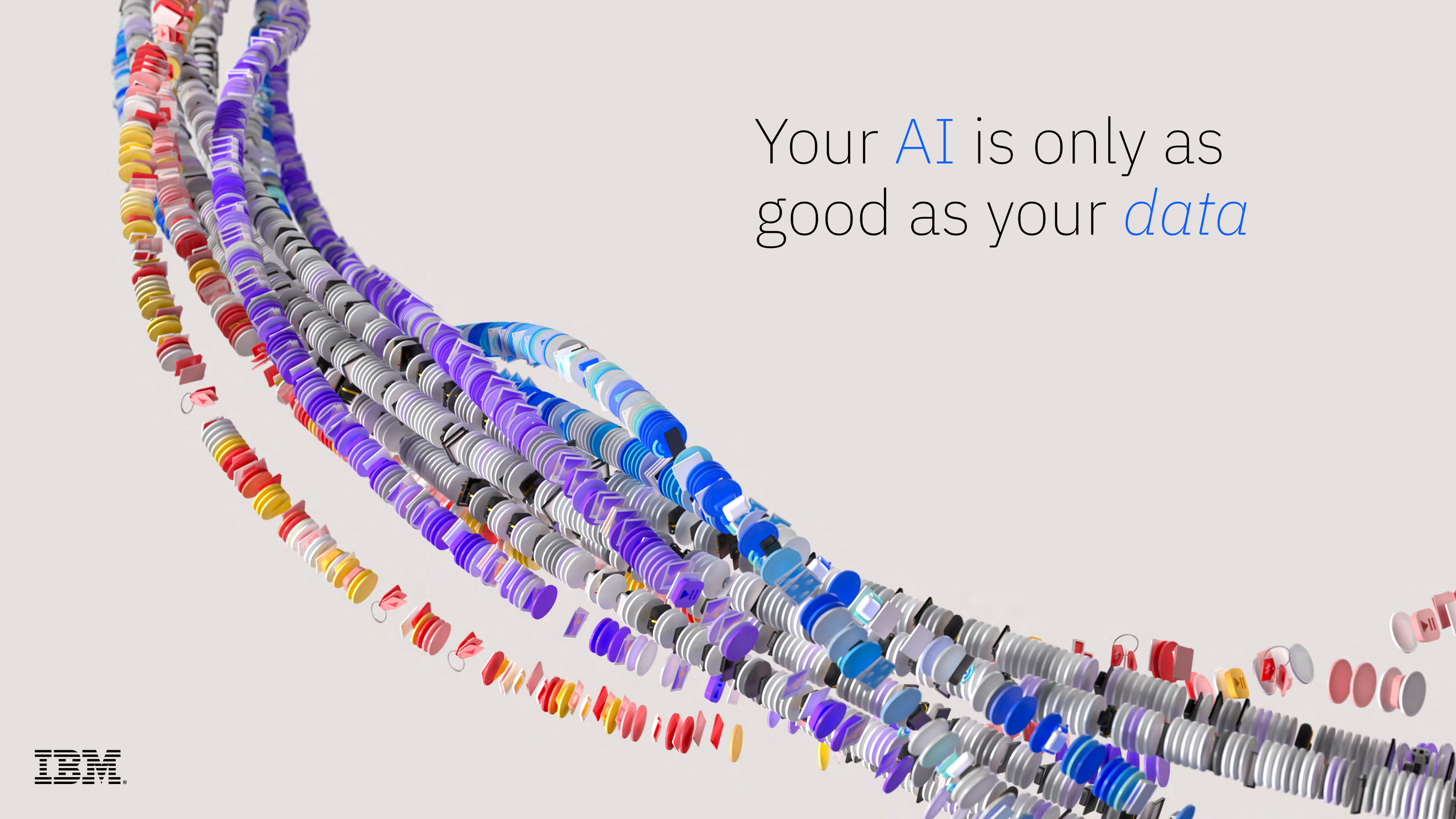


Your AI is only as good as your *data*

IBM


Every company that we work with is a data company. Whether or not they think that data is their main product or not, that is the thing that makes them unique.
Scott Brokaw
Vice President, Product,
Data Integration, IBM

Foreword

Your AI is only as good as your data.

In the AI era, no matter what industry you're in, you're in the data business.

At its heart, AI is only as good as the data that goes into it, which makes your organization's data its most important asset.

