Improve Data Protection Processes with Content Discovery, Monitoring and Filtering

Tom Olzak
November 2007

It's all about the data. Understanding this statement and implementing controls accordingly is a key element in protecting sensitive data. Yet, organizations are faced with a growing number of channels through which PII (personally identifiable information), ePHI (electronic health information), and intellectual property leak to targets not authorized to store or possess it.

The traditional methods of protecting data are no longer effective enough. Security managers must take additional steps to ensure data owners and security personnel know where the data are located and where they might be sent. At a high level, this is the function of content monitoring and filtering (CMF).

In this paper, I define the challenges facing organizations as they attempt to protect sensitive information from both unintentional and malicious activities, including what characteristics of data make them an easy mark. I then look at arguably the most important approach to meeting these challenges—CMF. Finally, I describe one of the top three CMF solutions as an example of how current technology can be applied to data leakage protection.

Challenges

There are two primary reasons why data leakage is becoming more prevalent. First, users are getting more and more access to databases and other data repositories so they can run queries and develop custom reports. On the surface, this is a big business benefit. Business users don't have to wait in an IS development queue for reports, and developers are released to work on more complex and often more strategic projects. However, access to large amounts of information outside of application controls often results in the storage of sensitive data in spreadsheets and other types of documents spread across the enterprise in repositories with questionable levels of trust, including:

- Local drives
- Network shares
- USB drives
- CD/DVD
- Memory sticks
- iPods or other media players

Further, once data is removed from secure repositories, it can be sent via any number of communication channels to countless destinations. This brings us to the second primary cause of data leakage—a growing number of user accessible communication channels.

Each day additional communication channels become available across various ports and from various online services. These include:

- Instant messaging
- Peer-to-peer networks, including GotoMyPC
- FTP
- Online file transfer services, such as transferbigfiles.com
- Email
- HTTP
- Online disk storage services, such as XDrive

Attempts to block all existing and emerging channels will only result in channel providers and users working to find new ways to circumvent controls. Encryption might be a possible answer, but according to Rich Mogull, a Gartner analyst,

> "Although encryption of sensitive information, including to the row or attribute levels in database tables should be an eventual control, the ability of businesses to achieve this level of encryption, and properly manage it can take years to accomplish" (2006)

Even if an organization already has a sophisticated, well-managed encryption solution in place, the biggest potential weakness in data leakage prevention is still information in the hands of those who actually have permission to access it. The data leak risk associated with specific data depends on four characteristics of that data: accessibility, significance, copyability, and detectability (Heiser, 2007).

Accessibility - The ease with which data is accessed is a big factor in whether it is a significant leakage concern. Data easily retrieved, manipulated, and stored in low trust repositories are excellent targets for cybercriminals. Methods of access to the data include

- Enterprise search solutions
- Data warehousing
- ERP
- Read-only direct database access for query functionality
- Access via other business intelligence systems

Significance - Data with no value puts the business at no risk if leaked. As value of information to an outside entity increases, so does the probability that a criminal will apply the effort necessary to acquire it. The attractiveness of information to an outsider depends on several things, including:

- How easy it is to sell and its market value
- Whether it provides the attacker, or the attacker's employer, with competitive advantage

- Whether the information has social or political significance that can be leveraged to advance an agenda
- Blackmail

Copyability – This is a no-brainer; the easier it is to copy data, the harder it is to control

Detectability – This measures an organization's ability to monitor for and react to anomalous use or movement of data. It further gauges the extent to which users are aware that the data is being monitored. According to Heiser,

> "The greater the expectation that data leakage would be noticed and acted upon, the less likely someone would be to steal it" (2006).

These four characteristics of data leakage risk can be depicted in a variation of the standard information security risk formula. The variation is shown in Figure 1.

$$\text{Risk} = \frac{\text{Accessibility} * \text{Significance} * \text{Copyability}}{\text{Detectability}}$$

Figure 1: Data Leakage Risk Formula

This is not meant to be a perfect model of data leak risk. It is simply a tool to help visualize the relationships between various data leak considerations. Using this model, we can see that risk is reduced by decreasing accessibility, significance, or copyability. It can also be mitigated by increasing detectability.

