CHECKLIST 2024

Five Pillars for the Comprehensive Governance of Data and Other Modern Assets

By Fern Halper, Ph.D.

snowflake®

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Five Pillars for the Comprehensive Governance of Data and Other Modern Assets

By Fern Halper, Ph.D.

In today's rapidly changing digital environment, data governance has emerged as a cornerstone for organizations seeking to manage their data more effectively. TDWI surveys consistently reveal that data governance is a top priority for organizations, underscoring its critical role in successful data management and analytics.

There are numerous reasons why data governance has become so critical. One prominent trend is the increasing complexity of data environments. Organizations are now grappling with numerous data types, ranging from traditional structured data to newer data types such as text data, machine data, image data, and audio data. TDWI research indicates that text data and machine data are already entering the mainstream phase of adoption. The use of text data is further compounded by the recent popularity of generative AI and large language models, which feed on and can generate vast amounts of text data.

Five important best practices for governing modern data- and AI-related assets:

1. Ensure integrity, trust, and transparency in data and other assets
2. Prioritize security and privacy
3. Focus on compliance and auditability
4. Balance discovery and accessibility with governance
5. Uphold ethical standards and data protection

Additionally, the shift towards hybrid and multicloud data environments presents unique challenges. As data storage and processing extend beyond on-premises infrastructures to include various cloud platforms, the need for robust governance frameworks that can operate seamlessly across these diverse ecosystems has become increasingly important.

The democratization of data and analytics also plays a significant role. With more users from different backgrounds—including business users, data analysts, data scientists, and developers—accessing, utilizing, and collaborating with this data for various purposes, there is an increased risk of data misuse and governance breaches. This environment necessitates a balance between accessibility and controlled governance to ensure data integrity and compliance.

Of course, the move to more advanced analytics is also driving the need to move beyond data governance to add support for the governance of AI/ML assets and applications. AI/ML models need to be governed to ensure that the input to and the output from these models is trustworthy.

AI/ML models also are core to many applications being put into production. These include recommendation engines, chatbots, AI assistants, robotics, self-driving cars, and other smart devices. The development of these applications will also need to be governed. AI/ML applications need to be versioned, documented, and certified.

Applications in production also need to be governed to ensure that the AI models that are a critical part of these applications don't go stale and that the applications meet compliance and other objectives.

In this complex and dynamic environment, traditional data governance objectives (e.g., ensuring data accuracy, consistency, and compliance), remain foundational. However, they now must be integrated with newer concerns, including the management of diverse and voluminous data, ethical data use, the governance of AI/ML assets and applications—all of which will involve technology considerations. This TDWI Checklist Report examines five important pillars and best practices for the comprehensive governance of modern data- and AI-related assets.

1 Ensure integrity, trust, and transparency in data and other assets

For organizations to make the best use of their data for management, analytics, AI/ML, and applications, they will need to trust the integrity of that data. Data integrity ensures that data is accurate, complete, reliable, timely, and reasonable. Trusted data needs to be transparent. That means that someone is accountable and responsible for it and the organization knows how it is used. Trusted data comes from trusted sources and is used as intended. These are core pillars for modern data governance.

As organizations move to modern platforms such as cloud data lakes and warehouses and collect new data types such as text data or machine-generated data, trust remains essential. Yet, only about half of respondents to a recent TDWI survey believe that their data is trustworthy and of good quality.[1] The rest are either on the fence or disagree that their data is trustworthy and of high quality.

These principles of integrity, trust, and transparency are foundational to any governance program that involves data, AI/ML assets, and applications. Organizations must ensure that people are accountable for ensuring data can be trusted, and

[1] Unpublished 2024 TDWI Data and Analytics survey.

those people will need tools to help them due to the complexity of the modern data environment. Some examples include:

Automated or augmented tools for data profiling and cleansing. These are tools infused with AI algorithms to help automate a task. For instance, these tools might use AI to automatically detect data quality issues such as duplicates, missing values, and outliers. For text data, they may match addresses or edit abbreviations. These automated tools are critical as data volume increases, and they are beginning to evolve to address newer data types (such as image data) and deal more fully with text data.

Additionally, with the advent of generative AI, new algorithms and tools are being developed to monitor the quality of text data used in large language models (LLMs)—to detect toxicity in model input and output as well as to avoid hallucinations (incorrect or irrelevant output). Examples include Perspective API, Detoxify, and Bert.

Tools to track data integrity. These tools can track data for accuracy, completeness, timeliness, and reliability and check if there are anomalies in the data. For instance, a tool may alert a user if data moving through a pipeline was not complete from one system to another. Tools may track data health metrics such as percent inaccurate data. Additionally, these tools should feed dashboards and alerts to help users understand their data integrity.

