CloudLock

Whitepaper

Using data classification to protect sensitive information from inappropriate sharing in Google Drive (Docs)

Solving the challenge of protecting information without deterring users

At a Glance

Intended Audience:

- C-level security, collaboration, and trust professionals at companies using or considering cloud collaboration in Google Drive
- Domain Administrators at organizations using Google Apps

Takeaway

- Readers will learn about best practices in data classification, and how information security measures can be implemented successfully in Google Drive, without deterring users from collaboration

In This Whitepaper:

Introduction

As large enterprises realize the efficiencies and value of collaborative computing, they face a growing exposure to careless, accidental or even malicious mis-sharing of sensitive information, between users within the enterprise, or worse with external collaborators such as suppliers and partners.

With over 600 million consumer and business users of cloud storage, collaboration in the cloud is now an integral part of B2B communication. While tremendously beneficial to employee productivity, unchecked cloud collaboration can become a vector for the externalization of sensitive information such as intellectual property, financial records, personally identifiable information, or other forms of regulated data, which puts the company brand at risk.

Enterprises must recognize that in the absence of an enterprise-sanctioned collaboration platform, users will find a way to share and collaborate, and in so doing, increase the exposure risk. Enterprises are advised instead to encourage collaboration on a sanctioned platform, take measures to monitor, manage and control how information is shared, and educate users on their role in safeguarding information assets. That is the subject of this whitepaper.

Data classification is key to information security in the cloud

Google Drive is emerging as the collaboration platform for many of the world's largest organizations. Organizations that "go Google" do so for a variety of reasons, as the platform represents a fusion of productivity tools, cloud storage, unified communications, and a robust third party ecosystem of applications that can extend and enhance the core offerings. With its seamless integration of users, apps and storage it provides a complete collaboration framework transforming how users work and interact, both with their colleagues and external parties.

But while Google Drive provides a robust platform for managing documents, it has no way of knowing who should be permitted to share what with whom. Google explicitly notes that they have no specific knowledge of the information security policies applicable to any one business.

This is the sole responsibility of enterprises themselves. Just as one would expect with on-premise solutions, internal collaboration and acceptable use rules must be established: employees should not be able to access payroll or accounting information, sales should not be able to access product roadmaps, and sensitive or regulated customer data such as credit card numbers should not be stored in spreadsheets. The same principles must be applied in the cloud, to protect the interests of the enterprise, the privacy of customers and the integrity of data. CloudLock provides a mechanism which overcomes the largest problem in implementing this type of security - the need to engage data owners and employees directly, without resulting in "data security fatigue”.

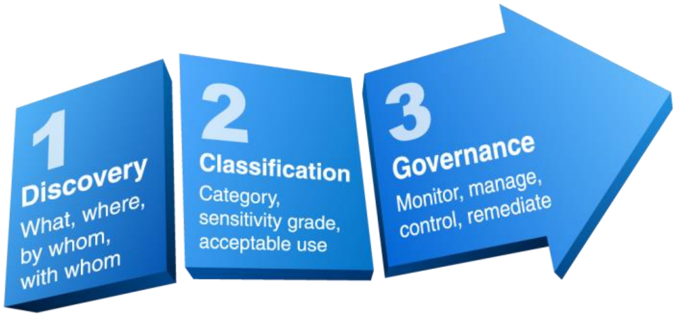
What is data classification and how should enterprises classify data?

Data classification is the activity of categorizing data to ensure that it is being used efficiently and appropriately within the organization. This typically requires users to tag every type of information they interact with, with the following attributes:

- Where the information resides
- The sensitivity of the data
- The type of access needed by users
- How compliance regulations dictate access

Good data classification practices lay the foundation for protecting important information within your organization, while reaping the benefits of collaboration in the cloud. Unfortunately, few businesses relying solely on manual classification succeed. Because eventually they run afoul of three classic impediments: Declining user engagement, inconsistent classification standards and excessive classification complexity.

