Data governance and privacy for ↪ data leaders

Contents

01
Introduction– IBM Data & AI on AWS

IBM and AWS are bringing together the #1 leading AI portfolio with the largest global cloud infrastructure, to help businesses find a better way to put their data to work.

In this guide, we'll look at the most common governance and privacy challenges modern organizations face, how IBM Cloud Pak for Data on AWS can help you create an effective solution and approach both prescriptive and predictive capabilities for model explainability, detecting model bias, fairness, decay and drift.

With IBM Cloud Pak for Data on AWS businesses can:

Create a robust hybrid cloud data architecture
With access to multiple data sources to choose from, including IBM & AWS, with IBM Cloud Pak for Data your business can create a robust hybrid cloud data architecture.

Connect all data
When using IBM Data and AI on AWS, you can connect to all your data sources, including Amazon S3, Amazon Redshift, Amazon RDS, Amazon Aurora, Snowflake, MongoDB, Teradata, Apache Hive, IBM DB2, Netezza performance server and more.

Additionally, data from hybrid cloud databases & services can be connected using inbuilt connectors available in Cloud Pak for Data.

Use single interface
Create a single interface for integrating data sources coming from multiple IBM and AWS infrastructures.

Get started with ease
IBM Cloud Pak for Data on AWS is available with an AWS Quick Start deployment, which ensures that secure, comprehensive analytics and AI platform is ready within four hours.

02

Introduction– a data fabric approach to governance and privacy

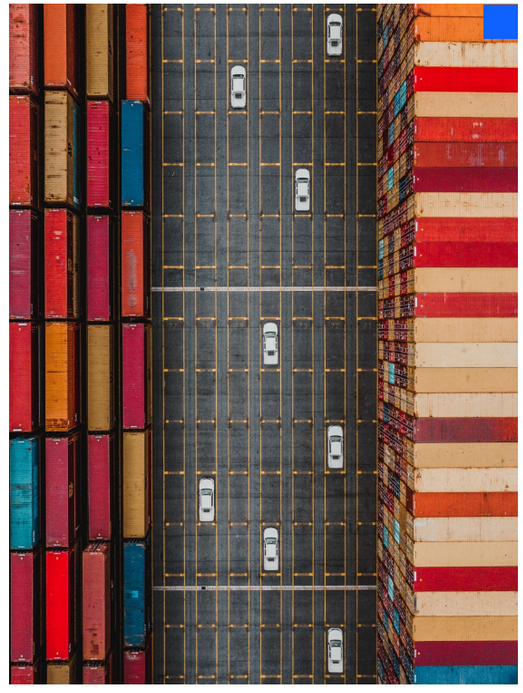
Data fabric is an architectural approach that helps ensure quality data can be accessed by the right people at the right time.

In addition to providing a strong foundation for multicloud data integration, 360-degree customer intelligence and trustworthy AI, the data governance and privacy capability of a data fabric strengthens compliance with automated governance and privacy controls, while maintaining regulatory compliance no matter where data resides.

Strong governance makes the right, quality data easier to find for those who should have access to it, while allowing sensitive data to remain hidden unless appropriate. 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, 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]

Strong privacy parameters help increase readiness for compliance and data protection anywhere, on-premises or across clouds. They allow businesses to understand and quickly apply industry-specific regulatory policies and governance rules on data wherever it resides.

In this guide, we'll look at the most common governance and privacy challenges modern organizations face, the building blocks of an effective solution/approach, and the technology components you'll need to build an automated, integrated data governance and privacy layer across all the data in your enterprise. We'll also provide helpful resources such as a data governance and privacy trial.



03
Why establish automated data governance and privacy?

As organizations strive to establish cultures of data-driven decision making, the ability to rely on quality data that is compliant with a dynamic regulatory environment is critical. Such an approach allows organizations to deal with challenges such as:

The need for data privacy at scale
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 2023 more than 80% of companies worldwide will face at least one privacy-focused data protection regulation.[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 privacy program across the organization.

The need to improve data access
Secure data sharing is a crucial factor when multiple teams require access to enterprise data. That data must be traceable and only visible to those who are authorized to use it. Yet 7 in 10 organizations are unable to secure data that moves across multiple cloud and on-premises environments.[3]

Without being able to ensure compliance at scale and from one environment to another, teams hesitate to share data between business units, deepening silos. This causes IT teams to have to protect and ensure each data repository on an individual basis and can lead to groups spinning up their own repositories (shadow IT), which only leads to more complexity.

EU data protection authorities have handed out a total of $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

Why establish automated data governance and privacy?