It will be important for organizations to consider what data integrity means with new kinds of data such as unstructured data being used in AI. For instance, if text data is converted to a vector embedding for use in a LLM, is that vector representing the right thing? Is the vector itself complete? This is an evolving area.

Tools to track who is using data and how it is being used. Monitoring tools track who is using the data and when they are accessing it. Additionally, monitoring tools should include how applications are consuming data. The tools should be capable of detecting whether data use complies with governance policies and whether access looks suspicious.

Data catalogs to ensure that data is understandable. Data catalogs have become popular for use in modern data governance. Modern catalogs typically have features that can help improve users' trust in the data on the platform. For instance, some data catalogs have rating features, so users can see how others rated the data. Some allow the data steward to certify or "badge" the data as trustworthy. Others allow users to post comments about the data and the use cases for which they used it.

Catalogs also provide metadata about the data—who created it, when it was created, how it is structured, where it originated, and how it has been used. Some catalogs provide visibility into data use across the cloud data platform and help track data lineage—metadata that describes where data originated and how it has been transformed, consumed, and shared. These catalogs serve as a central repository for all data assets, providing clear visibility into data origins, transformations, and usage, which is crucial for transparency.

Catalogs are also being used to store the features (i.e., derived attributes used to train AI models) used for AI models as well as data about the models. Although governance of AI/ML assets is not the primary focus of this checklist report, some catalogs can be used to store model version information as well, which is critical for governance. Additionally, there is some movement in the industry towards AI/ML catalogs.

Tools that include transparency and explainability. For governing data assets, explainability is important to understand the data used to train a model. This can be part of the metadata and data in the data catalog, described above. It is also important for communicating the output of an AI/ML model in a way a human can understand it. This is sometimes required for compliance purposes. For example, a customer should be able to understand why his loan application was rejected from an application that uses machine learning. This aspect of explainability is described in more detail in number 5.

2 Prioritize security and privacy

In the current landscape, security and privacy are critical for comprehensive and modern governance. As organizations handle increasingly sensitive and personal data, they need to protect this information while ensuring its utility for analytics. Security and privacy tools can help.

Some important technology components for comprehensive data-related security and privacy include:

- **Built-in platform security.** Traditional security mechanisms such as robust authentication, encryption, and masking will continue to be important for data both at rest and in transit. Data masking creates a structurally similar but inauthentic version of an organization's data using techniques such as replacing private identifiers with fake identifiers or pseudonyms or removing or modifying personal identifiers for anonymization. Differential privacy is another framework for dealing with sensitive data. It provides a way to maximize the accuracy of queries from databases while minimizing the chances of identifying individual entries.

- **Automatic and custom classification.** It will be important to implement tools with capabilities for automatic classification of sensitive data, along with custom classification options as data volume increases. These tools automatically identify and tag sensitive data, such as personally identifiable information (PII), across various data repositories. This includes any data that could create corporate risk or disadvantage if it were exposed, such as penalties for exposing personal data, leakage of corporate secrets, or descriptions of corporate security methods.

- **Role-based access control security.** At a minimum, it is important to implement robust role-based access control (RBAC) and policy-driven security measures. In role-based access, each user has a role and belongs to a group; this may be a department or a physical location or a user type. Some organizations are utilizing attribute-based access control (ABAC) which evolved from role-based access controls. In ABAC, access rights are granted based on the attributes of subjects (e.g., the user and their department, job role, management level), objects (what they want to access, e.g., a file), environment (the context, e.g., time accessed, location), and action (e.g., read, write, edit, delete). These are metadata-driven policies.

- **Data clean rooms.** Associated with the concept of protecting sensitive data is the data clean room. The data clean room is a secure environment

where organizations can share data while still maintaining data privacy. These clean rooms can be accomplished a few different ways such as via cloud platforms or third-party providers.

- **Continuous security monitoring.** Of course, it is important to adopt continuous monitoring tools that assess security risks across cloud environments. These tools should align with industry best practices and proactively identify potential security vulnerabilities.

3 Focus on compliance and auditability

The increasing complexity of modern data management environments make compliance and auditing extremely important. Effective compliance ensures that an organization adheres to external regulations and internal policies, safeguarding against legal and financial repercussions. These regulations, ranging from data privacy laws such as GDPR to industry-specific mandates, require careful handling of data, including how it is collected, stored, processed, and shared. Compliance might also include policies companies put in place for data retention. With the popularity of generative AI, companies should also expect new legislation to impact compliance. The EU has already passed the AI Act which establishes risk levels that classify AI systems in terms of their potential effects on people's rights, safety, or security. High-risk systems will include generative AI systems with foundation models.

An audit serves as an important way to verify and ensure that governance policies are properly implemented and followed. It provides an essential check on the organization's data practices, highlighting areas of risk, non-compliance, or inefficiency.