Employee fatigue in manual data classification systems is enough to kill most initiatives. For any solution to scale as usage grows, CloudLock has found that a combination of automation and human involvement is essential. To this end, a three step framework for data classification has been created that ensures efficacy and data security, without requiring excessive end-user effort.



Cloudlock's three step framework for successful data classification

Data Classification Step #1: Discovery

Successful Data Classification begins with Discovery. If you don't know what information is being shared, then you can't know who should be allowed to share it. Similarly, without visibility into how the information is being shared, acceptable use cannot be determined. Discovery is therefore the first step in enabling cloud information security policies to be defined and implemented. That means identifying what type of data is in use, where it is used, by whom and with whom.

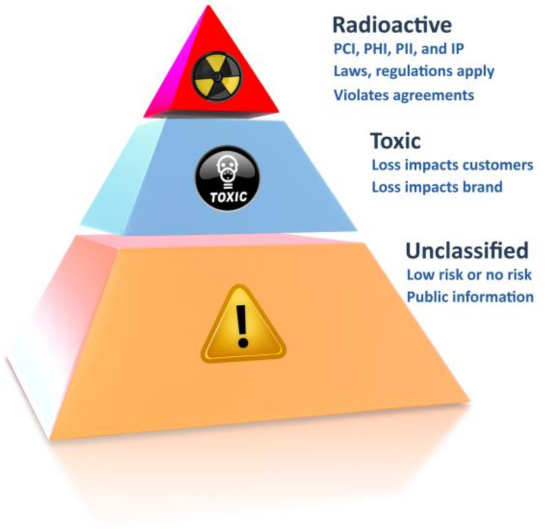
Different types of data fall into different categories. Some of those categories have a well-defined set of industry-specific statutes and legal obligations that businesses must comply with in handling such data. For example, financial institutions are subject to a range of regulations governing the disclosure of Personally Identifiable Information (PII) about customers. Accounting and finance departments are obligated to follow rules (GLBA, SOX, etc.) that dictate how they handle and disclose certain types of information regarding the accounts under their purview. But for other types of information, it is less clear cut. The importance and sensitivity of the information varies from business to business. From manufacturing processes, to chemical formulas, engineering specifications and product road maps, each business places different value on their intellectual property. No standard rules apply. The process of Discovery therefore lets enterprises identify the set of rules, statutes and policies that should be applied to protect the different types of information found to be in use.

From an auditing and compliance standpoint, discovering the *where*, is just as important as the *what*, because some types of highly sensitive information simply do not belong in the cloud environment at all. They are subject to strict statutory or regulatory compliance obligations. Depending on your industry you may have highly specialized systems for handling particularly sensitive information. These systems and the data in them are often subject to certain industry best practices and rules, such as PCIDSS for credit card transactions, HIPAA for healthcare information, SOX and GLBA for financial reporting, and so forth.

The goal is to ensure that such data is always contained in those systems and is subject to those applicable rules. Organizations should prevent that information from leaking beyond their established boundaries, by employing specific systems that meet regulatory requirements (such as encryption for PCI data) which satisfy their compliance requirements. If certain data from those systems is found in use outside of the system on a regular basis, not just in rare exceptions, it points to a need to extend the functionality of that primary system to better contain its data, and not perpetuate data sprawl by allowing repositories for this class of data to manifest in the cloud.

Data Classification Step #2: Classification

The next step is to classify the data, so that usage policies can be defined by a simple set of rules and algorithms. Some members of the organization should rightfully have access to certain types of data, and others not. Appropriate classification of this type of data allows organizations to control which groups of users can access different categories of information, and restrict what they can do with it. When users access a properly classified file or document, they should become conscious of the type of information it contains, and its acceptable use profile, without having to guess or review it. However, over-engineering the categories makes it too hard for users. CloudLock's broad experience in this space, as well as that of analysts, suggests that three broad categories are a manageable starting point.



Data should be classified into one of three broad categories.