But what does it mean to be a data leader? How can you take your company's data and turn it into the solid foundation you require for AI initiatives?

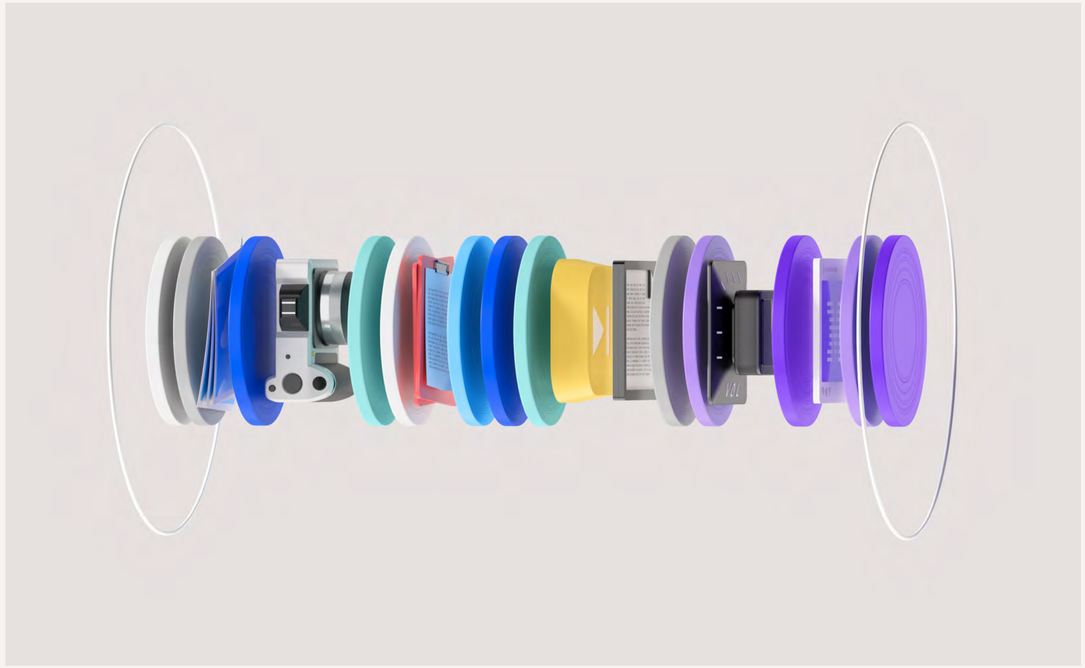
Now is a vital moment where success or failure can determine the future of your organization. Understanding how to properly manage and prepare quality data—both structured and unstructured—is a make-or-break prospect.

The reason is because your data alone isn't enough. For your AI initiatives to function properly, you need to be training them on quality data. Data quality is critical to AI success, playing a huge role in reliable analytics and decision-making. Organizations say data quality has a critical or very high impact on AI success, highlighting its critical role in reliable analytics and decision-making.

To turn your data into the kind of high-quality data that helps enable AI success, it's necessary to know how to prepare and manage all your data to get the best results.

Contents

What is high-quality data?



Not all data is equal.

Organizations will frequently hold onto all their data. There are a variety of good reasons for doing so, including the growing number of data consumers who drive new use cases, ever-changing regulations and a sense that unexpected circumstances might require the use of older data. However, it’s important to recognize that data which needs to be accessed every 10 years isn't the same as data that needs to be accessed every 10 seconds. Additionally, not all data will be relevant for each use case.

A data leader’s focus needs to be on data that’s meaningful and relevant for the use case and business objective at hand. And that data needs to meet a certain threshold of quality to be useful.

When data is incomplete, inaccurate or inconsistent, it can prove a detriment to your AI training and initiatives. That’s why it’s vital to focus on meaningful data you can use to create value.

To create that value from your data, you need to know 5 key pieces of information:

- What kind of data you have
- Where it’s located
- How good it is
- Whether you can use it
- If it’s relevant for your use case or business objective

But what makes for good or meaningful data? The following are a few examples:

- Data with a complete understanding of lineage, quality and service-level agreements (SLAs)
- Business and technical metadata that’s rich and descriptive
- The highest-value structured and unstructured data
- Data supporting regulatory compliance and reporting
- Data across multiple systems that feed into downstream apps

Using meaningful data is not, on its own, sufficient to ensure you avoid low-quality data.

29%

of tech leaders strongly agree that their enterprise data meets the standards to support scaling of generative AI.

