Data governance for → data-driven organizations

Contents

Data is useful only when it is protected and its context, content, and quality can be trusted.

01
Introduction:
A data fabric
approach to
data governance

Data fabric is an architectural approach that helps ensure quality data can be accessed by the right people at the right time. A data fabric architecture supports all data and analytics use cases. It enables the creation of a data governance foundation to support efficient delivery of business insights as well as address data privacy and regulatory compliance requirements.

Data governance is a foundational capability of a data fabric that optimizes the availability, understanding and trustworthiness of data to promote efficient data usage in an organization. In addition to providing a strong foundation for multicloud data integration, data science, MLOps, and AI governance, the data governance capability of a data fabric strengthens compliance with automated governance controls, no matter where data resides.

Central governance applies the right level of data governance across the data estate, making quality data easier to find for those who should have access to it, while allowing sensitive data to remain hidden. Centrally enforcing governance across a hybrid and multicloud environment simplifies data access with appropriate security, privacy, lifecycle management and adherence to ethics.

Having insights into your business and customers is a competitive advantage. The Forrester Analytics Business Technographics® Data and Analytics Survey, 2020, found that advanced insights-driven businesses are more likely to have a data governance strategy that involves defining, executing, training, and overseeing compliance than beginner and intermediate firms. They're also more likely to have an executive in charge of their data governance and to use AI to crowdsource and embed data stewardship in everyday data engagements.[1]

Learn more
about data fabric
→

Data is useful only when it is protected and its context, content, and quality can be trusted. Strong data privacy, data security and data protection parameters allow businesses to adhere to industry regulations and address data privacy requirements. The automated governance capabilities of a data fabric ensure the required level of privacy is enforced as sensitive data is consumed within key endpoints across a distributed data landscape. By combining both data fabric and data security, organizations can ensure that their data remains compliant, secure and their networks are protected.

In this guide, we'll look at the most common governance challenges modern organizations face, the building blocks of an effective solution or approach, and the technology components you'll need to build an automated, integrated data governance layer across critical data assets that enable your most strategic business outcomes. We'll also provide helpful resources such as a data governance trial.

02
Why establish automated data governance?

As organizations strive to establish cultures of data-driven decision-making, a foundational requirement to drive data consumption is embedding automated data governance to establish data ownership, simplify data access, improve data literacy, ascertain data quality, and enforce the required privacy and security. Satisfying the outlined requirements is especially critical in dynamic regulatory environments.

A strong data governance foundation helps activate business-ready data by helping to increase transparency, trust, and understanding of data, to accelerate time to insights, while helping address data privacy and regulatory compliance requirements.

Such an approach allows organizations to deal with challenges such as:

The need to know your data

Inadequate data literacy can hinder an organization's transformation into an insights-driven organization. Data practitioners should have a consistent understanding of data across the organization and be able to use data to derive business value. The foundation of a well-understood glossary of metadata helps users better understand the meaning of the data and determine how to use it.
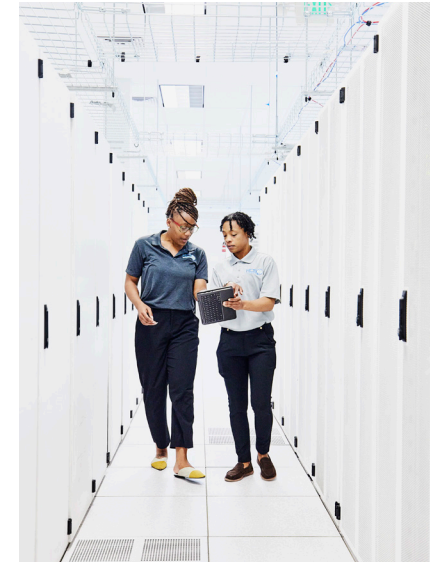
The need for data privacy and regulatory compliance at scale

Managing [data security](#), industry governance requirements and regulatory compliance across an increasingly hybrid and complex environment can be challenging. The risks of non-compliance such as legal penalties, loss of customer trust, and loss of reputation are real. More than 60 jurisdictions around the world have enacted or proposed privacy and data protection laws, and by 2024, more than 80% of companies worldwide will face modern data privacy and data protection requirements.[2] Rather than responding to each challenge individually, a proactive approach to privacy and data protection is an opportunity for organizations to build customer trust. But to do it, data leaders need to build a holistic data privacy program across the organization.

