boldonjames
A QinetiQ Company

Empower Your Users, Drive Your Business

How User-Driven Classification Can Empower Your Users And Protect Your Data

Introduction

With ever-increasing volumes of data to handle, the proliferation of communication channels, a wide range of security threats and the need to provide ready access to systems for customers and partners, organisations need to find more effective ways to protect their data.

Data classification enables organisations to add context to the information that they hold in messages, documents and files. This context (in the form of a visual and/or metadata label classification) allows the user to control which customers, partners or colleagues they share this particular information with and allows the rest of the IT infrastructure to gain efficiencies in business processes and information management by invoking Archiving and Access Control rules, thereby reducing cost.

In this whitepaper, we outline the key drivers for data classification and compare three different approaches to data classification – manual, automated and user-driven – that could be used to implement and enforce a data classification policy.

Why Classify?

In order to control access to data consistently and effectively, users and systems must understand the value of that data and who should be allowed to share that data. Key drivers for organisations to classify data include:

- **Growth Of Unstructured Data**

A quick Google search provides an indication of the current size and shape of the 'Big Data problem'. Gartner[1] predicts that enterprise data will grow by 800 percent in the next five years, with 80 percent of this being unstructured. Unstructured material spans all forms of content including email, documents, images, videos and text, making it difficult to manage with traditional business systems and security solutions.

Irrespective of volume, organisations must identify, manage and control the data that is crucial to their operation without systems becoming choked by data that is not relevant. They must also ensure that supporting technologies such as archiving and storage remain cost-effective and manageable. Data classification allows organisations to better manage their data by introducing a structure to their unstructured data and focusing time and resources on business-critical data.

Data classification allows organisations to better manage their data by introducing a structure to their unstructured data and focusing time and resources on business-critical data

- **Collaborative Environments**

The increase in collaborative working environments puts pressure on organisations to share data more widely both within and outside of their perimeters, for example with supplier organisations or partners. Additionally, the rapid adoption of collaboration tools, such as Microsoft SharePoint, has outpaced organisations' abilities to enforce consistent policies for data security, governance and access control. The ease with which SharePoint sites can be created means that they often sit outside the view of IT departments, security personnel and even dedicated SharePoint administrators. This presents a huge security risk for the business, as those handling data may not have sufficient knowledge to make informed decisions and the tools to be able to protect it. Collaborative environments can result in either over–protection (where not enough information is shared to get the job done) or under-protection of data (which may result in its loss or leakage).

Data classification means that the user's assessment of the importance of the data can travel with it, so that everyone handling that data is clear as to its sensitivity and safeguarding requirements.

[1] Gartner, Predicts 2013: Business Intelligence and Analytics Need to Scale Up to Support Explosive Growth in Data Sources, December 2012

- **Mobile and BYOD**

A more mobile workforce wants to be able to access an organisation's data on their own devices, which presents a greater-than-ever risk to the organisation's data. The organisation lacks the visibility to determine what happens to sensitive data on-device. When data is classified, policies can be enforced on where sensitive information is accessed and which devices can access it, and users can be educated on the potential impact of combining personal and organisational data on the same device.

When data is classified, policies can be enforced on where sensitive information is accessed and which devices can access it

- **Governance, Risk and Compliance (GRC)**

Many organisations have a strong governance or regulatory requirement for adding data classification to their business processes, in order to reduce business or financial risk. Regulations vary by sector and include:

- Financial Reporting (FCA, FINRA, Sarbanes Oxley)
- International Trade (UK Export Control Regulations, US ITAR)
- Retail (PCI - DSS)
- Healthcare (US HIPAA)
- Information Security (ISO27001)
- Data Protection (UK Data Protection Act, EU Data Protection Laws).

These regulations may be enforced with penalties such as fines which, although painful, pale by comparison with the financial and reputational damage caused by the news of the breach and subsequent market reaction.

