A Trend Micro TrendEdge Solution
Advanced Technologies and Techniques to Enhance Your Product

Data Discovery and Classification in Five Easy Steps

June 2009

Michael Stone
Senior Solutions Architect and Technical Marketing Manager
Trend Micro LeakProof DLP Product
Trend Micro, Inc.
Data Discovery and Classification in Five Easy Steps

Contents

Summary ............................................................................................................................................ 1
Data Loss Prevention Steps .............................................................................................................. 1
  1. Classify: Data Classification for Data Loss Prevention ............................................................... 1
     Identifying Sensitive Data ........................................................................................................ 1
     Questions to Ask During Data Discovery .............................................................................. 2
     Data Traversing Classifications ............................................................................................. 4
  2. Identifying Confidential Data .................................................................................................... 4
     Identifying Unstructured vs. Structured Confidential Data ................................................... 4
     Looking for Patterns in Unstructured Data ............................................................................. 5
     Structured Data Patterns ......................................................................................................... 6
     Combining Identification Markers to Improve Accuracy ......................................................... 8
     High Value Data vs. Highly Unique Data ............................................................................... 8
     Identifying Channels of Escape ............................................................................................. 8
  3. Discovery: Where is the Data? ................................................................................................. 9
     Where Does the Data Sleep? ................................................................................................. 9
  4. Developing Policies: Assigning Actions According to Classification ...................................... 10
  5. Monitor, Report, Refine .......................................................................................................... 10
Practical Tips for Discovering and Classifying Data ....................................................................... 11
Conclusion ....................................................................................................................................... 11
Appendix A – Data Classification .................................................................................................... 12
Classifications of Data ..................................................................................................................... 12
  Classification by Sensitivity ......................................................................................................... 12
  Other Classification Schemes ..................................................................................................... 12
    Top Secret ............................................................................................................................ 12
    Secret .................................................................................................................................... 12
    Confidential ........................................................................................................................... 13
    Restricted .............................................................................................................................. 13
    Unclassified .......................................................................................................................... 13
    Public Information .................................................................................................................. 13
About the Author ............................................................................................................................. 14
About Trend Micro Incorporated .................................................................................................... 15
Contacting Technical Sales Solutions ............................................................................................ 16

Copyright© 2009 by Trend Micro Incorporated. All rights reserved. No part of this publication may be reproduced, photocopied, stored on a retrieval system, or transmitted without the prior written consent of Trend Micro Incorporated. Trend Micro, the t-ball logo, and LeakProof are trademarks or registered trademarks of Trend Micro Incorporated. All other company and/or product names may be trademarks or registered trademarks of their owners.

Information contained in this document is provided "as-is" and subject to change without notice. This report is for informational purposes only and is not part of the documentation supporting Trend Micro products.

TREND MICRO MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS REPORT.
[TSS Part No: BP01LP50_090630US]

This document is a product of Trend Micro Technical Sales Solutions.
Summary

While Data Loss Prevention (DLP) holds the promise of securing sensitive information from leakage, the success of DLP projects frequently hinges on data discovery and classification. This guide is intended to provide guidance on how to quickly decide what data is important, effectively discover that data, and arrive at policies that can protect the data.

Experience gained through customer engagements has led Trend Micro to find that successful deployments typically involve the following steps:

1. Classify – What digital assets do you want to protect?
2. Identify – What is sensitive? How might sensitive information escape?
3. Discovery – Where is the sensitive data?
4. Develop Policies – What to do when you encounter sensitive data? What data should be ignored?

While in an ideal world, enterprises could establish comprehensive data classification regimes in a manner similar to how military or intelligence organizations protect their secrets, such regimes can require a cultural shift that can be difficult to make. This guide approaches the challenge of DLP by advocating an incremental or progressive approach. Find those key assets that might keep executives awake at night were they to leak, and grow from that success.

Data Loss Prevention Steps

1. Classify: Data Classification for Data Loss Prevention
   Defining the digital assets that you want to protect is frequently one of the major challenges in a DLP deployment. When asked, “What types of data do you want to protect?” many enterprises respond with “Everything!” However, protecting everything is neither desirable nor feasible. For example, do you want to protect family photos that employees accumulate on their PCs?

