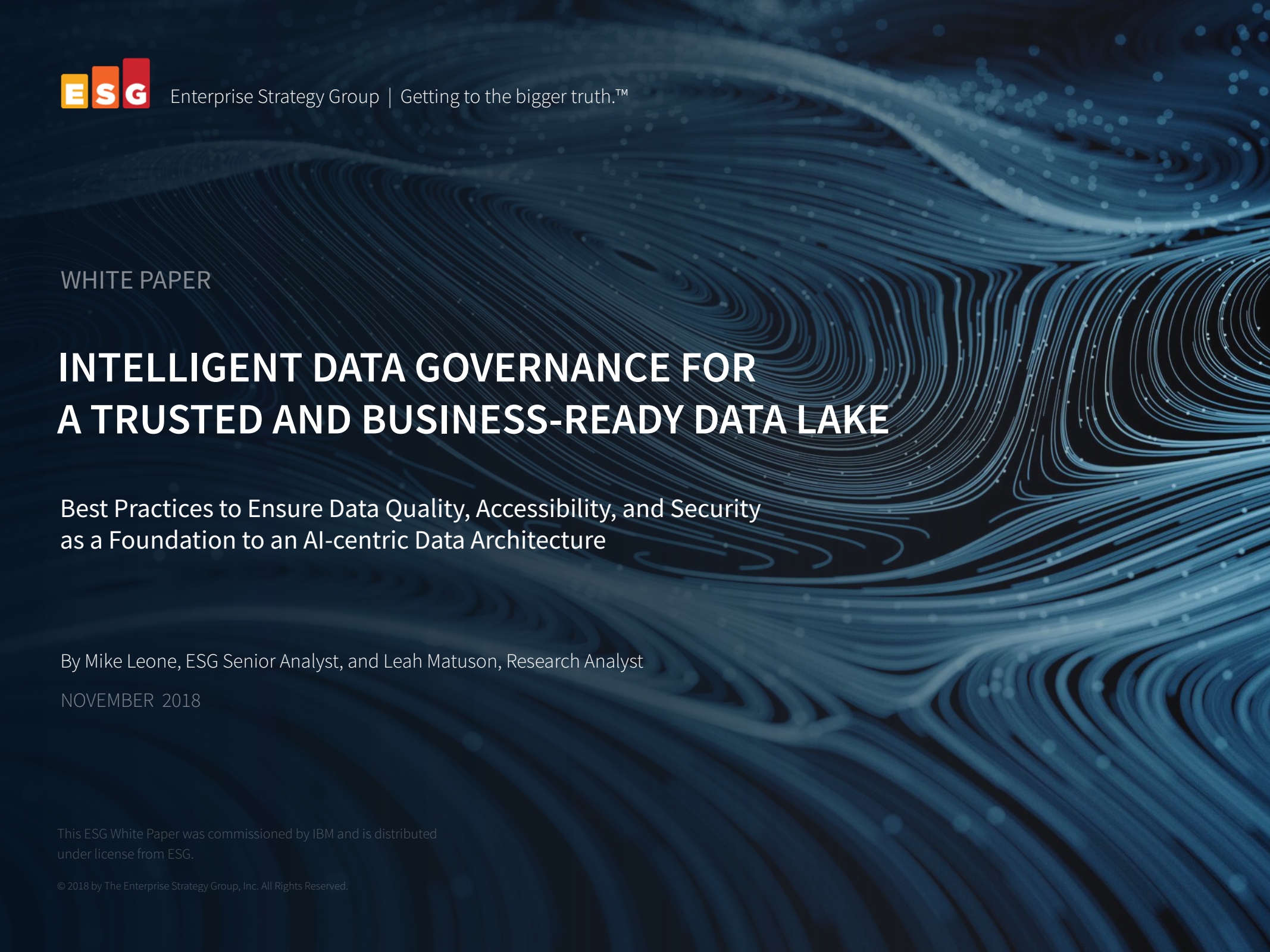


Enterprise Strategy Group | Getting to the bigger truth.™

WHITE PAPER

INTELLIGENT DATA GOVERNANCE FOR A TRUSTED AND BUSINESS-READY DATA LAKE

Best Practices to Ensure Data Quality, Accessibility, and Security as a Foundation to an AI-centric Data Architecture

