part of a regular process when thinking about the business impact and the volume of data that resides in the organization. Many organizations believe that the most efficient way to manage data in the future will be to classify it at the time of creation by often taking a set-it-and-forget-it approach, which is no longer good enough. Therefore, automation will be a key feature, in that DIU will manage reclassification based on both a standard set of policies and customized policies.

Actions for the Next Six to 18 Months

Actions for Product Developers

Increase integration capabilities — Successful data classification technology providers have a wide variety of integration technology partners, therefore partnering with BI, data discovery and analytics vendors, such as Tableau, SAS and IBM. As businesses seek to become more knowledgeable about their data and intelligence about the data is applied, the life cycle of the data be tracked and monitored.

Add product features for data security — Focus on adding automation and machine learning capabilities as an essential part of your product feature set.

Actions for Product and Marketing Managers

Know who the buyers are — Gaining a better understanding of who the buyers are within the organization will be key to future growth in your business. Gartner estimates that new spending on BI data discovery tools will come from buyers outside the IT organization. For example, data governance and data security governance are typically managed by different parts of the business,

such as the CDO and CISO, and other influencers of buying decisions will come from other departments, such as finance and human resources.

Focus on configuration flexibility and ease of deployment — Organizations no longer can rely on standard policies; they need to customize based on their specific use case (such as intellectual property protection). They need not only to support a multitude of file extension types, but to have specific policies.

Increase internationalization support — Expand capabilities to support multiple languages, such as Chinese, Russian, German and Korean, to increase visibility and support of countries outside your home region. As extended local languages are supported, marketing to these countries will become much easier and tie messaging back to ease of configuration.

References

"Forecast: Enterprise Software Markets, Worldwide, 2013-2020, 1Q16 Update"

"Market Guide for Data-Centric Audit and Protection"

"Hype Cycle for Data Security, 2015"

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"How to Overcome Pitfalls in Data Classification Initiatives"

"Automatic Classification and Tagging Make Metadata Manageable for ECM and Search"

"Forecast Snapshot: Data Discovery, Worldwide, 2016"

"Magic Quadrant for Business Intelligence and Analytics Platforms"

"Big Data Needs a Data-Centric Security Focus"

"Forecast: Information Security, Worldwide, 2013-2019, 4Q15 Update"

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp