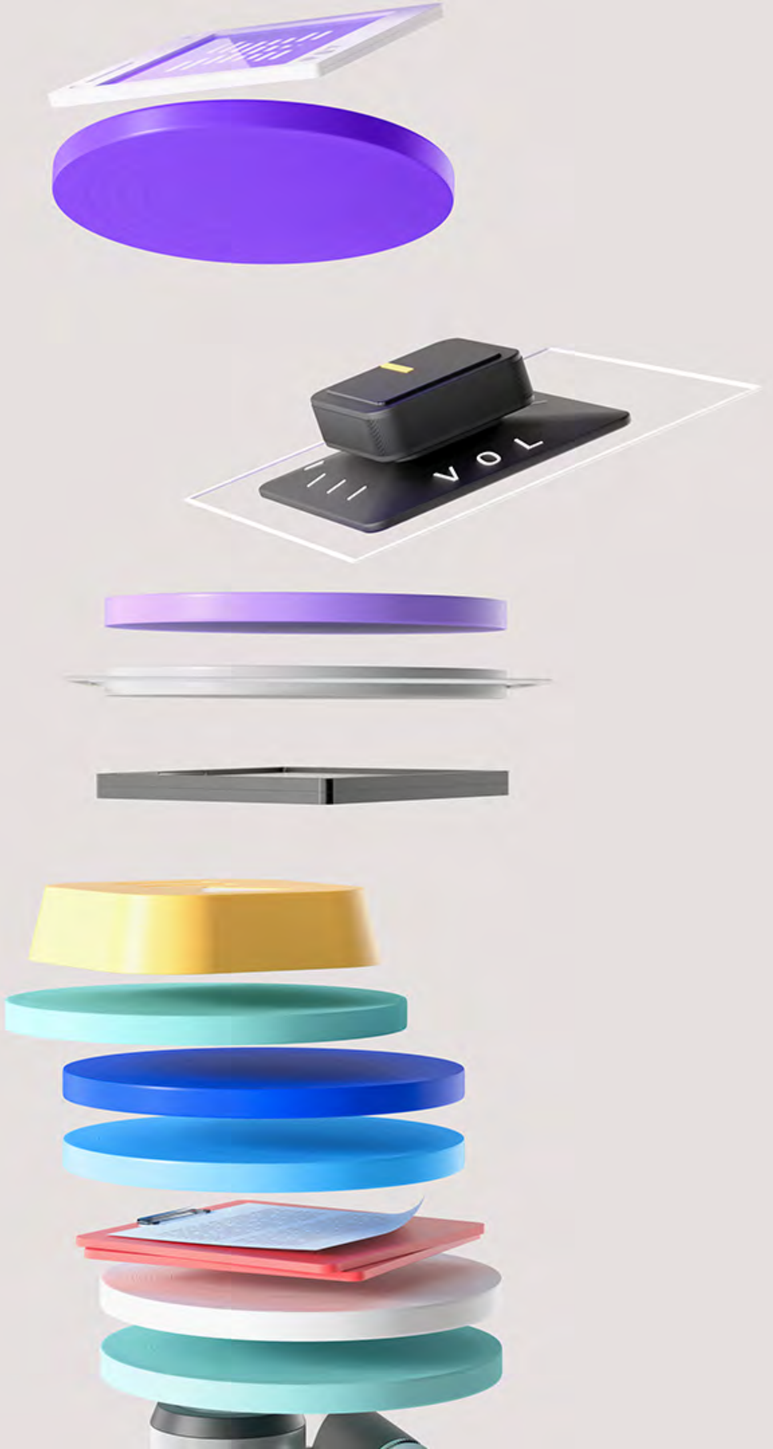
watsonx.data

The only hybrid, open data lakehouse for enterprise AI and analytics



Contents

01

Why your enterprise needs a gen AI-ready data lakehouse

Generic data produces generic results. If the data isn't specific to your business, the results won't be either. When everyone is using the same public data to train the same generative AI (gen AI) models, how can you outperform your competitors?

Your competitive edge lies within your unique data. Unstructured data—the type buried in emails, pdfs, images and videos—is especially valuable, but very difficult to operationalize. Organizations need a smart data architecture to access, prepare and deliver both structured and unstructured data to unlock the full potential of their enterprise data for AI.