The need to maintain data quality standards across the organization
Only 20% of business executives completely trust the data they get.[5] Every year, poor data quality costs organizations an average $12.9 million, according to a recent Gartner report. Gartner predicts that by 2022, 70% of organizations will rigorously track data quality levels via metrics, improving it by 60% to significantly reduce operational risks and costs.[6]

For all users throughout an organization to be able to fully understand and have confidence in the data they are about to use, a data governance foundation of business definitions and metadata is essential. This foundation includes business terms, data classifications, reference data, associated metadata, and the establishment and enforcement of data governance policies and rules.

The need for data lineage and traceability
Once analytics 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 consumption
To leverage the innovative and disruptive power of data, enterprises need to enable self-service data consumption. The ability to simplify data access and consumption is predicated on a robust framework and architecture that ensures data users in an organization can easily find and use the right data with a rich and metadata-driven index of cataloged assets. Data governance and privacy proactively enables enterprises to satisfy the need to drive innovation and meet business outcomes.

04
The building blocks of governance and privacy

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

Data cataloging
The quality of your data determines how confidently you can act on insights. If low quality data goes into AI models, it could lead to inaccurate, noncompliant or discriminatory results. Getting the best insights means being able to access data that is fresh, clean and relevant, with a consistent taxonomy. A data catalog can help users easily find and use the right data with a rich and metadata-driven index of cataloged assets.

#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 nonidentifiable 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.

04
The building blocks of governance and privacy

#3

Automated governance of data access and lineage
Data lineage shows how data has been accessed and used and by whom. 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, only the data users are authorized to access will be visible. Additionally, sensitive data can be dynamically masked so that models and data sets can be shared without exposing private data to unauthorized users. 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 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.

#5

Reporting and auditing
Enterprises 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.

05

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. This is why many organizations are adopting a data fabric.

A **data fabric** is an architectural approach to simplify data access and consumption in an organization. This architecture is agnostic to data environments, processes, utility and geography. It unites disparate data systems with end-to-end data management and governance, simplifying self-service data access and collaboration.

With a data fabric, enterprises elevate the value of their data by providing access to the right governed data at the right time regardless of where it resides. It 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 multicloud data integration, 360-degree views of customers, and MLOps and trustworthy AI (covered in separate ebooks).

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.

By 2024, data fabric deployments will quadruple efficiency in data utilization while cutting human-driven data management tasks in half.[7]

06

Data governance and privacy 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.

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 →

07
Consider these components

Employing a robust governance and privacy capability is dependent on a technology stack that is designed to gain end-to-end governance, deliver quality data, and ultimately accelerate collaboration. In the context of an enterprise, the value of data governance is amplified when this capability is integrated with data integration, providing a comprehensive view of clients and enabling maximum data utilization to drive business outcomes.

As part of a modern data fabric, the data governance and privacy capability 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 dynamically and consistently mask data at a user defined granular level
- The ability to create anonymized training data and test sets while maintaining data integrity

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 →

Get started with
Cloud Pak for Data on AWS →

06

Consider these components

IBM Watson Knowledge Catalog
IBM Watson® Knowledge Catalog provides intelligent cataloguing, with automated metadata collection and policy management to ensure the details of a model are automatically collected and stored for maximum transparency and repeatability. It ensures that models are impartial, address bias, are explainable and adapt to changing model parameters.

Learn more about
IBM Watson Knowledge Catalog →

Get started with Watson Knowledge Catalogue
on AWS with Cloud Pak for Data →

IBM Watson Studio
Within Watson Studio and Knowledge Catalog, data refinery tools save data preparation time by quickly transforming large amounts of raw data into consumable, high-quality information that's ready for analytics.