Therefore, to implement data classification, a classification schema is needed in order to identify the notation or characteristics which can be used to recognize and tag data as belonging to one category or another. Hence, the **Discovery** *(step 1)* of sensitive data within the Google Drive environment, thus prompts the **Classification** *(step 2)* of that information type, as something only certain individuals or groups inside and outside of the organization should be able access. This in turn, leads to defining rules that flag, control, and enforce acceptable use policy based on the type of data identified.

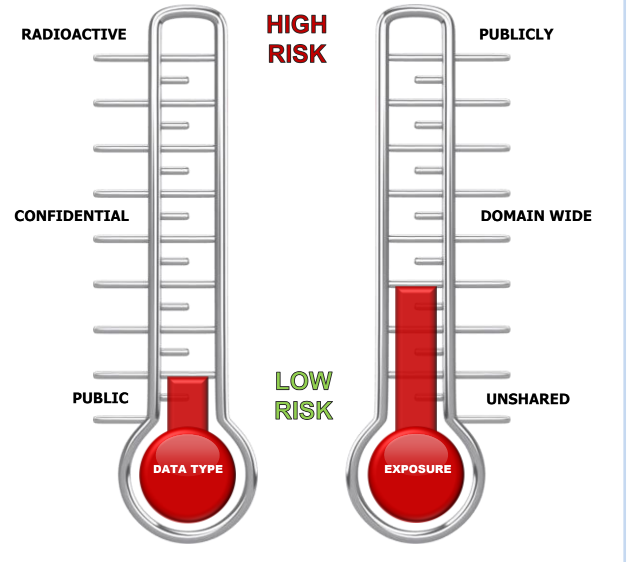
Classifying data is tricky, and unfortunately this is where most efforts of businesses fall short, for several reasons. First, getting each department in the organization to be consistent in the way they classify information is near impossible, because different stakeholders think differently. And, the bigger the organization and the more geographically dispersed it is, the more difficult it is to maintain consistency. Second, people often struggle with defining the sensitivity of information, especially when dealing with information that isn't regulated, but relates to intellectual property, business or financial planning, contracts and account details, often referred to in legacy data loss prevention schema as the company's "crown jewels". These are impossible decisions for employees to make at scale. Third, many organizations start out well with a good discovery and classification methodology, perhaps because they formed data classification task-force to start the process. Over time, however, maintaining the same classification acuity, breaks down. As new players get involved, inconsistencies arise and classification criteria morph.

Finally, the most common problem is over-complexity. It is all too easy to overdo it, and come up with classification taxonomy and schema that are simply too hard for employees to follow. Too many *if, then else's* become debilitating handcuffs that get in the way of work. This last observation, leads to the most important conclusion about the Data Classification challenge. User engagement and cooperation is everything. The classification schema must be easy for ordinary users to follow and execute. Otherwise it simply won't fly, and all the effort of discovery and classification is for naught. With too many obstacles in their way, users will go around them, refuse to use the system, and may eventually self-select insecure private alternative systems through which they will share and expose sensitive information.

Data Classification Step #3: Governance

The final step is to define the internal governance and legal actions that should be taken to A) minimize the possibility of violations occurring in the first place, and B) remediate a violation promptly should it occur. However, remediation actions are not dependent solely on the characteristics of the data. It is more complicated than that, because what someone does with that data is also a factor.

There is one risk continuum based on the sensitivity of data, and another continuum based on its exposure. Remediation actions should consider a *risk index* that combines both elements.



The risk index is a factor of both sensitivity and how it is shared

For example if an employee stores PCI data (e.g. credit card numbers for vendor purchasing) in a spreadsheet that is not shared, it may represent a compliance issue, but no immediate harm occurs! However, if that spreadsheet is shared with people outside the company, it is a serious breach, with a higher *risk index* than a possible regulatory infraction, calling for immediate remediation.

