Create a strong data foundation for AI

Contents

01

Unlock your data for AI

Data is the lifeblood of every modern organization, and it's being created, stored and analyzed at an unprecedented rate across industries. By 2025, global data creation is estimated to reach a staggering 180 zettabytes.[1] This explosion of data presents enormous opportunities for businesses that are prepared to take advantage of it—from generating new offerings and revenue streams to protecting against regulatory risk. Those organizations that fail to address their data management will find this influx of data to be more a challenge than a chance for innovation.

But it isn't enough to simply collect and store large volumes of data; businesses need to seamlessly and securely access, govern and use this data to drive digital transformation and successful AI adoption.

To create a robust data foundation for AI, businesses need to solve data management complexity in 3 key areas:

Access
Given the growing volume, variety and velocity of data, organizations need quick access to data spread across complex hybrid cloud environments that include multiple clouds, data stores, locations and vendors, as well as disparate data types.

Governance
They must contextualize and classify this data to ensure they're getting relevant, high-quality information to the right people at the right time, with self-service access to reduce the time to value for AI initiatives.

Privacy and compliance
They also need to identify and protect personal and confidential information while ensuring regulatory compliance so that sensitive data can be safely used for AI and analytics.

Addressing common challenges in your data foundation helps set the stage for more accurate AI outcomes and successful AI implementations.

The DataOps methodology and the modern architectural pattern of a data fabric can help businesses address data access, governance and privacy throughout the AI lifecycle. Together, these approaches enable self-service data consumption and automation of complex and tedious data engineering tasks to help you unlock the full value of your data estate and establish a strong data foundation for AI success.

- According to Forrester, 80% of firms expect the number of AI use cases to increase within the next 2 years.[2]

- 90% of companies have difficulty scaling AI across their enterprises.[3]

02

Transform data management with DataOps and a data fabric

Every organization is looking for ways to unlock its data to innovate and stay competitive, but most face significant challenges with collecting, storing and analyzing data for AI. Data environments have grown increasingly complex, with disparate data types and sources dispersed across the data landscape.

To fuel AI with the most relevant, trustworthy information, business data will need to be discovered, cataloged and transformed, and organizations will need protocols for ensuring data privacy and regulatory compliance. A robust data management strategy and the supporting information architecture can help you unlock business data and discover insights to transform your business.

Establishing a data foundation built on DataOps principles

DataOps is a methodology that enables businesses to strategically design their information architecture and use AI-powered automation to derive maximum value from their data.

DataOps principles can help your organization achieve:

- Data cataloging to provide definitions that facilitate self-service management
- Data preparation to transform raw data into consumable information
- Data quality assessment to ensure the best business-ready data
- Data integration to meet data access and delivery needs
- Data privacy and compliance definition and enforcement

A DataOps approach helps drive agility, speed and new data initiatives at scale, empowering your business to use AI while ensuring proper governance and security controls.

Modernize your information architecture with a data fabric

A strong foundation for AI success requires more than just a methodology or set of principles; organizations also need to modernize their information architecture technology. That is, you need an architecture designed for AI—one that will help you optimize and automate data access and availability, deliver high-quality governed data and manage privacy and compliance.

Data fabric is the technological connective tissue between data endpoints that enables the full range of data management capabilities: discovery, integration, governance, curation and orchestration. It equips data citizens with access to the right data at the right time.

- 75% of larger companies report drawing from more than 20 different data sources to inform their AI, business intelligence and analytics systems.[4]

- More than 50% of firms struggle with data integration when it comes to both data science and machine learning platforms and analytics and business intelligence platforms.[2]