By Mike Leone, ESG Senior Analyst, and Leah Matuson, Research Analyst

NOVEMBER 2018

Contents

Back to Contents

Introduction

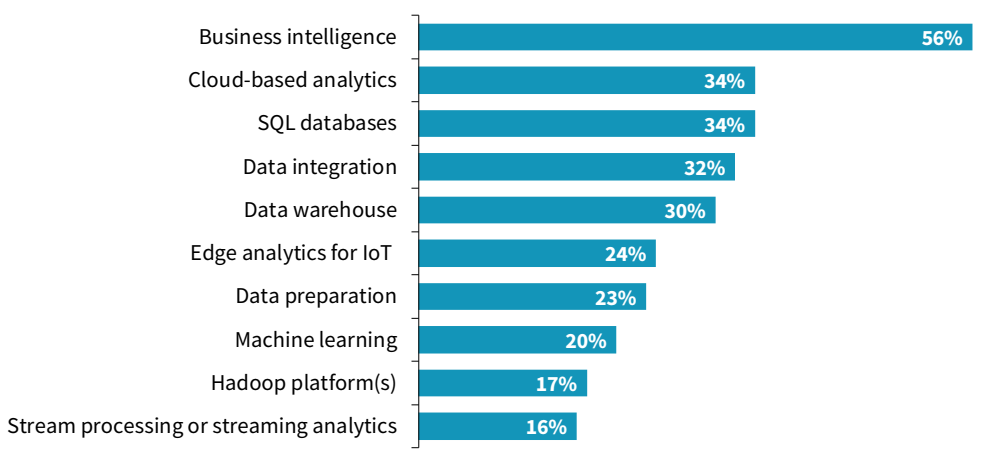
Organizations continue to rush down the digital transformation path. Whether by modernizing their IT infrastructures, leveraging the cloud, or becoming data-centric and data-driven, organizations must become more agile in their business practices and within their IT infrastructure stack to effectively compete in today's dynamic business environment. Between the speed and distributed nature of modern businesses, as well as the expectation of instantaneous access to data from everyday users, it's not surprising that nearly one in three organizations are looking into ways to improve data analytics for real-time business intelligence and customer insight.[1]

So how do organizations get there? Governance! It starts with creating a foundational inventory of data by discovering, cataloging, and integrating data when needed to help ensure data quality and protection, and leads to enabling improved data accessibility in a trusted and secure way. This is understandably a daunting task for many organizations, especially once factoring in the challenges that come with a globally distributed organization and workforce that relies on that data to gain as close to real-time insights as possible regardless of where it is being accessed from. And while this requires investment in personnel, infrastructure, and time, organizations recognize they must make the investment to remain competitive and keep their edge. As such, among respondents with purchase influence or authority for business intelligence and analytics solutions used to support big data initiatives, 56% of organizations said they would make the most significant investments in business intelligence.[2]

It starts with creating a foundational inventory of data by discovering, cataloging, and integrating data in an ongoing way that ensures data quality and protection, and leads to enabling improved data accessibility in a trusted and secure way.

Figure 1. Top Ten Areas of Data Analytics Investment

In which of the following areas of data analytics will your organization make the most significant investments over the next 12-18 months? (Percent of respondents, N=207, five responses accepted)



Source: Enterprise Strategy Group

[1]Source: ESG Brief, *2018 Data Analytics Spending Trends*, February 2018.
[2]Ibid.

Back to Contents

The Path to AI

While becoming more data-driven is clearly a priority for most organizations, it's not just about having an agile data platform for analytics and BI that addresses the real-time needs of the business. It's about applying automation, intelligence, and self-service across the entire organization to eliminate manual processes and accelerate business processes to empower not just one or two business units, but all of them such that data-driven becomes the norm throughout the firm. This transformation to an intelligent, data-centric culture requires a solid technological foundation where trusted and business-ready data is made available to the masses and leads to applying machine learning (ML) throughout the technology stack to automate the manual tasks. Once organizations master an agile and trusted data pipeline to serve ML, they can then deploy artificial intelligence initiatives that will accelerate their paths to transformation.