   An optimal way of approaching the issue of what to protect is to speak with line of business owners to understand what information they want protected, watched, and ignored. These conversations can serve multiple purposes by engaging the people who will be paying for the project as well as identifying the pertinent data flows within the organization.

   Appendix A provides an overview of best practices for data classification schemes. While such schemes provide an ideal framework that can be leveraged by DLP solutions, they may be problematic for enterprises to adopt quickly. This is because information classification schemes frequently require top management sponsorship and entail organization-wide changes.

   Most line of business owners or top management know what information preoccupies them, the loss of which would cause sleepless nights.

Identifying Sensitive Data

Identifying sensitive data is one of the single biggest challenges when implementing a DLP solution. The challenge usually results from a lack of knowledge on the part of the IT professionals who are typically responsible for implementing the solution. The majority of IT professionals may have a general knowledge of the information that various groups within the organization use but they may not know the specific data and business processes that must be protected. The data owners in the organization have
that knowledge along with the information about how that data is used on a day-to-day basis. However, these data owners usually lack the expertise from a technological standpoint to make decisions about data protection and leak prevention.

Sensitive data may differ by business segment or region. Some businesses have specific regulations that define what is sensitive. This is certainly the case with PII, PCI, and HIPAA data. Other cases could include groups within an organization that may be affected by regulations even though the primary business does not specifically fall into a regulated category. For example, a manufacturing business may closely handle its own insurance claims. They are not directly in the insurance business, but they could be considered a covered entity according to HIPAA regulations, and efforts should be made to handle that type of data appropriately.

In all cases though, someone within the organization will have an understanding of what data needs to be protected and how that data is used. It is imperative that a DLP project include finding these groups and individuals to understand fully which data is at risk and how that data is at risk.

Questions to Ask During Data Discovery

When a data owner responds with “Everything!” to the question “What types of data do you want to protect with DLP in your environment?” this usually means that you must spend several hours asking the person questions such as:

- What types of information does the organization consider sensitive?
- What does this data look like?
- Where does it “live”?
- Where does it “sleep”?
- How does it move?
- Who uses the data?

By asking the correct questions, you can quickly limit the list to just the important bits of data the organization needs to protect instead of the unmanageable “everything.”

As previously discussed, the type of data that an organization must protect can vary by business type, the regulations involved, or the region. Usually a few basic questions such as those above will usually vet out some simple answers and will give you a better idea of what truly needs to be protected.

What is the Primary Business of the Organization?

Most IT personnel will be able to answer this question in general terms but specific information may be lacking. To complicate matters, the organization may also be involved in multiple lines of business, which is perfectly acceptable. During the initial stages of data discovery, it is important to think in high-level terms and not get mired in details too quickly. A good overall profile of the organization is necessary before diving into the details.

In almost all cases, the business will be primarily concerned with five general categories of data. These are:

- Intellectual property
- Financial information
- Legal information
- Regulatory compliance
- Personal privacy
It is safe to assume that any organization will be interested in at least one of these areas. Every organization has something to protect. It is important though to focus only on one or two of these areas initially as the task of discovering and classifying multiple categories of data simultaneously can be overwhelming.

**Who has the Information?**

Identifying who has the information that must be protected can be challenging in a large organization. Finding the right person with the right answers can require some investigation, but there is always someone or some group who has the knowledge. In some cases, this step can be a great time saver. Persons who are intimate with data that must be protected can shed a great deal of knowledge about the data itself including information on any relevant regulations and industry requirements. Identifying any additional information makes performing the next steps much easier.

**What Form Does the Data Have?**

For each of the categories above, the data will always fall into one of two forms. These forms are **Structured** and **Unstructured**. These terms may be familiar to some who have worked with databases, data models, and schemas.

**Structured data** is anything that has a definable structure or elements that contain structure. A structure can be any combination of words or numbers that appear in various patterns. Identifying the relationships between structured data elements can show us how each element affects the value or importance of other elements. Structured data can occur in a document even if the entire document is unstructured. Some examples of structured data include credit card numbers, national ID numbers, and even postal addresses.

