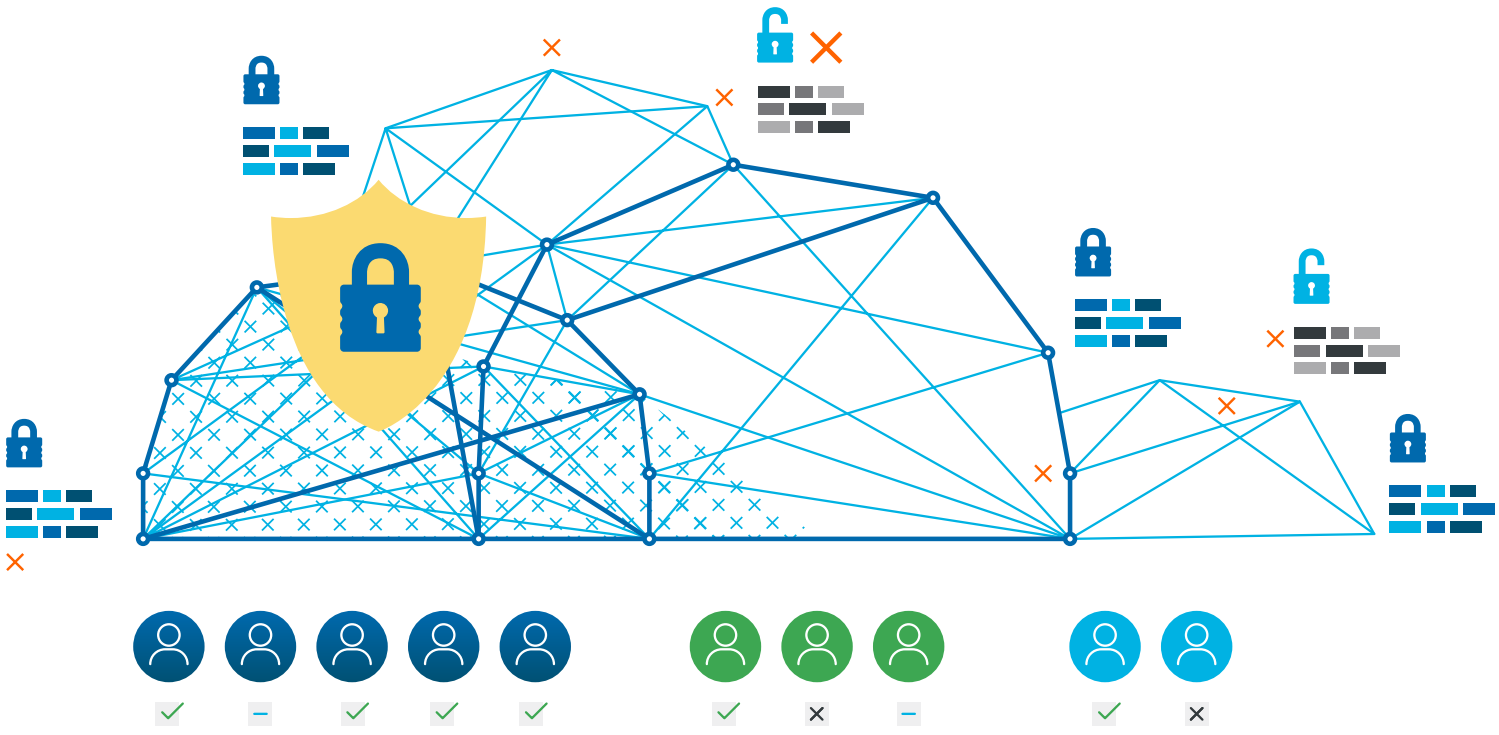
SPIRION

Data Classification for Data Privacy Management, Compliance and Security



Redefining Data Classification

Data classification is not a new concept. In a way, it's existed for millennia. When the point is keeping sensitive data out of the hands of the wrong people, historically there's no better example than in Greece in 678 AD. To defend Constantinople from the sea, the Greek army created a type of fire that could burn on water. The "Greek fire" formula was kept a state secret from all enemies. Their "top secret" classification of this intellectual property worked so well that it died with the Byzantium Empire.