65%

said that AI and ML will be one of the forces driving their digital transformation efforts

It is more apparent than ever that to achieve a market leading position within any industry, artificial intelligence and machine learning must become part of future plans, whether leveraging the technology as a feature within a product or deploying a separate infrastructure to support strategic AI and ML initiatives. In fact, according to ESG research, when asked how they viewed AI and ML as part of their organization's digital transformation strategy, the majority of respondents (65%) said that AI and ML will be one of the forces driving their digital transformation efforts, with 22% citing it as the driving force behind digital transformation.[3]

But it is important to note that organizations are not simply snapping their fingers and leveraging AI and ML—it's an evolutionary process. ESG research shows that 17% of organizations are depending on AI/ML to deliver significant measurable business outcomes immediately, 52% expect these outcomes to be achieved in the near-term, and another 25% expect business value eventually.[4] While this alludes to the fact that organizations are at different stages of the AI journey, it is becoming clear that most of them understand that partnering with the right vendors is essential to success.

This transformation to an intelligent, data-centric culture requires a solid technological foundation where trusted data is made available to the masses and leads to applying machine learning (ML) throughout the technology stack to automate the manual tasks.

[3]Source: ESG Survey, *Machine Learning and Artificial Intelligence Trends*, June 2017.
[4]ibid.

Back to Contents

AI is an evolutionary process that requires organizations to ensure the right technology and processes are in place before advancing too far into the AI journey.

Evolutionary, Not Revolutionary

Jumping into AI without preparation isn't realistic, nor wise. Why? Simply put, AI is an evolutionary process that requires organizations to ensure the right technology and processes are in place before advancing too far into the AI journey.

Organizations must ensure an agile data architecture is in place—one that not only meets the current demands of the business but also enables them to succeed in the future. What they require is an architecture that not only simplifies access to all organizational data, regardless of its source, size, and structure, but ensures the person or system accessing that data understands what it means and how it applies to their needs. Once in place, organizations should look to leverage the appropriate modern technology as a foundation to ensure secure and governed data access, while meeting the dynamic needs of a modern business.

When making data accessible, organizations should not overlook security and governance, especially as they apply to AI and ML, as these are both required for conformance—with industry regulations and security best practices. In fact, according to ESG research, organizations most frequently reported data security/data compliance and data governance as the weakest links in their technology stacks as they look to properly meet the demands of their AI/ML initiatives.[5] At the organizational level, conformance is the most important requirement.

End-users, on the other hand, care most about performance. By leveraging tools that enable self-service, business analysts are empowered to achieve more without requiring IT intervention. This leads to improved operational efficiency, faster insights, and an overall improvement in the performance of their businesses in their respective markets. By fostering collaboration, organizations allow users to more easily share insights. This performance improvement rolls up to the organizational level, leading to a business that can make faster decisions and better respond to changes in the market and business landscape. This improved company performs optimally, drives greater changes, and disrupts markets, forcing others to follow. Organizational conformance leads to improved and evolving performance.

[5] ibid.

Back to Contents

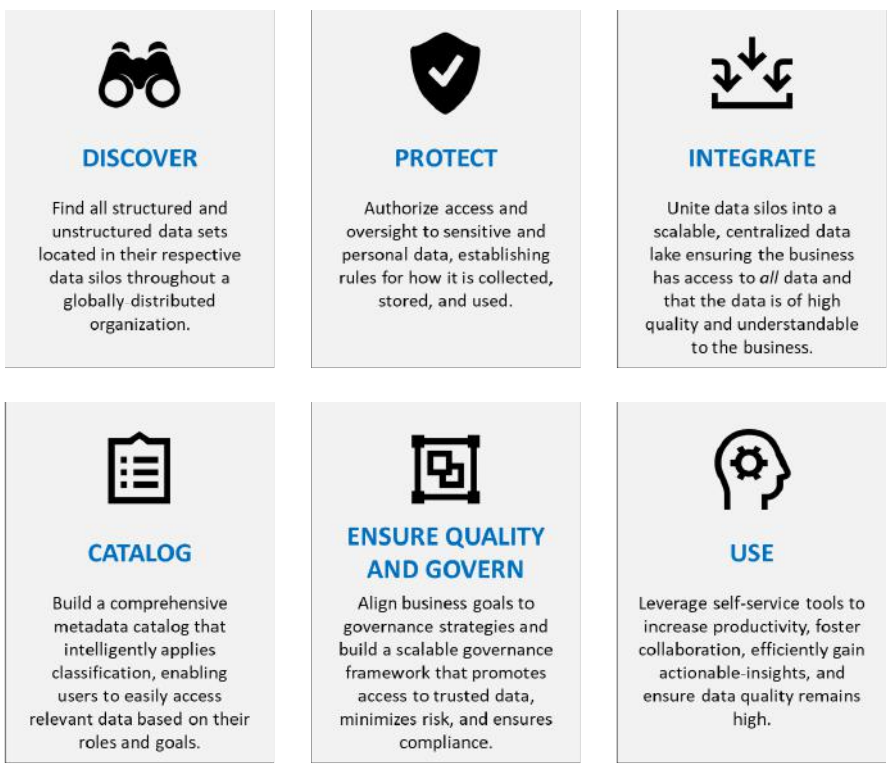
Anchoring Your Analytics Foundation with a Trusted Data Lake