Accessibility can be decreased by improving identity and access management processes and technology or by encrypting sensitive data. Denormalizing information stored in databases and storing it in special-use data warehouse repositories can limit the amount of information business users can access when running their queries.

Copyability is addressed through administrative controls that are well-known to the user population. Technical controls to control or prevent the use of removable storage devices also reduce the copyability factor.

Reducing the significance of the data is very difficult. If the data have no value to an attacker, resources applied to protecting them are better utilized somewhere else.

Finally, detectability is increased through the use of tools that detect sensitive data stored in low-trust repositories and, in general, monitor the use of that data anywhere on the network. The rest of this paper deals with steps organizations can take to strengthen detectability across their enterprise networks.

Content Monitoring and Filtering (CMF)

What is CMF?

CMF is not a preventive control. It is a detective safeguard that determines whether your preventive policies, standards, guidelines, processes and technology are working effectively. Stated another way,

> "Enterprises should use CMF/DLP technologies to develop and enforce better business practices in the handling and transmission of sensitive data, and vendors should recognize that this is where their products' true value lies" (Proctor, Mogull, and Ouellet, 2007).

Not only is CMF a key element of detectability, it is considered by Gartner to be the number one step to take to prevent data loss (Mogull, 2006 July (b)).

CMF solutions typically perform two basic tasks on a network. First, they discover and classify sensitive data on local or network storage. Classification is performed by using user-defined rules. After discovery, business policies defined by management and configured into the CMF software determine whether the data are stored in a location with controls commensurate with its classification.. If it's determined that they don't, then the software can either move the data to a secure location, send an alert to an administrator, or both.

Second, a CMF solution should be able to look inside packets to identify sensitive data in transit. If sensitive data are detected being copied or sent over a monitored communication channel, user-defined business rules determine whether to allow the data to continue to its destination, alert an administrator, or both.

Both tasks described above enable organizations to know where its information is located, where and how it's traveling, and whether controls protecting it comply with business policies.

How does CMF locate/detect sensitive data?

Anyone who's used filtering software knows that false positives are an ever-present problem. Ratcheting down the ability to identify text reduces errors, but it can also significantly reduce filtering effectiveness.

A true CMF product meets the challenge of false positives by employing multiple methods for detecting target data: key terms, key phrases, policies, and filters (Bowers, 2007).

> **Key terms and phrases –** Using key terms and phrases alone will most certainly cause a high number of false positives. In addition, getting the rules set so as to eliminate them can take months and result in a crippled filtering solution.
> **Key policies and filters –** Vendors typically provide a significant number of policies and filters out-of-the-box, including compliance sets for regulations like the HIPAA. By themselves, policies and filters can result in a lower false positive rate. Used in isolation, however, they can miss key information bites that terms and phrases filtering might detect.

The best solution is to use a layered approach in which terms, phrases, policies, and filters work together to provide effective monitoring with a low number of false positives. According to Bowers, this is the best way to clear away the 'noise' so that management can see the true data leaks (2007).

An Example of a CMF Solution – Vontu

The process I used to select an example solution was simple; I looked at Gartner's magic quadrant and picked one of the vendors listed. It wasn't a big list, and I had no special reason to select Vontu. Before looking at the Vontu approach to CMF, I'm going to describe Gartner's definition of an adequate solution.