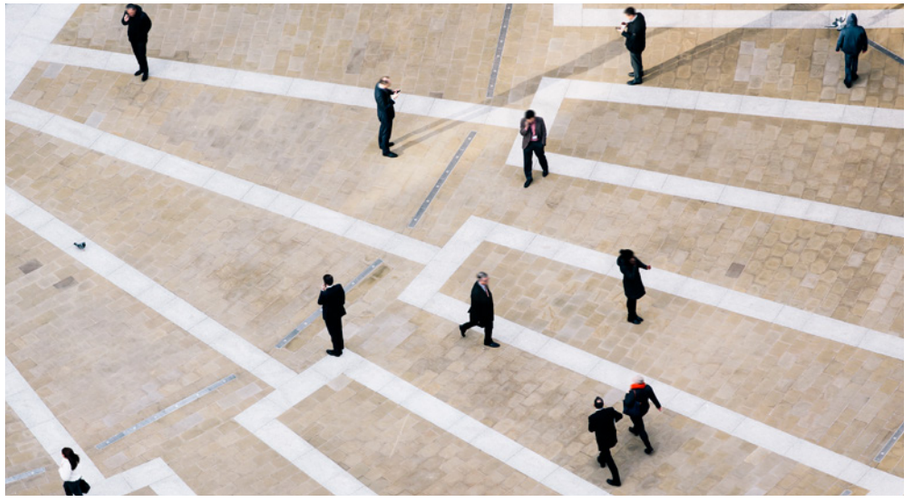
A data fabric is an architecture that dynamically orchestrates disparate sources across a hybrid and multicloud landscape to provide business-ready data for AI.

A data fabric employs emerging technologies such as machine learning (ML), data virtualization, a semantic layer, metadata management and automated data cataloging to break down the boundaries separating applications, data, clouds and people.

A data fabric offers 3 key benefits:

- Self-service data consumption and collaboration
- Automated governance, data protection and compliance
- Data integration across a hybrid and multicloud data landscape

By implementing an information architecture designed for AI—and underpinned by DataOps principles—your business can eliminate data silos, govern the data and AI lifecycle, and run anywhere with agility. Ultimately, the right architecture and principles can help you operationalize AI with trust and transparency.



- DataOps is the orchestration of people, processes and technology to deliver trusted data to digital citizens fast.

03

Simplify, unify and connect all your data

Most organizations are dealing with an overwhelming volume of data from disparate sources. It's often siloed and sprawled out across multiple clouds, data stores, locations and vendors—making data access time-consuming and complicated for data scientists, business analysts and other stakeholders.

The challenges of data consolidation

Many organizations depend on outdated data architectures that attempt to consolidate disparate, siloed data into data warehouses or data lakes. Consolidation approaches that use extract, transform and load (ETL) procedures to copy data into a single data store are time-consuming and costly, adding complexity to the data landscape. And most organizations still end up with multiple repositories. As a result, data scientists struggle with lengthy data preparation cycles and difficulty organizing data to achieve a single view.

Unifying the data landscape for easier access and universal queries

A data fabric is the architectural answer to these challenges, helping businesses unify and simplify their information architecture for AI and empower their users with faster access to the business data they need.

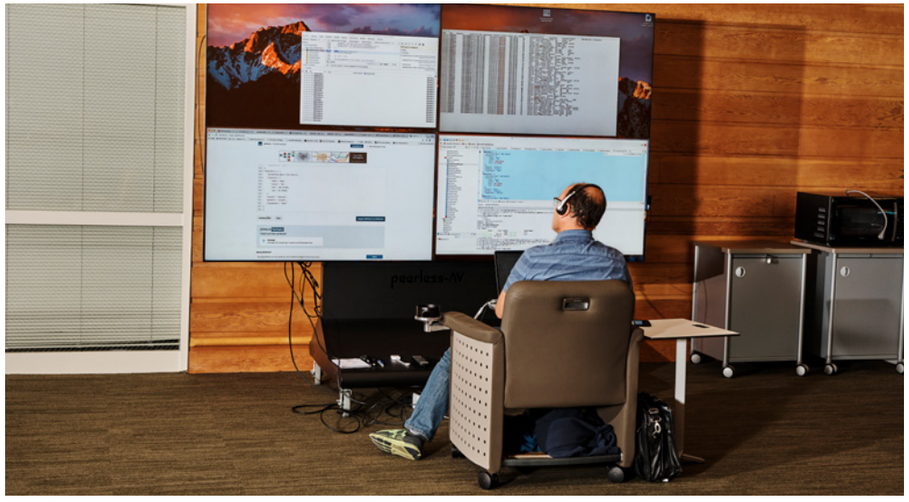
A data fabric uses data virtualization to access disparate sources across an environment, so you no longer need to move data or create duplicate sets. Data virtualization provides a single, seamless view without copying data. As a result, a data fabric enables self-service access, so users can simply query the data where it resides. This capability gives data scientists and other data citizens faster access to information—no matter what platform or geography it's located in. Self-service access also means users can start querying the data immediately, without waiting on a data engineer to find and prepare it first.

