Guide

How to Identify and Classify Sensitive Data

QOHASH

Each day, an organization creates thousands if not millions of files and records containing corporate data. Some of this data could escape into the wild and nobody would particularly care. But some of the data created each day could lead to regulatory fines, the release of corporate secrets, or even a public relations scandal that could hurt or destroy the company if it fell into the wrong hands.

The danger is real, too. From phishing to incidental data exposure as a result of human error, data breaches are becoming increasingly common among organizations of all types. Roughly 73 percent of businesses admit that they have encountered at least one sensitive data leak in the past year. According to Microsoft Research, In the first half of 2020 alone, more than 36 billion corporate records were exposed by mistake.

That's why roughly 68 percent of business leaders feel the cybersecurity risks for their company are increasing, according recent research to by Accenture.

> **68% of business leaders feel that cybersecurity risks for their company are increasing.**

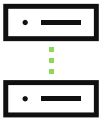
Without understanding the sensitivity of data, it is hard to properly secure it.

Part of the problem is that most companies still struggle with identifying the full scope of the data generated by their organization, and fail to classify it based on sensitivity level. More than half of all data within the typical company is unclassified or untagged. Without understanding the sensitivity of data, it is hard to properly secure it.

Sensitivity classification is a foundation for proper data security, and the good news is that establishing a robust classification system is not as hard as it was even a decade ago. This guide outlines the advantages of data classification, the method for developing a classification scheme, and the process for actually putting that classification scheme into practice.

Why You Need Data Classification

Security is of course a main reason for data classification. But there also are other reasons that businesses should classify their data. Global technology consultancy, Gartner, outlines four basic use cases for data classification.



Risk Mitigation

This includes limiting access to personally identifiable information, controlling location and access to intellectual property, reducing attack surface area for sensitive data, and integrating classification into security and policy-enforcing applications.

Governance/Compliance

Identifying data governed by regulations such as GDPR, HIPAA, CCPA, PCI-DSS, SOX and those not yet developed, applying metadata tags to protected data for additional tracking and controls, enabling quarantining, legal hold, archiving and other regulation-required actions, and facilitating “Right to be Forgotten” and Data Subject Access Requests (DSARs).

Efficiency and Optimization

Enabling efficient access to content based on type, usage, etc., discovering and eliminating stale or redundant data, and moving heavily utilized data to faster devices or cloud-based infrastructure.

Analytics

Enabling metadata tagging to optimize business activities, and informing an organization on the location and usage of data.

How to Make a Data Classification Plan

The first step toward data classification is developing a classification plan–a document that includes a data classification framework and descriptions of the various classification levels.

The process is not complex, consisting of three basic steps. But it does require understanding all the data within a company and how it is created, as well as working with senior leadership and key stakeholders throughout the company.

Step 1: Define Classification Objectives

The first step is establishing the reasons for classifying a company's data. Will classification be used solely for risk mitigation, or will it encompass other use cases such as compliance and improving operational efficiency?

From these objectives will flow additional considerations such as defining the specific regulations that apply to the company, and the classification needs for analytics or operational efficiencies.

In terms of risk mitigation, most security teams share common goals when it comes to security.

> **Data classification is both a security essential and a key component of the processes and technology that enable digital transformation.**

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

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

Step 2: Categorize Data Types

Once objectives have been defined, the second step is determining what kinds of data exists and will be created within the business, as well as where the data for each type is typically located. Each type of data should be defined and categorized so classification can be applied in the next step.

One common data type is personally identifiable information (PII). A business might define the PII data type category as follows, for example:

For the purpose of meeting security breach notification requirements, PII is defined as a person's first name or first initial and last name in combination with one or more of the following data elements:

• Social security number

• State-issued driver's license number

• State-issued identification card number

• Financial account number in combination with a security code, access code or password that would permit access to the account.

As part of this process, businesses also should define which data is public, which is proprietary, and which data is regulated, along with where the data typically lives.

Does the data live in a database on a company server, in a spreadsheet on a public cloud service such as Dropbox, in email, etc.?

Step 3: Define Classification Levels

With both objectives and data types now defined, the final step in creating a data classification plan is specifying the classification schema that the business will use, and which data type categories will map to each classification.

For data sensitivity classification, there are many ways that organizations can define levels of sensitivity. The U.S. government has seven levels of classification, for instance, including Restricted Data, Top Secret, and Controlled Unclassified Information, among others.

Each organization will want to develop a classification scheme that best meets its needs, but generally most corporate data classification schemes include a minimum **four high-level sensitivity categories:**

1. **Restricted:** The highest level of sensitive data. This includes the data that, if compromised, could put a firm at risk for financial, legal, regulatory or reputational damage.

2. **Confidential:** Exposed data would inflict a moderate risk to the organization or one of its employees. Unintentional access would bring consequences greater than short-term embarrassment, and could possibly have a negative impact on company operations or long-term reputation.

