IDC TECHNOLOGY SPOTLIGHT

Leveraging Metadata Framework Technology to Take Control of the Information Explosion

September 2010

Adapted from *Worldwide Governance, Risk, and Compliance Infrastructure 2010–2014 Forecast: Increased Regulatory Oversight, Privacy, Cloud Computing, and Smart Cities Drive Emerging GRC Obligations* by Vivian Tero, IDC #222214

Sponsored by Varonis Systems

This Technology Spotlight examines the role of Varonis Systems solutions in addressing governance, risk, and compliance, as well as operational objectives for semistructured and unstructured information.

Managing Governance, Risk, and Compliance Obligations in the Age of the Information Explosion

The explosive growth in digital information, combined with continued budget constraints and expected increases in compliance and risk management obligations, highlights the need for effective information governance, risk, and compliance management in the following areas:

- IDC research on the digital universe finds that by the end of 2010, the digital universe volume will reach 1.2 million petabytes. Total volume will increase by a factor of 44 in 2020. During this period, the number of IT professionals will grow by a factor of 1.4. Organizations will need to do more with less and will increasingly rely on automation to execute information management and governance policies and effectively utilize business-critical information in a timely fashion. There are three major data categories:
  - **Structured data** has data models that define its relationships to other data and is managed by technology that allows for querying and reporting against predetermined data types and understood relationships. Data in application databases belongs to this category.
  - **Unstructured data** (or unstructured information) has no defined, standard structure that would enable convenient storage in automated processing devices. Unstructured data cannot be defined in terms of rows and columns or records, and the data cannot be examined with standard access. Examples of unstructured data are email, spreadsheets, and documents. Some of the most valuable information in a corporation resides in unstructured data.
  - **Metadata** is data about the data. At a basic level, metadata describes how, when, and by whom the data is collected and how it is formatted and stored. Information about who is able to access the data is also another kind of metadata. Technologies today allow for additional custom metadata tags that define the data's business, compliance, legal, and security profile. The aggressive volume growth in unstructured data, which is fueled in part by the combination of an increasingly mobile and distributed workforce and increases in compliance and security requirements, is driving the metadata volume growth.

IDC research finds that unstructured data and metadata are the fastest-growing categories, increasing at an average annual growth rate of 62%. Up to 90% of the digital universe is unstructured data and metadata. Up to 70% of the information is created by individuals, while corporations will be responsible for the security, privacy, reliability, and compliance of 85% of the information.

- Statistics on Varonis customer installations reveal that a single terabyte of data in file systems (NFS/CIFS), SharePoint, and NAS devices averages 50,000 containers (such as folders and SharePoint sites) and 1 million files. The Varonis study also finds that, on average, 5% or 2,500 of these containers have unique permissions. An average organization with 1,000 users has over 1,000 Active Directory groups. Each unique container's access control list has an average of 4 Active Directory groups, with an average of 10 members per group. The functional relationships between users, groups, folders, and files quickly reach the tens of thousands — without taking into account access activity and the many organizational and data structure changes that occur. With the inclusion of these temporal variables, the functional relationships skyrocket.

  Examples of temporal variables include the multiple changes to an employee's role, user access rights and permissions within an organization, records of actual data usage, and retention and disposition policies on multiple versions of a working file. Keeping track of the dependencies across files, folders, users and user groups, access rights, and user and group activities presents management and computational challenges. The scale of the computational challenge puts it into the category of other NP-Complete computational problems. Manual approaches to managing and tracking complex functional relationships and enforcing the appropriate policies are unwieldy and error prone. Tasks such as creating permissions reports, remediating a permissions error, or finding a data owner can be automated (which saves IT security administration time) and properly documented (which enables organizations to mitigate errors and meet audit requirements).

- The percentage of data that is subject to security, compliance, and retention concerns continues to increase. IDC research concludes that in 2008, 22% to 33% of the digital universe was high-value information, also known as data and content that are governed by security, compliance, and preservation obligations. IDC forecasts that high-value information will make up close to 50% of the digital universe by the end of 2020. The majority of this information is unstructured and semistructured content.

Collaboration in the Metadata Era

The widespread use of collaborative content technologies fuels the aggressive growth of unstructured and semistructured information. Unstructured information is a complete information product and a critical organizational asset.

Collaboration produces highly valuable information, but it also introduces significant risk due to the increasingly complex and dynamic access control requirements. Organizations can manage these risks by understanding the security, privacy, and compliance profile of this information and its containers (folders, SharePoint sites, etc.); evaluating the current approach for tracking relationships between the information, the authorized users and user groups, and authorized activities within and across partners; and using the risk profile and analytics to facilitate secure collaboration and data sharing.