Modern data environments require the following components to ensure successful compliance and auditability:

- **Platform certifications.** It will be important for any platforms your organization utilizes to have compliance certifications in place for data security. These include International Organization for Standardization (ISO) 27001 for risk assessment, disaster recovery, and incident management; PCI DSS for handling credit card information; HIPAA for sensitive healthcare data, FedRamp for public sector data; and SOC II audit reports to attest to compliance with security, integrity, confidentiality (among others).

- **Compliance monitoring.** Implementing effective compliance frameworks involves establishing policies, procedures, and controls tailored to the specific regulatory requirements an organization may face. This includes data classification, access controls (described earlier), and data retention policies.

  In addition to traditional systems, it will be important to have a way to classify AI systems and applications according to risk and other parameters that may emerge with new regulations. Some vendors are already addressing this space, tracking new regulations and providing tools and templates to ensure that new systems and applications are in compliance.

- **Support for data quality with compliance-focused metrics.** Data quality is key to ensure regulatory and internal obligations are met. This may include metrics that are compliance

focused, such as inaccurate customer credit score. Organizations should have developed these kinds of metrics, although many have not. These metrics will need to evolve to support diverse data types as well.

- **Ability to generate audit reports.** Organizations need to provide documentation to auditors. Some vendors can provide automatically generated documentation and reporting for auditors. This may include access history, schema tracking changes, and reports for auditors supporting regulatory compliance. The kind of reports will need to evolve as well. For instance, if organizations are building new AI apps, the models feeding the applications need to be versioned and monitored and reports should be issued for audit about the model.

  Note that there has been work by auditors on continuous auditing of AI systems. This builds on earlier work on the continuous audit of online systems which examines risk controls and compliance on an ongoing basis rather than performing audits at a single point in time.[2]

- **Support for data lineage.** Many regulations, such as GDPR in the EU, CCPA in California, and other data protection laws, require organizations to know where their data comes from, how it's processed, and where it's stored. Data lineage provides a comprehensive map of this information. Understanding the flow of data also helps in identifying and mitigating risks associated with data handling and processing. It allows organizations to demonstrate to auditors or regulatory bodies how data is managed, transformed, and protected. As mentioned, in addition to specific data lineage tools, data catalogs often have data lineage features.

- **Business continuity and disaster recovery.** Many organizations also have recovery time objectives (RTO) and recovery point objectives (RPO) to safeguard mission-critical accounts and data sets to maintain uptime. This includes the ability to replicate and synchronize databases, accounts, and pipelines for resiliency, durability, and failover, by choice or in a stressed event.

4 Balance discovery and accessibility with governance

Modern data management supports data democratization, allowing individuals with different personas to access the data they need for analytics and insights. That means that the ability to balance discovery and accessibility with governance is a key pillar of modern governance for data and related assets.

Organizations are looking to access new data types from new sources to enrich their data sets for analytics. They want to collaborate on their analytics across the organization or with partners. Vendors are providing platforms, such as internal and external marketplaces to help easily discover and access data and related assets. With the advent of generative AI, some organizations are looking to use natural language interfaces, powered by large language models, that can help them easily find information and take action.

For example, according to TDWI research, more organizations are using marketplaces. A marketplace is an online, transactional store that facilitates the buying and selling of data, apps, ML assets, and more. It can help organizations streamline access to external assets. Data providers offer all kinds of

[2] Miklos A. Vasarhelyi and Fern Halper, "The continuous audit of online systems," *Auditing: A Journal of Practice & Theory* Vol. 10, No. 1 (1991).

data-related products through these marketplaces including weather data, demographic data, and industry-specific data. They even offer products such as machine learning models or notebooks. Consumers can purchase or subscribe to data sets and products.

The marketplace itself may be hosted by a third party such as a cloud provider or other service provider that offers a secure platform for such exchanges to happen smoothly. For instance, if an organization wants to do logistical planning, it can access weather data to help better plan its supply chain. It can purchase industry-specific data for risk analysis.

All of this means that there needs to be a balance between discovery/accessibility and governance. In other words, there should be controls in place that still enable people to access the assets they need for insights. The marketplace makes it easy to discover and share data. However, this data needs to be governed.

The good news is that some cloud platform providers that offer marketplaces can provide a governed and collaborative environment. For instance, they may provide object tagging technologies to track sensitive data across environments or provide direct access to data without requiring ETL.

Additionally, new technologies for governance, such as data catalogs, help organizations better understand data, which aids in democratization. Catalogs also help with governance by storing metadata, providing features for identifying sensitive data, and tracking lineage. Modern platforms need to easily integrate with these tools and others such as data observability tools, which measure the health of data assets. That integration can be through technologies that support technical catalogs such as those for Apache Iceberg tables, or it can be through a seamless integration with a platform vendor and its partner.