In some organizations, the meaning of the form in which data appears is specialized knowledge held only by a few stakeholders. For example, a credit card number is divided into four sections. The first digit identifies the major industry of the business entity that issued the card, and the next five digits define the issuer of the card. The remaining \((n – 1)\) digits are the customer or card number and the last digit is a check digit. If we want to find data leaks within a credit card company or bank, it is not enough to look just for credit card numbers. Instead, if we speak with members of the account operations department, it may be possible to prevent leaks of credit card information from specific providers by using DLP to block the release of all data containing specific Issuer Identifier Numbers.

**Unstructured data** is information that does not follow any recognizable pattern or have any consistently identifiable characteristics. Examples of unstructured data include legal documents, technical manuals, written reports, press releases, images, video, and other types of free-form documents or data that do not fit into a structured data model.

While it may be tempting to look at everything as unstructured data, this approach quickly can lead to false positives in detection unless careful steps are taken to identify and protect only sensitive data. Without the support of executive leadership, considering everything as unstructured data can turn a DLP project into a massive data discovery effort that may prove both cumbersome and have a high risk of failure. A better approach is to spend a little time to identify structured elements within the data and to classify the data accordingly.
Where Does the Information Live?

Understanding how an organization uses data on a daily basis is an important step in understanding how that data could be leaked. Perhaps the best questions to use to determine how data is handled by an organization include:

- “Where does the information live?”
- “How is the information handled on a day-to-day basis?”
- “What channels are used to move data around the organization?”

These questions all deal with identifying potential loss vectors. For example, if a sales group uses a database for customer and order information, it is not adequate to say, “I want to protect my database.” The real issue is the form the data takes when it is used since that will be the likely form a leak takes. If the sales operations group downloads information from the database into a spreadsheet, then that spreadsheet becomes the medium for a leak over any number of channels. In some cases, it may be adequate simply to put in place a policy to block all spreadsheets from leaving the organization. This may seem like an oversimplification, but approaching the problem from the top down will lead to a much simpler implementation of a DLP solution from a policy perspective. In many cases, the key to data discovery is not to find the minutia (bottom up), but to grasp the overall picture (top down).

Data Traversing Classifications

There is an old tongue-in-cheek saying among plumbers concerning leaky pipes. Simply put, if the rate of evaporation is greater than the rate of a leak, then there is no leak. Now it may seem to some that any data leakage is bad. Determining the affect of a breach can be an excruciatingly difficult question. Sometimes a single piece of data moving within an organization or even outside of an organization may have modest consequences. However, if the same piece of data were multiplied by 500 times, the loss would be considerable. Take, for example, a single US Social Security Number (SSN). In an environment like a university administrative office, one SSN leaking is important but not critical. If we apply the standard classifications in Appendix A, this information probably is considered “Restricted”. Would leakage cause an undesirable effect? If the data were intercepted by the wrong person and used maliciously, the effect would be bad for the person whose SSN was leaked and for the organization that compromised the information. Would this breach have a meaningful impact on the entire organization? Probably not.

Now let us escalate the scenario. Assume someone has a spreadsheet in this same university office with a list of 500 student names and their SSNs. The user inadvertently sends this spreadsheet to an incorrect external email address or perhaps copies the file to a USB memory stick and consequently loses the stick. The result is that the information leaks. While the loss of a single SSN was bad (and deserving of a “Restricted” classification), the loss of a spreadsheet containing 500 SSNs and names deserves a “Confidential” classification. This is because the implications for the organization could be much more severe in terms of regulatory fines, reputation, customer trust, and recovery costs if the lost data were discovered and used maliciously.

2. Identifying Confidential Data

Identifying Unstructured vs. Structured Confidential Data

One of the principles we use in identifying confidential data is the idea of Structured and Unstructured data. For the purposes of Data Loss Prevention, it is possible to find some types of structured data even within unstructured documents. In a legal court document, for example, we can identify specific case number patterns, legal terminology, or other data elements that are common to these types of documents.
Looking for Patterns in Unstructured Data

Continuing with the example of a legal document, below is an excerpt from one of the ENRON court cases that we will use to illustrate the use of unstructured data in Data Leak Prevention.