Can your sensitive data protection program stand up to that standard? If not, you're not alone. According to The Value of Data study[1], which surveyed 1,500 IT decision makers and data managers across 15 countries, on average over half (52%) of all data within organizations remains unclassified or untagged. This means that businesses have limited or no visibility over vast volumes of potentially business-critical and private data.

The study further reported that three in five (61%) organizations admit they have classified less than half of their public cloud data.

It's even worse for small businesses. A GetApp study[2] reported that only 16% of small businesses have a data classification policy that provides different levels of access based on data sensitivity.

Unclassified data is putting organizations at risk, because there is no way to ensure that it is safeguarded. The result is a lot of potentially sensitive data that sometimes may be insecure, and other times may be too secure. Being just as secure as is required should always be the goal. Data that is too unsecure is at risk of breach and noncompliance. Data that is too secure can hinder day-to-day business operations and become more difficult to search, share among applications and databases, and use for its intended purposes.

Stricter Regulations Require Data Classification

Data classification has been used in the modern world for several decades, but in fairly rudimentary forms. Today it's generating a new wave of interest from the business world. The reason is likely the increasing attention being paid to data classification by today's more progressive compliance regulations — the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA).

These much-stricter rules are intensifying the demand for stronger privacy and security measures — and forcing organizations' hand toward adoption. They are making new demands on organizations, including requiring them to respond to consumers' requests for their personal data to "be forgotten."

Right now, if organizations deal with any personally identifiable information (PII) from European Union or California citizens, that data must be classified. What's more, it's highly likely that new or updated national and state regulations will also require data classification to add a vital added layer to data privacy and security.

1 Vanson Bourne, "Realizing the Power of Enterprise Data" The Value of Data Study https://www.veritas.com/form/whitepaper/realizing-the-power-of-enterprise-data, IDG, February 2019

2 Maria Gitanjali & Zach Capers, "How to Select Data Classification Levels for Your Business" Get App https://lab.getapp.com/data-classification-levels/ Get App July 10, 2019

Traditional Data Classification

Classifying data according to its risk levels, enables organizations to quickly scan and tag the data to ensure that sensitive, or risky information is properly managed and protected. IT teams and data managers can ensure its handled and protected appropriately by, for example, applying proper data access controls, keeping permissions up-to-date, and implementing the right backup and remediation capabilities.

There are many schemas that organizations can use for classifying data, but most categorize data as variations of four general categories — public, private, confidential, and restricted. A simple example of a four-level data classification schema includes:

- **Public** — Information that is freely available and accessible to the public without any restrictions or adverse consequences, such as marketing material, contact information, customer service contracts, and price lists.
- **Internal** — Data with low security requirements, but not meant for public disclosure, such as client communications, sales playbooks, and organizational charts. Unauthorized disclosure of such information can lead to short-term embarrassment and loss of competitive advantage.
- **Confidential** — Sensitive data that if compromised could negatively impact operations, including harming the company, its customers, partners, or employees. Examples include vendor contracts, employee reviews and salaries, and customer information.
- **Restricted** — Highly sensitive corporate data that if compromised could put the organization at financial, legal, regulatory, and reputational risk. Examples include customers' PII, PHI, and credit card information.

Redefining Data Classification Today

While data classification has been with us for a few decades, in the past few years it's changed considerably — becoming much more sophisticated and delivering much greater value. What used to be a simplistic process that involved applying data to a few buckets to streamline data management, is now a much more sophisticated process that meets organizations' intensifying data privacy and security demands.

Along with traditional approaches to data classification there is a new way to look at it — one that is updated to address today's more sophisticated data privacy, regulatory and business needs. This new methodology adds an additional layer to data classification that accounts for three critical variables: data processing, purpose, and privacy. Simply stated, these sub-categories are:

- **Data Processing** (aka, consent) — New and evolving data privacy regulations require individuals' consent for how organizations use their private data, in particular, GDPR and CCPA.
- **Purpose** (aka, access) — GDPR requires organizations that process European Union citizens' personal data to clarify the purposes for which they are collecting data. As a result, companies now have to manage their data according to what purpose or purposes it serves within their organizations.
- **Privacy** (aka, compliance) — Both GDPR and CCPA are laser focused on data privacy. Complying with these stricter regulations requires more advanced data classification schemas.