Government departments are required to add visual classification markings to information to reflect its sensitivity by legislation including:

- US Government Directive on Controlled Unclassified Information (CUI)
- UK Government Public Services Network (PSN) Code of Connection (CoCo)
- UK Government Protective Marking System (GPMS)/UK Government Security Classifications (GSC) policy
- Australian Government Information Management Office (AGIMO)

Users are mandated to add these visual marks to every document and email message as a matter of national security. The security classification of documents is generally related to the level of impact that disclosure of the information would have to the security of individuals, commerce and the nation.

- **Improving Complementary Security Technologies**

Many organisations have invested in security technologies from which they have yet to derive the benefits they had hoped for or, indeed, been promised. A good example is a Data Loss Prevention (DLP) solution. Organisations using DLP solutions commonly experience frustration with poor end-user acceptance, event management overload and a constant need to refine policy and rules.

Organisations using DLP solutions commonly experience frustration with poor end-user acceptance, event management overload and a constant need to refine policy and rules

When data is classified by the users that understand its context, through the application of metadata in the message or file, DLP tools can act on this additional contextual information to provide more effective results, with fewer false-positive errors, improved user experience and greater overall risk reduction.

- **A Layered Security Approach**

Data classification solutions bridge the gap between the more traditional perimeter IT security solutions (such as firewall protection) and information management solutions. Increasingly, data classification is becoming a best-practice part of a layered security approach, which may include DLP, encryption and Rights Management.

Data classification is becoming a best-practice part of a layered security approach

- **Empowered and Trained Staff**

Enterprises and their employees understand that they now have a responsibility (and in some cases a legal requirement) to protect their customer's data, as well as their own intellectual property, and that taking proactive responsibility for data will in turn protect their own shareholder and brand value.

Organisations that use data classification solutions place users at the heart of their data security approach, which helps increase their awareness of the value of information they handle and determines how it should be protected.

How to Classify

In order to classify information you will first need to establish the schema or taxonomy by which you can categorise information, together with the labelling or marking formats that your users will be expected to recognise. The next step is to decide the method by which classifications are to be derived and applied. The choices broadly fall into the following categories, which we will consider in turn:

- Manual Data Classification
- Automated Data Classification
- User-driven Data Classification

Manual Data Classification

Once the need for data classification is established and a policy defined, some organisations attempt to implement that policy without the aid of classification software. Policy guidance might be circulated and training given, but users are left to manually type classifications into their emails, documents and files. This approach has a number of pitfalls, the first and most significant being consistency of application, as each individual may execute the policy slightly differently (if indeed, at all). The second issue with a manual implementation is enforcement, as each user is left to remember to classify. All it takes is a distracting phone call whilst a user is writing an email, the label is forgotten or misspelt and the email is in breach of policy.

Using data classification software removes the need for manual workarounds, helps organisations enforce a classification policy and ensures their employees are following the same guidelines in a consistent manner

Perhaps the most critical point to consider here is that without a solution in place an all-important metadata label cannot be applied to the message or file, and as a result any security solutions that might be dependent on a label (e.g. an Encryption Gateway or DLP solution) will be unable to apply the correct controls. Finally, there is no clear audit trail, which is vital when proving compliance and also when tracing a security incident.

Using data classification software removes the need for manual workarounds, helps organisations enforce a classification policy and ensures their employees are following the same guidelines in a consistent manner, thereby avoiding any labelling errors and reducing the risk of accidental data loss.

Automated Data Classification

Automated data classification is often provided by Data Governance or Content-Aware DLP solutions and uses software algorithms to select a classification for a file or message. These algorithms may be based on keyword or expression matching of the content, or may involve more complex statistical analysis or fingerprinting techniques. The less processing-intensive techniques may, for example, be delivered by endpoint DLP solutions, allowing for direct policy feedback to be given to the end-user. Otherwise, the more processing-intensive techniques are more commonly applied by server or gateway systems, where processing speeds are not noticeable by the end-user.

