

# **Putting the Cat Back in the Bag**

## **A Whitepaper on Sensitive Data Detection**

By Bill Pierce – TSF Program Manager

July, 2006



TERACLOUD

Storage Analytics: the value of knowing

## Contents

Overview .....	3
Problem .....	3
Solution Space .....	3
TSF Solution.....	8
About Us.....	10
References .....	11

## Overview

High-profile loss of sensitive data can create serious problems for victims and reams of legal, monetary and public relations issues for businesses. Low-profile loss of internal data and intellectual property can also be damaging. Meanwhile, a cottage industry in data theft and trading has developed.<sup>1</sup> Securing your sensitive data behind a wall of controls can help prevent its escape, but how will you know if these controls have been circumvented or the data was copied before these controls were put in place? Distributed policy-based scanning can monitor for illicit copies of sensitive data across an enterprise.

## Problem

In the last year, government, industry, and academic institutions have been rocked numerous times by high profile, preventable data losses,<sup>2,3,4,5</sup> many involving stolen laptop computers. As the kinds of records we keep on computers have grown, so has data's mobility. No longer locked within the bowels of mainframe systems, powerful databases can be run on inexpensive and increasingly mobile platforms. When those platforms are stolen or cracked, the data they contain (usually worth much more than the hardware itself) is at risk of exposure.

It is altogether too easy to make copies of data that are every bit as sensitive as the originals. Databases may be copied for a variety of purposes such as backup, offline access, data mining, legal quarantine and development projects. Sensitive documents such as financials, strategic plans and intellectual property, usually kept in less structured file systems, can also be copied and put at risk.

In response to this risk, many organizations are putting policies and Information Technology (IT) systems in place to identify, secure and monitor for the unauthorized proliferation of sensitive data and are able to demonstrate to auditors that these steps have been taken.

## Solution Space

As with most IT issues, the solution starts with corporate policy.<sup>6</sup> You need to identify the cat (your sensitive data) and the bag you want to keep it in (where it should be stored). Then you establish layers of security to keep the cat where you want it. Data classification and data placement are areas where Storage Resource Management (SRM) tools can help, but that is the subject of another whitepaper. This paper will assume you know where your cats are and how to identify them.

Next, you need to define corporate policies surrounding access to that information and translate those policies into processes--such as training--and IT mechanisms--such as software--that help you implement them. Employee training on data policies is an important component. Employees must know what constitutes sensitive data and where it should be kept. IT mechanisms can be put in place to help employees comply.

## Layers of Sensitive Data Security



The layers of security must be as multi-faceted as the attack vectors.<sup>7</sup> The lower layers are aimed at keeping the cat in the bag by tightening down access to sensitive data. The upper layers start from the assumption that the lower layers have failed, were circumvented or were implemented after the escape of the data. They help you put the cat back in the bag or declaw the cat so that it is harmless if it gets out.

### Keeping the Cat in the Bag

One way to reduce the risk of sensitive data proliferation is to limit access to the data in its raw, file form so that it cannot be copied in total. This involves making the parts of the data available only through a networked application. This part of the solution is pretty

well understood and involves three layers of data security: Application, Operating System (OS), and Storage.

### Application Security

Application security works by restricting access only to parts of a database at a time, allowing it to reside only in volatile storage (memory) on the clients. Data can be kept behind a networked service application and requested and used on an as-needed basis. This makes wholesale copying of large amounts of data much more difficult. Applications can implement their own security mechanisms such as authentication and authorization controls and server-to-client encryption.

Even with application-level security, there's still a cat. The data must exist as files on volumes that are either directly attached to the server or accessed via a Storage Area Network.

### OS Security

OS security involves making access to the raw data (or a copy of it) difficult if the person gaining access has circumvented the protections the application provides. Perhaps the best understood, OS Security involves logical and authenticated access control and file system-level encryption. File access auditing could be considered another layer, although this is more of a detection function than a prevention function.

## Storage Security

The lowest and perhaps the least-understood layer of data security is storage security, which involves providing physical security to the hardware. If this is breached, so are layers of logical security above. Storage security includes SAN device security for switches, hubs and storage devices, and SAN access control with fabric zoning or Logical Unit Number (LUN) masking and mapping providing logical firewalls that control which devices are allowed to communicate on the SAN. More recently, Fibre Channel standards have added password, key and certificate based authentication mechanisms.

Together, these measures help you keep the cat in the bag and can serve as your first line of defense. But what if it is already out, or what if these mechanisms fail? How will you know?

## Knowing When it Gets Out

Preventing proliferation of sensitive data is somewhat like the process of preventing virus infections. Just as scanning incoming email and web content (protecting the perimeter) is augmented by local scanning of hosts (policing the interior), detecting unintentional proliferation of sensitive data by scanning augments the techniques suggested above for securing that data behind a wall of security.