As business transactions move online, sensitive information about customers, products, employees, company financials, and intellectual property is increasingly shared and housed across business partners and third-party service providers. These third parties have a vested interest in the security and compliant information management of this data. High-profile data breaches in the past three years demonstrate that organizations that fail to protect sensitive data will incur serious regulatory and legal liabilities, revenue and market share declines, as well as, in many cases, an unrecoverable loss of confidence by the public. Digital integrity is a critical business differentiator for any organization.

Visibility, Actionable Intelligence, and Automation Are Critical to Managing the Explosion of Unstructured and Semistructured Content in Distributed Systems

Unstructured and semistructured data in distributed systems presents challenges for comprehensive data governance, including protection and management, data security, data classification, data migrations and consolidations, compliance management, and worker productivity objectives. These challenges include:

- **IT budgets.** IT budget constraints underscore the need for automation of analysis and data management protocols. IDC's digital universe research notes that corporate IT budgets are, on average, growing at less than one-fifth the forecast annual growth rates of digital information. At the same time, manual approaches to managing and protecting the information become unwieldy, error prone, and ineffective. Organizations have an urgent need to address the costs, time, and service disruptions associated with manually verifying data entitlements and remediating compliance violations. Relying on manual approaches prevents the majority of IT organizations from reliably or efficiently answering critical questions about their data, such as the following:
  - Who has access to a data set?
  - Who should have access to a data set?
  - Who has been accessing the data?
  - Which data is sensitive?
  - Where is my sensitive data overexposed, and how do I fix it?
  - Which folders need an owner?
  - Who is the likely data owner?
  - Who has unnecessary permissions to each data set?
  - What data is unused?
- **Data access and classification.** There is a need for a programmatic approach to determining the business relevance, security, and risk profile of the data that does not disrupt existing business processes. The ever-increasing volume of information and its associated functional dependencies present challenges in the identification of exposed sensitive content, remediation of excessive permissions, and proper alignment of users and groups with data. These challenges also highlight the need for more solid analysis of user activity in order to automate entitlement reviews and authorization processes, detect active and orphaned data, identify inappropriate data, find lost files, and facilitate forensic investigations. The IT function needs the same kinds of visibility, auditing, actionable intelligence, and automation to manage data that exists for other corporate assets. IT needs the same automated decision-making assistance available in other realms, like search engines for the Internet, credit card fraud identification, and shopping recommendations on online Web sites. Here, automation will be used to enable the organization to determine who should and shouldn't have access to data. Too often, users have access to significant amounts of data that isn't relevant to them. The WikiLeaks story is a top-of-mind example of the disastrous results when this happens. It also underscores the need to track how and why specific materials are classified, how access to these sensitive materials is controlled, and how these documents are handled by authorized users.
- **Data ownership.** There is a need for a programmatic approach for determining data ownership for active and dormant data, as well as ensuring ongoing access rights management. IT needs automated analysis of the permissions structure to determine which containers require ownership, and analysis of actual access activity, to identify likely data owners. According to the

Ponemon Institute, approximately 90% of organizations have no process to identify the owner of file data containers and 76% are unable to determine which individuals and roles are authorized to access the data. IDC studies conclude that up to 60% of accounts on most systems are expired. Organizations therefore have a need to ensure that users and roles are aligned with the correct groups and that the groups enable access to the appropriate data containers.

- **Data loss.** It is necessary to mitigate the risks of data leakage due to excessive access rights and permissions. IDC research notes that organizations average 14.4 unintentional data losses in 12 months, where 52% are considered unintentional data losses through employee negligence. Excessive and/or out-of-date privilege and access rights are considered as having the most financial impact on organizations. A fragile economic recovery is compelling most organizations to use contractors and temporary employees, increasing the risks of data loss. When faced with changing business objectives, mergers, consolidations, and divestitures, organizations have a pressing need to ensure that controls are in place to mitigate the risks of data leakage, theft, loss, and integrity arising from excessive access rights and permissions and nonexistent audit trails.

- **Stale data.** There is a need for a solid and legally defensible approach to cull stale data and execute multiple IT initiatives. Stale, unused data burdens the IT infrastructure and adversely impacts backup windows and overall storage operational efficiency objectives. In many instances, inactive and orphaned folders can account for as much as 70% to 85% of the data in distributed systems. Keeping and moving stale data onto tape without understanding its value also increases the organization's risk profile and future legal liabilities. Also, IT consolidation and migration efforts could become prohibitively expensive and inefficient when an organization fails to consistently enforce compliance, data management, and security policies during the platform and content migration process. In addition, a messy migration and consolidation project exposes the organization to future security, compliance, legal, and regulatory burdens. A solid metadata framework defines the data management, storage, security, compliance, and business attributes of unstructured and semistructured data. Organizations can employ this metadata information to support critical IT initiatives such as entitlement reviews, compliance audits, data ownership identification, records management, domain and data migrations, consolidations, archiving, and retention projects. Metadata information improves transparency, thus alleviating the inherent conflict between the IT and legal compliance functions that typically occur during server/storage consolidation projects.