Making data classification manageable for every-day users

Data Loss Prevention (DLP) solutions which focus on up-front controls, often stifle collaboration and are rarely fully implemented. In contrast, CloudLock's approach gives businesses a means to automate data classification, and remediate violations through policy. Say for example, engineering part numbers are an identifier for documents related to engineering specs. Once how to recognize a part number is codified, documents with these part numbers can be classified as engineering related, automatically, eliminating manual effort at document inception.

Suppose a company's engineering plans should only be shared within a certain department, C-Level employees and three named suppliers. If a user shares an engineering plan with someone outside of this group, the user can be warned and the file share can optionally be disabled. This enforcement mechanism facilitates collaboration, while still protecting the organization. And it serves to educate users at the same time. Continuous education and awareness building is crucial as more users collaborate in the cloud, and more data is at risk of inappropriate sharing. As cloud computing erodes traditional lines of demarcation, enterprises must place greater emphasis on making information security and awareness a high priority in company culture.

Is data classification really worth the trouble?

In deciding to implement a data classification system, an organization must consider the effort and up-front costs in light of the potential repercussions of not doing so. Proving to auditors that sensitive data in the cloud is safe, when you really have no governance over it might be a very costly exercise. Cleaning up the damage to your brand, if the wrong information leaks out, is costlier still. According to the 2013 Cost of Data Breach Study by the Ponemon Institute, the per-record breach cost, for breaches due to negligence, averaged $150 in 2012. The right way to look at this is: This is a necessary investment to safely unleash collaborative computing in the cloud. The sooner it is done the better.

In practice, with the right methodology, user training, and tools, creating and maintaining a robust data classification strategy in Google Drive is not as daunting a task as it may seem. CloudLock's unique data classification approach via policy eliminates much of the manual work that was required in older on-premise solutions. Of course, human intervention and decision making is still necessary for initial data classification, definition of the enforcement actions and verification of potential violations in corner cases. But a combination of human intervention and powerful data classification tools can enable organizations to march forward with collaborative computing. In confidence that their information can be properly safeguarded and that their cloud data management practices are auditable.

Effective strategies for data classification in Google Drive

User engagement is a critical success factor

It is a mistake to underestimate the role of every-day users in successful data classification. Training, awareness and practical involvement are the prerequisites of a successful deployment. The key here is recognizing what constitutes practical involvement for different users. Do not expect users to assume the role of compliance officers. Collaborative computing is supposed to be a productivity enabling tool, not a barrier. So what is the scope of engagement you can realistically hope for, from them?

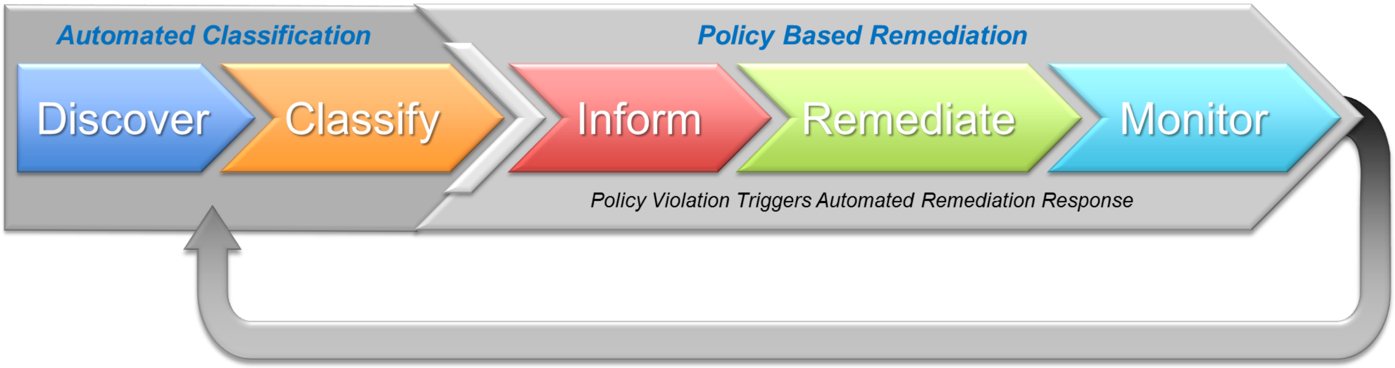
As Gartner implies, any data classification strategy that relies on users to do a lot of manual classification is doomed to failure. What is needed is a solution which balances between identifying and classifying sensitive information and not driving users away from the collaboration environment into which data is being moved.