We find that a common problem in SRM is that customers do not have a very good handle on what data is stored, where it is stored, and who might have access to it. Corporate file systems have grown unchecked for years. Copies of sensitive information are made, used and forgotten. On top of the resource consumption concerns, there are now security and liability risks created by unchecked data and storage growth.

One reason this condition exists is that customers do not have adequate tools in place to give them visibility into their data storage. Tools are available that are capable of scanning file systems across an enterprise for files meeting certain criteria. For the purposes of this whitepaper, we will call them policy-based scanners. Policy based scanners allow a set of

**"Corporate filesystems have grown unchecked for years. Copies of sensitive information are made, used and forgotten."**

policy criteria to be defined centrally and evaluated in a distributed fashion. These kinds of tools have been developed for Storage Resource Management (SRM), and Information Lifecycle Management (ILM) purposes, but they can be applied to the problem of sensitive data detection.

Policy-based scanners allow you to build a set of metadata criteria (a policy) that are evaluated against files that are scanned. When one of these policies is violated by a file, that file is associated with the policy and a list of files violating the policy is built. Metadata scanners look at file metadata such as name, owner, or last modified time. UNIX's *find* command is an example of a metadata scanner. Some scanners are capable of looking at not only file metadata, but also matching patterns within the content of each file.

## Putting Policy-Based Scanners to Work

So let's consider how a policy-based scanner would be used to detect sensitive data. In many cases, it may be as simple as looking for files with a given name. Suppose you keep an SQL database with customer information called CRM.mdb that you know is greater than 100MB in size. One policy might be:

"All files that end with mdb"

This would catch anyone using a Microsoft Access database, for which there might be many legitimate cases resulting in false positives, so you could refine the policy.

"All files that are larger than 100MB AND begin with CRM AND end with mdb"

Once a policy like this has been defined for each type of sensitive data you wish to detect, the scanner will distribute it to the agents deployed on each host you wish to scan. Hosts can include servers, workstations and laptops. To protect laptops, you create an IT policy that ensures each laptop on the network is available for scanning at least once a week. Most solutions should be able to track how well that scanning policy is followed. The next time a scan is triggered, files meeting these criteria will be identified and their locations sent back to the central policy engine, where they may be viewed in one location and actions taken. The files might be immediately deleted, auditing might be turned on, or the owner might be notified by email.

Rather than checking these policies for violators regularly, commercial solutions should allow you to automate actions, such as sending email when a violation is detected. Policy based-scanners can find intentional or unintentional copies of sensitive data and give you peace of mind that other security methods have been effective.

While home-grown solutions based on scripts and free software such as the *find* utility are capable of detecting sensitive data, they also come with their own development and maintenance costs and lack many convenient features commercial solutions provide.

## Encryption

Another layer of security at your disposal is encryption. Encryption is something like declawing the cat so you can safely let it out of the bag. Encryption is your last line of defense when business requires that sensitive data reside on mobile platforms. Simply locking the data behind OS password authentication is insufficient, as a wide variety of mechanisms can circumvent this, particularly when physical access to the computer is available. The security provided by encryption is only as good as the algorithm it uses and the security of the encryption keys. For encryption to be effective, you must have secure key management mechanisms in place, and more importantly, the employees who use these keys must be trained to treat the keys with as much reverence as they would the data the keys protect.

## Vulnerabilities of Each Layer

With any security scheme, it is important to understand the scheme's weaknesses and vulnerabilities. All of these approaches have weaknesses.

### Application/OS/Storage Security

Locking your data behind an application is not always practical. There may be traveling and offline users who have a legitimate need to access the data without a network connection to the server. Furthermore, the data may be more vulnerable on the wire than it would be encrypted and at rest on a properly secured, strictly-offline laptop.

Even with Application and OS security, the data is still stored in files and volumes, and copies will be made for a variety of reasons. If Application/OS security is circumvented or copies proliferate at the file level, there will be no way to know.

### Encryption

Encryption can be broken. Algorithms once thought to be secure are later defeated. Effective encryption also requires effective key management and is utterly ineffective if a key is obtained. Conversely, a risk of using encryption is that the data can be lost if the key is lost. Encryption also comes with a performance penalty for accessing the data, which may or may not be a concern. Lastly, encryption does not address the problem of encrypted or unencrypted copies being made of the data.

### Data Detection

Relying solely on the detection of sensitive data also has its drawbacks. There may be legitimate reasons for copies of the data to exist, even on laptops. A window of opportunity may exist between the time when the data is copied and the time when the copy is detected. Metadata scanning can be easily circumvented by renaming the data, although this may render it unusable to applications that are required to access it. It can also be circumvented by copying the data to removable media that are not scanned.