Data is the fuel for AI, but it is estimated that less than 1% of enterprise data is being leveraged by AI models.[1] And 90% of enterprise data is unstructured.[2]

A hybrid, open data lakehouse can help you supercharge your AI and infuse your data into a wide variety of modern use cases. It unifies all data types, simplifies integration and makes managing and governing data—structured and unstructured—easier than ever before. No wonder it is emerging as the preferred long-term data architecture for analytics and new gen AI workloads.

90%

of enterprise data
is estimated to
be unstructured.[2]

02

Current and emerging data challenges

As the importance of enterprise data has increased, so have the challenges associated with it. The sheer volume of data has become overwhelming, and it's often siloed across an organization. Additionally, new data varieties have complicated integration, and low-quality data has compromised the impact on AI. Gen AI can help solve these problems, but a resilient and flexible data architecture is essential for it.

Today, limitations of retrieval-augmented generation (RAG) prevent enterprises from realizing the value of unstructured data for gen AI. So, how do you bring unstructured data into gen AI and traditional analytical workloads? By adopting a hybrid, open data lakehouse.

03

Data lakehouse defined

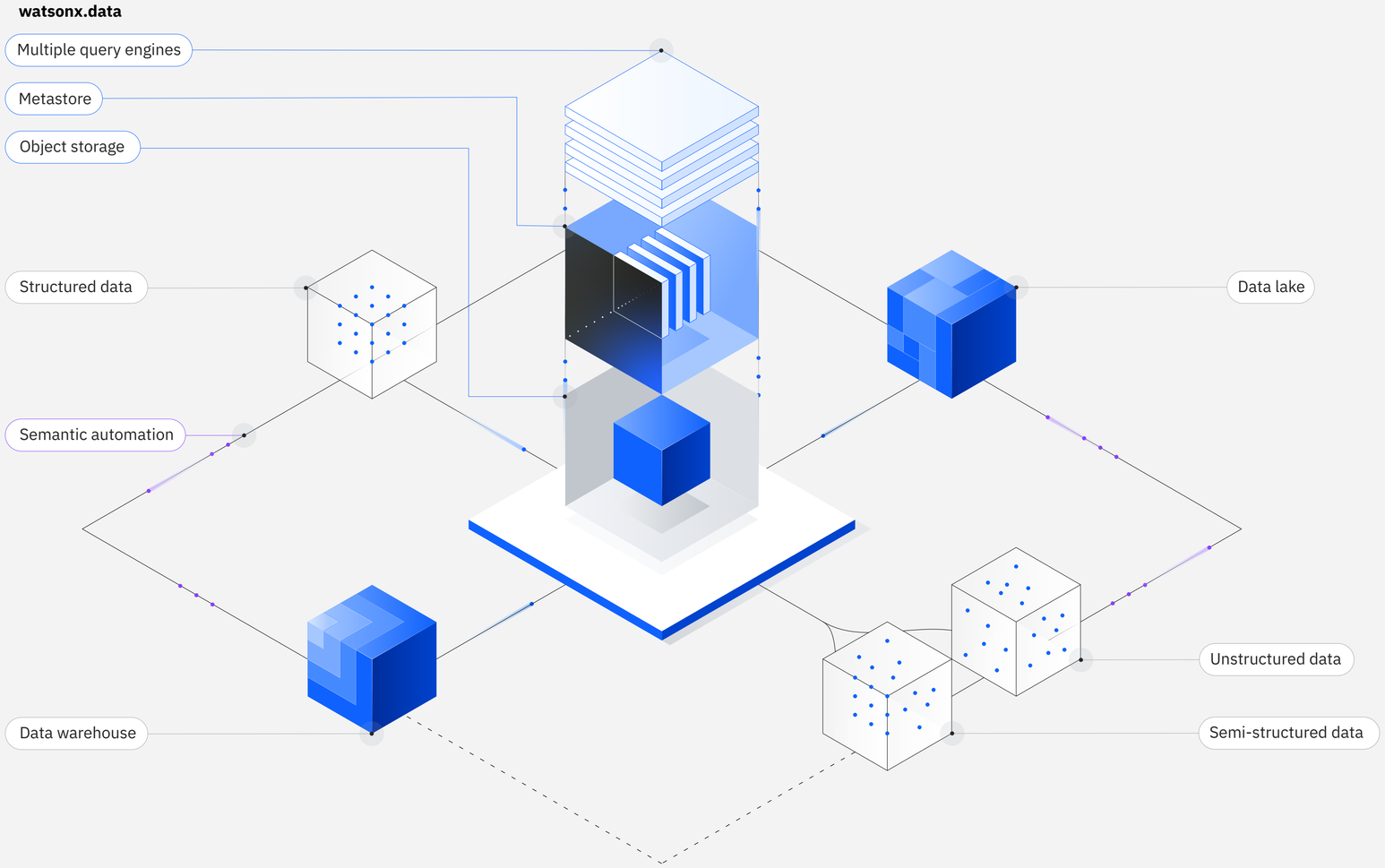
Enterprise data is at the heart of any AI architecture. Organizations pursuing gen AI tools in 2025 will need to make sure their data is AI-ready. A lakehouse solution provides a modern foundation to scale AI-ready data.