The magic quadrant I used for this paper was released for the 2nd quarter of 2007. Written by Gartner analysts Paul E. Proctor, Rich Mogull, and Eric Ouellet, it describes the requirements the authors believe should be met for an application to be an affective CMF solution (2007). These requirements include,

- Ability to perform content aware, deep packet inspection on outbound packets using a variety of communication channels.
- Ability to extend beyond individual packet analysis to complete session analysis.
- Ability to use linguistic analysis to supplement simple word matching, including advanced regular expressions and document fingerprinting. ("Linguistic Analysis is the process of breaking down a document to extract the important concepts and meanings it contains" (http://corp.infocious.com/tech_tech.php)).
- Ability to detect content in accordance with policy-based rules.
- Ability to monitor traffic for email (minimum requirement) and other common communication channels (e.g. HTTP, IM, and FTP). Analysis must be performed across multiple channels with results accessible via a single management interface.
- Ability to block policy violations over email.

Using these requirements, and other criteria, Gartner produced the magic quadrant shown in Figure 2. There are few players in this space with only three doing well in both ability to execute and completeness of vision. Again, I chose Vontu as an example as an example system.

How Vontu works

Vontu is a collection of modules that meet all the requirements listed by Gartner as well as those covered in *Content Monitoring and Filtering (CMF)*. Figure 3 depicts the elements of the Vontu solution I'll be discussing in the remaining sections of this paper.

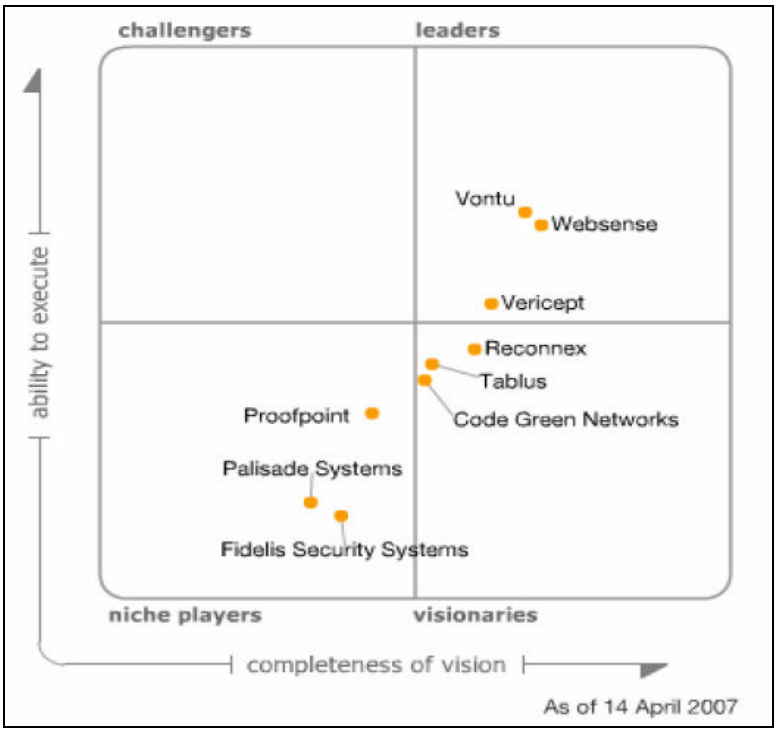

Figure 2: Gartner 2Q2007 CMF Magic Quadrant
(Proctor, Mogull, and Quellet, 2007)