The data governance and privacy capabilities of a data fabric help organizations stay compliant with ever-changing regulations by applying governance and data protection rules through a unified, consistent approach across an organization. Data policy enforcement at the granular level can also be automated.

The need to maintain data quality standards across the organization

Only 20% of business executives completely trust the data they get.[3] If you want all users throughout your organization to fully understand and have confidence in the data they are about to use, you need to establish a data governance foundation of business definitions and metadata to ensure data quality. Your teams can automate this process through metadata-driven automatic data profiling, classification and business term assignment.

By 2024, more than 80% of companies worldwide will face modern data privacy and data protection requirements.

Why establish automated
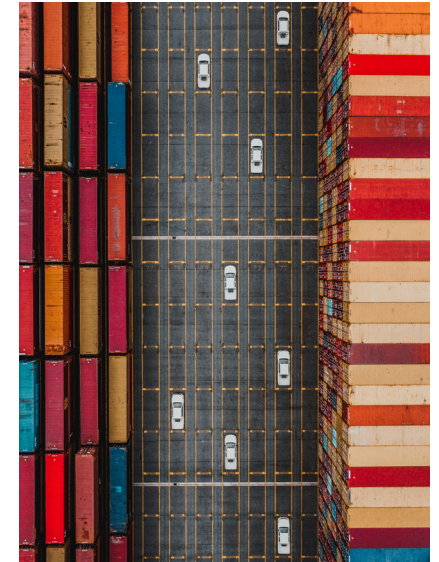data governance?

**The need for data lineage
and traceability**

Once analytics, operational, line of business or
AI teams have built and deployed data products
(such as dashboards, reports, and machine
learning models), they need to be able to look
back and see where the data product came
from. For auditability and compliance use cases
(often in regulated industries), an analytics team
may be required to show all the steps taken in
the life of the data as it has been transformed
from the transactional system where it was
originally created into its final form as it is used
to support business decision-making.

And for end users, being able to see the
data sources and transformations can save
a great deal of time as they build their own
customized version of the dashboard.

**The need to facilitate data access
and consumption**

To take advantage of the innovative and
disruptive power of data, organizations need
to enable self-service data consumption.
The ability to simplify data access and
consumption requires a robust governance
framework and architecture. This enables the
right data users in an organization to easily
search for, find and use governed, quality
data in a self-service manner with a rich and
metadata-driven index of cataloged assets.

Without being able to ensure centrally
managed governance across their
environment, teams hesitate to share data
between business units, deepening silos.
This forces IT teams to protect each data
repository individually, spinning up their own
repositories—shadow IT—leading to more
complexity.

The EU has handed out $1.2 billion in fines over breaches of the bloc's GDPR law since Jan. 28, 2021, according to law firm DLA Piper.[4]

03
The building blocks of data governance

Ultimately, the goal of governance is knowing where data comes from, what it is, who can access it and when it should be retired. Several key technology building blocks exist to meet the need to integrate and improve data privacy, access, quality and traceability for all the data in an organization.

Let's look at what you'll need.

1. Augmented data cataloging

A data catalog can help users easily find and use the right data with a rich and metadata-driven index of cataloged assets. It can act as a system of record to capture and manage the portfolio of governed data and analytics assets, policies, and other governance artifacts throughout the data product lifecycle. Using robust search methods, AI recommendations and user reviews, an intelligent data catalog can provide a strong marketplace experience for well-described and governed data assets.

Read "A comprehensive guide for the modern data catalog" →

2. Automated metadata generation

Metadata tracks the origin, privacy level, age and potential uses of your data. Manually generating metadata is cumbersome, but with machine learning, data can be automatically tagged with metadata to mitigate human error and dark data. Automatic tagging of the metadata allows for policy enforcement at the point of access, so that more sensitive data can be used in a non-identifiable and compliant way. In addition, metadata is used to establish a common vocabulary of business terms that provide context to data and to link data from different sources. This context adds semantic meaning to data so that it becomes more findable, usable and consistent within the organization, a key factor when seeking data for analytics and AI.

