EGNYTE

Understanding Data Governance:

What It Is, Why It Matters, and How To Get Started

Table of Contents

01

What is Data Governance?

Data governance is a commonly used and misunderstood term. It is often confused with concepts like data cataloging, master data management, regulatory compliance, and data security, which are parts of the greater whole.

At its most basic, data governance is a set of formal processes that ensure that data within the enterprise meets expectations. Most companies have embarked on one or more data governance initiatives – sometimes without knowing it. These can range in complexity, from simple written policies outlining who can access sensitive files to robust technology suites that monitor every data point inside a corporate network.

02

Why Data Governance **Matters**

While data governance, as a concept, has been around for a long time, interest in the topic is reaching an all time high for several reasons. Specifically:

Digital Business Is The Norm

Today's workforce is flexible and mobile, untethered to a single location.

More remote workers and hybrid workplaces means data spread across more devices, clouds, and apps than ever before. This decentralization has pushed the traditional data security stack, which was built for the in-office era, to the breaking point.

- **62% of remote workers** confess to using rogue applications for work that their IT department doesn't know about. (NetMotion)
- **28% of employees** are using unsecured home networks to access corporate documents
- **Half of IT leaders** say remote work has made their business less secure.

Expanding Global Privacy Regulations

With the rise of digital business came an awareness of the need to protect individual privacy online. A slate of high-profile data breaches, ambiguity around who owns personal data, and new regulations on industry have prompted a new set of privacy regulations including GDPR, CCPA, NYDFS and more. In the past, industries were either "regulated" or "unregulated" but the rise of privacy laws has changed that. According to the United Nations Conference on Trade and Development, more than 128 out of 194 countries have data and privacy legislation that regulates the use and storage of data, each with their own terms and requirements. Understanding and maintaining compliance with various laws and regulations has become a herculean task for most businesses today.

The Rise of AI Technology

Harvard Business Review reported that 60 percent of companies see their future success dependent on successfully implementing AI, and another 36 percent said their organizations were in some stage of deploying machine-learning technology. AI depends on large amount of clean data, readily available.

Expanding Cyber Threats

Historically, the emphasis has been on perimeter security, by means of firewalls and infrastructure hardening, and device security, with a focus on locking down individual endpoints like laptops and mobile devices. As more workloads shifted to the cloud, this expanded to include network security and application security. Despite decades spent building security infrastructures that protects endpoints, firewalls, networks, and applications, in most companies, data itself is still largely unprotected. Data security tools like DRM and DLP come closer to protecting data wherever it moves, but are financially and administratively out of reach for all but the largest organizations, and have significant limitations for mobile workforces.

For all these reasons and more, robust data governance is no longer a "nice to have" – it's a "must have". While the old pillars of data governance like data quality, data catalog, data security and data life cycle management haven't changed, their meaning is rapidly evolving to address the heightened requirements of day.

Organizations large and small are reinventing data governance in a way that is:

- Sensitive to a diverse set of risks, including internal and external attacks, compliance, loss of data;
- More scalable than ever, leveraging AI and automation to minimize human error and maximize worker productivity, identify sensitive data and remediate issues dynamically;
- Built for the world of distributed work, which is hybrid and multi-cloud; and,
- Able to protect data, both in transit and at rest, working in tandem with the authorization and authentication layers of the security stack, so protections can follow wherever and however data flows throughout the organization.

03

Getting Started With **Data Governance**

The first step in any data governance program, after defining its goals, is to clearly identify the nature of the data to be governed. Data can be vary in terms of volume, velocity, source type and location. At the highest level, there are two types of structures that must be considered.

Structured Data

Structured data conforms to a tabular format with relationships between the different rows and columns. Common examples of structured data are Excel files or SQL databases. These datasets reside in structured rows and columns that can be sorted. Structured data is critical to protect as this often includes personally identifiable information (PII) as well as credit card information and accounts. It's a treasure trove for today's cybercriminals looking to resell stolen identities or credit cards on the black market.