Data Classification for Processing

To effectively and consistently manage the consent rules for both GDPR and CCPA — and any new and stricter regulations that come on board in the future — organizations would benefit greatly by classifying their data according to how they intend to process it.

Classifying data for processing begins with asking simple questions: What data are you collecting? How are you processing it? The answers are critical because organizations need a lawful reason for processing the data they collect.

In order to classify data according to processing and consent, various tags can be applied to the data, for example:

- Personally identifiable data (PII)
- PII for order processing
- PII for marketing analytics
- PII for selling to third parties

These additional labels help organizations to manage both data privacy programs and compliance. When data is classified with tags like these, understanding how it's processed is clear to the organization and it's easy to access when individuals request information or for their data to be deleted.

Two classification sub-categories that fall under the process-based category are purpose-based classification and privacy-based classification.

Internal Data Access Rights

Another aspect of classifying data according to its purpose is controlling who exactly has access to the data internally. The only employees who should have access to personal data are those who need to access it for specific business purposes.

For example, by its very nature, human resource departments retain a lot of PII data on a company's employees, including credit reports, medical information, and benefits information. However, there are restrictions on specific reasons for retaining data. Access should be granted to employees — such as sales, marketing, finance — based upon the data's actual purpose.

A sophisticated data classification policy allows organizations to conquer their internal data access challenge by facilitating the enforcement of data access rights. As such, the internal access restrictions are informed by data's purpose-based classifications.

Achieving these two critical business goals requires a data classification schema to be sophisticated enough to allow organizations to organize the data at a granular enough level to comply with the privacy laws, while still conducting business.

Classification Schemas

In March 2020, Amazon Web Services (AWS) published its Data Classification for Secure Cloud Adoption[3]. The paper provides insight into data classification categories for public and private organizations to consider when moving data to the cloud.

Among AWS's recommendations to businesses and governments is undergoing "a quality assurance assessment to ensure that assets and data sets are appropriately labeled in their respective classification buckets. Additionally, it may be necessary to create secondary labels for data sub-types to differentiate particular sets of data within a tier due to privacy or other compliance concerns."

Owing to the popularity and rapid ascension of AWS as the leading cloud provider,[4] its recommendations should influence many organizations to either launch or beef up their data classification programs. The first step is determining the best data classification schema for each organization, because every business will have different needs.

3 AWS, "Data Classification for Secure Cloud Adoption", https://d1.awsstatic.com/whitepapers/compliance/AWS_Data_Classification.pdf, AWS, March 2020

4 Gary Dignan, "Top Cloud Providers in 2020: AWS, Microsoft Azure and Google Cloud, Hybrid, SaaS Players", https://www.zdnet.com/article/the-top-cloud-providers-of-2020-aws-microsoft-azure-google-cloud-hybrid-saas/, ZDNet, February 6, 2020

Traditional Data Classification Schemas

According to Garter,[5] many organizations have failed or stalled data classification deployments because they attempted to assign classification labels to content without first understanding their data, its characteristics, and its usage pattern. Understanding these factors is critical when considering the launch of a data classification program. It will help organizations determine the ideal schema for their needs.

Here are six common classification schemas used by today's organizations, including their angle of approach and major categories:

1. PII: Non-sensitive PII, Sensitive PII, Non PII
2. GDPR: Special category personal data, Personal data, Non-personal data
3. HIPAA: Critical, Restricted, Public
4. ISO 27001: Confidential, Restricted, Internal use, Public
5. U.S. Government: Top secret, Secret, Confidential, Sensitive but unclassified (SBU), Unclassified
6. NATO: Cosmic Top Secret, NATO Secret, NATO Confidential, NATO Restricted, NATO Unclassified (copyright), Non-sensitive information releasable to the public

There are many examples of sophisticated data classification systems in the world of business and government. Two examples include the University of Michigan and U.S. Dept. of Information Technology.

Schema Example — University of Michigan

The University of Michigan has created a detailed and well-organized data classification program that it makes available on its website. It states: "As a member of the U-M community, you share in the responsibility for ensuring U-M complies with data protection and privacy laws, regulations, and industry standards, as well as U-M policies and standards that require security safeguards around sensitive institutional data. You are expected to learn about compliance requirements and make use of the tools, safeguards, and information the university has put in place. You may also be responsible, depending on your role at the university, for compliance in your unit. Lack of compliance can result in significant consequences for the university and individuals, including fines, reputational damage, and harm to individuals whose data is exposed."