5 Uphold ethical standards and data protection

Data ethics is becoming an important extension of the governance process for data associated with AI and application development. Ethical data practices refer to the responsible collection, use, and management of data, ensuring that it aligns with moral principles of right and wrong. There are numerous ethical issues associated with data management, such as whether data could be used to target, profile, or prejudice people; unfairly restrict access (e.g., exclusive arrangements); or whether the data collected for one project could be used for an entirely different purpose.

The ethical and responsible use of data also involves bias, fairness, and explainability. Yet, in TDWI surveys, although data trustworthiness and governance rank high in importance, this is not the case for data ethics. For example, in a 2022 survey, less than 20% of respondents rated avoiding unintended consequences, fairness, or avoiding bias as important practices in data management.[3] Issues such as data privacy ranked higher.

Data privacy ensures the ethical and legal handling of sensitive information, maintaining trust among stakeholders and complying with stringent regulatory standards. It is essential in safeguarding an organization's reputation and operational integrity, preventing data breaches that can lead to significant legal, financial, and reputational damages. Organizations are utilizing techniques such as masking data (described earlier) to ensure privacy.

In terms of fairness, bias, and explainability, it will become important for organizations to develop frameworks that include policies, guidelines, and

[3] See the 2022 *TDWI Best Practices Report: Responsible Data and Analytics*, online at tdwi.org/bpreports.

procedures focused on fairness, accountability, and transparency. This involves regular training for staff, periodic audits, and the incorporation of ethical considerations in the design of applications. Some organizations are establishing data ethics committees or AI councils to help with these kinds of issues. These may become extensions to data governance organizations and programs.

Some organizations are also moving forward with new tools to help identify data bias in the input for machine learning and other AI models. Many data sets carry historical or societal biases. Transforming these into high-quality data involves identifying and mitigating these biases, either by rebalancing the data or by applying techniques to adjust for known biases. There is work underway to help identify bias. One example is FairML. FairML is an open-source Python tool that audits ML models to detect fairness. The algorithm examines dependence on inputs by changing them. If changing an input feature dramatically changes the output, that means it is a sensitive feature and the model is sensitive to it. This is an evolving area, but organizations should be paying attention to it.

Additionally, organizations are starting to utilize tools that help with explainability of model output. Two popular techniques are LIME (local interpretable model-agnostic explanations) and Shapley values. In some instances, vendors have incorporated these into their software to produce charts where users can determine the most important features in predicting the outcome of interest.

Recently, some organizations have begun to develop and advocate for AI model cards. Just as the government requires nutritional fact labels on packaged goods in the U.S. for promoting overall health, AI model cards provide relevant documentation about an AI asset or application. These cards provide a standardized overview of the model's key characteristics, capabilities, and limitations. They include model details, data sources, performance, intended use, and bias and ethical considerations.

Concluding thoughts

As data environments grow in complexity, including a wide range of data types and extending across hybrid and multicloud infrastructures, the demand for comprehensive governance has become increasingly important. Governance is critical for maintaining data integrity, trust, and transparency, which are foundational to leveraging data for analytics, management, and application development.

With the democratization of data and the advent of advanced analytics, including generative AI and large language models, the challenge of balancing accessibility with controlled governance has intensified. Ensuring data accuracy, consistency, and compliance while supporting analytics and application data governance has become imperative. Prioritizing data security, privacy, compliance, and auditability are key.

In addition to people and processes, organizations will need to utilize tools to help them govern data and data assets. These automated and augmented tools, alongside platforms equipped with security and privacy features, will play a pivotal role in upholding data quality and integrity. Companies must also begin to consider ethical standards, focusing on the responsible use of data to prevent biases, ensure fairness, and maintain transparency.

Already, organizations are re-examining some of their roles and processes for governing modern data and AI assets. Some are forming AI centers of excellence that address governance and trust. Others are starting to consider the role of the AI steward who provides operational support for AI quality, governance, and trust.

About our sponsor

Snowflake enables every organization to mobilize their data with Snowflake's Data Cloud. Customers use the Data Cloud to unite siloed data, discover and securely share data, power data applications, and execute diverse AI/ML and analytic workloads. Wherever data or users live, Snowflake delivers a single data experience that spans multiple clouds and geographies. Thousands of customers across many industries, including 639 of the 2023 Forbes Global 2000 (G2K) as of July 31, 2023, use Snowflake Data Cloud to power their businesses.

Learn more at snowflake.com.

About the author

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. She has taught at both Colgate University and Bentley University. Her Ph.D. is from Texas A&M University.

You can reach her by email (fhalper@tdwi.org), on X/Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

About TDWI Checklist Reports

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

About TDWI Research

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi
Transforming Data With Intelligence™

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367

E info@tdwi.org

tdwi.org