The use of keyword or expression matching can benefit from the solution vendor providing industry-specific templates, however unless highly-specific patterns can be used throughout (e.g. a credit card number) then false positives will inevitably occur. Likewise, the accuracy of algorithms based on statistical techniques will depend on the body of data that has been used to train the system, with the added challenge that there is no easy way to understand and eradicate unwanted behaviour that the system has ‘learnt’.

Unlike the process of manual classification, an automated approach will normally be able to add metadata to record the result of its classification decisions. However, it is less likely that an automated system will be trusted enough to add the correct visual markings to the body of document, as the document owner is not in a position to review the resulting changes to the content.

Whilst many automated data classification systems offer an attractive range of algorithms that can be used to classify data, the challenge is to tune these algorithms to provide an acceptable balance between the false positive results that frustrate users and business processes alike, and the false negatives that risk exposing sensitive data to loss. There are many examples of projects failing to deliver on original expectations as operational systems have to be detuned in order to strike an acceptable balance between user acceptability and accuracy.

A number of challenges are generally experienced by organisations when using Automated Data Classification solutions alone:

- **Detection Errors**

An automated classification system is expected to automatically identify and categorise. However errors in automated decision-making are inevitable, resulting in either search rules miscategorising data (known as false positives) or failing to identify sensitive data (known as false negatives).

False positive scenario: an automated classification tool might use simple pattern matching to detect a social security number (SSN) of the format ‘nnn-nn-nnnn’ in order to then categorise the content as ‘Personal’. However the presence of the text “order part number 987-65-4320” would result in a misidentification of the content, as the part number has the same format as the SSN.

False negative scenario: an email containing an incorrectly formatted SSN of “987_65_4320” might evade the pattern detection resulting in the content failing to be categorised.

- **Missing Context**

Many automated classification systems focus on categorising the content of individual data containers, such as files or emails, but will fail to understand the overall context of an aggregated set of data, such as an email body with a series of attachments. This is where a user is best placed to add in their knowledge to provide the true context for that information. A common scenario is where a message is sent with an attachment which may not be sensitive individually, but when combined with content in the email itself requires a more sensitive classification. For example, this might apply if a message contains a customer account number, but the email attachment contains personal address details.

- **Lower Trust Within User Community**

Automated data classification tools are normally configured to provide little feedback to the user on their classification actions for any particular email or file – deliberately so, as they were designed to have as little impact on the user community as possible. However, this lack of information may leave the user unaware of mistakes made by the system or frustrated when such mistakes interfere with their work, appearing to many as a computer error that they may try to find a way round, in the interests of ‘getting the job done’.

Lack of routine engagement may leave the user unaware of mistakes made by the system or frustrated when such mistakes interfere with their work

- **More Resource-Intensive**

A common problem experienced by users of automated classification systems that operate at the desktop is the apparent slow-down in machine processing speed. The increase in processing caused by running complex detection rules creates a drain on system resources and takes up memory, resulting in the user experiencing noticeable delays in some of their most frequently-performed tasks, such as sending an email.

User-Driven Data Classification

User-driven data classification captures the user’s knowledge of the context and business value of the data they create and handle, so that informed decisions can be taken about how it is managed, protected and shared.

Data classification empowers user communities who create and handle data to assign value to it, in a language they understand. These values are then stored as visual & metadata labels on messages and documents, and can range from as simple as ‘Confidential’ labels to complex national security driven data classifications.

User-driven data classification captures the user’s knowledge of the context and business value of the data they create and handle

Advantages of User-Driven Data Classification:

- **Captures the user’s knowledge of value of the data**

No one knows more about a piece of data than the person who created it. Capturing the user’s knowledge and insight and applying it to this data is vital if an organisation is to correctly identify the value of information.

- **Reduces errors, increases trust and improves system performance**

User-driven classification involves the user in the categorisation process from the outset, giving them more trust in the solution and the business processes that depend upon the classifications. The classifications applied by the users can be incorporated into the rules of automated detection systems such as DLP, allowing more accurate and predictable decisions to be made which increases user trust in the system, avoiding workarounds and improving consistency of approach. Fewer errors also means fewer help desk calls with consequential savings in support resources and costs.

