

# Guide to Comprehensive Data Discovery and Classification

Data is a valuable asset for organizations, and its protection is paramount in today's digital landscape. Without a clear understanding of the data they possess and its associated risks, organizations are more vulnerable to data breaches, unauthorized access, or misuse.

Data discovery and classification provide the foundation for a robust cybersecurity strategy by enabling organizations to identify, prioritize, and protect their most sensitive data. By implementing these practices, organizations can significantly enhance their ability to detect and respond to potential security incidents, safeguard customer information, maintain trust, and ensure regulatory compliance.

**A well planned data discovery and classification strategy can help mitigate the risk of data breaches and other security incidents.**

Data discovery is the process of locating and identifying data within an organization's systems, networks, and repositories. Its primary purpose is to gain visibility into the organization's data landscape, including structured and unstructured data, to understand what data is stored, where it resides, and how it is being accessed and used. Data discovery helps organizations identify data assets, assess their risk posture, and establish a foundation for effective data governance, protection, and compliance.

Data classification is the process of categorizing and labeling data based on its sensitivity, confidentiality, regulatory requirements, or other predefined criteria. The goal of data classification is to enable organizations to identify and prioritize their data assets, understand their level of protection needed, and implement appropriate security controls. By classifying data, organizations can tailor their security measures, ensure compliance with relevant regulations, and apply risk management practices effectively.

**Without knowing where your data is and understanding the sensitivity of that data, it is hard to properly secure it.** This guide provides comprehensive strategies to implementing data discovery and classification processes.


# Benefits of Implementing Data Discovery and Classification

**Enhanced visibility:** Data discovery and classification provide organizations with a comprehensive view of their data landscape. This visibility helps in improving data governance and management practices.

**Identification of high-value and sensitive data:** By implementing data discovery and classification, organizations can effectively identify and prioritize high-value and sensitive data. This identification enables the allocation of appropriate resources for data protection.

**Alignment of security controls and investments:** Data discovery and classification assist in aligning security controls and investments with the specific needs of different data types. This targeted approach ensures that security measures are tailored to the specific requirements of each data category.

**Streamlined compliance efforts:** Implementing data discovery and classification allows organizations to streamline their compliance efforts. By focusing on the data that falls under various regulatory requirements, organizations can optimize their compliance activities and reduce the risk of non-compliance.



**Implementing data discovery and classification enables organizations to protect sensitive data, comply with regulations, manage data efficiently, mitigate risks, and make informed decisions. It establishes a solid foundation for data-driven strategies, enhances data governance practices, and supports responsible and secure data management throughout the organization.**

## Challenges to Consider

**Sheer volume of data:** With the exponential growth of data, organizations face the challenge of dealing with vast amounts of information. Data discovery and classification can help manage this overwhelming volume by providing a structured approach to categorizing and organizing data.

**Diversity of data sources and formats:** Data can originate from various sources and exist in different formats, making it challenging to identify and classify consistently. Implementing data discovery and classification requires addressing this diversity and establishing effective processes for data classification across different sources and formats.

**Constant evolution of data repositories:** Data repositories evolve over time, with new data sources and systems being added or existing ones being modified. Organizations need to ensure that their data discovery and classification framework can adapt to these changes and accommodate new repositories effectively.

**Ongoing maintenance and updates:** Data discovery and classification require continuous maintenance to reflect changes in data landscape, business priorities, and regulatory requirements. Regular updates to the classification framework are necessary to ensure its accuracy and effectiveness.

**With the right data discovery and classification tools, organizations can establish robust data governance frameworks.**

The U.S. National Institute of Standards and Technology (NIST) has defined a framework that makes a good starting point for establishing security objectives for classification.

## The top three most common security objectives according to the NIST include:

- 1. Confidentiality** Preserving authorized restrictions on information access and disclosure, which includes protecting personal privacy and proprietary information.
- 2. Integrity** Guarding against improper information modification or destruction, which includes ensuring information's authority and authenticity.
- 3. Availability** Ensuring timely and reliable access to and use of information.

**Table 1. Summarizes the potential impact definitions for each security objective: confidentiality, integrity, and availability.**

SECURITY OBJECTIVE	LOW	MODERATE	HIGH
<p><b>Confidentiality</b> Preserving authorized restrictions on information access and disclosure including means for protecting personal privacy and proprietary information.</p> <p>[44 U.S.C., SEC. 3542]</p>	<p>The unauthorized disclosure of information could be expected to have a <b>limited</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The unauthorized disclosure of information could be expected to have a <b>serious</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The unauthorized disclosure of information could be expected to have a <b>severe or catastrophic</b> adverse effect on organizational operations, organizational assets, or individuals.</p>
<p><b>Integrity</b> Guarding against improper information modification or destruction, and includes ensuring information non repudiation and authenticity.</p> <p>[44 U.S.C., SEC. 3542]</p>	<p>The unauthorized modification or destruction of information could be expected to have a <b>limited</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The unauthorized modification or destruction of information could be expected to have a <b>serious</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The unauthorized modification or destruction of information could be expected to have a <b>severe or catastrophic</b> adverse effect on organizational operations, organizational assets, or individuals.</p>
<p><b>Availability</b> Ensuring timely and reliable access to and use of information.</p> <p>[44 U.S.C., SEC. 3542]</p>	<p>The disruption of access to or use of information could be expected to have a <b>limited</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The disruption of access to or use of information could be expected to have a <b>severe</b> adverse effect on organizational operations, organizational assets, or individuals.</p>	<p>The disruption of access to or use of information could be expected to have a <b>severe or catastrophic</b> adverse effect on organizational operations, organizational assets, or individuals.</p>

Source: Federal Information Processing Standards (FIPS) publication 199, National Institute of Standards and Technology (NIST).



