Sensitive Data Identification and Protection in the Enterprise

Sensitive data is produced and stored across a large variety of devices on the enterprise network. Financial data, operation plans, R&D projects and software source code, intellectual property, client and partner confidential information, restricted and controlled government information, personnel and personal identifiable information are prime targets of cyber security, counterintelligence and privacy threats.

Identifying what is sensitive in the organization and the associated access context can significantly help security analysts differentiate normal from unusual behavior, which could indicate risk. Knowing ***Which*** data is sensitive, ***Where*** is it stored, ***Who*** access it, ***When*** and ***How*** (flow patterns) is fundamental for implementing a robust enterprise information protection (EIP) posture that achieves risk management security goals[1].

Currently, the EIP posture is primarily based on perimeter control solutions for protecting systems and networks from outside attacks. However, it has been estimated that over 60% of data breaches are a result of insider threats[2] and 85% of those breaches took approximately two weeks until discovering that sensitive data had been compromised.

Content-Aware Data Loss Prevention (CA-DLP) techniques are emerging with a focus on detecting and preventing data breaches. The challenge of CA-DLP is the robust and scalable implementation capable of effectively monitoring large and heterogeneous unstructured data volumes stored on network storage devices.

---

[1] National Institute of Standards and Technology (NIST), Discussion Draft of the Preliminary Cybersecurity Framework, August 28, 2013

[2] http://www.computerlinks.de/FMS/22876.magic_quadrant_for_content_aware_data_loss_prevent.pdf

File monitoring systems used for file activity auditing, generate a large volume of alerts, which are not correlated to the context of the access and the content of the files. Thus a lot of valuable analyst time is consumed exploring non-threat activity.

Bolero can automatically identify sensitive content and can track its flow across the enterprise. Working with file monitoring systems, Bolero can filter out normal activity allowing analysts to focus on threats. The scalable expandable architecture can cope with large data volumes. Bolero mitigates insider threat and improves the cyber situation awareness preparing the organization for a rapid defensive response. Bolero identifies threats in minutes, a significant improvement from the state-of-the-art, which takes weeks.

The Bolero Approach

Bolero protects sensitive files by: Determining file sensitivity; correlating file activity events with file sensitivity; monitoring user access patterns to sensitive files; and detecting suspicious file activity.

Bolero learns which files are sensitive, and filters the reports produced by file activity auditing tools so the security analyst can immediately focus on the abnormal activity that may be indicative of inside threats or an external advanced persistent threat. This reduces the "noise" caused by non-relevant activity and increases the efficiency of detection of threats.

Unlike other Content-Aware Data Loss Prevention systems (CB-DLP), Bolero uses a comprehensive processing pipeline of similarity algorithms that results in high accuracy automatic recognition of sensitive content. The approach substantially increases the quality of detection of sensitive information and ensures that the technique is reliable not only in detecting exact matches but also near duplicates and cases of simple changes to file contents.

Bolero's approach is shown in Figure 1. Bolero scans files on network drives, extracts text from a variety of native file formats, and builds file clusters based on content similarity. Sensitivity tags are propagated across files within a cluster so if a file is clustered with a known sensitive file the file inherits the sensitivity tags of the reference file. Additional tags are added by parsing the file contents matching terms to those listed in dynamically managed lists. Bolero is able to process different file types, such as Microsoft documents, Adobe files, pictures, video files, email files and archives. Encrypted files are recognized and tagged. The system can keep track of all metadata of files such as file sizes, names, creation/modification/access dates. File size can be logged to track changes. Tags can be added automatically to track when was the last time the files were accessed and perform intelligent retention management to remove Redundant, Outdated and Trivial content (ROT).

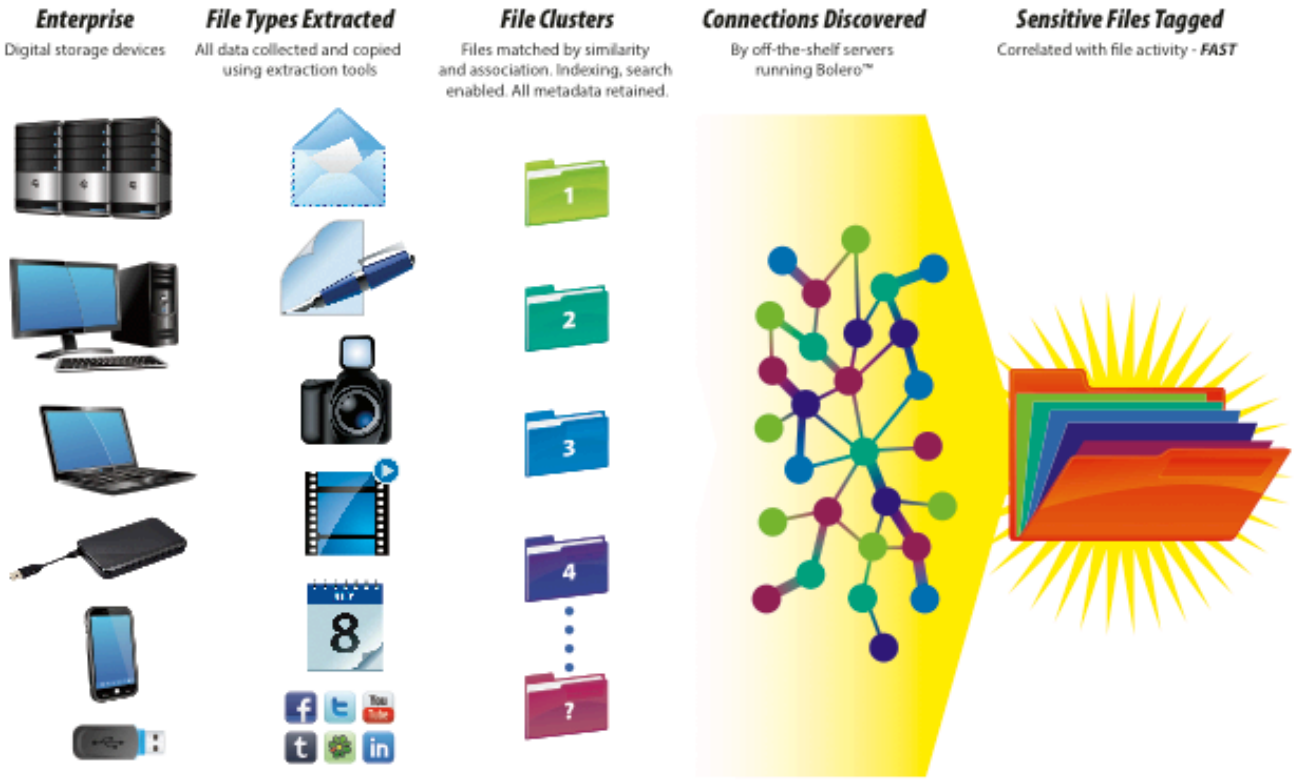

Figure 1. Bolero's file sensitivity engine.

Semandex
CONNECTED INTELLIGENCE.

57 Hamilton Ave., Suite 303, Hopewell, NJ 08525
(Ph) 609-309-5168; (Fax) 609-309-5207
www.semandex.net