- One-third of organizations cite data complexity and silos as the top barrier to AI adoption.[4]

By 2023, organizations using data fabrics to dynamically connect, optimize and automate data management processes will reduce time to integrated data delivery by 30%.[5]

Another key element of a data fabric is the ability to perform universal queries with a query semantic layer that essentially abstracts or translates queries into a universal language. No matter what query engines your organization uses, a data fabric's semantic layer makes it possible to query distributed data much faster than a standard data warehouse and with a higher degree of relevancy.

Simplifying, unifying and connecting data across complex, dispersed environments is critical to building a foundation for successful, timely AI initiatives in your business.



- Data virtualization allows direct access to disparate data sources without data movement; helping to reduce storage and replication costs.

04

Ensure data quality and governance for AI

Making sure data can be quickly and easily accessed by data scientists is crucial for successful AI projects, but equally important is ensuring that the data they're working with is relevant and of high quality. Given the enormous volume, variety and velocity of data enterprises are dealing with, they need effective tools for cataloging, tagging and organizing data, and governing access levels.

According to IDC research, 80% of worldwide data will be unstructured by 2025, and difficulty analyzing semi-structured and unstructured data is a growing challenge to achieving the analytics at scale required for AI and ML.[7]

In addition, business users in different roles—from HR to data scientists to lines of business—need access to different data. Governance policies are needed to safely determine appropriate data entitlements in accordance with data privacy and regulations.

Contextualizing data and the use of data Governance, data quality and data cataloging capabilities are at the heart of data fabric's value. A data fabric automatically profiles data so you understand it in its current form and then classifies it to be fit for purpose—making it easier for people with different roles and experience levels to put organizational data to use.

- Organizations report approximately 90% or more of their time is spent preparing data for advanced analytics, data science and data engineering.[6]

The cataloging function of a data fabric offers several beneficial capabilities:

- Detecting sensitive data for internal and external regulatory purposes
- Creating enforcement policies to protect data through masking, redaction and substitution of values
- Profiling data to make its initial form comprehensible to data consumers
- Recommending data based on a user's history and search patterns
- Discovering data by inventorying assets and applying governance rules
- Assigning business terms from your business taxonomy to new assets

These cataloging features solve many of the historically manual and time-consuming aspects of data preparation and management.

As a result, data consumers can more quickly derive value from relevant organizational data to make informed decisions while adhering to appropriate governance policies.

Discover how global banking and financial services institute ING uses a data fabric in a hybrid cloud environment to improve data access and governance.

Accelerate the journey to AI
To take full advantage of the promise of AI, organizations need well-organized, trusted data that's business-ready for analytics and AI model building. A data fabric introduces automation technologies that help solve many of the challenges and inefficiencies of data management—from access to preparation to governance—getting you one step closer to putting AI to work in your business.

Read the blog

ING got more out of their centralized governed data lake and hybrid cloud environment with the help of data fabric on IBM Cloud Pak for Data.

Discover how →

Nearly 90% of an organization's time is spent preparing data for advanced analytics, data science and data engineering.

05

Manage privacy and compliance of sensitive data

As a digital enterprise, your business likely stores and manages sensitive data sets, including customer and employee personally identifiable information (PII) and intellectual property. Sensitive data can be a significant portion of your data landscape and, therefore, can and should be used to power your AI models. However, it's critical to identify and protect sensitive data and adhere to relevant regulatory requirements. Failure to do so can result in hefty fines, loss of business and loss of reputation.

Many organizations struggle with identifying and managing sensitive customer data. Research by IBM and the Ponemon Institute found that 80% of data breaches in 2020 included customer PII. Out of all the types of data exposed in these breaches, customer PII was also the costliest to the businesses studied.[9]

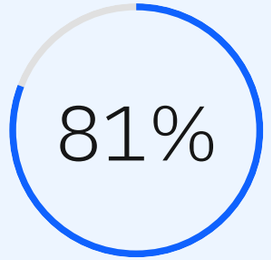
To effectively manage data privacy, organizations need full visibility into data located across their hybrid cloud environment—anywhere in the data and AI lifecycle. Establishing trusted data pipelines equips your teams with the ability to share data safely across the enterprise ecosystem to achieve faster time to compliance, innovation and AI deployment.