While most organizations collect and store massive amounts of raw data (structured and unstructured) from numerous sources across an organization, much of the data offers partial business value. Sure, organizations can gain insights from a specific data silo, but organizations would agree that cohesively gaining a complete perspective into all available data sets offers more potential for the most valuable insights. With business processes that span multiple organizational silos a major source of inefficiency, organizations need to improve or streamline these processes by intelligently and efficiently connecting the representative siloed data sets. As such, it makes sense that bringing these silos together to improve data analysis, data modeling, and eventually the use of AI/ML-enhanced processes creates major efficiency gains.

With the rising demand for instantaneous, actionable insight, organizations have found that the adoption of data lakes where these data silos are stored is an effective approach for anchoring modern data platforms that look to serve multiple business units and processes that cross business unit domains.

Organizations can use data lakes as a repository or collection that unites disparate data silos into a virtual single entity, regardless of structure, size, type, or dependency. However, data carelessly thrown into a data lake does not automatically solve the findability of data, nor provide an understanding of what meanings lie within the data. To create an effective data lake, data must be properly inventoried via automated cataloging, integrated or transformed as needed, organized, and linked to business terms so that every user can understand what it means in a consistent and easily accessible way. Additionally, to ensure value is derived from the data within the data lake, applying intelligence and self-service is essential for a successful end-user engagement.

Figure 2. Pillars of an Effective Data Lake Governance Strategy



Back to Contents

The Importance of a Complete Data Governance Strategy

While a data lake serves as a solution for exploratory data projects led, for example, by data scientists, enterprises should not overlook the importance of having an overall data strategy across the organization with data in a data lake being part of that strategy.

Improper Governance, or Non-governance

There's a great deal at stake when data governance is not properly applied and embedded into the daily business practice. An organization can lose the ability to manage and protect growing volumes of data, leading to wasted time and resources across the organization trying to locate the right data. This, in turn, can negatively impact the ability of an organization to be audit-ready, and even diminish the ability to maintain compliance.

Governance the Right Way

Governing a data lake takes discipline, good policy, and collaboration between the people who manage data access and the people who access the data. The right way to govern requires the alignment of the corporate data strategy to regulatory requirements, ensuring no conflicts exist between the two. Governance principles must serve as a guide to how data should be managed, defining the rules from both a corporate and regulatory standpoint around managing, accessing, and sharing the data, as well as incorporating insights derived from the data in the overall strategy of the business. And this should all be done with approval across the entire organizations, including the definition of best practices and tools.

A properly governed data lake should offer granular data management capabilities, enabling teams across an organization to easily and securely access data. And in theory, making the data more accessible and available will instill confidence from data owners and users in the quality of the data itself, providing more value to the business based on more meaningful analytics. By ensuring the right people have access to the right data, organizations increase the potential for faster, more valuable insights—providing quicker time to value and accelerating innovation. But to stay relevant to the changing needs of the organization, it's essential that data lake governance is applied comprehensively, intelligently, and eventually, automatically through machine learning technology.

Governance principles must serve as a guide to how data should be managed, defining the rules from both a corporate and regulatory standpoint around managing, accessing, and sharing the data, as well as incorporating insights derived from the data in the overall strategy of the business.

Back to Contents

The Value of a Governed Data Lake

The foundational business benefits of a data lake are derived from governance, which serves as the key pillar to ensuring an ongoing data-driven culture within an organization . Governance produces efficiency by providing clarity—allowing users to take ownership of data, agree on policies and rules, and possess a common understanding of standards and definitions, which saves both time and expense by eliminating misunderstandings and misconceptions. Thus, the overall business value of a data lake grows from data awareness and importance to real-time analytics.