The data lakehouse is an emerging architecture that offers the flexibility of a data lake with the performance and structure of a data warehouse. Most lakehouse solutions offer a high-performance query engine over low-cost storage in conjunction with a metadata governance layer. Intelligent metadata layers make it easier for users to categorize and classify unstructured data such as video and voice, and semi-structured data such as XML, JSON and emails.

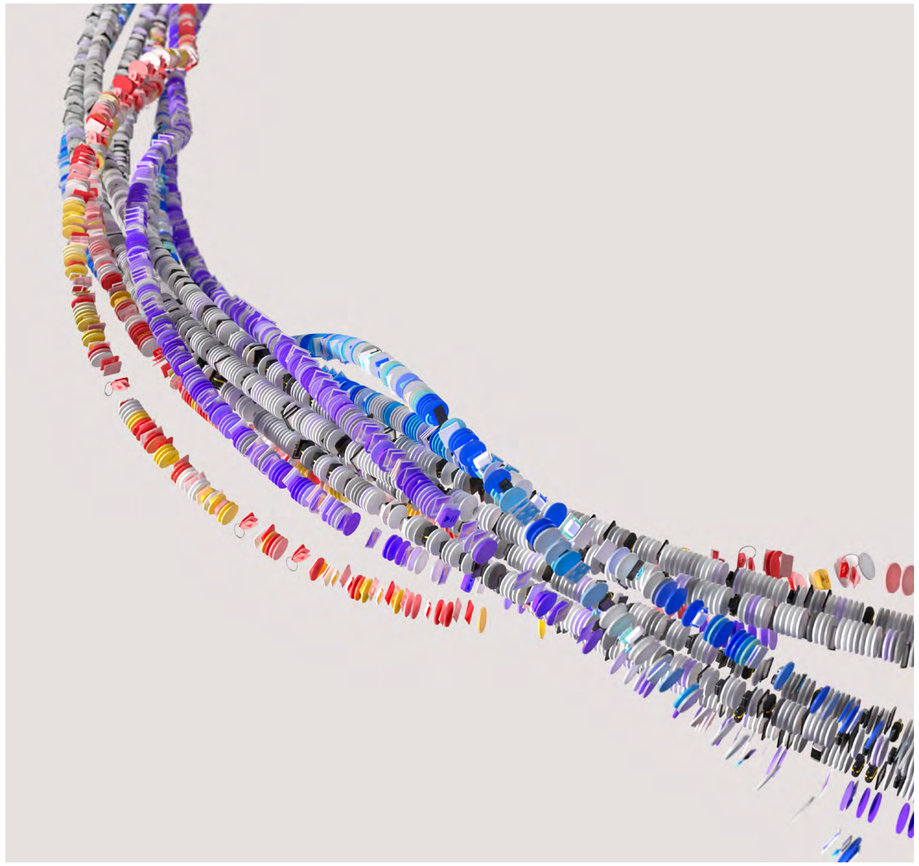
The best data lakehouse will offer open-source technologies that reduce data duplication and simplify complex pipelines. However, some first-generation lakehouses have extract, transform, load (ETL) constraints that limit their ability to address the challenges of cost and complexity. For example, a single query engine that's designed for business intelligence (BI) or machine learning (ML) workloads could be ineffective when it's used for another workload type.

The IBM data and AI teams believe that every workload is unique and should be optimized with the most suitable environment that keeps cost at a minimum and performance at a maximum. Ultimately, turning raw data into actionable insights at scale requires a lakehouse that delivers cost-effective performance and format flexibility.