A holistic data privacy and security framework helps businesses avoid a piecemeal approach to securing data across disparate sources and using disparate point solutions. This approach enables enterprises to discover, audit and govern sensitive data. With a data catalog—an essential capability in a data fabric—organizations can automate the detection and governance of sensitive information, simplifying the effort to manage privacy and compliance.

Identify sensitive data

The first step to managing data privacy and regulatory compliance is to know what data you're working with.

- The average cost of a data breach is USD 3.86 million.[8]



- 81% of consumers would stop engaging with a brand online following a data breach.[9]

- 80% Of data breaches in 2020 included customer PII, which can cost businesses financially and reputationally.

Consider the following questions:

- Does your team know where your customers' private data is located?
- Do you have control of who can access it?
- How ready is your team to respond to customers' data access requests?

A modern data catalog automates metadata curation, including automatic detection and classification of sensitive data. It also automates core data governance services, including data lineage—a real-time visualization showing the data's complete journey across the lifecycle, policy and rules enforcement, and reference data management.

By automating the ingestion and classification of sensitive data, a data catalog gives you critical visibility of your sensitive data, which is step one in managing privacy and compliance.

Anonymize sensitive data for AI

Once you've identified sensitive data, it must be anonymized to safely use it for AI. In a data fabric, automated privacy features help anonymize PII and other confidential information using the best-fit techniques for any given data set, such as encryption, tokenization, masking techniques and statistical noise. A data fabric also delivers auto-enforcement of an organization's data protection rules. These automated privacy capabilities help reduce risk without excluding sensitive data from AI projects.

Stay compliant and audit-ready

As regulatory requirements evolve and introduce complexity into data management, businesses need a solution that helps ensure compliance and streamline the auditing process. A data fabric's automated metadata and governance layer speeds time to compliance by automating routine, manual governance activities. It provides a PII taxonomy for regulatory standards such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), as well as industry reference data to support data mapping for compliance.

By helping businesses identify and anonymize sensitive data and simplify compliance, a data catalog is central to building a strong foundation for AI using a data fabric architecture.

06

Start building a data foundation for AI with a data fabric

There's no doubt that data is key to digital transformation and the foundation for AI and machine learning success. It's therefore imperative that businesses modernize and deploy an information architecture that's designed to support their AI initiatives.

According to a 2021 study conducted by IBM Institute for Business Value, AI is one of the top 3 technologies CEOs expect will most help deliver results. The study found that 84% of organizations surveyed expected to maintain or increase their level of focus on AI, with nearly a third boosting their AI investments as a direct result of the pandemic.[10]

With an enterprise data fabric architecture and DataOps methodology, your organization can build a foundation of trusted, compliant data that supports self-service access to enterprise data—residing essentially anywhere—without having to move or copy it. With governance, security and regulatory compliance built into the fabric, your organization will be able to address data quality and privacy challenges arising from a hybrid and multicloud data landscape.

Take the next step

Want to learn more about data fabric or talk about how it can support your AI journey? Schedule a complimentary one-on-one consultation with an AI expert.

Talk to an expert →

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
April 2022

01 Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025, Statista, 7 June 2021.

02 Overcome Obstacles To Get To AI At Scale: Invest In And Scale AI To Become An Industry Leader, Forrester, January 2020.

03 Proven concepts for scaling AI: From experimentation to engineering discipline (PDF, 301 KB), IBM Institute for Business Value, September 2020.

04 Global AI Adoption Index 2021 (PDF, 5.9 MB), IBM and Morning Consult, 2021.

05 Magic Quadrant for Data Integration Tools, Gartner, 9 March 2021.

06 The State of Metadata Management: Data Management Solutions Must Become Augmented Metadata Platforms, Gartner, May 2021.

07 80 Percent of Your Data Will Be Unstructured in Five Years, Solutions Review, 28 March 2019.

08 Cost of a Data Breach Report 2021, IBM Security and the Ponemon Institute, 2021.

09 81% of Consumers Would Stop Engaging with a Brand Online After a Data Breach, Reports Ping Identity, Business Wire, 22 October 2019.

10 2021 CEO Study: Find your essential, IBM Institute for Business Value, 2021.