3. Automated data access and lineage

Data lineage shows, traces and analyzes how data is moved and consumed across all your applications and data sources. Knowing where data comes from is useful not only for compliance reporting but also for building trustworthy and explainable AI models. And it can be automated without complicating access. With restrictions built directly into access points—for example, dynamic data masking—only data that users are authorized to access will be visible. This clarity around what data can and can't be used supports self-service data demands and allows organizations to be nimble in responding to line of business needs.

4. Data privacy management

With an open and intelligent approach to accessing, curating, and sharing data across your organization, you can not only increase compliance with data privacy and industry regulations, but also enhance your ability to make data-driven decisions. First, you should identify where sensitive data lives across your data estate.

You must then be able to protect such sensitive data from unauthorized users with data policy management, which describes how data should be handled and automated through data protection, data quality, and automation rules. As sensitive data is consumed within key endpoints across your distributed data landscape, data protection rules to de-identify data or deny access can be automatically applied according to user and data attributes.

5. Data quality management

The quality of your data determines how confidently you can act on insights. If low quality data enters your AI models, it could lead to inaccurate, non-compliant or discriminatory results. Low quality data also impacts operational and analytical use cases. Getting the best insights means being able to access data that is fresh, clean and relevant, with a consistent taxonomy.

Address data quality issues with your critical data elements by assigning data quality scores to assets and simplify curation through AI-driven data quality rules with an automated, metadata-driven foundation. Mitigate the potential impact of bad quality data on downstream processes using data observability capabilities to monitor and identify data quality issues closer to the source.

6. Data virtualization

Data virtualization connects data across all locations and makes the disparate data sources appear as a single database. This helps you ensure compliant access to the data through governed data access, regardless of where it lives, without movement. Using the single virtualized governed layer, user access to data is defined in one place instead of at each source, reducing complexity of access management.

7. Reporting and auditing

Organizations must comply with a wide variety of changing regulations that differ according to geography, industry and data type. They need to be broken down into a catalog of requirements with a clear set of actions that businesses must take. Regulatory information should be automatically ingested, deduplicated, and applied to workflows.

The secret to harmonizing all these data privacy and governance needs with business opportunity is aligning the technology components with a global data strategy and an open and holistic architecture.

04
Data fabric: A holistic approach

To harness data for insights and business growth—and ultimately create a data-driven culture—you need a holistic approach to data architecture and strategy that is efficient and doesn't involve manually patching together many solutions. Therefore, many organizations are adopting a data fabric.

A modern data architecture ensures data is accessible to relevant data users based on their unique workflows. Data fabric is an architectural approach that simplifies data access in an organization and facilitates self-service data consumption. Teams can use this architecture to automate data discovery, governance and consumption, through integrated end-to-end data management capabilities. Whether data engineers, data scientists or business users are your intended audience, a data fabric delivers the data needed for better decision-making.

With a data fabric, you can elevate the value of your organization's data by providing the right data at the right time regardless of where it resides. A data fabric brings together capabilities like those listed previously as part of a unified architecture, avoiding the cost and complexity of integrating a plethora of point solutions. Instead of a fragmented group of products that have been stitched together, a data fabric offers a single, holistic solution that is built to work seamlessly.

In addition, a data fabric can address three separate use cases beyond data governance and privacy. These include data integration, data science and MLOps, and AI governance.

A data fabric can elevate the value of existing data by providing access to the right governed data at the right time regardless of where it resides.

Get a more detailed view →

By 2024, data fabric deployments will quadruple efficiency in data utilization while cutting human-driven tasks in half.[5]

05

Data governance success story

Financial services: ING ↷

ING is a Dutch bank with over 57,000 employees serving around 39.3 million customers, corporate clients and financial institutions in over 40 countries. To bring his vision of data governance to life at ING, Ferd Scheepers, ING's Chief Architect, wanted to implement a data fabric approach in the company's hybrid cloud environment.

ING needed to govern its data in the cloud consistently with its on-premises environment. As the data leader, Scheepers had specific goals:

– Empower ING's data citizens with fast and simple access to governed data and toolsets
– Ensure strong governance and privacy parameters across a complex global ecosystem
– Comply with business policy and multi-jurisdiction regulations with changing requirements

ING created a data fabric solution to help implement a single corporate operating model and streamline data management and applications across all operational countries.