- **Cloud initiatives.** There is a need to address security and compliance challenges associated with an organization's cloud initiatives. IDC research also finds that security and compliance are among the top 3 challenges to cloud computing. Without adequate information on the security and compliance profile of the data, including its ownership, access controls, audits, and classification, cloud initiatives are amorphous and imprecise. Understanding the data owners and the authorized users and user activity is critical to garnering organizational input, which, in turn, is critical to defining the security and compliance profile of the data. For example, CFOs and CIOs are hesitant to move critical data and processes into the cloud when there is very little visibility into access and ownership, traceability, and data segregation. Organizations must have data governance in order to provide secure collaboration and data protection for their customers, partners, and employees. Without it, it will be virtually impossible to conduct digital business.

- **IT operational efficiencies.** When asked how long various manual data management and protection activities take, IT staff members commonly report that a single permissions report or access request can take 30 minutes of a system administrator's time, safely remediating a permissions error or finding a data owner can take hours, and discovering who moved or deleted files is often impossible. These are some of the many data management and protection tasks that IT must perform every day.

Metadata provides IT with critical information on the attributes that define the security, compliance, and business profile and functional relationships of unstructured and semistructured content. Without this critical metadata, IT cannot effectively manage and protect data.

Benefits: Leveraging Metadata Framework Technology to Address Risk and Compliance in Distributed, Virtualized, and Cloud Services Environments

A metadata framework provides visibility, auditing, and actionable intelligence to better manage and protect data.

Security, risk management, compliance, and operational objectives are inexorably intertwined. Data stores can no longer be treated as independent silos. They require a common set of controls and procedures for their protection and management. Organizations have to take steps to align the relevant controls and mechanisms in order to avoid compliance conflicts and cost inefficiencies arising from siloed projects. Maintaining and tracking these interdependent physical relationships such as replication and archiving, as well as the logical relationships between users, groups, containers, and views, is becoming exponentially more complex.

Organizational mergers and divestitures and the introduction of new technologies and nontraditional devices into corporate networks exacerbate this challenge. Take, for example, virtualization and cloud services: Employing these technologies adds layers of abstraction across the infrastructure and IT infrastructure stack. The on-demand self-service nature of cloud services makes it relatively easy to circumvent traditional IT governance processes for the provisioning of IT assets. In this scenario, having real-time abilities to compare access rights against policies and detect and address orphaned data and accounts is paramount.

As unstructured content is propagated across an organization's shared storage, virtualized networks, and applications (e.g., SharePoint, Exchange, NFS global namespaces), a metadata framework functions as the string that tracks the dependencies across the information infrastructure and the underlying IT infrastructure. Sound information governance, risk, and compliance management practices would take advantage of metadata technology to ensure that information is managed consistently across on-premise, virtualized, and cloud-based environments.

Information workers and business partners utilize unstructured and semistructured content in numerous platforms at the center of the IT infrastructure. While these business processes are taking place, the metadata framework tracks and analyzes the myriad functional relationships and dependencies across the data infrastructure, including access rights, data usage, and file content. To do this effectively, organizations need automation to enumerate and analyze the relationships across data storage platforms, the nature and value of the content (with respect to compliant retention, security, privacy/confidentiality, and legal considerations), and the users' and data owners' access activities and their ongoing entitlements.

In addition, organizations have to consider having the ability to quickly remediate potential exposures and policy violations. Creating a full, effective index of advanced metadata structures by capturing a broad range of metadata fields across the entire corporate data repositories and directory services will yield consistent, measurable results; systematically reduce risk; and perfect data governance practices.

Metadata is the fastest-growing category in the digital universe. Organizations need to be smart about how and which metadata to capture, or they could end up taxing computing, database, network bandwidth, and storage resources. For unstructured and semistructured content in distributed endpoints, creating a full content index and capturing a broad range of metadata fields *across the entire corporate network may be overkill.* Organizations need smarter and more practical approaches to tackling risk identification of exposed sensitive content. To enable content, identity and owner, and

physical and logical location awareness, an organization should consider tracking relationships across four kinds of metadata:

- Information about owners' and users' group memberships (as provided by Active Directory and LDAP)
- User access activity
- Unstructured and semistructured permissions/entitlements
- Classification results/tags indicating the value of the content (such as privacy, confidentiality, records compliance, legal preservation requirements)

Considering Varonis Metadata Framework Technology

The Varonis Metadata Framework is the foundation for all Varonis products within the Varonis Data Governance Suite. It creates and manages a metadata layer that enables IT and the business to work together to protect data and keep pace with its growth through automation. The technology nonintrusively collects critical metadata about unstructured and semistructured data and generates metadata where existing metadata is lacking (for example, the file system filters and content inspection technologies). The metadata is also preprocessed, normalized, analyzed, stored, and presented to the IT administrators in an interactive, dynamic interface — all of these functions are accomplished without impacting server performance. Four types of metadata are collected, synthesized, processed, and presented: permissions information, user and group information, access activity, and sensitive content indicators.