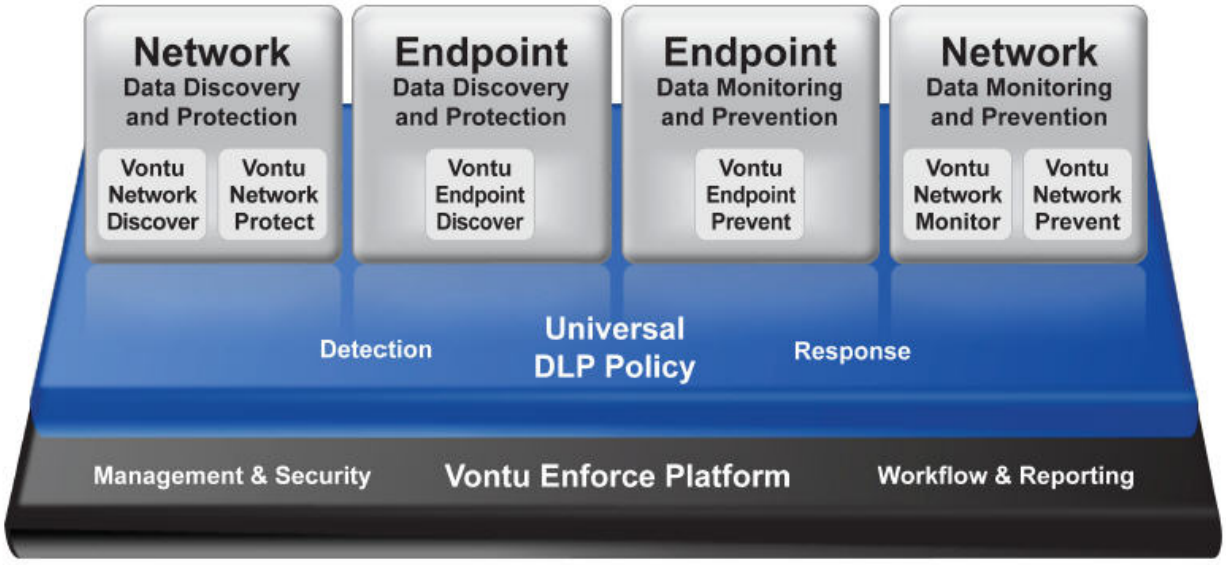

Figure 3: Components of Vontu Solution
(Vontu, 2007)

The Vontu suite is divided into four functional areas – network data discovery and protection, endpoint data discovery and protection, endpoint data monitoring and preventing, and network data monitoring and prevention. Within each of these areas, one or more software modules drive CMF functionality. Underlying all modules is the universal DLP (data loss protection) policy and a management console/platform.

Monitoring and filtering within the functional areas combine three approaches to identifying sensitive information that, when layered, provide for results with few or no false positives (Heck, 2007). See Figure 4.