> ”Organizations are changing radically - tearing down and redefining traditional boundaries via collaboration, outsourcing and the adoption of cloud-based services - and information security must change with them.”
>
> *Tom Scholtz, Vice President and Gartner Fellow*

Automation of data classification is the key

Since user adoption of Data Classification is the weakest link, CloudLock has found, in working with more than 500 of Google's largest customers that the right approach to this problem is to automate detection and classification through a combination of pre-defined algorithms and customer-defined criteria. Doing so can limit the scope of day-to-day manual intervention, to dealing only with remediation or exception processing. While it requires some up-front investment in initially defining the classification criteria, once it is implemented, users are relieved of the burden of classifying documents on ingress to the system. Moreover, when users are only involved in remediation and response, they are more inclined to cooperate and act in an intelligent and informed way, rather than to ignore or circumvent the system.

Essentially, this approach largely eliminates the hardest and most laborious part of data classification - The very thing which causes initiatives, even those that initially go well, to eventually grind to a halt because the necessary level of manual intervention is simply unsustainable.



CloudLock's approach automates discovery and classification, and reduces user involvement only to dealing with remediation.

It is CloudLock's unique ability to combine established data fingerprinting techniques, with user-defined criteria that makes it possible to look at data content, identify what it is, and then tag it automatically in accordance with the classification schema defined for the organization. Metadata, which may include anything from content keywords, to age, ownership, and usage, can then be examined in real-time whenever a user performs any action on a document, in order to determine if any type of misuse is occurring. Thus a violation can be detected and trigger warnings or alerts to individuals or groups affected by it.

This enables users to be informed immediately, when a violation occurs, so they can take corrective action, such as fixing the document, or perhaps request and exception to the rules because they are too broad. Also, risk and compliance administrators can look into the cause of the violation and adjust policies as necessary to handle the situation better in the light of real-life use cases. Having a mechanism to inform users in real-time about their own use of sensitive data whenever it is inappropriate or even just borderline, provides an invaluable feedback loop. It can enable the data classification system to continually educate its users about their role and the correct behaviors expected of them to protect sensitive data.

Automated data classification through deep content analysis

The table below provides examples of the data types that can be automatically discovered, classified and tagged through content analysis for industry standard notations and customer-defined attributes indicative of a class of information.

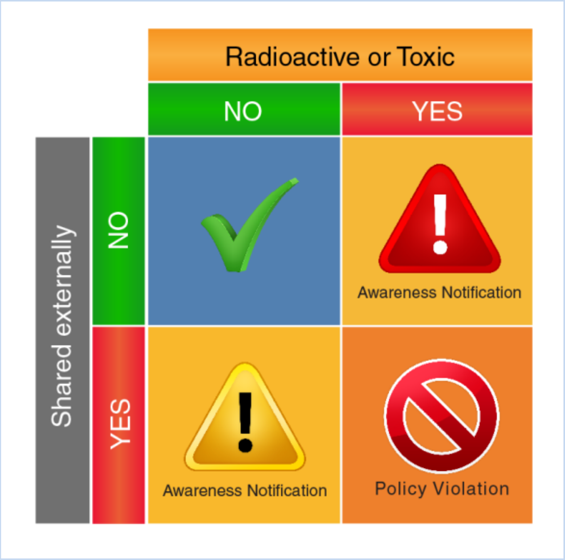
Industry standard content fingerprinting algorithms	**Customer-defined classification criteria**
Credit card information (PCI) Personally identifiable information (PII) Protected Health Information (PHI) International trade regulations	Engineering schematics, Chemical formulas, Part numbers Financial data, Pricing, Contracts, M&A Customers and contact information