Figure 1 Example Order of Forfeiture Document from ENRON Proceedings

IN THE UNITED STATES DISTRICT COURT FOR THE
SOUTHERN DISTRICT OF TEXAS
HOUSTON DIVISION

UNITED STATES OF AMERICA

- against -

KENNETH RICE,

Defendant

Fed. R. Crim. P. 32.2(c)(2)

Cr. No.: H-03-93-01

INAL ORDER OF FORFEITURE

WHEREAS, on June 18, 2007, this Court entered a Preliminary Order of Forfeiture, ordering Defendant Kenneth Rice to forfeit the following items ("the Subject Property") in which he had an interest:

(a) real property known as 7207 Last Dollar Canyon, located in Telluride, Colorado, including lot 15A titled in the name of Summit Canyon Qualified Personal Residence Trust I and Summit Canyon Qualified Personal Residence Trust II,

(b) Lot 11A titled in the name of Summit Canyon, LLC;

(b) a platinum, sapphire and diamond necklace with 16 diamonds (total weight approximately 3.38 carats) and 226 sapphires (total weight approximately 15.05 carats) and a platinum, sapphire and diamond bracelet with approximately 6.65 carats of sapphires and approximately 1.68 carats of diamonds, purchased from Bonhime's Jewelry on June 13, 2000;

(c) One 1995 Ferrari F355 Challenge, VIN no. ZZPPR41A2S01043478, registered to Ken Rice;

(d) $55,914 currently on deposit with the United States Marshal Service in substitution for an asset identified as one 1999 Shelby, VIN no. 5CXSX1810XL00027, registered to Kenneth Rice;

(e) $219,112.03 in Ameritrade account no. E240-052559, in the name of Kenneth D. Rice and Teresa K. Rice;
Typically, a document like that in Figure 1 would be considered unstructured data; however, in the context of a courtroom, an attorney’s office, or a legal clerk’s office, this document is easy to identify by some simple structural data elements.

When such a document is normalized (content is extracted), we can find patterns of data in the text. For example, we could search for “UNITED STATES OF AMERICA –against-“ and identify a case that has or is scheduled to occur in any US Federal court.

This document also contains some uncommon legal terminology that also can be considered structured data. Examples from this document include:

- Final Order of Forfeiture
- WHEREAS
- The subject property

These are not terms commonly used by most people on a daily bases. For example, not many people use the word WHEREAS in casual conversation. If we can find terms, words, or patterns such as these in this type of data, it is not necessary to prevent this specific document from leaking. Instead, policies can be developed that limit data leakage by finding documents that contain just the terms, patterns, and other structured data elements used in this type of publication.

**Structured Data Patterns**

Finding structured data elements or patterns is quite simple if we have a sample of the data. For example, a form can contain field identifiers can be easily identified. Then by identifying those elements in a document, policies can be enforced with a high degree of confidence that the data contained in the form is confidential and should be protected.

Figure 2 shows a sample of a common medical insurance claim form.
Figure 2  A Common Medical Insurance Claim Form
In the form in Figure 2, there are over 80 fields that must be completed to file a medical insurance claim. Some of these fields are required; others are situational depending on the claim. In all cases, the field identifiers, such as PATIENT NAME, HCPCS/RATE/CIPPS CODE, NPI, HEALTH PLAN ID, etc., will always be present. Searching for a list of these identifiers in a document will accurately identify this claim form in its electronic format.

Combining Identification Markers to Improve Accuracy

In the previous example, we looked at a medical claim form and found we could identify the form by using its field identifiers. In some cases, this may be adequate if the organization’s policy is to block ALL instances of this form from leaving the enterprise. However, in some cases, it may be possible to improve DLP scanning accuracy by identifying just the data elements that occur in particular fields within the form.

Again, using the example claim form in Figure 2, we can identify particular fields that contain specific patterns of data. For example, the NPI number has a particular pattern. Dates, tax ID numbers, addresses, ZIP codes, and the like also have patterns that can be identified. We can improve detection accuracy by using some simple logic. If we find both the field identifiers, AND some data in the fields that match the type contained in a pattern, we have high confidence that we have a match for that data.

High Value Data vs. Highly Unique Data

When we start to combine structured data elements, we must consider two characteristics of that data. These characteristics are Value and Uniqueness. In many cases, the data will be considered high value but will not be unique enough to stand on its own as a protected data element.