Content

The second type is unstructured data, better known as content. Content is the human-readable form of data that lives in files of various kinds. This type accounts for the vast majority of corporate data created (>80%), accessed and shared by end-users. As opposed to structured and semi-structured data, this data may pose a more significant risk, because the systems where it is stored (e.g., public email folders, collaboration systems, cloud storage services) are not well governed. And further, the content of the data is not easily identified.

For most organizations, the best place to get started with data governance is with content governance – because content is typically the more exposed of the two data types.

To demonstrate the unique risks associated with enterprise content, consider the most common use case: the corporate file share. In contrast to "closed" systems (e.g. payroll) and structured databases with fewer authorized users and higher baseline levels of security (e.g. multi-factor authentication) the relative risk exposure in an ungoverned file share can be considerably larger. Take the example below: even when equalizing for security and access controls like SSO, the relative impact of an ungoverned file share is significantly higher:

Despite the relatively high risk exposure, content governance is often overlooked or deprioritized in favor of endpoint, network, or application security. While all of these are important elements of a comprehensive security strategy, there is a strong argument for reprioritizing content governance as a security starting point, not an afterthought:

1. Unstructured content is typically a company's largest overall data source, and by extension, the largest source of sensitive and regulated data.
2. Content-rich file-shares, email, and productivity apps are accessed by more end-users than any other IT system.
3. Content is the only kind of corporate data that is routinely shared with third-parties with little oversight.

04

Top Signs You Need a **Content Governance Strategy**

Most organizations are already applying some form of governance to their information systems, be it for records management, GDPR compliance, or simply training employees on how and when they are authorized to share certain data. Determining when and how to mature a data governance program is different for every organization, but there is always a tipping point. **Below are the most common red flags that drive companies to pursue data governance projects:**

1. Manual Audit and Compliance

In order to comply with legal and contractual requirements, internal teams are spending time manually tracking down where data originated, where it is located, who has access to it, how it is being used, and when and how it should be decommissioned (archived or deleted). In structured environments, this can be straightforward – especially in well-organized systems with fewer authorized users. When data is spread across cloud apps and devices, it can be much more difficult. If manual DSAR response, usage auditing or breach reporting are diverting security and compliance teams away from more proactive pursuits, content governance may be needed to reduce manual workload and improve the quality of results.

2. Perpetually Dark Data

Because sensitive data is typically embedded internal files (e.g. contracts and invoices) it can be much harder to find than in structured databases. Despite corporate and regulatory requirements to locate and protect personally identifiable information (PII) this data is often difficult to find, spread across many apps and devices, or reliant on manual processes (like end-user tagging for sensitive data) to track and protect. If more than 10% of your data is consistently unclassified, user-driven security practices will not offer sufficient protection for most kinds of businesses, and automated classification should be a strong consideration.

3. Widespread Shadow IT

Unsanctioned applications, shadow apps that are usually SaaS, pose a significant information security threat. Shadow apps include all applications that are not part of the application suites approved by an organization's IT group. These applications are dangerous, because they are not monitored or secured by the organization, which increases their vulnerability to internal and external threats. In the event of employee departure, it can be impossible to deprovision these accounts or revoke access to the data shared there. Unsanctioned cloud storage applications and sending files to personal email are very common workarounds for employees looking to escape legacy IT security systems that weren't built for the remote era.

4. Uncheck Content Sprawl

You may identify with the nearly half of IT leaders (49%) say the needs of different departments lead to content sprawl and weaken overall IT security. Remote work has driven exponential adoption of tools like Microsoft Teams, which are natural sprawl creators. Organizations have difficulty implementing and enforcing consistent data access permissions and rules for handling sensitive content when it spans multiple data sources. This includes formal repositories (file-shares, cloud storage) and informal repositories (apps that create and store files).