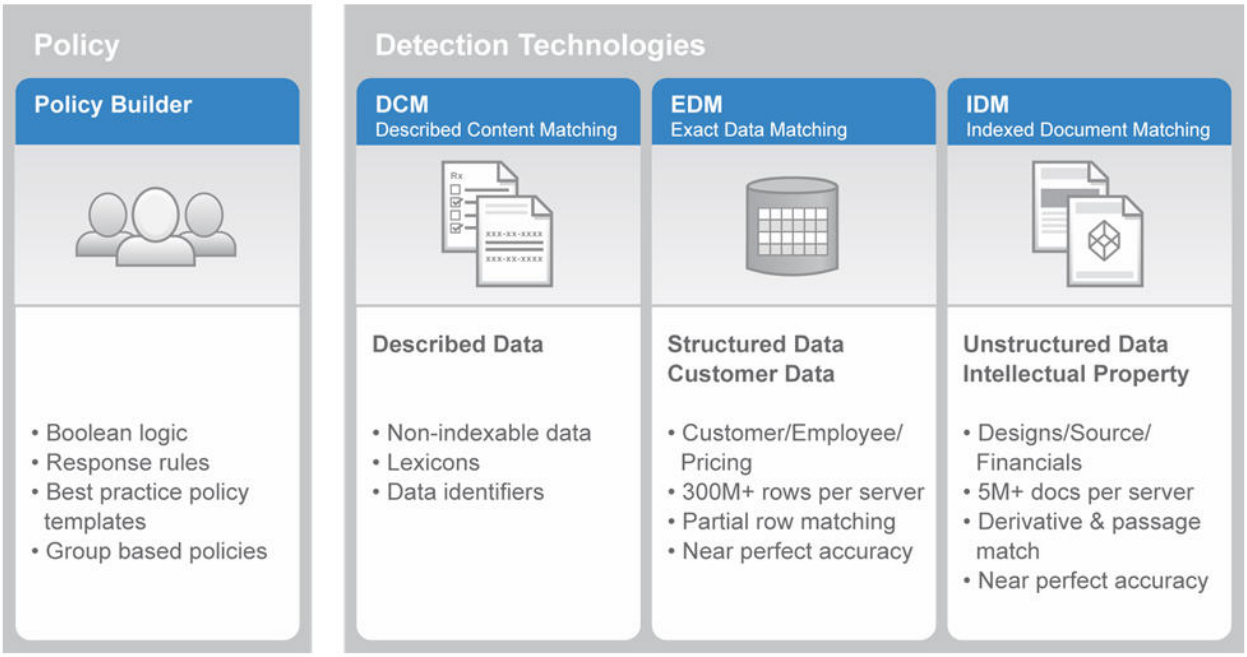

Figure 4: Vontu Layered Data Detection
(Vontu, 2007(b))

Described content matching uses lexicons, Boolean logic, and data identification patterns in an attempt to identify sensitive information contained in non-indexable documents, such as email messages.

Exact data matching uses extracts from your data repositories to create indexes of your sensitive data. This means that instead of looking for a regular expression like <999-99-9999>, Vontu will look for actual social security numbers that exist in your employee database.

Indexed document matching uses linguistic analysis and fingerprinting to determine if a document contains sensitive information.

Once sensitive information is discovered, business rules created with Policy Builder determine its classification and what steps to take to protect it, including blocking communication, moving it to a location at an appropriate trust level, or alerting security personnel and administrators.

Now that we have a general picture of how sensitive data is identified, let's dive a little deeper into each of the Vontu functional areas.

Network Data Discovery and Protection

The purpose of Network Data Discovery and Protection is to discover sensitive data stored in network-based repositories and take steps to protect exposed data according to established business policies. Figure 5 depicts this process.

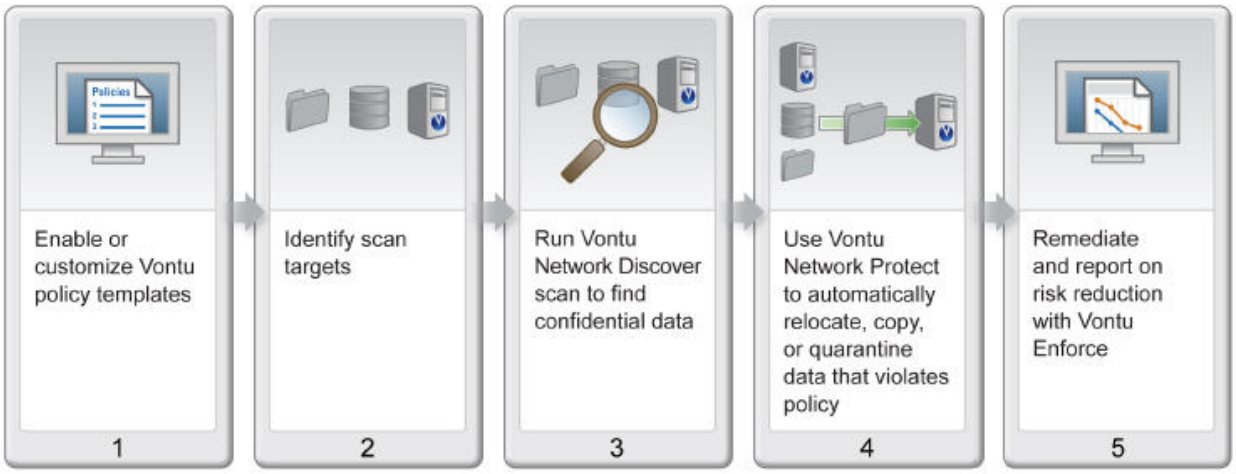

Figure 5: Network Data Discovery and Protection Process
(Vontu, 2007 (c))