# Types of Data Classification

Data classification can be performed based on various criteria, depending on the organization's specific needs and requirements. Some common types of data classification include:



## Sensitivity

Data can be categorized based on its sensitivity level, ranging from public or non-sensitive to highly sensitive or confidential. This classification helps identify the degree of protection required and dictates access controls and security measures.



## Confidentiality

Data can be classified based on its confidentiality requirements, such as whether it contains personally identifiable information (PII), trade secrets, or intellectual property. This classification helps determine who can access the data and under what conditions.



## Regulatory Compliance

Data can be classified based on the regulatory requirements that apply to it. This includes data subject to specific regulations like the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), or Payment Card Industry Data Security Standard (PCI DSS). Regulatory classification ensures adherence to legal obligations and data handling practices.



## Business Impact

Data can be classified based on its impact on the organization's operations, financials, or reputation. This classification helps identify critical data that, if compromised or unavailable, could significantly affect the organization's ability to function.

# Steps for Implementing Data Discovery and Classification

To effectively implement data discovery and classification, organizations follow a systematic approach that encompasses several key steps. **This process enables them to gain insights into their data landscape, identify valuable and sensitive information, and establish robust governance frameworks.**

The following steps outline the typical journey of data discovery and classification: planning, data source identification, data discovery, data classification, data inventory and documentation, and data governance and protection. Each step plays a crucial role in ensuring the successful implementation and maintenance of a comprehensive data management strategy. Let's delve into each step to understand their significance and how they contribute to leveraging data effectively while mitigating risks.

## Step 1: Planning and Objective Mapping

The planning phase sets the foundation for a successful data discovery and classification initiative. It involves defining goals, objectives, and scope. Key activities include:

**Understanding business requirements:** This step involves engaging with stakeholders to identify their specific needs related to data discovery and classification. It helps determine the desired outcomes, such as improved data governance, regulatory compliance, or risk mitigation.

**Defining the scope:** Clearly defining the scope ensures that the initiative focuses on the most relevant data sources, systems, and processes. This helps in efficient resource allocation and prevents unnecessary complexities.

**Establishing project timelines:** Planning includes creating a timeline that outlines the stages, milestones, and deadlines for the data discovery and classification initiative. This helps in managing the project effectively and monitoring progress.

## Step 2: Data Source Identification

Identifying data sources is crucial for comprehensive data discovery and classification. This step involves:

**Conducting data source mapping:** Organizations need to identify all data sources within their infrastructure. This includes databases, file systems, cloud storage, endpoints, and third-party applications. Mapping these sources helps create an inventory of data assets.

**Collaborating with relevant stakeholders:** Engaging with various departments and teams is essential to gather information about potential data sources. This collaboration ensures that no data sources are overlooked, and all relevant stakeholders are involved in the process.

**Documenting data source details:** Creating a detailed inventory of data sources, including information such as source type, location, data format, and accessibility, provides a comprehensive view of the organization's data landscape.

## Step 3: Select your Data Discovery and Classification Tool

Research and select appropriate data discovery and classification tools based on your organization's needs, infrastructure, and budget. Consider factors such as scalability, integration with existing systems, and ease of use. Other key factors include:

**Automation capabilities:** To streamline the process and ensure consistency, organizations can leverage automated tools to discover and classify data based on predefined rules. These tools can analyze data attributes, keywords, or content to assign the appropriate classification labels while saving teams a considerable amount of time.

**Granular visibility:** Organizations should not limit their sensitive information visibility to the file level; instead, they should go one level deeper and opt for a tool that tracks the information itself. By getting more granular, organizations can achieve better control, contextualized results, and actionable insights.



**Extensive scanning:** Organizations should ensure that, when selecting their data discovery and classification tool, it is capable of discovering PII, PCI, PHI, and any other sensitive information that is unique to their organization. This will guarantee complete visibility over their sensitive data. It is advisable to choose a tool that can scan files of any type and size and that supports optical character recognition.

## Why double down on unstructured data?

### Surge of unstructured data

Unstructured data is growing **3x faster** than structured data.

### Extent of unstructured data

**80-90%** of all new enterprise data is unstructured.

### Volume of unstructured data

The volume of unstructured data is expected to reach **175 zettabytes** by 2025, a 430% increase from 2018.

*It is critical to select a tool that provides complete visibility and control over unstructured sensitive data.*

Source: European Society for Opinion and Market Research (ESOMAR) publication, Global Market Research 2022.

## Step 4: Maintaining Data Compliance and Auditability

To ensure ongoing compliance and maintain auditability, organizations need to establish processes and practices that align with regulatory requirements. This involves:

**Monitoring and enforcing data classification policies:** Organizations should implement mechanisms to monitor adherence to data classification policies. Regularly reviewing and enforcing these policies ensures that data is consistently classified and protected according to the defined criteria.

**Conducting regular audits and assessments:** Regular audits and assessments help identify any gaps or non-compliance with data classification practices. These audits evaluate the effectiveness of data discovery and classification processes, assess data protection measures, and identify areas for improvement.

**Demonstrating compliance with regulatory requirements:** It is essential for organizations to be able to demonstrate compliance with relevant regulatory requirements. This involves maintaining documentation, records, and evidence that showcase adherence to data classification policies and the protection of sensitive data.

## About Qohash

Qohash is a leader in data security software development, delivering innovative and user-friendly security technologies that provide businesses with visibility and control over their sensitive customer data. Since its founding in 2018, Qohash has rapidly scaled its operations to offer solutions that empower organizations to maintain continuous oversight of their security posture. Currently available in the U.S. and Canada, Qohash's mission is to protect the world's most sensitive data.

**To learn more about Qohash's data discovery and classification capabilities, visit [qohash.com](https://qohash.com)**