5. Loose Permissions and Access Controls

Failure to control access is a leading cause of data breaches as compromised user access can expose sensitive data. Close to half of all IT leaders report that broken permissions models lead to other significant issues:

- **46% of employees** have access to files they shouldn't
- **40% of employees** don't have access to files they should

Permissions in unstructured data environments are more complex than most structured databases, which usually follow a one-person, one login scheme. Alternatively, nearly anyone in a company can create or share enterprise content with anyone else, and it is up to the user to establish permissions on a fle. It is often widely shared outside the boundaries of the company, with customers, suppliers, and partners, many of whom need at least partial access to the repository. Because of this, permissions hygiene is often the first thing to slip in cloud environments.

6. No Rules for Data Archive or Disposal

Most data privacy regulations like GDPR and CCPA contain specific provisions about how to dispose of an individual's personal data, and how long it can and should be kept prior to deletion. Sales and services contracts often contain similar language, requiring the archival or retention of data for certain periods of time. In the event that retention periods overlap, a strong content governance program must be able to detect, prioritize, and solve for discrepancies. Most companies that attempt this at all will spend significant time and energy developing protocols and training employees on disposal practices, and even more manual time enforcing these rules. Due to the high volume of enterprise files and email, content lifecycle automation is most often necessary to execute archiving and data disposal at-scale.

In an ideal world, businesses should start with sound governance practices to avoid ever encountering one of the scenarios above. Unwinding years of lacking governance can be daunting, so whenever possible, it helps to lay a solid foundation of tools, technologies, people and processes that encourage **smart data management, visibility and control** from day one.

05

How To Assess Your Current State

A healthy governance program should have both policies and technologies that combine to address the following needs:

Discover

- What high-value or high-risk data exists on my system?
- What regulated data do I own?
- What personally identifiable information am I obligated to protect?

Define

- How is data access granted or revoked?
- Who has access to sensitive data types?
- Who are my organization's data stewards?

Alert

- What happens to data when it leaves my environment?
- Can I identify attacks in progress?
- What is normal or abnormal user behavior and can I spot the difference?

Remediate

- Are there mechanisms to identify sensitive and regulated data?
- Can I detect and prevent attacks on my data stores?
- Can I trust my data users?

Report

- Can I be sure this data is current and accurate?
- Can I be sure this data hasn't been altered or manipulated?
- If compromised, can we meet breach reporting deadlines?
- Is there a repeatable process for running a DSAR?

Optimize

- Can my systems distinguish between different data types and owners?
- Can I use data-use patterns to find efficiencies and cost savings?
- Can I identify and remove duplicate data?
- Am I able to adhere to data retention requirements under GDPR?

Retire

- How old are my oldest files?
- Is there data that can potentially be archived or eliminated?
- Am I complying with data residency laws?

Companies that can effectively address these questions through processes that don't slow down users, overburden IT, or rely on outdated, manual methods can count themselves among the lucky few. Those who can't should begin looking for opportunities to strengthen their governance programs to better protect their data and their business.

06

Conclusion: Adopt a **S.M.A.R.T.** Approach to Data Governance

Whether updating existing content governance systems or implementing new ones, the SMART Data Governance Framework can help get organizations on the right track:

Specific	Start by triaging the most exposed data type – typically, that's content
Motivating	Be able to track progress. Set goals. Connect goals to the core mission/purpose of the organization – not just risk avoidance.
Achievable	Be achievable with no incremental in headcount or skills and without dropping other competing responsibilities
Relevant	Get everyone in the organization involved, by making data governance relevant to them
Time/cost bound	Do it fast – see value quickly – fund by cost savings/reallocations

S.M.A.R.T. Data Governance Framework

Most importantly, this process needs to be done with full recognition that governance is not a one-time project, but rather a cultural skill. Data governance needs to be built into an organizations' culture and technical backbone. This means empowering business leaders and educating end users to be effective data guardians. Complement a culture of governance with systems that support the way organizations run and people work.

07

Data Governance Glossary

Access Management: Policies and procedures that define, track, and control the data an individual can access in systems or applications.

Anonymize: Deidentify data by stripping Personally Identifiable Information (PII) from it.

Audit Trail: An electronic log used to track computer activity—at a system or individual level.

Authentication: The process of verifying the identity of a user or process when accessing a computing system.