Network data repositories supported include,

- File servers
- Databases
- Microsoft SharePoint
- Lotus Notes
- Documentum
- Live Link
- Microsoft Exchange
- Web servers

Once sensitive data is discovered that is not located in a repository with a sufficient level of trust, one of the following steps can be taken automatically, as determined by policy templates:

- Auto-quarantine
- Relocation of data to a more secure repository
- Alerting data or security administrators
- Support for integration with third party security solutions, including,
  - Data classification
  - Storage tiering
  - Archiving
  - Encryption

- Digital rights management

Endpoint Data Discovery and Protection

Endpoint data discovery and protection is very similar to the network process. Using an agent, the endpoint discovery process looks like Figure 6.

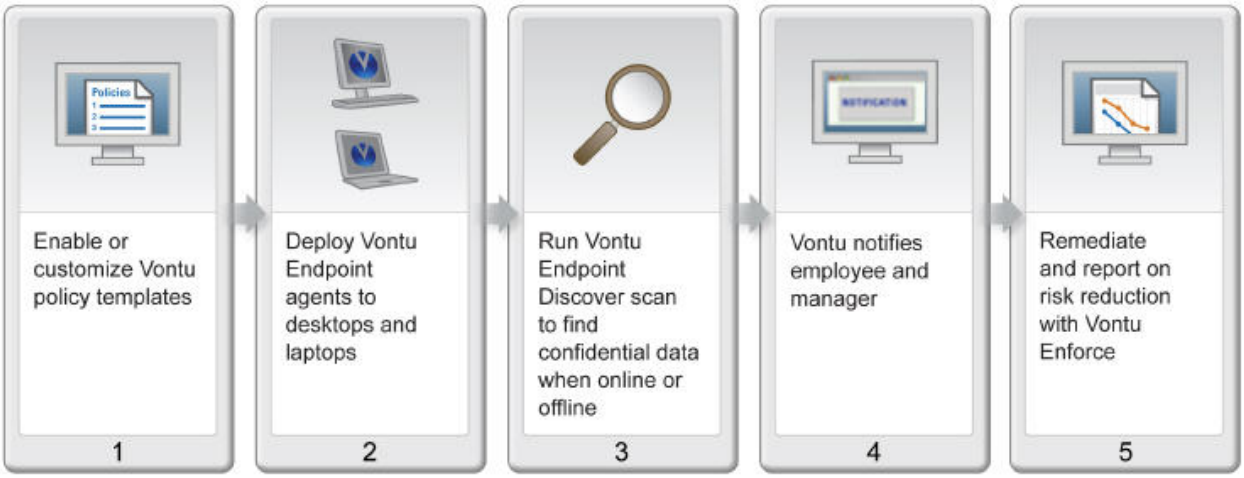

Figure 6: Endpoint Data Discovery and Protection Process
(Vontu, 2007 (d))

Scanning desktops and laptops for unprotected data is continuous, even when the device is not connected to the network. All agents are managed centrally and reporting is accessed via the same management console.

Network Data Monitoring and Prevention

Discovery and protection addresses the security of data at rest. Monitoring and prevention ensures data in transit is protected. Figure 7 steps through the Vontu monitoring and leakage prevention process.