According to an IBM Institute for Business Value 2024 survey of technology leaders, less than 1/3—only 29%—of surveyed tech leaders strongly agree their enterprise data meets the quality, accessibility and security standards that support the efficient scaling of generative AI (gen AI).[1]

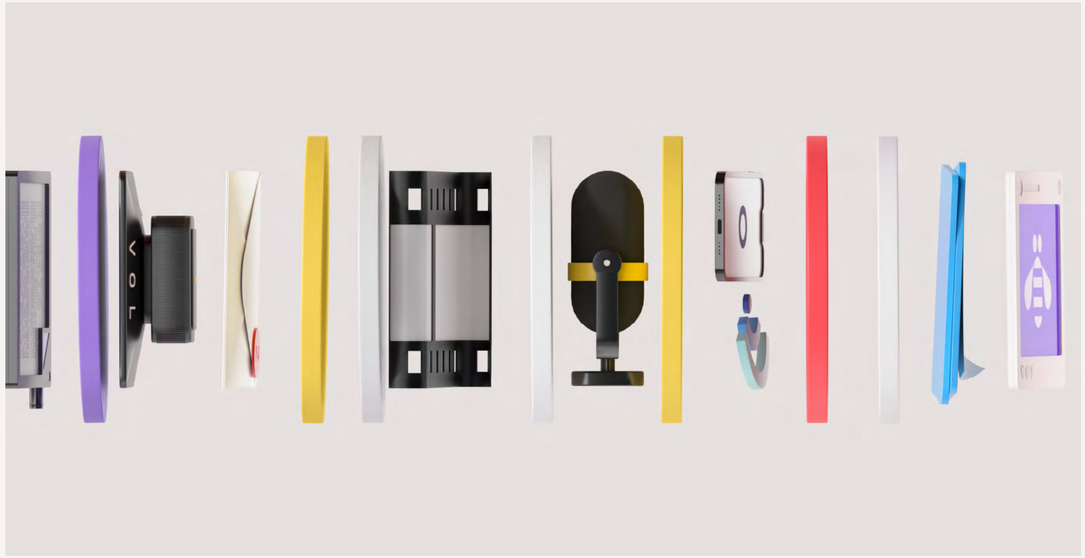
Why are these leaders so worried? Because poor data quality can cause a negative impact on ROI, increased occurrence of incidents, duplication of data, fines for regulatory noncompliance, long-term reputational damage, erosion of customer trust and lost revenue.

Low-quality data can take many forms, including data that's untrustworthy, inconsistent, irrelevant, siloed, inaccurate, repetitive, or ungoverned and unregulated. These types of datasets can directly hinder AI effectiveness.

And while data quality has always been important, in the era of gen AI and agentic AI it's even more critical for an organization's success.

However, differentiating high-quality from low-quality data is easier said than done for many organizations. The obstacles that stand in the way may not be obvious and can often offer a daunting level of complexity.

Barriers to high-quality data



Many of the barriers to high-quality data result from the ways in which an organization’s data and workflows are interrelated.

One of the biggest challenges organizations face in this regard is understanding the linkage and the relationships between datasets and within workflows. These relationships help organizations understand where data is used across an enterprise, which can lead to understanding whether a dataset is of high quality or not.

The barriers to this understanding, though, are multiple, including:

Data integration
As teams patch together individual tools to pursue solutions that better align with their use case, data teams find themselves wrestling with complex, unintegrated pipelines, thus accumulating technical debt and diminishing the long-term value of their data assets.

Fragmented and evolving data environment
A scattered landscape of ever-evolving data coming from an increasing number of sources across SaaS, multicloud and hybrid environments can create silos, making it hard to maintain consistency, control and visibility.

Limited data visibility and lineage
Difficulty tracking the origin and flow of data across various systems can compromise the reliability and usability of that data.

Lack of support for structured and unstructured governance
Without organizational support, there will be significant challenges in establishing any governance framework, hindering the effective use of data.

Talent gap
Many organizations that attempt to prepare their data for AI simply don’t yet have the talent and bandwidth needed to solve all these problems. They need to make changes to their organizational process, structure and culture to succeed.