Data detection may be enhanced against these risks by scanning content in addition to metadata, though this can also be defeated through various forms of encoding (for example, compression, encryption) Content scanning is also significantly more resource intensive than metadata scanning. While these drawbacks may seem severe, keep in mind that many of the data losses to date were not due to malicious intent, but ignorance or unintentional exposure of the data. This is where data detection techniques can help.

The weaknesses inherent in each of these approaches by itself mean that layered security that combines them will be most effective.

# TSF Solution

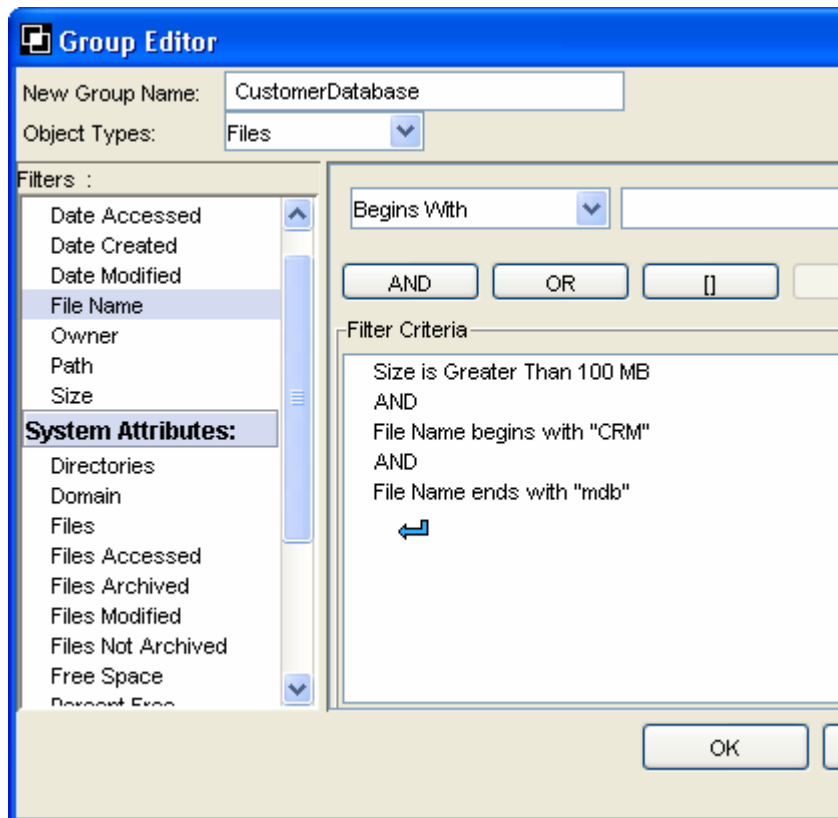
TeraCloud Storage Framework (TSF) is a tool developed for SRM. One of the early strengths of TeraCloud's technology in the marketplace was its support for file-level detail, including the ability to define groups of files based on metadata criteria. Policies can further be defined and applied to the members of those groups. TSF includes a feature called ProActivity, which allows user-supplied scripts to be executed on the members of a group or set of policy violators.

## A Hub of Protection

Simply install the TSF Agent (which supports Windows, Solaris, AIX and Linux) on the systems you wish to protect, and create groups and policies for the file criteria you wish to monitor. A single TSF server can provide a hub of protection for up to 20TB of storage with industry-leading pricing for Small and Medium size Businesses.

## Centralized Policy Definition

File and directory criteria can be specified through a powerful "query-builder" interface allowing logical expressions to be built based on metadata attributes.