For example, in a financial institution like a bank, a national ID number may be used in conjunction with an account number to identify a customer’s account. Standing alone, the national ID number may not be unique. A US Social Security Number (SSN) is nine digits with no specific pattern. If the number is preceded by a term like SSN, then it is likely that the number that follows is an SSN. However, since there is no exact specification for an SSN, it is also likely that the number could be a phone number or some other identifier. Over all, the data is not very unique.

Now continuing with this example: an account number could be a simple string of digits—perhaps in a pattern, or perhaps with a validation algorithm—but likely without such validation. Finding this type of information alone can be difficult and lead to false positives unless it is found to be unique. In the context of a bank though, this information is unique. This means we can combine this data with other less unique data to get a positive match.

In this example, neither the bank’s account number, nor the SSN can stand by itself. However, when combined, these two data elements increase both their value and uniqueness, and thus improve the accuracy of the match.

Identifying Channels of Escape

Identifying channels of escape is really a function of how data is used in an environment. How data is moved around in an organization, and the business processes the organization has created around that data both give us an idea of how data might leak.

Here are a couple of real-world examples:

1. In our first case, we can look at an order management system. Behind the system is a database containing confidential information that must be protected. This database system is protected by firewalls and various access controls. The problem is not the database itself but how that database is accessed and how that data can be transformed into a leakable form. For example, an employee working in Operations may need to generate a report of the last quarter’s product
orders. To do this, the employee may run a query against the database and extract information about each order into a Microsoft Excel spreadsheet. Then using tools within Excel, the employee develops an appropriate chart that they send to the person or persons who requested the report. Now that the data is in a spreadsheet, it can be leaked over a number of different channels. The most likely accidental escape channel would be email in this case since the employee uses it to distribute the report and the accompanying spreadsheet. It is important to remember that it is possible for the spreadsheet to leak through any number of channels, including webmail, instant messenger, USB drives, or other devices. The important point here though is not to be concerned primarily about possible vectors that can be used to leak the data, but to determine when the data first takes a leakable form.

2. In our second case, let us look at a manufacturing and design firm. This organization’s sensitive data consists of design drawings and specifications for products to be manufactured. To be sure, this data is critical, and could possibly even belong to other organizations, such as the design firm’s customers.

An example of how this data could leak is the following: A design engineer may need to spend some extra time to complete a project over a weekend. Rather than come into the office to complete the work, the engineer may decide to complete the work at home. This would require taking the design drawings out of the office in some form. Since most design drawings are rather large, the likely escape channel is a DVD burner or large external USB drive. Again, it is possible that such data could leak through other channels, but the most likely vector is some channel that facilitates a quick transfer of the data.

3. Discovery: Where is the Data?

Where Does the Data Sleep?

Discovering where sensitive data lives is most important when dealing with unstructured data. Frankly, if data has structure, then locating that data is only necessary for risk assessment. If the data can be detected using structured patterns on a server, file system, document repository, or other system, then that information can be discovered through a loss vector. However, with unstructured data, the information must be located first so it can be identified when it leaks.

Some possible places to look for sensitive data include:

- File server shares
- Document repositories
- Databases
- Email archives

One particular challenge is file servers. The use of file servers and shares always starts with the best intentions of keeping things organized, but unless the data users are diligent in placing files in the appropriate shares, it will be difficult to identify which shares need to be protected and which can be ignored. Ideally, each group in an organization will have shares assigned around job functions and data classifications.

Document management systems are less of a challenge since they impose a certain degree of organization on their content by virtue of their structure. Browsing through the structure of a data repository and identifying the administrators for various sections should allow you quickly to discover what documents are sensitive and which are not.
4. Developing Policies: Assigning Actions According to Classification

What movement is OK? What should be restricted? What should be denied?

While it may be possible to apply the same policy to all data and all persons, this is not practical. If the information cannot move, the organization cannot run. Thus, it is important to understand what movement should be allowed, restricted, or denied. Should a mostly closed policy be in place with exceptions, or should a targeted approach be taken? These are questions that the organization must answer as part of forming a policy.

There are many cases where legacy business processes have inherent possibilities for loss. In these cases, the process could be changed, but this could be a long and painful process. An understanding of the process though will usually lead one to possible areas where loss risk could be minimized.