– ING sought an approach to orchestrate data governance that could scale for a global operation, reducing friction for data scientists.
– IBM and ING piloted a common, smart metadata tier to easily discover, govern, and secure data, enabling users (e.g., business analysts, data scientists) to collaborate with data stewards who declaratively write data access policies.
– IBM and ING leveraged an orchestrator that ensures proper data protection, privacy, and compliance policies are consistently enforced.

It runs across an open hybrid cloud environment that adapts to ING's multi-platform, heterogeneous landscape. Applying data virtualization across existing on-premises investments, it removes data silos, enabling just-in-time access to the right data across any cloud and on-premises, at the optimum cost, with the appropriate level of governance.

Using their data fabric, ING can provide a consistent user experience to increase collaboration, streamline application management, and optimize licensing and IT costs.

Read the case study →

Consider these components

An effective data governance strategy is dependent on a technology stack designed to gain end-to-end governance, deliver quality data, and ultimately accelerate collaboration. The value of data governance is amplified when this capability is combined with data integration and entity resolution capabilities.

As part of a modern data fabric, data governance creates an end-to-end user experience rooted in metadata and active policy management that allows users to view, access, manipulate and analyze data without the need to understand its physical format or location, and without having to move or copy it.

The technology components of the IBM data fabric approach allow companies to automatically apply industry-specific regulatory policies and rules to their data assets, securing across the enterprise, with:

– An AI-augmented data catalog allowing business users to easily understand, collaborate, enrich and access the right data
– A metadata and governance layer for all data, analytics, and AI initiatives increases visibility and collaboration on any cloud
– The ability to mask data dynamically and consistently at a user defined granular level
– The ability to create anonymized training data and test sets while maintaining data integrity

Learn how Standard Bank Group built an industry-recognized data management and governance practice to deliver a single source of quality data across all systems and users.

Read the case study →

IBM Cloud Pak for Data

IBM Cloud Pak® for Data is a platform built specifically with a data fabric architecture in mind to predict outcomes faster and allow you to collect, organize and analyze your data, no matter where it may reside. The platform thus helps to improve productivity and reduce complexity by building a data fabric that connects siloed data distributed across a hybrid cloud landscape.

Learn more about
IBM Cloud Pak for Data →

IBM Knowledge Catalog

IBM Knowledge Catalog enables intelligent governance through advanced data discovery, data quality management, automated data lineage, data cataloging, and data protection capabilities across a hybrid distributed data landscape. The solution is powered by active metadata and is designed to enable self-service access to trusted data for insights generation as well as for regulatory compliance use cases. Access, curate, categorize and share data, knowledge assets and their relationships, wherever they reside.

Also, integration between MANTA's end-to-end Data Lineage platform and IBM Knowledge Catalog's business-friendly native data lineage provides a more complete picture of technical, historical, and indirect data lineage.

Learn more about IBM Knowledge Catalog →

IBM Watson Query

Applying sweeping governance rules across data lakes, databases, and warehouses is time consuming, and often leaves users with long delays to get access to the right data. Watson Query enforces governance policies when data is accessed across multiple sources, quickly providing data to your end applications through one view without manual changes, data movement or replication.

Learn more about
IBM Watson Query →

Databand

Data observability capabilities offered by Databand, an IBM company, allow customers to identify and fix data and machine learning pipelines errors, pipeline failures and poor quality. Databand technology allows engineers to tackle challenges associated with bad or incomplete data at the source.

Learn more about IBM Data Observability by Databand →

IBM Match 360 with Watson

IBM Match 360 with Watson helps consolidate data from disparate sources across domains and systems to provide a broad view of persons, organizations, and custom entities. This is done with the assistance of ML and governance capabilities, for a simplified experience for business users.

Learn more about IBM Match 360 with Watson →

IBM Security Guardium Insights

With out-of-the-box compliance capabilities and workflows such as compliance policy creation, audit process definition, access privilege assignment, and user activity reports, IBM Security Guardium Insights allows the enterprise to quickly meet its data regulatory needs.