Manage the entire lifecycle of data for AI



Meet watsonx.data

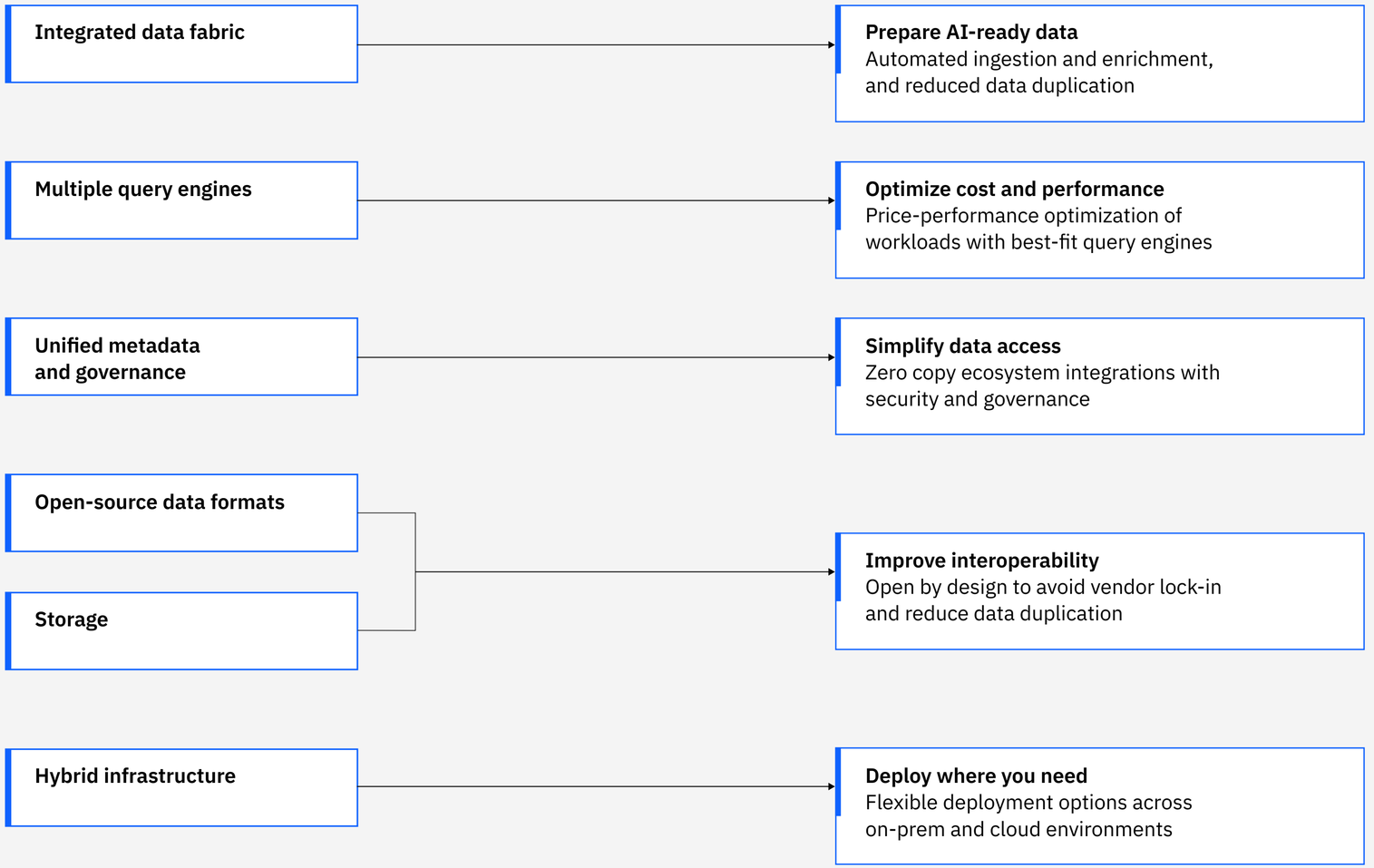


IBM® watsonx.data® is the only hybrid, open data lakehouse for enterprise AI and analytics.

It helps you manage the entire lifecycle of enterprise data for AI within your data lakehouse, to power the next generation of gen AI and AI-infused BI applications.

With watsonx.data, you can simplify and scale structured and unstructured data integration, management and governance across on-premises, cloud and multicloud environments. It’s part of an open, modern data stack utilizing innovative open-source technologies that integrate with your existing data estate, with no vendor lock-in.

IBM watsonx.data seamlessly integrates the capabilities of a data lakehouse and data fabric.



What are the components of watsonx.data?

A hybrid, open data lakehouse architecture provides a unified platform to address the key data challenges businesses face today—large volumes of data, multiple locations and formats, data quality and governance.