Referring back to our discussion about identifying channels of escape, we can make some good assumptions about policies. More importantly, we can use this information as a starting point to begin developing a policy around data leak prevention. In the examples we used in Section one above, it may be perfectly acceptable for a person in Operations to handle and even transmit company data, whereas people from other departments should be denied that privilege.

It is important to mention that most policies in a DLP implementation start out in a monitor only mode. In some cases, an organization may opt to stay in this mode. For these purposes, the denial or restriction of movement is not as critical, but it is still critical to record who is taking the action.

5. Monitor, Report, Refine

While the above steps enable a project to get “off the ground” and begin showing results, long-term effectiveness relies on a process of monitoring, reporting on events, and refining policy rules to improve effectiveness. Initial rules for endpoint policies may be adequate, but they may result in some unintended consequences, as some legitimate events may be identified inadvertently as violations. Monitoring the DLP system can flag such events and enable policy refinement. Reporting on policy violations also helps administrators to understand potential points of leakage as well as communicate the positive impact of the DLP system to management.

Refining policies enables you to cut through the chaff by eliminating events that, while correctly captured by the system, are not actual endpoint DLP policy violations. You might initially decide to deny “XYZ”, but over time, you can simply deny “Y” and achieve a more precise result without flagging the benign “X” and “Z”. Increasing the accuracy of policies can result in a reduction of security events and facilitates the wider deployment of a DLP system to cover more endpoints. Endpoint policies typically will not be perfect, but the policies can be refined over time to become more precise.

In the absence of other factors, violation events typically decline over time as users begin to understand that endpoint security policies are being enforced and users start to self-regulate.

Practical Tips for Discovering and Classifying Data

With the above “nuts and bolts” of executing data discovery project, what follows are some practical tips that we have learned while implementing Data Loss Prevention solutions. Many of your major obstacles will be more political than technical.

- Understand what is practically achievable. Rather than perfection, aim for what is achievable. This can yield a short-term impact. Like any project, you need to set yourself up for success and meet or exceed expectations. Rather than discovering and classifying every piece of potentially sensitive data, you might focus on high-risk data like credit card information, CAD diagrams, and customer data.
Data Discovery and Classification in Five Easy Steps

- **Crawl, walk, run.** Get the most important policies implemented and then slowly expand. This sort of gradual, progressive approach allows you to understand internal workflows, optimize policies, and better understand the capabilities of the technology. In other words, trying to boil the ocean on Day 1 has a high risk of failure.

- **Involve key players early.** Involving key stakeholders early in the process increases the likelihood that they will support it during implementation. Define what is and is not achievable so that you can meet or exceed expectations.

- **Beat the regulatory requirements drum.** Staff recognizes that they have to comply with regulatory requirements, and such awareness can both galvanize support for a data classification initiative and support the business. Saying "Sarbanes-Oxley", "Graham-Leach-Blilly" or "HIPAA and Electronic Protected Health Information" generates more awareness and internal support than "because the IT department says so."

- **Address people, processes and technology.** Although you might like to deploy a silver bullet technology to solve your problems, success in data discovery and classification involves people, processes, and technology. Your people need to be trained in how the system works. You need to ensure that processes are streamlined to ensure security while not reducing productivity. Technology needs to be configured and implemented to ensure that it acts as a security enabler instead of as a business inhibitor.

**Conclusion**

Data discovery and classification is a prerequisite to a successful deployment of a Data Loss Prevention solution. Understanding the data flows and classifying information enables organizations to protect sensitive information while avoiding relatively benign information like family photos or grocery shopping lists. This guide provides a quick, practical framework for organizations who want to:

- Begin discovering sensitive information
- Start protecting this information
- Begin evaluating and refining DLP policies and rules once they gain knowledge about the nature of their organization’s internal and external information flows

Traditionally, security is part preventative and part reactive. Data Loss Prevention should be completely preventative but this is not always possible. With proper planning and some understanding of the business processes and data structure, it is possible to cast a net that will catch violations without knowing the identification of a piece of information or the location of a piece of data.
Appendix A – Data Classification

One of the basic elements of a data protection program is creating and adhering to a data classification scheme. Such a scheme facilitates data security because it communicates the level of protection required for a piece of data as well as identifies what audience may view it. Without such a data classification scheme, all data would be treated the same. The problem with an egalitarian treatment of data is that it increases the possibility both that sensitive information will lack security controls and that the sensitive data itself will be compromised. A good DLP solution rests on a foundation of data discovery and classification. This section of the document is intended to provide practical advice to organizations during the data discovery and classification process.