Its robust capabilities help enterprises build and automate compliance policy enforcement and stream and centralize data activity across a hybrid-cloud ecosystem. Guardium Insights is architected to provide data security specialists with features such as automated compliance, audit and reporting, and real-time monitoring.

Learn more about IBM Security Guardium Insights →

07
Create your
ideal governance
solution

If you're ready to embrace a unified strategy and architecture approach to improve the accessibility, security and compliance of your data of all types and sources, we encourage you to take advantage of a few free IBM resources below:

**Get hands-on experience
with data governance**
Check out the trial →
Explore IBM data governance solutions →
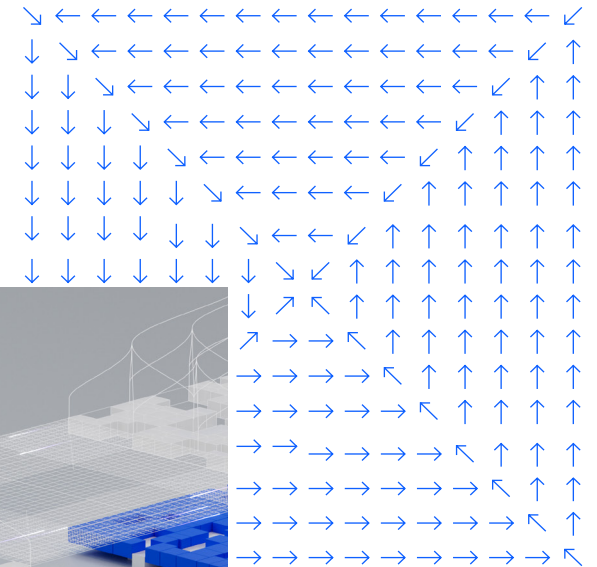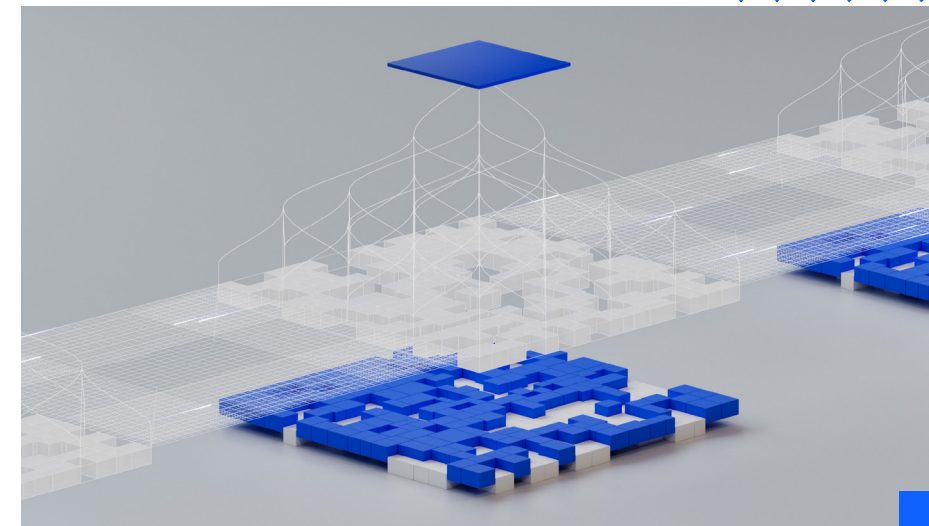Connect with an IBM expert →

Check out related use case ebooks
Data integration →
AI governance →

Build a strong data foundation for AI

Explore the guide for data leaders
The Data Differentiator →

01   Break Through Data Governance Fatigue,
     A Framework for Effectiveness and Sustainability,
     Forrester, 25 February 2021.

02   Gartner® Hype Cycle™ for Privacy 2022.
     GARTNER is a registered trademark and service mark,
     and HYPE CYCLE is a registered trademark of Gartner,
     Inc. and/or its affiliates and are used herein with
     permission. All rights reserved

03   The data-powered enterprise, Capgemini Research Institute.

04   Fines for breaches of EU privacy law spike sevenfold to $1.2
     billion, as Big Tech bears the brunt, CNBC, 18 January 2022.

05   Top Strategic Technology Trends for 2022: Data Fabric,
     Gartner, 18 October 2021.