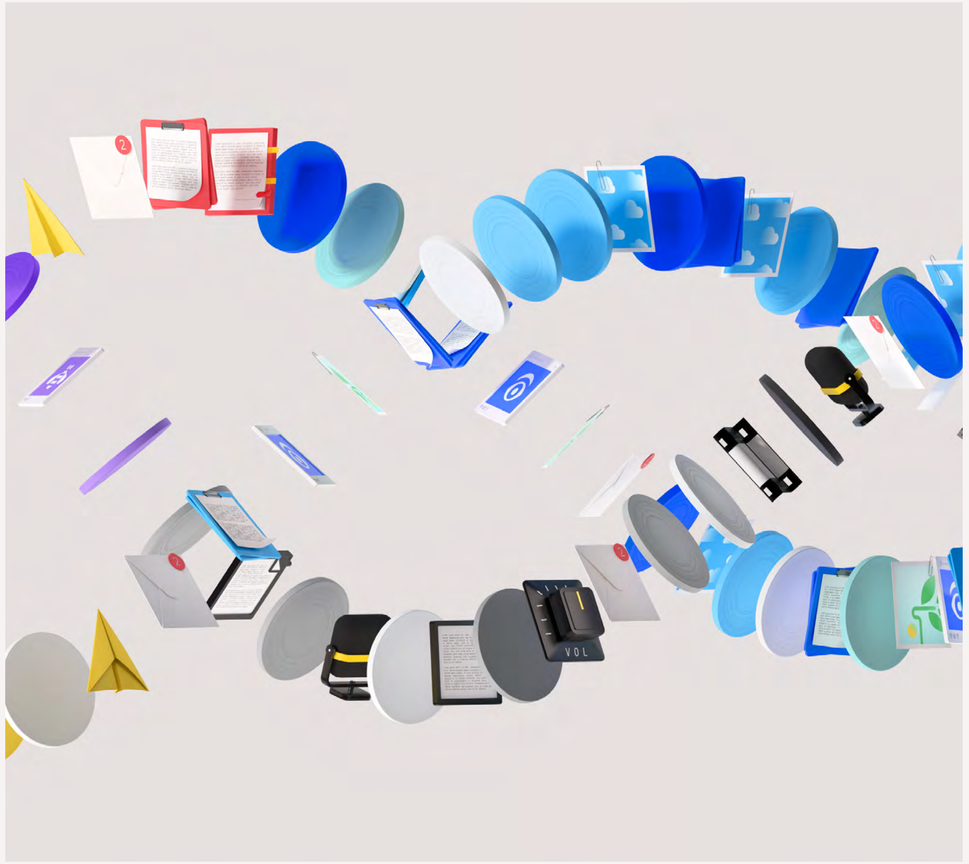
What’s more, high-quality data will not only draw from structured sources, but also unstructured ones.

Unstructured data, such as data inside emails, documents, presentations and visuals, can be immensely difficult to harness. The data locked inside these diverse formats is highly distributed and dynamic, lacks neat labeling and often requires added context to be fully understood.

Yet in order to obtain a complete understanding of its data and a strong data foundation for AI, an organization needs both structured and unstructured data.[2]

Fortunately, there are several approaches available to organizations to ensure their data is—and remains—high quality. The right strategy for any organization will vary, depending on the current state of their data, the root causes of poor-quality data, the skills available and the processes in place. However, a general 3-step framework has shown to be relevant for many organizations.

How to manage and prepare high-quality data



Ensuring you’re managing and preparing your high-quality data can be a complex and daunting task, but IBM recommends a 3-step framework that focuses on data intelligence and integration.

1.

Know your data
Understand what data you have—especially your technical metadata and its relationships to business metadata—to build a foundation.

2.

Score your data
Create a consistent methodology to assess the quality of your data, scoring it so you can compare it to other datasets across the organization, and continue to observe and assess that data over time.

3.

Observe your data
Establish the relationships and lineage of data and workflows to their consuming patterns, data pipelines, reports, data products, AI products and so on. This way you can build out a heat map showing which datasets create the greatest risks if they’re of poor quality.

It’s not enough to just *have* high-quality data. That data also needs to be carefully managed and prepared for use in your AI and gen AI initiatives.

1.

Know your data

Understanding your data includes not just knowing what data you have, but also knowing which data is high-quality. For data to reach this threshold, IBM uses and recommends a consistent set of criteria:

Accuracy: Data values are as close as possible to real values and correctly represent real-world entities or events.

Completeness: All required data fields, records and attributes are present with no gaps or missing values.