The payoff of a properly governed data lake is invaluable. Applying data governance to a data lake allows organizations to keep up with a continuous stream of data, follow regulatory requirements using industry-specific compliance tools, and accelerate master data management adoption and information integration. With high-quality data, organizations improve the accuracy of their insights, more easily respond to compliance audits, reduce risk, and increase their ability to more effectively protect their data.

Regardless of the data source, all data in the data lake is considered trusted. This further encourages IT and end-users to continue down a path of enabling the much-needed benefits of self-service together and using the data they need with confidence. In turn, organizations view data as being business-ready at all times and, through self-service capabilities, the increased speed of insights leads to increased business agility, new opportunities, and digital transformation, thereby helping organizations to monetize data.

The foundational business benefits of a data lake are derived from governance, which serves as the key pillar to ensuring an ongoing data-driven culture within an organization.

Back to Contents

From Cataloging to Effective, Comprehensive, and Automated Data Governance

For effective governance, it is key to ensure that a working enterprise data catalog is in place. Why? Because you can't effectively apply governance if you don't have organized data with proper metadata tags and lineage for a complete understanding of that data. Data organization includes detailing each data object—documenting data properties, ownership, business context, origin and structure, evaluating data quality, and properly classifying data so it can automatically be used to define and continue to refine not just an organization's data governance strategy, but its overall data-centric strategy. The catalog can serve multiple stakeholders in the organization, eliminating inefficiencies associated with "lost in translation" issues. It can serve as the single, trusted source of a company's inventory of knowledge assets. This includes data sources, business intelligence reports, machine learning models, business terms, and regulatory compliance and governance assessments.

Aligning a Governance Strategy to Business Goals

While cataloging is the first step to help ensure successful data governance, it's essential for a company's data to be aligned with the overall business goals of the organization. For instance, meeting compliance guidelines with government or other organizational regulators involves demonstrating the appropriate use and management of certain types of data based on the given regulation. This means organizations must ensure proper alignment between the company's data strategy and regulatory requirements, while also enabling proper access to all data needed to gain complete and accurate insight. The concept of meeting compliance requirements through proper governance is just one example of ways organizations look to tie governance to business goals. Part of aligning a governance strategy to business goals requires organizations to identify key performance indicators (KPIs) that pertain to something as grand as the overall business strategy or as granular as an individual task that is part of a specific business analyst's daily tasks—something as simple as a particular set of data getting accessed by a particular person. Each of the tasks can be tied directly to governance.

You can't effectively apply governance if you don't have organized data with proper metadata tags and lineage for a complete understanding of that data.

Back to Contents

Providing Control and Applying Automation

Once a data governance strategy is in place, organizations must ensure the proper level of controls are in place across the data pipeline. These controls not only define the processes, rules, data collection, and procedures performed on an organization's data, but also provide access points into the workflows when applicable. This level of control ensures certain groups (IT, architect, business analyst, etc.) have the access they need to all information across any workflows, pipelines, etc., at any time. More importantly, if something goes wrong, controls can be leveraged to rapidly respond to an issue, whether flagging sensitive data, identifying and remediating issues, or collecting information in response to an audit.

These controls are defined in terms of data classification, describing how sensitive personal data is managed in a variety of circumstances. With a clear and limited number of classifications, users can more easily learn and properly use the classifications, making governance that much easier.

But data lake governance would not work smoothly without automation. With data being constantly added and accessed, automation is vital. Various runtime engines (think workflow and security tools, data movement engines, and data access APIs) pass through access services that execute governance rules in the governance process, supporting enforcement as well as verification points. By automating the stages of data governance and eliminating labor-intensive manual processes, organizations will not only save time and money, but also minimize, and in some cases eliminate, human error-prone tasks.

Empowering the Whole Organization

Once an organization has a data lake strategy that aligns with its business goals, using a comprehensive metadata-centric data catalog with defined data classification and automation, it's time to ensure every person across the organization is aligned with and empowered to gain easy access to the data lake, as well as take responsibility to ensure data quality remains high. This speaks to accountability of the end-users that interact with data, who should be mindful of data inaccuracies and notify those responsible when an issue arises.

Generally, personnel can be broken into two groups: technical and business users. And it's not just about one role or the other. While group and role may have individual requirements based on the ideal data interaction cadence or style, organizations must also take into account the need to easily communicate and collaborate across roles, teams, and groups. This ensures that data accessibility, data quality, and data governance are central to the organization's culture.