Key features of the solution include the following:

- Bidirectional permissions visibility. The application exposes the users and groups that have access to a folder, site, or mailbox, and their associated permissions, as well as the folders to which a user or group has access. The application also exposes the folders, sites, and mailboxes to which users or groups have access, their level of access, as well as an explanation of how they are accessing said folders. The application supports Windows and Unix/Linux file servers, NAS devices, Microsoft SharePoint, and Microsoft Exchange.
- Reports on exposed sensitive data and prioritized list of folders with excessive permissions, as well as highly utilized, dense, and/or sizable file populations *and* concentrations of sensitive content.
- Reports and an audit trail on events and activities (found in the log area of DatAdvantage) allow an IT administrator to view every open, create, delete, and move event that any individual generates on the file system and every email sent, received, and opened. The analytics allow the correlation of file events and activities with the classification tags.
- Analysis and recommendations on restricting access to sensitive and overexposed content without affecting normal business activity by combining the permissions data, the access events, and sophisticated data analysis.
- Simulates and automates the change and cleanup of permissions and entitlements. Varonis allows organizations to simulate changes outside of production and, by reviewing past access activity, measure what the impact would have been had a change been made earlier.
- Activity reports on user access behavior, including most/least active users and most/least active directories. Built-in reports list inactive directories to facilitate disk cleanup projects.
- Monitors and alerts for anomalous user behavior by looking for statistically significant deviations on each user's normal day-to-day activity. Access deviations may signify a worm or an automated process running under a user's credentials or an employee's potential departure from an organization.

- Analysis and reports to correlate files and folders with data owners and enable automated entitlement reviews.
- Workflows to get data owners involved in the authorization process and to identify, define, and enforce "ethical walls" within an organization.

IT administrators use these reports to identify and communicate with data owners. Once data owners are identified, they are empowered to make informed authorization and permissions maintenance decisions that are then programmatically executed — with no IT overhead or manual back-end processes.

With these metadata streams collected, synthesized, processed, and presented intelligently by the Varonis framework, organizations are now able to quickly fix their data governance issues with minimal IT involvement and without breaking business processes – enabling them to easily answer the following critical questions:

- Who has access to a data set?
- Who should have access to a data set?
- Who has been accessing the data?
- What other data have they been accessing?
- Who is the likely data owner?
- Who has unnecessary permissions to each data set?
- Which data is sensitive?
- Where is my sensitive data overexposed, and how do I fix it?
- What data is unused?

Challenges

Varonis does face market challenges, however. IDC has identified the following:

- Budget constraints mean that multiple IT projects are competing for limited dollars. As a result, Varonis works closely with multiple functional stakeholders within each prospect to educate and build the business case and solicit buy-in. This is easier said than done for a relatively small and lean organization.
- The Varonis Metadata Framework is a relatively novel approach. The target audience may need further education on the differences and benefits of the Varonis Metadata Framework relative to solutions offered by access governance and search and text analytics applications. In many instances, organizations may have existing investments in these tools. Varonis' field sales organizations will have to do a solid job of highlighting the limitations of the competitive products and articulating the benefits of its comprehensive and normalized analytics and workflows.

Customers looking to better manage compliance, collaboration, and security objectives in their distributed systems will need to be smart about selecting the appropriate solution and consider the following:

- Depending on an organization's business and risk profile, identity and access management (IAM) solutions can be costly and require extended periods of time for deployment. Identity and access management solutions are group centric. Properly aligning groups with data is a prerequisite to the success of any IAM initiative and requires the metadata streams listed earlier (users and groups, permissions, content and activity).

- Full content indexing requires considerable network bandwidth, storage, and computational cycles. True incremental scanning is not possible without actual usage metadata, nor is it possible to effectively use classification results for data governance without the other metadata streams — permissions, user and group relationships, and actual data usage.

Conclusion

IDC believes that the combination of increasingly stringent global security and compliance regimes, continued budget constraints, and explosive growth of unstructured and semistructured information in distributed storage devices and applications, as well as remote offices, will compel organizations to take advantage of the metadata layer to manage their business, operational, security, and compliance objectives.

Organizations need to capture the metadata required for effective governance nonintrusively — so as not to introduce service disruptions. Further, they need to effectively normalize the data so that analysis, computation, and storage are possible while presenting meaningful, actionable information for decisive action and review. The Varonis Data Governance Suite utilizes the Metadata Framework to facilitate this process.