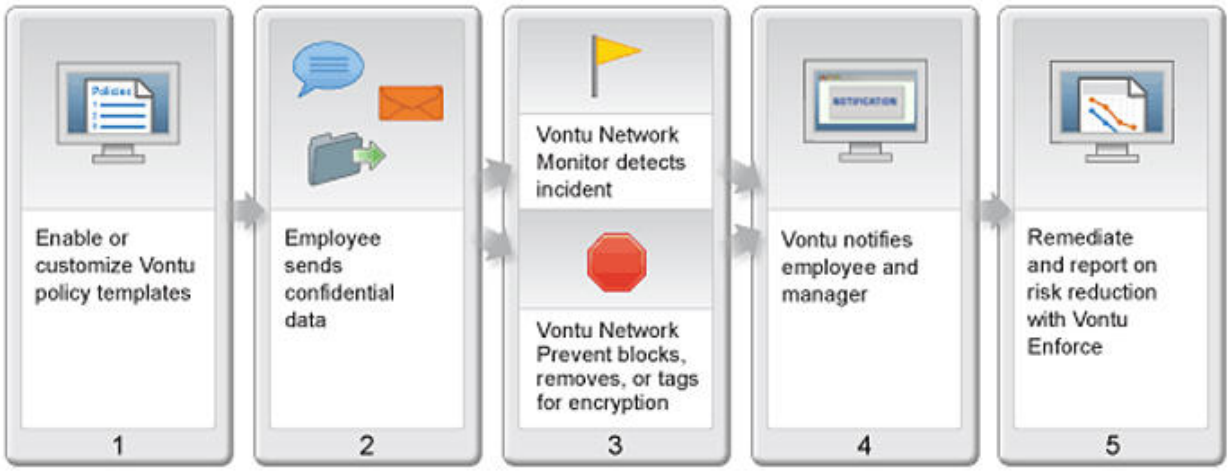

Figure 7: Network Data Monitoring and Protection Process
(Vontu, 2007 (e))

Vontu inspects packets traveling internally as well as those heading out of the network via any communication channel. Using policy templates, it looks for and reports on confidential data in motion. This simple monitoring alone provides organizations with information necessary to both qualify and quantify the risk of data leakage.

If configured to do so, Network Prevent blocks network communication based on the data found in packets. It can also tag traffic for encryption by a third party encryption solution. For example, an email message found to contain sensitive information might be tagged so that an email encryption gateway encrypts the message prior to its travels over the Internet.

Endpoint Data Monitoring and Prevention

One of the biggest challenges facing security managers today is the movement of data to mobile storage devices, such as USB drives, memory sticks, and CDs. Figure 8 shows how Vontu's endpoint protection solution helps monitor and control data as it travels through endpoints.

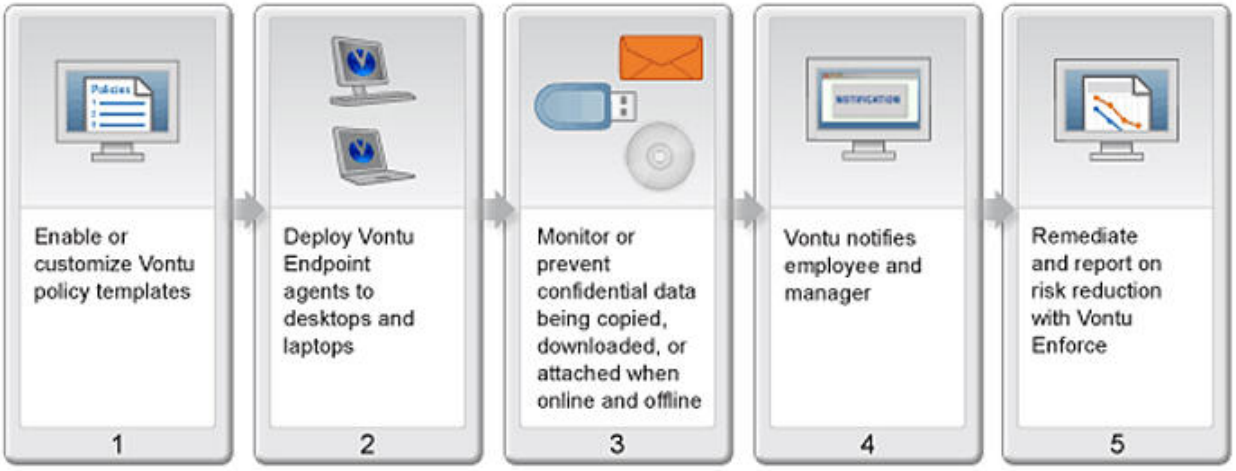

Figure 8: Endpoint Data Monitoring and Prevention Process
(Vontu, 2007 (f))

Vontu's Endpoint Prevent monitors data passing through an endpoint device and takes predefined actions when confidential data is detected passing to external or mobile locations, such as,

- USB devices
- CD/DVD
- Web Mail
- IM
- Peer-to-peer connected devices

Detection is effected whether the endpoint is attached to the network or in standalone mode.