Here's how IBM watsonx.data addresses these challenges.

Infrastructure
The infrastructure is where your lakehouse will be deployed—fully managed across any cloud or on-premises environment.

Storage
The storage layer is where the data is physically stored as files or in open data formats such as Apache Parquet and Avro. Open data formats are file specifications and protocols made available to the open-source community so that anyone can ingest and enhance them.

Fit-for-purpose query engines
Multiple, fit-for-purpose query engines enable queries to be efficiently executed for large datasets. With workload optimization across multiple query engines and storage tiers, you can minimize the cost of your data warehouse while providing fast, reliable and efficient big data processing at scale.

Open table formats
Open table formats such as Apache Iceberg help provide structure and deliver the reliability and simplicity of SQL with big data. These formats allow different engines to access the same data, at the same time, which helps avoid vendor lock-in. Data can also be shared across multiple tools and data repositories such as your data warehouse. A single copy of data lets you reduce data duplication and break down silos.

Metadata store
The metadata store in watsonx.data is based on the open-source Hive Metastore, which helps ensure interoperability and allows external tools to easily access your metadata.

Governance
Metadata is also stored in open table formats. It serves to define the file formats for any tool that can read or write open data formats.

Technical metadata service
The technical metadata service component is required to understand what data is available in the storage layer. The query engine uses metadata to provide full lineage information so you know where your data is located, what it looks like and how to read it.

Data catalogs
Data catalogs help users find the correct data for the job and deliver semantic information for policies and rules. Business metadata such as business terminologies and tags are stored to support search and data protection.

Policy engine
The policy engine enables users to define data protection policies and enables the engine to enforce those policies. To create a governance framework that's scalable, a policy engine is often deployed with the technical metadata service and the data catalog.

Query engine
The query engine, often known as the compute component, is at the heart of the open data lakehouse. It can be open source or proprietary, and it accesses data in open table formats.

In an open lakehouse architecture, the query engine is fully modular, which means that the engine can be dynamically scaled to meet workload demands and concurrency. Query engines can also attach to any catalog and storage system.

What differentiates IBM watsonx.data from the rest?

Hybrid and open
Watsonx.data is built with foundational open-source technologies for community innovation and support—with no vendor lock-in. You can access data and deploy anywhere across hybrid cloud environments, in-client VPC, multicloud and on-premises environments, and as SaaS.

Workload-optimized
Underpinned by integrated data fabric capabilities, including data governance, quality and lineage, watsonx.data helps you manage growing data volumes with a unique multi-engine architecture optimized for price-performance.

Gen AI-ready
Watsonx.data enables consistent end-to-end user authorization from document source systems through retrieval.

How does it all come together?
Watsonx.data enables organizations to unlock the business-critical information in their unstructured enterprise data for more accurate, powerful gen AI and analytics.

40%

more accurate AI than conventional RAG with watsonx.data[3]

Ingest at scale

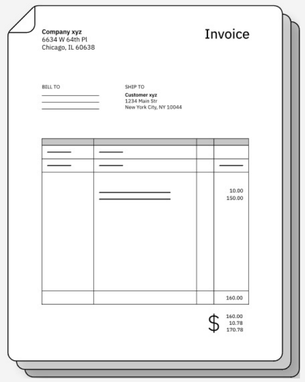
Curate and govern

Deliver to AI

Document systems

- IBM FileNet®
- Box
- SharePoint
- Google Docs

Unstructured data



watsonx.data integration

Unstructured data preparation

- Extracted entities/values <JSON> [Iceberg]
- Extracted text [Iceberg]
- Vector embeddings _Milvus _AstraDB