The university's U-M Data Classification Levels[6] define four sensitivity levels for all U-M institutional data. The examples illustrate what levels of security controls are needed for certain kinds of data. The site also includes examples of data by a person's U-M role.[7]

5 Eric Ouellet, "How to Overcome Pitfalls in Data Classification", https://www.gartner.com/en/documents/2813233/how-to-overcome-pitfalls-in-data-classification-initiati, Gartner, July 31, 2014

6 The University of Michigan, "Examples of Sensitive Data by Classification Level", https://safecomputing.umich.edu/protect-the-u/safely-use-sensitive-data/examples-by-level, University of Michigan, 2020

7 The University of Michigan, "Examples of Sensitive Data by U-M Role", https://safecomputing.umich.edu/protect-the-u/safely-use-sensitive-data/examples, University of Michigan, 2020

Schema Example — U.S. Dept. of Information Technology

The U.S. Department of Information Technology[8] published a document that serves as a data classification guide for several agencies within the U.S. Federal government. Its purpose is: "To establish protection profiles and assign control element settings for each category of data for which an agency is responsible. Security categorization provides a vital step in integrating security into the state agency's business and information technology management functions and establishes the foundation for security standardization among its information and information systems.

"Security categorization starts with the identification of what information and information systems support which government lines of business, as defined by the Federal Enterprise Architecture (FEA). Subsequent steps focus on the evaluation of the need for security in terms of confidentiality, integrity, and availability. The result is strong linkage between missions, information, and information systems with cost effective information security."

The department's data classification security control levels are based on the National Institute of Standards and Technology (NIST) guidelines published in the NIST Special Publication 800-53, which associates recommended minimum security controls with the Federal Information Processing Standards (FIPS) 199's three security categories:

- **Low** — The potential impact is low if the loss of confidentiality, integrity, or availability could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals, such as minor financial loss or minor harm to individuals.
- **Moderate** — The potential impact is moderate if the loss of confidentiality, integrity, or availability could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals, such as result in significant financial loss, or significant harm to individuals that does not involve loss of life or serious life-threatening injuries.
- **High** — The potential impact is high if the loss of confidentiality, integrity, or availability could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, or individuals, such as major financial loss or severe or catastrophic harm to individuals involving loss of life or serious life-threatening injuries.

After organizations choose their basic data classification structure, they should, as AWS advises, "create secondary labels for data sub-types to differentiate particular sets of data within a tier due to privacy or other compliance concerns."[9] To better comply with the new, more sophisticated, and stricter compliance regulations, organizations should consider adding the subcategories of process, purpose, and privacy as described previously in this paper.

8 Michael Varney, "Data Classification Methodology", http://www.ct.gov/opm/lib/opm/policies/data_classification_methodology_2_8_10.pdf, Department of Information Technology, March 30, 2010

NIST, "Guideline for Mapping Types of Information and Information Systems to Security Categorization Levels, SP 800-60", https://csrc.nist.gov/CSRC/media/Events/ISPAB-SEPTEMBER-2003-MEETING/documents/Barker_SP800-60.pdf, 2019.

9 AWS "Data Classification for Secure Cloud Adoption", https://d1.awsstatic.com/whitepapers/compliance/AWS_Data_Classification.pdf, AWS, March 2020

Classifying Sensitive Data

The U.S. General Accounting Office estimates that the identity of 87% of Americans can be determined using a combination of the person's gender, date of birth, and ZIP code. When taken separately, these details might not seem sensitive. However, a breach of those three elements would likely also compromise the individual's name, home address, social security number, and other personal data. As a result, those elements should be labeled sensitive.

Data that needs to be classified as private is often called "sensitive data." This means that if it is exposed inside or outside of the organization it presents risks to an individual's privacy and security, and it risks falling out of compliance with leading data protection regulations.

To make matters worse, both the regulatory environment and what's considered sensitive data is changing. For example, the California Consumer Privacy Act (CCPA) defines personal information that "could reasonably be linked, directly or indirectly, with a particular consumer or household." The word "household" is not found in General Data Protection Regulation (GDPR), however. It implies that personal information does not have to be tied to a specific name or individual, for example, home address, home devices, geolocation data, and home network IP addresses.