Handling the remediation response through policies

The appropriate remediation action for each violation is driven by its risk index, which combines the data sensitivity and the level of exposure of that data. Depending on the severity of the incident, there could be a variety of responses and corrective actions. In some cases, the logic to trigger the optimal response is complex. CloudLock believes remediation cannot be accomplished efficiently without a Policy Engine to detect violations and automate remediation responses. Without a Policy Engine, remediation will be unacceptably arduous for your security and risk team. It would be hard to distinguish the severity of violations, or even detect them. Consequently, small infractions may go unnoticed, leaving uninformed users to become lax about the information security standards they are expected to uphold.

To illustrate how a remediation policy might work, imagine the case of a user saving a credit card number in a spreadsheet on Google Drive. If the user doesn't share the document, the policy engine might simply inform the user, this is an incorrect data type to use, ask them to remove it, and refer them to the correct system they should be using for handing credit card data. It could also add PCI metadata tag information to the document and place it in the PCI Data folder in the user's own Google Drive account. Thus CloudLock helps to automate and enforce a governance model, in which the user is both corrected and educated at the same time. For this incident, the security and risk team might receive a low level alert.

But what if the spreadsheet was shared with parties external to the company? This is a high risk exposure, which would trigger a different response, such as immediately disabling the file share, and escalating a high-priority alert to the user's manager and compliance officer. All this could occur automatically, without user involvement. So the only user involvement necessary is the corrective action. This is manageable for both users and security and risk professionals alike, and provides a framework for making cloud collaboration safe and auditable.



		Radioactive or Toxic	
		NO	YES
Shared externally	NO	✓	Awareness Notification
	YES	Awareness Notification	Policy Violation

In simplistic terms, detecting a policy violation boils down to examining two criteria. How sensitive is the data, and how widely has it been exposed? If sensitive data has been exposed inappropriately and it represents real harm, whatever those metrics might be for your organization, then do something immediately to first contain the risk, and then re-educate the user.

Exception handling through whitelist policies

There will always be exceptions. Data classification systems must accommodate the element of human context. For example a user may share a set of credit card numbers in a spreadsheet with external users. But those credit card numbers may be fake numbers that the user is asking an external test team to use in order to verify merchant account functionality of a new e-commerce system. Under normal circumstances this would be deemed a violation. But in practice, in this case there is no risk. So there needs to be a mechanism such as whitelisting to allow exception scenarios based on special business justifications. Ideally, when the system triggers an alert and remediation response, it should also give the user a means to request an exception. This exception request could pass through an approval cycle, involving the right stakeholders for that type of data, ultimately resulting in an exemption rule being added to the metadata for that document.

Providing an administrative view of data and policies

Finally, there must be an overarching administrative view that allows risk and compliance management to maintain metadata, rules and policies, while also monitoring the behavior of users. For example, it is not sufficient only to warn a user of a potential violation, and hope they take the remedial action prescribed. There should be some check and balance to verify that a corrective action was taken, or send reminder notifications if not. Or, to escalate incidents through a relevant chain of command in the event that notifications are being continually ignored or when individual users become repeat offenders. Exception handling is another important area that can provide a means for valuable refinement to the rules and policies.

Conclusion

Automated data classification in Google Drive enables enterprises to protect sensitive information from inappropriate sharing in a sustainable way as usage scales

Current and potential users of Google Drive should not rush into widespread cloud collaboration without putting in place the procedures and mechanisms for data classification and the subsequent detection and remediation of information security violations. But they should also be wary of making those procedures so cumbersome that they drive users away from the platform into rogue behaviors that represent even higher risk for the organization. While traditional DLP systems focus on blocking inappropriate sharing, at the expense of collaboration, CloudLock encourages collaboration. CloudLock's systemic approach enables enterprises to automatically classify data in Google Drive, educate users to be better data custodians, and prevent security breaches, all without creating barriers to cloud collaboration adoption.