Ensure every person across the organization is aligned with and empowered to gain easy access to the data lake, as well as take responsibility to ensure data quality remains high.

Back to Contents

Technical Users

Technical data lake users comprise those who are heavily involved with the infrastructure and data pipeline from an operational and execution standpoint. From beginning to end, this incorporates those involved with data ingestion to the data lake, data preparation to optimize and prepare ETL jobs, development of applications based on insights into the supporting infrastructure and workflows, development of end-user-based applications based on insights from business user interaction, and the creation of models to incorporate artificial intelligence and machine learning. Some of these roles include IT, software engineers, data engineers, and data architects. Data engineers, for example, are responsible for building transportation systems so new data source requests are fulfilled. They aim to ensure data is delivered efficiently and reliably, that it meets operational models and targets, and fuels AI objectives.

Business Users

Whether it's the actual owner of the data or someone accessing it for analytics to gain actionable insight, business users expect access to the data to feel as fast as possible and to come with no constraints, ensuring they can complete their jobs as quickly and effectively as possible. These roles consist of everyone from knowledge workers and business analysts, through PhD-level personnel like data scientists, to line-of-business executives, who are responsible for setting business-level goals. And these business-level goals come from various aspects of the business, whether related to product management establishing a new product, customer support ensuring top-notch support, marketing looking to break into a new market, or even supply chain hoping to manage goods and services that are constantly on the move.

Governance Roles

A key role in the organization, regardless of official title, is played by individuals like data stewards who are responsible for data governance—those who build governance and security strategies, manage metadata, ensure data quality, and ensure compliance across the broad technical user group, as well as within specific business units based on the needs of that group. The data steward ensures data assets are inventoried and documented, and that they follow governance rules. While this role works mostly with the technical users, business users must be considered when implementing policies that will directly impact the level of eventual and desired self-service.

Back to Contents

The Bigger Truth

With clean, trusted, and business-ready data turning into arguably their most valuable asset, organizations are prioritizing business strategies that are data-centric. But to achieve the ultimate goal of gaining actionable insights that meet the real-time and dynamic needs of the business by leveraging artificial intelligence and machine learning, organizations must evaluate the complete data architecture to ensure it aligns to the business goals.

This starts with understanding all of the data in an organization. And not just where it is stored, but where and how it is generated, accessed, and used. Organizations will continue to look for ways to best integrate and manage that data, and governed data lakes have served and will continue to serve as an ideal approach to doing that and keeping the data business-ready for AI. But it's not just simplifying access to the right data by integrating data in a centralized and governed data lake—that's just the start. It's about ensuring the people who are accessing the data can easily understand what they are looking at in business terms. It's about minimizing risk associated with improper access, lost data, or inaccurate and incomplete insight due to data quality issues. It's about making data business-ready by helping organizations evolve through conformance with regulatory and security best practices, improved performance through end-user adoption of AI and ML insights, and alignment of the value being gained directly to business strategy. All these factors tie back to one concept: intelligent data governance.

Once they understand the who, what, when, where, why, and how of an effective data governance strategy, organizations are set on a path of enabling self-service for everyone in the organization to properly govern, manage, and analyze the data most relevant to their jobs.

- **Who** – Personnel responsible for implementing a data governance strategy and those impacted by one.
- **What** – The data itself and the actual policies that will be put in place to govern it.
- **When** – The frequency at which governance is applied (as close to real time as possible), as well as how long it should be retained, and if/when it should be disposed.
- **Where** – The stages across the data pipeline where governance policies must be applied, including discovery, integration, preparation, quality checks, access, and analysis.
- **Why** – The tie between the overarching data governance strategy and the overall goals of the business.
- **How** – Automatically and intelligently discovering and classifying data, and applying the right level of control, access, and governance across the data pipeline.

A properly governed data lake serves as the trusted foundation of an organization's data-driven and more specifically, insight-driven goals. It not only ensures security and reduces risk, but also promotes access and fosters enterprise-wide collaboration. And it should do so at scale to meet the dynamic needs of the business.

LEARN MORE

For more resources to help you on your data lake journey or to talk to an expert, visit https://www.ibm.com/analytics/use-cases/governing-data-lake

Back to Contents

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.

Back to Contents

© 2018 by The Enterprise Strategy Group, Inc. All Rights Reserved.