3. **Internal Data:** That is not meant for the public, but has a relatively low impact if exposed. The company wouldn't want this data leaked, and it might cause some short-term embarrassment or reputational damage. But access to this data wouldn't have regulatory or significant lasting repercussions.

4. **Public Data:** That anyone can see, and is not of a personal nature. Exposure of this data would result in little or no risk, and doesn't need encryption or significant protection.

While these four basic classifications have been in use for decades, privacy regulations and more advanced data management systems have led many organizations to adopt three additional sub-layers.

These include:

• **Data Processing layer (consent)** Many data privacy regulations now require an individual's consent for how their private data can be used by an organization.

• **Purpose layer (access)** Some privacy regulations, most notably Europe's GDPR, require organizations to specify the purpose for which specific data was collected.

• **Privacy layer (compliance)** Some privacy regulations, including California's CCPA, make additional demands on organizations that keep an individual's data. To ensure compliance, organizations therefore often add a specific and advanced set of data classifications around privacy.

Once a data classification plan is in place, the next step is implementing it through data discovery and automated classification.

Steps for Implementing Data Classification

Armed with a plan, the next step is implementing the plan for comprehensive data classification. This is where proper planning will pay dividends. The process also can be eased by having the right technology in place for faster data discovery and classification.

There are four main steps for implementing a data classification plan within an organization.

There's no point in making classification a manual process.

Step 1: Define the Automated Classification Process

There are three ways that each piece of corporate data can get classified. Data can be manually classified by employees, it can be classified automatically by data discovery and security systems, or organizations can take a hybrid approach that combines both automated and manual classification

As the heading for this step implies, most organizations will want at least some measure of data classification automation because complete manual classification is functionally impossible for the data volumes generated by the typical business.

Most businesses that successfully classify data according to sensitivity take a hybrid approach that combines comprehensive automated data discovery and classification with an ongoing manual audit of data classification performed by the system.

There are several robust data discovery and classification solutions on the market, including Qohash's cloud-based Qostodian platform that can uncover sensitive corporate data even on cloud drives and personal computing devices within an employee's home.

As part of this step, businesses also will define what data should be scanned first, the frequency of scanning, and the resources used for maintaining data classification.

Step 2: Specify Classification Criteria and Review Process

With a data classification solution in place, the third step is defining the criteria for classification, and the process for verifying correct classifications.

In terms of classification criteria, an organization will want to define the classification patterns and labels within the automated solution that will be required for correct classification. This will be easy or hard depending on the organization's knowledge of its complete data footprint, the schema developed during the classification planning phase, and the automated data classification solution that has been chosen.

Businesses also should specify the process for reviewing the automated data classification as part of this step, and the process for validating its ongoing accuracy.

Classification accuracy typically is broken down into two measures, according to the Association for Information and Image Management (AIIM):

• **Precision:** What percentage of the automated solution's data classifications are the same as a human would assign? Perfect precision means that all data is classified in the same way as what would be achieved by manual classification.

• **Recall:** A distinct but related metric, what percentage of all valid data is the automated solution able to classify correctly for a given data category? Perfect recall means that all corporate data intended for inclusion is classified correctly by the solution.

Step 3: Establish Overall Outcomes and Classified Data Usage

Data classification is not an end unto itself, so the next step is defining and setting up the analytics and security processes that will take place as a result of data classification.

This is where the classification usage and overall outcomes are defined and implemented. From a data security perspective, this also is the payday for all the preparatory classification work that came before.

Where possible, outcomes based on data classification should be automated so they can take place in real-time with minimal opportunity for latency between data creation and the use cases that rely on classification.

This is especially critical for security processes built around sensitive data classification, because data such as intellectual property and personally identifiable information require immediate security action when a variance is discovered.

Along with establishing an ongoing data classification system, businesses also should define processes for periodically reviewing data classification categories and the classification system overall.

Step 4: Monitor and Maintain Classification

The final step is the most important. Data classification is not one and done, so establish a process for discovering and classifying corporate data resources on an ongoing basis. This includes specifying the frequency of discovery and classification if it is not performed automatically in real-time.

Along with establishing an ongoing data classification system, businesses also should define processes for periodically reviewing data classification categories and the classification system overall.

If regulatory compliance is one of the objectives for the classification effort, this should include specifying a process for monitoring regulatory changes on an ongoing basis for continued data classification relevancy.

Make all your sensitive data secure

Identifying and classifying data is not instant. But at the same time, it is easier than many businesses realize. All it takes is awareness of a company's data landscape, a little structured planning, and the right automation technology in place. With a little work and the right data classification system, even businesses with small cybersecurity departments can ensure that sensitive data remains safe.

Qohash is a leader in data discovery and classification security software. It blends innovative and easy to understand security technologies, allowing businesses to gain visibility on sensitive data. Founded in 2018, with solutions available in the U.S. and Canada, Qohash provides customers with solutions tailored to meet today's hybrid and remote work environments in the financial services sector.

To learn more about Qohash products, schedule a demo.
Write us at info@qohash.com