Consistency: Data values remain uniform and free from contradictions or conflicting values across different datasets or systems.

Uniformity: Data within a data asset is uniform and consistent over time, with all data points sharing similar characteristics, formats or structures.

Timeliness: Data is available when needed and up to date to reflect the current situation at the time of its use.

Uniqueness: Each distinct value or transaction appears only once within a column of data, with no repetition.

Validity: Data adheres to predefined formats, business rules and constraints, ensuring it’s usable and meaningful.

2.

Score your data

To ensure that your data scores well on each of these dimensions, you need a thorough understanding of good governance and the implementation of good governance policies. Good governance of data requires you always know the following:

- The roles and personas that interact with the data, including their responsibilities and authorities
- The source for critical data and master data
- The location of the most critical data in the organization and the meaning of that data
- The standard processes to acquire, maintain, change, use and dispose of data through its lifecycle

Furthermore, sustaining data quality requires change management, so you can continually understand how good your data is and how good it could be.

Change management means recognizing every change in organizational design, resources and technology. This understanding may require revisiting the assumptions and implementation of your governance process, as well as the configuration of the supporting architecture and tooling in the data environment.

Thus, it’s not enough to just *have* high-quality data. That data also needs to be carefully managed and prepared for use in your AI and gen AI initiatives, which means breaking down silos between that data, as well as being able to access and manage unstructured data.

3.

Observe your data

Managing and preparing data requires observability and quality control with continuous and automatic scoring, data incident detection and remediation, plus the ability to ensure that data is ready for its intended purpose or use case.

A good data framework for data management and preparation includes a focus on data catalogs, data integration, data observability and data lineage, with the ability to:

Centralize pipeline metadata and analyze it with machine learning (ML)-powered detection to enable continuous monitoring for data anomalies.

Use a data catalog to organize and understand the data that you have.

Catalog and profile alerts around data incidents, creating and routing custom alerts to data stakeholders in real time.

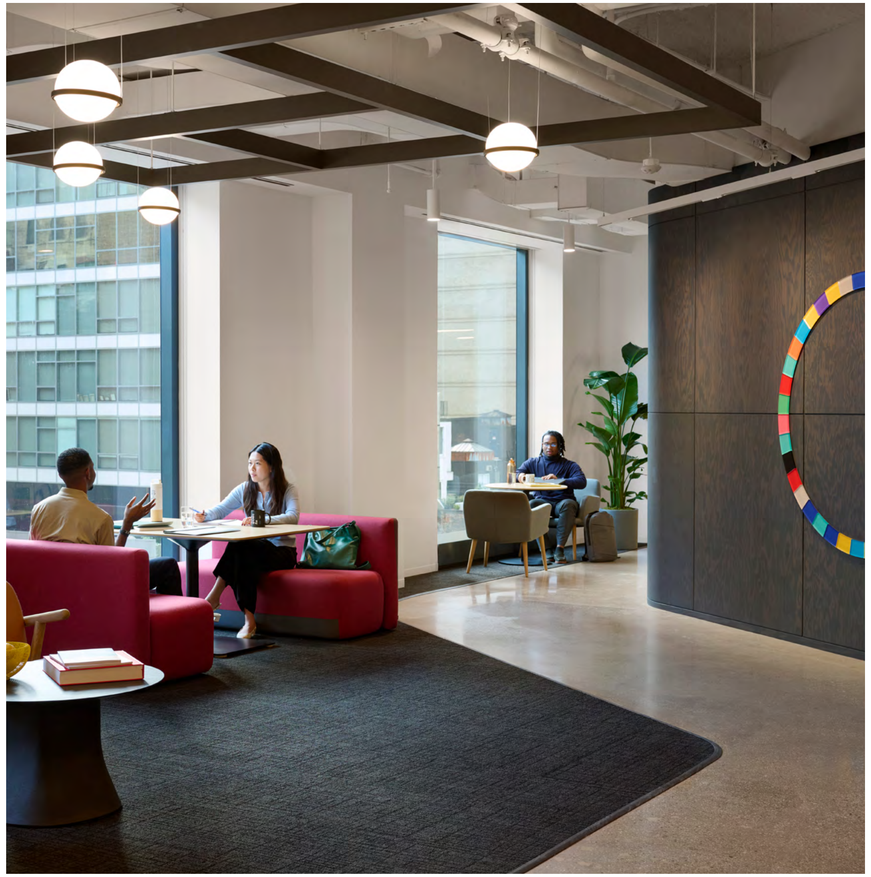
Deliver trusted data to data consumers by detecting data incidents earlier and resolving them faster with continuous data observability.