Actions taken when confidential data is detected moving to a mobile device or target outside the control of the responsible organization are determined by the system administrator, and include reporting and blocking.

Conclusion

In this paper, I described the various data leakage challenges facing businesses today, including enterprise-wide distribution of data and the sending of confidential information over a growing number of communication channels.

We also walked through the four characteristics of data that make them susceptible to data leakage/theft: accessibility, significance, copyability, and detectability.

Focusing on detectability, we looked at the importance of CMF in decreasing data leakage risk and the various requirements of an effective CMF solution. This was followed by a walkthrough of the functionality of one of the top CMF application suites on the market, Vontu.

Deployment of a CMF solution is not an easy task. It requires planning as well as management and workforce support. However, implementation doesn't have to be an all-or-nothing proposition. Organizations should apply risk management principles to define and prioritize their greatest business risks associated with data loss. Using tools like Pareto charts, decisions should be made to tackle the top risk producing vulnerabilities immediately while placing the remaining issues on a long and short term roadmaps.

Tom Olzak, MBA, CISSP, MCSE, is President and CEO of Erudio Security, LLC.

He can be reached at tom.olzak@erudiosecurity.com
Check out Tom's book, Just Enough Security
Additional security management resources are available at http://adventuresinsecurity.com
Free security training available at http://adventuresinsecurity.com/SCourses

---

Works Cited

Bowers, T. (2007, February). Getting started with content monitoring. *Network World.* Retrieved October 24, 2007 from http://www.networkworld.com/columnists/2007/020507insider.html

Heck, M. (2007, September). Vontu 7 covers your end point. *InfoWorld.* Retrieved October 24, 2007 from http://www.infoworld.com/article/07/09/20/38TC-vontu-7-covers-your-endpoint_1.html

Heiser, J. (2007, August). Understanding data leakage. *Gartner Research*, research article #G00149979, retrieved October 25, 2007 from http://www.gartner.com.

Mogull, R. (2006, July). Database activity monitoring is a viable stop gap to database encryption for the payment card industry data security standard (and beyond), *Gartner Research*, research article #G00141630, retrieved October 25, 2007 from http://www.gartner.com

Mogull, R. (2006, July (b)). Top five steps to prevent data loss and information leaks, *Gartner Research*, research article #G00141829, retrieved October 25, 2007 from http://www.gartner.com

Proctor, P. E., Mogull, R., & Ouellet, E (2007, April). Magic quadrant for content monitoring and data loss prevention, 2Q07. *Gartner Research*, research article #G00147610, retrieved October 25, 2007 from http://www.gartner.com

Vontu (2007). *Security for a wide open world.* Retrieved October 31, 2007 from http://www.vontu.com/uploadedfiles/global/Vontu_8_Overview.pdf

Vontu (2007 (b)). *Data loss prevention*. Taken from Vontu marketing presentation dated October 31, 2007.

Vontu (2007 (c)). *Vontu network discover*. Retrieved October 31, 2007 from http://www.vontu.com/products/network-discover.asp

Vontu (2007 (d)). *Endpoint data discovery & protection.* Retrieved November 5, 2007 from http://www.vontu.com/products/endpoint-data-discovery-protection.asp

Vontu (2007 (e)). *Vontu network monitor.* Retrieved November 6, 2007 from http://www.vontu.com/products/network-monitor.asp

Vontu (2007 (f)). *Vontu endpoint prevent.* Retrieved November 6, 2007 from http://www.vontu.com/products/endpoint-prevent.asp