Authorized Access: Also known as Permissible Access, allowances made for internal and external users to view and process PII on a need-to-know basis.

Big Data: Extremely large amounts of data processed with powerful systems and analytics tools to find trends and patterns that lead to insights.

Business Intelligence: Insights, resulting from data analysis, that are delivered as reports, dashboards, and visualizations (e.g., charts, graphs).

Chief Information Security Officer: The executive-level manager who directs strategy, operations, and the budget for the protection of the enterprise information assets and manages that program.

Classification: The process of labeling and sorting data assets based on predefined criteria, such as sensitivity level or data owner.

Cloud-Access Security Broker: An on-premises or cloud-based security policy enforcement point that is placed between cloud service consumers and cloud service providers to combine and interject enterprise security policies as cloud-based resources are accessed.

Compliance: The practice(s) of ensuring that sensitive data types are organized and secured in such a way as to enable organizations to meet legal and governmental regulations. Examples of common data privacy laws include GDPR, HIPAA, CCPA and FDA regulations.

Content Governance: A system of tools, policies, people, and processes defining who within an organization has authority and control over unstructured, human-readable files, commonly known as enterprise content. Common examples include documents, PDFs, email, and images.

Cloud Content Governance: A class of technology that uses cloud-first architecture and machine learning to analyze large volumes of unstructured, human-readable files and automatically apply protections.

Content Management: Processes and technology used to collect, deliver, retrieve, and manage data in a variety of formats. It is also used for data governance.

CRUD: An acronym for the four basic functions of persistent storage—Create, Read, Update, and Delete—that are interfaces to databases to allow users to create, view, modify, and alter data.

Cyber Hygiene: the practices and steps that users of computers and other devices take to maintain system health and improve online security. These practices are often part of a routine to ensure the safety of identity and other details that could be stolen or corrupted.

Data Analytics: Processes and algorithms used to examine raw data and extract meaning. Data analysis systems transform, organize, and model data to draw conclusions and identify patterns.

Data Architecture: A framework of rules, policies, standards, and models that govern what data is collected then how it is used, stored, managed, and integrated across an organization.

Data Breach: An incident that involves the intentional or unintentional viewing, access, retrieval, or removal of data by an individual, application, or service.

Data Classification: The organization of data based on its level of sensitivity and the impact should that data be used, shared, altered, or destroyed without authorization.

Data Discovery: The process of detecting and organizing data by identifying key characteristics and applying a distinct class to make it easier to locate, track, and retrieve. Once undergoing discovery, data is then tagged, often with the specification of its access restrictions.

Data Custodian: An administrator responsible for the appropriate storage, transportation, and access of data as well as the technical environment and database structure.

Data Flow: The path that data follows through a system—from source to final instantiation (e.g., report, database).

Data Governance: A system of tools, policies, people, and processes for defining who within an organization has authority and control over data assets and how those data assets may be used and shared.

Data Hygiene: the collective processes conducted to ensure the cleanliness of data. Data is considered clean if it is relatively error-free. Dirty data can be caused by a number of factors including duplicate records, incomplete or outdated data, and the improper parsing of record fields from disparate systems. Poor data hygiene can lead to improper classification that causes increased risk exposure.

Data Integrity: The completeness, validity, reliability, accuracy, and consistency of data.

Data Loss Prevention (DLP): a strategy for making sure that end users do not send sensitive or critical information outside the corporate network. The term is also used to describe software products that help network administrators control what data end users can share. It's most effective for preventing data loss, either purposeful or accidental.

Data Minimization: the practice of limiting the collection of personal information to that which is directly relevant and necessary to accomplish a specified purpose.

Data Residency: The physical or geographic location of an organization's data or information. Some privacy regulations, such as GDPR, require that certain kinds of data physically reside in the same geographic location as the individual they reference.

Data Ownership: Assignment of formal accountability and legal ownership of data—a single piece or set of data. This comes with a list of owner rights and responsibilities.

Data Privacy: Defines whether or how data is shared, with whom data is shared, and how data is legally collected or stored.