Learn more about
IBM Watson Studio →

Get started with Watson Studio on
AWS with Cloud Pak for Data →

08
How it works

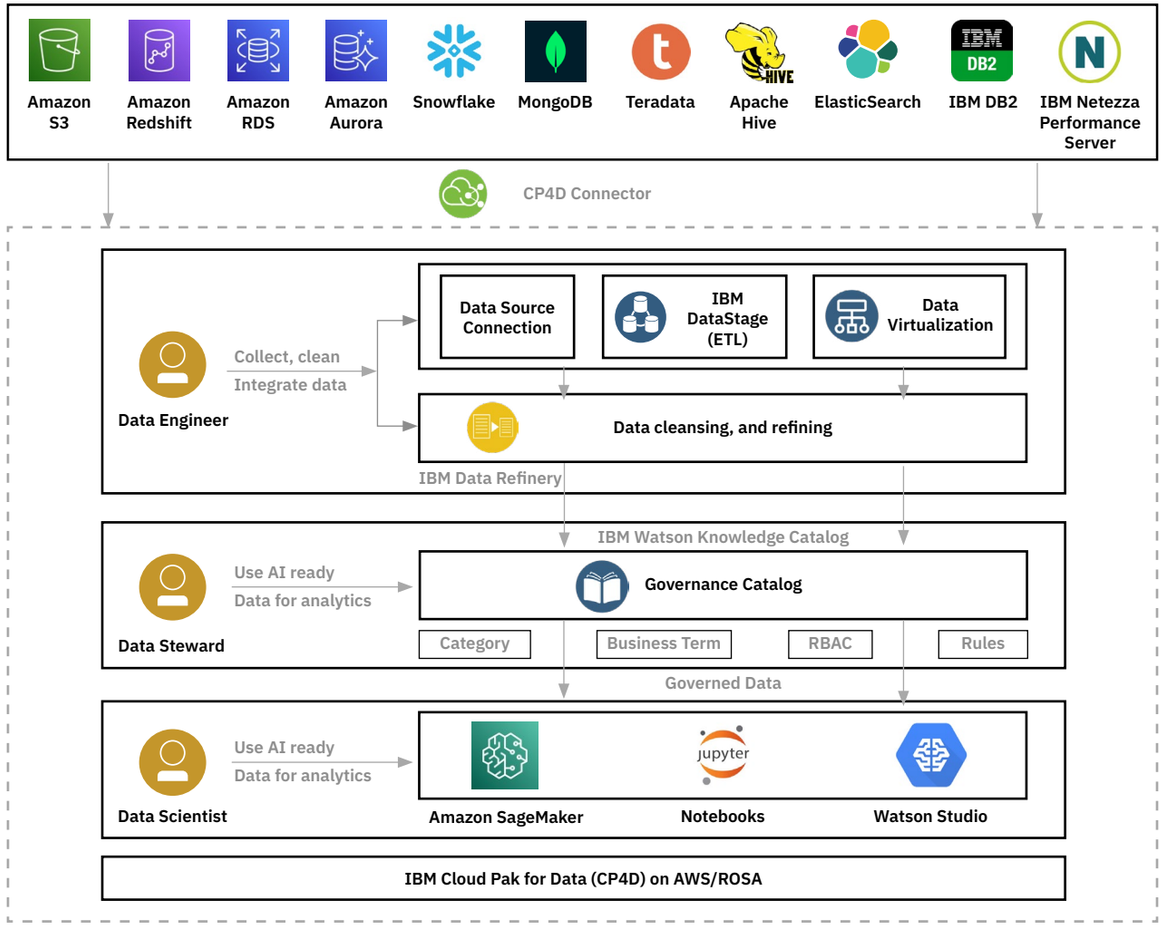
Data governance, quality, and policy enforcement

With IBM Cloud Pak for Data, Watson Knowledge Catalogue enables organizations to organize, define, and manage an AWS data footprint. Automatically apply data protection rules to AWS data sources.

Cloud Pak for Data lets you connect to your data no matter where it lives. It supports a variety of data sources such as Amazon S3, Amazon Redshift, Amazon RDS, Amazon Aurora, Snowflake, MongoDB, Teradata, Apache Hive, IBM DB2, Netezza performance server, etc.

Enable your organization to:

- Determine workflow management for proper review, approval and publication processes by using governance artifacts
- Improve decision-making by identifying, understanding and correcting data
- Access visibility across the data lifecycle
- Easily find data with metadata-driven index



09

Create your ideal governance and privacy 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 resources.

Get started on AWS Marketplace.
Schedule a consultation with a data fabric on AWS expert.

Check out this data fabric use case ebook:

MLOps and trustworthy AI

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
May 2022

01 Forrester, “Break Through Data Governance Fatigue, A Framework For Effectiveness And Sustainability,” February 25th, 2021

02 Gartner, “Hype Cycle™ for Data Privacy,” 2021

03 McCurdy, Chris, Shue-Jane Thompson, Lisa-Gaine Fisher, and Gerald Parham. “Getting started with zero trust security: A guide for building cyber resilience. ”IBM Institute for Business Value. August 2021. https://ibm.co/zero-trust-security

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

05 The data-powered enterprise, Capgemini Research Institute

06 Gartner, “How to Improve Your Data Quality,” July 14, 2021 https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality

07 Gartner “Top Strategic Technology Trends for 2022: Data Fabric,” Mark Beyer, Ehtisham Zaidi, Guideo De Simoni, October 18, 2021.