In the simplest form, data classification is the notion of applying labels to various forms of data based on some criterion. The purposes for data classification (DC) can be numerous. For some organizations, a data classification project can be a route to analyzing and improving business processes. For others, the focus of a project could be refining storage requirements for data or information lifecycle management. In the context of this paper, DC will refer to classifying data for information security with the objective of minimizing potential data loss.

Classifications of Data

There are several ways to classify data. This section discusses two of them: classification by sensitivity and other miscellaneous classification schemes:

Classification by Sensitivity

Data classification for security and loss prevention can use many different methods to rank data. One method is to rank data by the sensitivity. In this case, we define sensitivity by the affect the data would have on the business should it become public or fall into the hands of someone who might cause harm to the organization.

Businesses can take a cue from the governments of the world and use the same rankings with some slightly modified definitions. In government, classification of data is primarily concerned with restricting access. In a similar fashion for data loss prevention, we can use such classifications to determine the policies around data and the actions to be taken in the event of a breach.

Other Classification Schemes

The basic levels of classification that most governments use are Top Secret, Secret, Confidential, Restricted, and Unclassified. Within each of these classifications, it is possible to compartmentalize data further—with limits—by department, project, committee, or specific individual. This type of data classification scheme is organized around a “need-to-know” basis and is more of a concern for access control.

Top Secret

Typically, this is the highest category of classification. The loss of such material would cause exceptional damage to the organization should the information be made public. The risk of loss is high and the damage would be irreparable.

Secret

The loss of secret material would cause damage to the organization, disrupt business, and cause potential loss of revenue either directly or indirectly. Losses could be in sales, brand damage, lawsuits, or company image.
Confidential
If information in this classification were disclosed, there might be damage to the organization. This is especially true if the data got into the wrong hands.

Restricted
The disclosure of restricted information would cause undesirable effects. While this classification is close to confidential, the damage resulting from disclosure would likely be limited in scope.

Unclassified
This classification is almost self-explanatory. This type of information will typically start its life as data with a higher classification but will be degraded to unclassified for some reason. An example here would be a press release. The draft of a press release will not be considered restricted or confidential after the release date.

Public Information
Unlike the unclassified category, this information is considered public knowledge. Like the unclassified category above, the information put into public knowledge may start out in some higher classification, but eventually becomes publishable and publicly consumable.
About the Author

Michael Stone has been with Trend Micro for 2 years and is currently a Senior Solutions Architect and Technical Marketing Manager for the LeakProof DLP Product. Mr. Stone joined Trend Micro through the acquisition of Provilla, the original developer of the LeakProof endpoint-based, data leakage prevention application. With over 20 years experience in the network and system security, Mr. Stone has focused the last five years specifically on data leak prevention and regulatory compliance. Prior to Trend Micro/Provilla, Mr. Stone held positions at Reactivity, Proofpoint, Sun Microsystems, and Cobalt Networks.
About Trend Micro Incorporated

Trend Micro Incorporated, a global leader in Internet content security, focuses on securing the exchange of digital information for businesses and consumers. A pioneer and industry vanguard, Trend Micro is advancing integrated threat management technology to protect operational continuity, personal information, and property from malware, spam, data leaks and the newest Web threats. Its flexible solutions, available in multiple form factors, are supported 24/7 by threat intelligence experts around the globe.

Founded in 1988, Trend Micro provides individuals and organizations of all sizes with award-winning security software, hardware, and services. With headquarters in Tokyo and operations in more than 30 countries, Trend Micro solutions are sold through corporate and value-added resellers and service providers worldwide. For additional information and evaluation copies of Trend Micro products and services, visit our Web site at http://www.trendmicro.com/.
Contacting Technical Sales Solutions

The Trend Micro Technical Sales Solutions team is always seeking to improve our publications. Have a question or comment about this document? We would like to hear from you. Contact us at:

sav@trendmicro.com