4 Key Regulations Protecting Sensitive Data

Most sensitive data living in today's enterprises is regulated by several compliance regulations, including local, state, and national agencies. Among the many rules that cover data privacy, there are four main regulations to which organizations must adhere — **GDPR, CCPA, HIPAA, and PCI**.

General Data Protection Regulation (GDPR) — This regulation protects the PII of European Union residents. The GDPR defines personal data as any information that can identify a natural person, directly or indirectly, such as:

- Names
- Identification numbers
- Location data
- Online identifiers
- One or more factors specific to a person's physical, physiological, genetic, mental, economic, cultural, or social identity

To comply with the GDPR, organizations must classify data within an inventory structure, including the following:

- Type of data (financial information, health data, etc.)
- Basis for data protection (personal or sensitive information)
- Categories of the individuals involved (customers, patients, etc.)
- Categories of recipients (especially international third-party vendors)

California Consumer Privacy Act (CCPA)

This new regulation, launched on January 1, 2020, brings the key data privacy concepts of Europe's GDPR onto America's shores — specifically to California residents. It requires businesses that interact with California residents to adhere to a new set of obligations around consumer rights related to personal data that is collected, processed, or sold by companies that are covered by the law. The obligations include:

- Giving citizens the right to request information about what types of data a company has collected, the purpose of collecting it, and the names of companies to whom the data was sold.
- The right to opt-out of data collection or sale.
- The right to request deletion of personal data.

Health Insurance Portability and Accountability Act (HIPAA)

This regulation protects individuals' personal health information (PHI). HIPAA has up to 18 identifiers of sensitive data that must be protected, including medical record numbers, health plan and health insurance beneficiary numbers, and biometric identifiers, such as fingerprints, voiceprints, and full-face photos. The HIPAA Privacy Rule requires organizations to ensure the integrity of electronic personal health information (ePHI).

HIPAA classification guidelines require organizations to group data according to its level of sensitivity, such as:

1. **Restricted/confidential data** — Data whose unauthorized disclosure, alteration, or destruction could cause significant damage. This data requires the highest level of security and controlled access in accordance with the principle of the least privilege.
2. **Internal data** — Data whose unauthorized disclosure, alteration, or destruction could cause low or moderate damage. This data is not for release to the public and requires reasonable security controls.
3. **Public data** — Data that doesn't need protection against unauthorized access but does need protection against unauthorized modification or destruction.

Payment Card Industry Data Security Standard (PCI-DDS)

This regulation protects individual's payment cared information, including credit card numbers, expiration dates, CVV codes, pins, and more. The PCI-DSS regulation has one identifier of sensitive data that must be protected: cardholder data.

Data classification is requested in terms of regular risk assessment and security categorization processes. Cardholder data elements should be classified according to its type, storage permissions, and required levels of protection to ensure that security controls apply to all sensitive data, as well as confirmation that all instances of cardholder data are documented and that no cardholder data exists outside of the defined cardholder environment.

Fulfilling the requirements of these four standard data privacy compliance regulations is nearly impossible without intelligent data classification policy underpinning data privacy programs.

Adding Data Discovery

Data discovery goes hand in hand with data classification. Basically, one cannot exist without the other. Data discovery is the process of collecting data from databases and silos and consolidating it into a single source that can be easily and instantly accessed.

Together, data discovery and classification make data more secure by providing the critical first step in a comprehensive data privacy and security program. In fact, data discovery and classification are the first phases of **Forrester's Data Security and Control Framework**,[10] which breaks down data protection into three areas: 1) defining data, 2) dissecting and analyzing data, and 3) defending data.

According to IDC's **The State of Data Discovery and Cataloging** report,[11] the ability to locate, understand, access, and trust data is a key enabler of business in the era of digital transformation. "No one driver stands out. Data discovery is important for business. Period."

Organizations surveyed said that data discovery supports these five business drivers:

1. **82%** ✓ Operations and efficiency
2. **81%** ✓ Policy compliance
3. **80%** ✓ Risk reduction
4. **78%** ✓ Regulatory compliance
5. **77%** ✓ Increasing revenue

The report also found that 30% to 50% of organizations aren't where they want to be when it comes to data discovery. As a result, data professionals say they are wasting on average 30% of their time because they cannot find, protect, or prepare data.