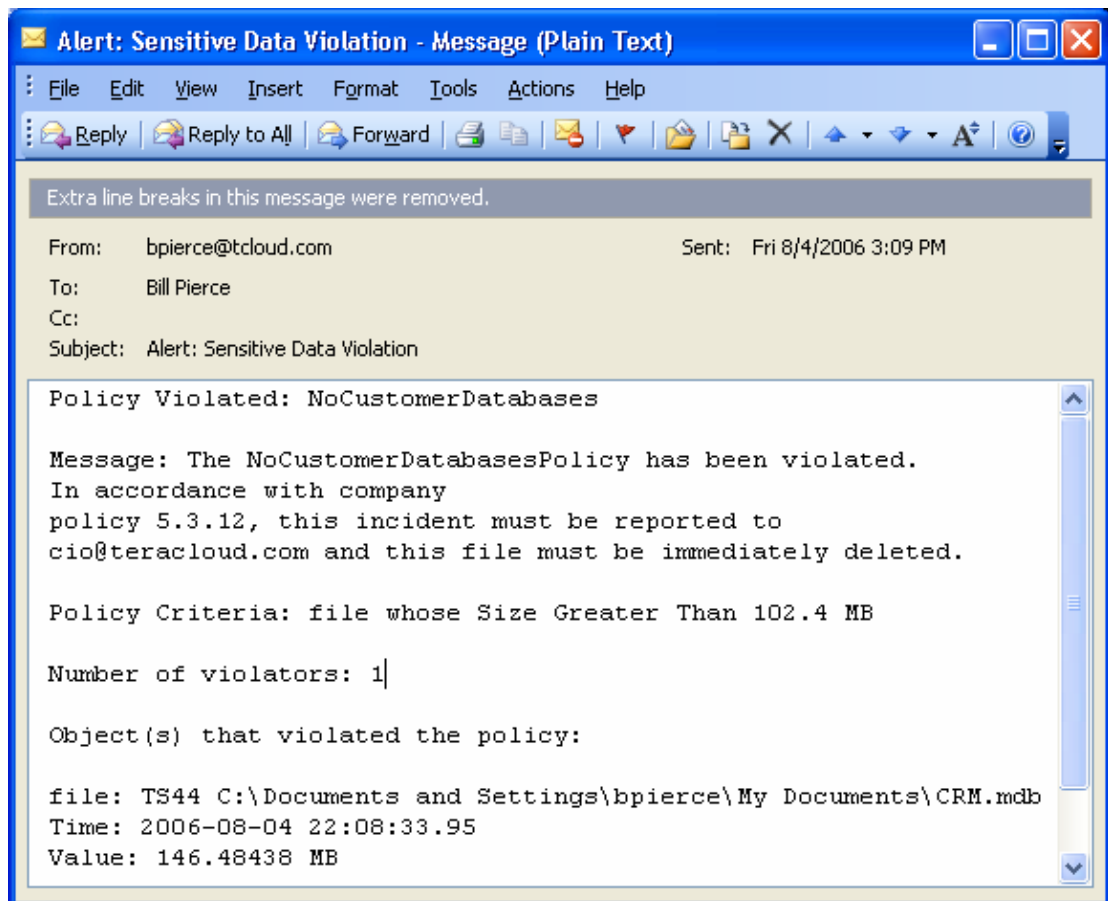


## Distributed Scanning

The group and policy criteria are pushed to the agents on a regular interval and file systems are scanned for files matching the criteria. The server maintains records of when each agent last successfully scanned.

## Centralized Reporting, Alerts and Actions

The location of files that match the criteria are returned to the server where the results can be viewed, sorted, searched, graphed and printed in custom reports. Policies allow different users to be notified by email when a violation is detected.



ProActivity allows arbitrary scripts to be executed on the group members and policy violators.

## Storage Analytics: The Value of Knowing

TSF provides a layer of scanning security on top of other measures you may have implemented. It can serve as part of an effective plan to protect against costly exposure of sensitive data. TSF is also a fully functional SRM tool that has considerable benefits when deployed in your environment. Reduce your overall storage needs and rate of storage growth. Prevent outages. Reclaim space. Forecast and plan capacity growth. Implement chargeback. Improve levels of service. Helping to put the cat back in the bag is just one example of Storage Analytics: The Value of Knowing.

## About Us

TeraCloud Corporation continues to define the storage analytics market serving global Fortune 2000 and small and medium business (SMB) customers. TeraCloud's groundbreaking Storage Analytics (TSA) technology provides storage utilization analytics to pin-point conditions and predictive events where storage utilization is sub-optimized. Regardless of platform, the TeraCloud Storage Framework (TSF) v2.1 provides enterprise customers with instant knowledge for predictable storage management from a single view. TSF Lite is the only product to provide SMB's with a powerful, enterprise-level and cost-effective tool for predictable storage management. TeraCloud's SpaceFinder Suite v4.4.1 gives customers the knowledge for predictable storage management on the Z-series platform by providing automating storage administration that monitors, detects, analyzes and proactively resolves issues threatening storage availability.

Founded in 1991, TeraCloud Corporation is a privately held company based in Bellevue, Washington.

## References

1. Zeller, Tom Jr. "U.S. Arrests 7 on Charges of Credit Data Trading." *New York Times*, 29 March 2006.
2. Lee, Christopher "Top VA Officials Criticized in Data Theft." *Washington Post*, 12 July 2006, A13.
3. "Equifax Inc.: Company Laptop Containing Employee Data is Stolen." *Wall Street Journal*, 21 June 2006, Eastern Edition.
4. "Aetna Reports Theft of PC with Personal Data." *Los Angeles Times*, 27 April 2006, C4.
5. Conkey, Christopher, "FTC Reports Laptop is Stolen in the Latest U.S. Data Breach." *Wall Street Journal* 23 June 2006, B2.
6. Limoncelli, Thomas A., and Christine Hogan, *The Practice of System and Network Administration*. New York: Addison-Wesley, 2002.
7. Britt, Phillip "Data Security: An Ounce of Prevention." *Information Today* June 2006: pg. 1.