Data Security: Measures to protect data, residing in systems or applications, from unauthorized access, corruption, or theft.

Data Silos: a collection of information isolated from — and not accessible to — other parts of the organization due to incompatible systems or permissions. Silos restrict visibility of data and content across the organization.

Data Steward: Role within an organization focused on high-level policies and procedures for the monitoring, security, and management of data use according to data governance rules related to access, accuracy, classification, and privacy.

Data Stewardship: Tactical coordination, implementation, and enforcement of data governance policies and procedures across an organizations' data stakeholders.

Data-Centric Audit and Protection (DCAP): an approach to information protection that combines extensive data security and audit functionality with simplified discovery, classification, granular policy controls, user and role based access, and real-time data and user activity monitoring to help automate data security and regulatory compliance.

Database: An organized collection of structured data that can easily be accessed, managed, and updated.

De-Identification: The process of removing or obscuring personal data in a document or record.

Encryption: The process of converting information or data into a code, especially to prevent unauthorized access.

Enterprise Password Management: practices and software that use security controls to prevent internal and external threats from capturing master passwords, credentials, secrets, tokens, and keys to gain access to confidential systems and data. These centralized password management systems can be on-premises or in the cloud. Most important is that they provide password security for all types of privileged accounts throughout your enterprise.

File Versioning: the digital practice of storing more than one version of a file simultaneously with the goal is to provide access to previous iterations of important documents for a number of potential scenarios, including mitigating ransomware attacks.

Identity and access management (IAM): is a framework of policies and technologies for ensuring that the proper people in an enterprise have the appropriate access to technology resources.

Indirect Identifiers: Information that can be combined with other information to identify individuals, such as date of birth, race, education, occupation, marital status, and zip code.

Information: The result of processing to data to provide context and meaning.

Insider Threat or Insider Data Breach: an incident where sensitive, protected, or confidential personal information and personal data has potentially been accessed, stolen, or used without authorization due to negligence or malice by an employee or contractor.

Masking: A data security technique in which sensitive data is obfuscated for testing or training purposes.

Metadata: a set of data that describes and gives information about other data.

Nonpublic Personal Information (NPI): Personal data that is already widely available, such as data obtained through Internet collection devices or cookies.

Personally Identifiable Information (PII): information that can directly identify an individual when used alone or with other relevant data. PII includes name, address, social security number or other identifying number or code, telephone number, and email address.

Policy: A rule or set of rules that outlines how companies and their employees are intended to interact with corporate data.

Pseudonymization: a data management and de-identification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers. The GDPR defines as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information."

Ransomware: a type of malware that threatens to publish the victim's data or perpetually block access to it unless a ransom is paid to regain control of the content.

Risk Management: The identification, analysis, assessment, control, and avoidance of risk through precautionary steps that reduce or eliminate threats.

Sensitive Data: Data that is classified as information that requires elevated protection and tightly managed access.

Shadow IT: the use of computer or network hardware or software by a department or individual without the knowledge of the IT or security group within the organization.

Structured Data: Data that resides in a fixed field within a file or record. It is easily organized and searchable. A SQL database containing customer records is a common example of structured data.

System of Record: A storage system that is the authoritative data source for a given data element or piece of information.

Tag: A label attached to a data asset for the purpose of identifying, grouping, or providing context.

Threat Detection: the practice of analyzing the entirety of an information-security ecosystem to identify any malicious activity that could compromise the network.

Unstructured Data: Information, in many different forms, that doesn't fall into conventional data models and thus typically isn't a good fit for a mainstream relational database. Most enterprise content is unstructured data, including email, documents, spreadsheets, images, and PDFs.

Zero-Day Detection: The deployment of behavior or activity-based AI to detect suspicious actions indicative of an attack in near-real time.

Zero-Day Vulnerability: An exploit for an unknown vulnerability previously in a software program or operating system. ■

Contact Us

+1-650-968-4018

1350 W. Middlefield Rd.
Mountain View, CA 94043, USA

www.egnyte.com