10 Heidi Shey, "Rethinking Data Discovery and Classification Strategies Strategic Plan: The Data Security and Privacy Playbook", Forrester, July 10, 2018

11 IDC Commissioned by Alteryx, "The State of Data Discovery and Cataloging", https://community.useready.com/wp-content/uploads/2019/08/The-StateofData-DiscoveryandCataloging-IDC-Infobrief.pdf, IDC, January 2018

The Data Sprawl Problem

According to two studies, the volume and variety of data is exploding across today's business world.

IDC released a report[12] on the ever-growing data-sphere. It predicts that businesses will generate 175 zettabyes of data by 2025 at a compounded annual growth rate of 61%. A zetabyte is a trillion gigabytes. Multiply that 175 times to gauge the scope of the challenge.

This unprecedented data sprawl means that data is stored in every nook and cranny of enterprises' IT infrastructures. As data volume and variety continues to grow, resources become more ineffective because data and information assets are harder to find. No one knows which files contain personally identifiable information (PII), information about particular projects, intellectual property (IP), or other valuable or regulated content. As a result, organizations struggle to protect information adequately, comply with legal mandates, weed out duplicate and redundant data, and empower employees to find the content they need to do their jobs.

Based on its prediction, the report urges CEOs to act now to ensure their data strategy is focused on storing data in small sets (categories) according to their business impact — rather than trying to analyze and use anything and everything. In other words, the more data organizations have to manage, the more data classification becomes a critical necessity.

The Dark Data Problem

Among the mountains of data already stored and arriving daily in today's businesses is a secret hiding in plain sight — dark data. This is unknown and unused data, and it comprises more than half the data collected by companies, creating a considerable issue, said a Splunk report, The State of Dark Data.[13]

According to the report, 55% of all data collected by companies is dark data. Within this category lies two subcategories — data that they know has been captured, but don't know how to use, and data that they are not even sure with certainty that they have. Further, 85% of companies say they aren't using dark data, because they don't have the tools to find it, capture it, and analyze it.

Among the dark data could be important customer information, for example, about a transaction, but it's missing location or other important metadata because that information sits somewhere else or was never captured in a useable format. The implications are vast. When companies don't know where all the sensitive data is stored, they can't be confident they are complying with consumer data protection measures. At the same time, data that is misused or improperly protected makes businesses vulnerable to legal action or theft from hackers.

12 David Reinsel, John Gantz, John Rydning, "The Digitization of the World from Edge to Core", https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, IDC, November 2018

13 Splunk, "The State of Dark Data Industry Leaders Reveal the Gap Between AIs Potential and Today's Reality", https://www.splunk.com/pdfs/dark-data/the-state-of-dark-data-report.pdf, Splunk

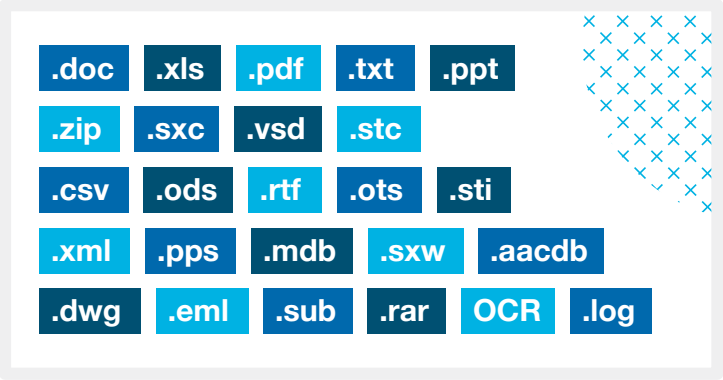
Leveraging Data Discovery Automation

Serious questions arise when it comes to vast quantities of data floating around within enterprises, unbounded, unrecognized, unused, and unprotected:

- If all of that data is not accurately classified and tagged, how can enterprises know whether it contains sensitive information?
- If an enterprise doesn't have an automated process for tagging data intake, how much of the problem is compounding each day?
- How will an organization determine who has access to that data, who's making changes to it, what those changes are, and whether or not the surrounding environment is secure?

Today's organizations don't have to be in the dark about their data. Intelligent data discovery automation can tell them exactly what data they have and where it resides.