Transform high-velocity data into actionable intelligence, in real time, by integrating streaming data across hybrid cloud environments.

Select a sample subset of data for use cases and split it into training and test datasets.

The ultimate goal of this process isn’t to ensure all your enterprise data is high quality before you ever implement any AI. Rather, managing and preparing your data is an ongoing process that needs to become part of your permanent workflow.

By starting small, focusing on a particular workflow or domain, you can establish a use case that will provide value, have an impact and serve as an origin point for building repeatable patterns. This method will get you support for the next use case so you can create momentum, implementing new data quality processes through incremental improvements that provide value along the way.

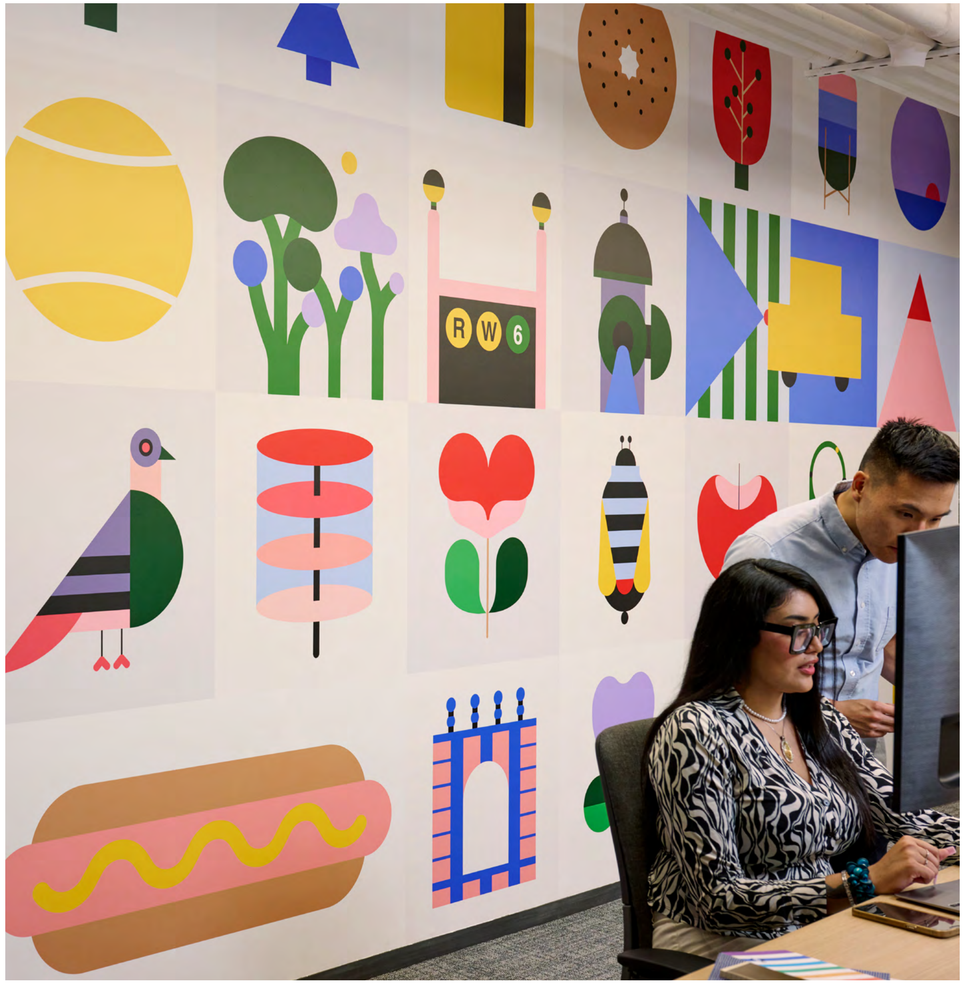



The job of data quality
is never done.
Scott Brokaw
Vice President, Product,
Data Integration, IBM

Solutions

Fortunately, IBM offers a variety of solutions that can help with managing your data, preparing it and helping ensure that it’s high quality.

Explore them now to learn more.



1. 6 blind spots tech leaders must reveal. How to drive growth in the generative AI era

2. Agentic AI has an unstructured data problem; IBM is unveiling a solution