Document library

watsonx.data intelligence

Multi-engine optimization

- DataStax
- Milvus
- Spark
- Presto
- IBM Db® & IBM Netezza®

Better AI and analytics

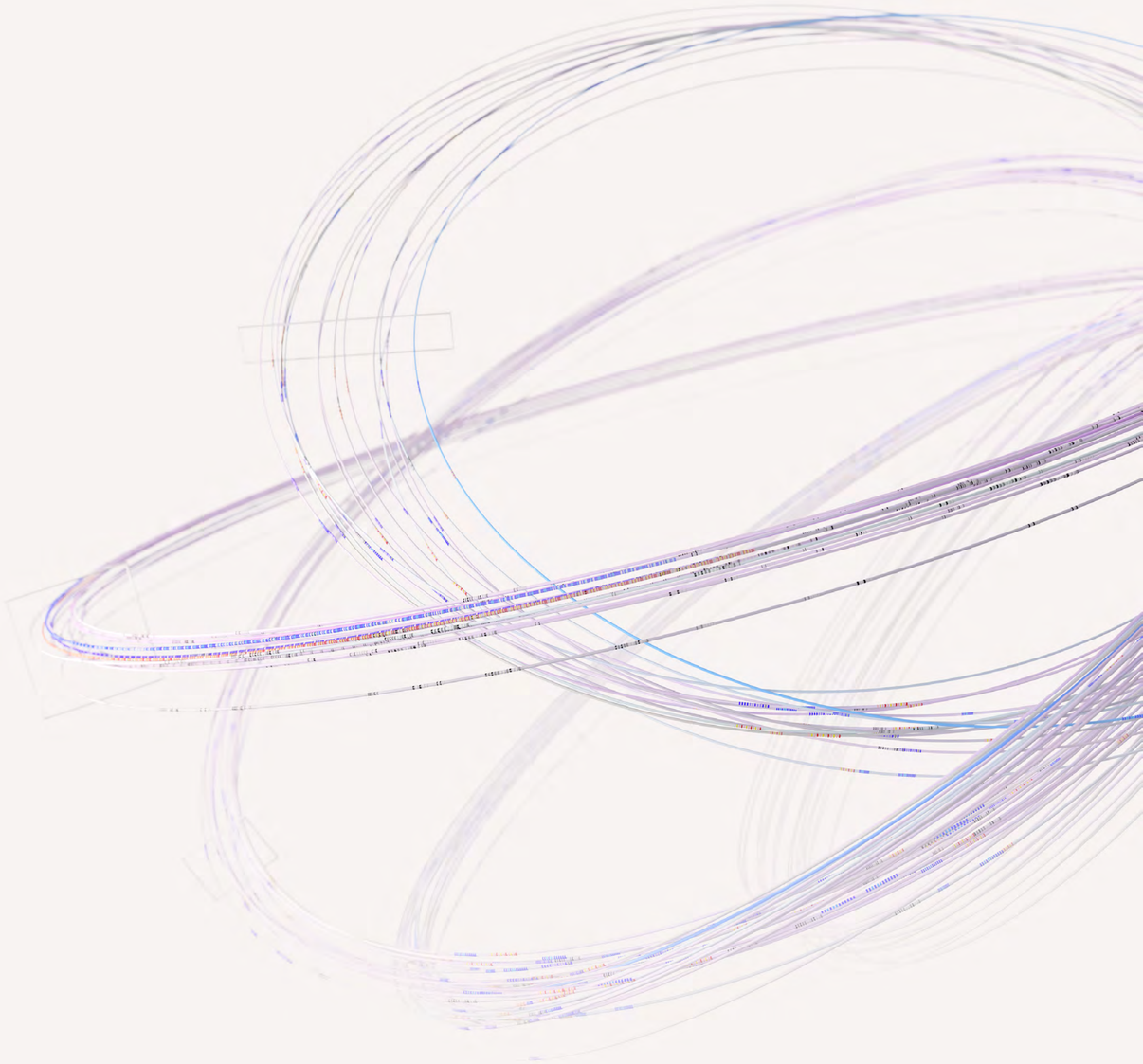
- Gen AI applications
- Gen AI assistants
- Data engineering and AI/ML
- Business intelligence
- Data warehouse

Consistent user access controls

05

Why choose watsonx.data

- To unify structured and unstructured data for enhanced AI accuracy and ROI
- To unlock untapped data for conversational analytics
- To simplify and scale structured and unstructured data integration, management and governance
- To empower all data users with high-quality, trustworthy data
- To protect your existing data investments with an open, modern data stack



06

Next steps

Take advantage of the IBM data management and optimization knowledge honed by decades of handling the world's most demanding data workloads. See how quickly you can maximize the ROI of your data with watsonx.data.

Start your free trial →

Request a demo →

1. The Future of AI is Open, IBM, 2024

2. IDC, UNTAPPED VALUE: What Every Executive Needs to Know About Unstructured Data, 2023

3. Based on internal testing comparing the answer correctness of AI model outputs using watsonx.data Premium Edition retrieval layer to vector-only RAG on three common use cases with IBM proprietary datasets using the same set of selected open-source commodity inferencing, judging and embedding models and additional variables. Results can vary.