- **Lowers overall risk of sensitive data being lost**

User-driven classification can provide a more robust and consistent approach to data protection, reducing the likelihood of detection errors by automated systems causing inadvertent data loss. A more consistent and accurate approach to data classification driven by the knowledge workers who create and routinely handle content ultimately means organisations can be more confident that data is appropriately protected.

User-driven classification can provide a more robust and consistent approach to data protection, reducing the likelihood of detection errors by automated systems causing inadvertent data loss

- **Increasing awareness of the value & sensitivity of data**

Involving a user in the process of identifying and classifying sensitive or valuable data increases their understanding of the nature of such content and its safeguarding needs. One large Financial Services client explained that they had specifically looked for a data classification solution that would be visible to their employees, so that they would be involved in and aware of data classification and their obligations around protecting valuable data. They actively sought to use their employees to safeguard their data, rather than be passively defended by a host of background data security products. In turn, this ensures that organisations can benefit from…

- **Exponential increase in security resource/team/capability**

Organisations who implement a user-driven classification solution can expand their security team exponentially by making all users aware of their security responsibilities. Not only does this mean information security policies and guidelines are more consistently adhered to, but each user becomes an active part of a company’s defence against data loss. This results in an increase in security capabilities and ensures resources are fully maximised.

Organisations who implement a user-driven classification solution can expand their security team exponentially as each user becomes an active part of a company’s defence against data loss

Conclusion

Whether it's to improve management of an ever-growing volume of data, the need to protect sensitive data whilst sharing it with both partners and an organisation's own workforce, or ensuring compliance with regulations, both to avoid penalties and the reputational damage associated with a breach, many organisations are looking to data classification as the foundation of their security approach.

There are many ways to implement any IT security policy, but most methods require some knowledge about the content that is being processed in order to be effective. Automated classification techniques can be used to complement a user-driven classification solution, with user-driven classification providing context to automated content detection which act as a backstop to check that no other sensitive information has been overlooked.

However, the only way for data classification to meet the host of challenges outlined in this paper is for it to be driven by users and informed by their knowledge of the business value of information. If organisations follow this best practice, then they stand a much better chance of protecting their data, their employees and the value of their business.

About Boldon James

For almost 30 years, Boldon James has been a leader in data classification and secure messaging solutions, helping organisations of all sizes manage sensitive information securely and in compliance with legislation and standards, in some of the most demanding messaging environments in the world.

Our Classifier product range extends the capabilities of Microsoft core infrastructure products to allow users to apply relevant visual & metadata labels (protective markings) to messages and documents in order to enforce information assurance policies, raise user awareness of information security and orchestrate multiple security technologies.

Our customers range from commercial businesses to Government, Defence & Intelligence organisations and we are a Microsoft Global Go-To-Market Partner and a Gold Application Development Partner. Boldon James is a wholly-owned subsidiary of QinetiQ, a FTSE 250 company, with offices in the UK, US, Australia and Europe and channel partners worldwide.

www.boldonjames.com

More Information

For more information about how you can transform your data security with a user-driven approach using Boldon James Classifier, please contact us at sales@boldonjames.com or call +44 (0)1270 507800.

Alternatively please register here for a free 14 day trial of Classifier.

Boldon James does not provide legal advice. Nothing in this document constitutes legal advice, nor shall this document or any software product or other offering made reference to herein serve as a substitute for the reader's compliance with any Laws (including but not limited to any act, statue, regulation, rule, directive, administrative order and/or executive order) made reference to in this document. If necessary, the reader should consult with competent legal counsel regarding any Laws made reference to herein. Boldon James makes no representation or warranty regarding the completeness or accuracy of the information contained in this document.

This document is provided "as is" without warranty of any kind. All express or implied representations, conditions and warranties, including any implied warranty of merchantability or fitness for a particular purpose, are disclaimed.