Automated data discovery and inventory tools work by scanning endpoints or corporate network assets to identify resources that could contain sensitive information, such as hosts, database columns and rows, web applications, storage networks, and file shares. Sophisticated systems can find data located in all file types, including .doc, .xls, .pdf, .txt, .ppt, .zip, .sxc, .vsd, .stc, .csv, .ods, .rtf, .ots; .sti, .xml, .pps, .mdb, .sxw, .aacdb, .dwg, .eml, .sub, .rar, OCR images and .log.



To help security professionals adapt to the new data economy, Data Privacy Officers and data security stewards can deploy a data discovery platform, and apply data security and control frameworks to:

- Define their data by identifying their data
- Classify that data
- Dissect and analyze their data to understand its implications
- Create data security and control policies to defend and protect the data.

In this way, data discovery can transform data security policies from ineffective to effective and a help to business processes.

A Gartner report[14] on overcoming the pitfalls in data classification stated: "Most data classification implementations continue to be unexpectedly complex and fail to produce practical results. CISOs and information security leaders should simplify schemes, leverage tools, and allow for implementation flexibility to make classification valuable for the entire organization."

The leading tool companies need to overcome this complexity by an intelligent automated data classification platform. The right automation system can aid in streamlining the data classification process, automatically analyzing and categorizing data based on pre-determined parameters persistently in real-time.

14 Eric Ouellet, "How to Overcome Pitfalls in Data Classification Initiatives", https://www.gartner.com/en/documents/2813233/how-to-overcome-pitfalls-in-data-classification-initiatiti, Gartner, July 31, 2014

Organizations often struggle with their data classification programs because they approach them mostly as manual processes. But classifying data manually is simply too labor-intensive, time-consuming, and error-prone to be a practical solution for all but the smallest companies. In particular, manual data classification suffers from the following issues:

- **Inconsistency** — Different people classify similar documents in different ways.
- **Inaccuracy** — Busy employees often fail to classify data at all or simply pick the first tag in the list to expedite the process.
- **Inflexibility** — As companies' sensitive-data requirements and regulations change, no one has the time or inclination to update the tags on terabytes of existing data.
- **Failure** — As users realize that data is not classified correctly, they will quit trusting the process and the whole project will fail.

Automating data discovery and classification overcomes these limitations by making the process reliable, accurate, and continuous (aka, persistent). For example, the platform can spot personally identifiable information (PII) by looking for data patterns, such as names, dates of birth, addresses, phone numbers, financial information, health information, and social security numbers. Importantly, automated systems can also re-classify data as needed, such as for updates and changes within the business or compliance regulations.

Leveraging Data Classification Automation

Some companies have taken a completely automated approach to data classification. Some have taken a completely manual approach. Others have chosen a hybrid approach. However, even small organizations that haven't considered full automation, realize it's a godsend after seeing how everything they are doing manually can be fully automated — thereby, eliminating the man hours and human errors inherent in manual processes.

Today's automated data discovery and classification applications vary in terms of usage, access, and enforcement capabilities. But common features include:

- Pull-down menus of available data classification selections for user or file types
- Content-aware capabilities to suggest a classification for review and change or confirmation by the user
- Automatic selection of the appropriate classification level based on content analysis engines
- Classification lifecycle policy enforcement, such as preventing a user action unless the file is classified, or preventing an unauthorized classification change
- Some limited integrated DLP functionality, often around specific use cases, such as email

Conclusion

Every organization should assess the options and determine which solution provides them with the capabilities they need. Classification of data is critical to understand the landscape of your environment. By understanding where the sensitive data is located across an enterprise, Spirion can mitigate your security risk footprint. Data classification delivers this insight with a consistent process that identifies and tags all sensitive information in structured and unstructured forms across your business- such as in networks, sharing platforms, endpoints and cloud files.

Talk to a Spirion data security and compliance expert today: expert@spirion.com

Spirion is the leader in data discovery, persistent classification, and protection of sensitive data on-premise and in the cloud. Since 2006, thousands of organizations worldwide have reduced their sensitive data footprint and proactively minimized the risks, costs and reputational damage of successful cyberattacks. Spirion provides greater command and control of sensitive data to leading firms across all industries from financial services to healthcare to public sector. Visit us at **spirion.com**