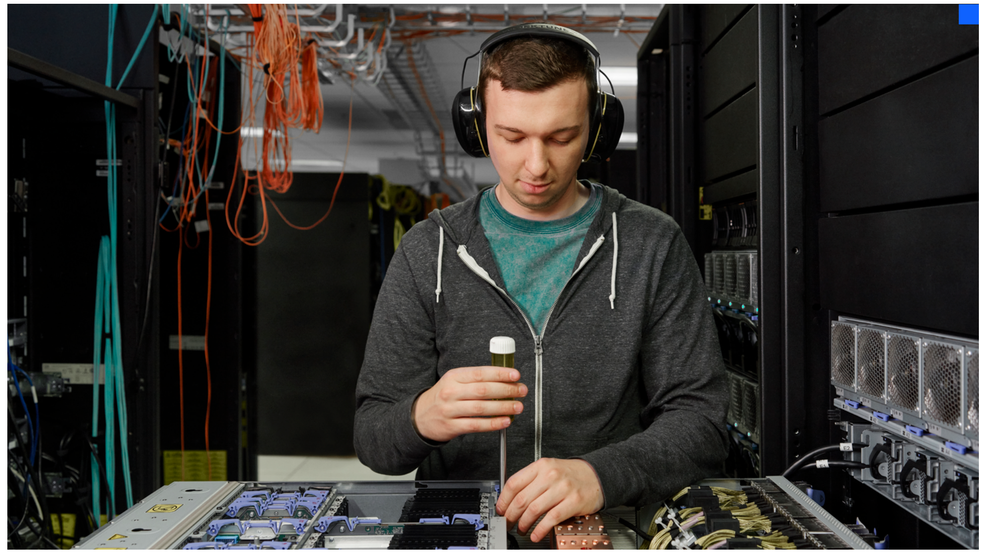
watsonx.data™

The data store
↳ for AI

IBM

Contents



01

Introduction

This ebook will examine the latest open data management solution for data and analytics leaders who want to significantly reduce cost, simplify data access and automate unified governance to scale AI. It's time for the data lakehouse.

Data is at the center of every business. It keeps applications running, powers predictive insights and enables better experiences for customers and employees. But the full benefit of data is elusive because of the way that data is stored and accessed for analytics and AI.

You're not alone if you rely on monolithic repositories with multiple data warehouses and data lakes, on premises and on cloud; 82% of organizations are inhibited by data silos.[1] And it's about to get worse: according to IDC, the amount of stored data is expected to grow 250% by 2025.[2]

The data lake was supposed to fix all these issues; just land your data in a centralized place and process it. But it's not so easy to update the lakes, properly catalog data or ensure good governance—and the skillsets required for these tasks are specific, rare and expensive. As a result, data lakes have proven costly to build and maintain. A data warehouse does offer high performance for processing terabytes of structured data. But warehouses can become expensive, too, especially for new and evolving workloads. Most organizations run analytics and AI workloads in ecosystems that are complex and cost inefficient. It's time for a change.

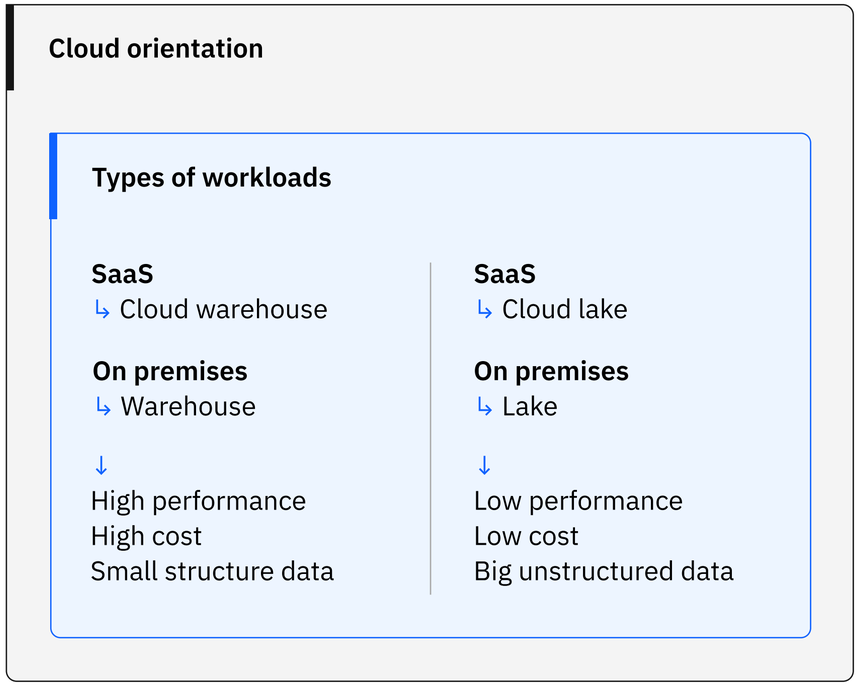
↑250%

The amount of stored data is expected to grow 250% by 2025.[2]

02

The current state of data architecture

A combination of on-premises and cloud-native warehouses and bespoke data lakes is common for enterprise architecture today. You likely find that juggling cost, siloed data and data governance are constant challenges.



The data lakehouse is an emerging paradigm shift in how enterprises surface insights.[3]

03

The data lakehouse defined

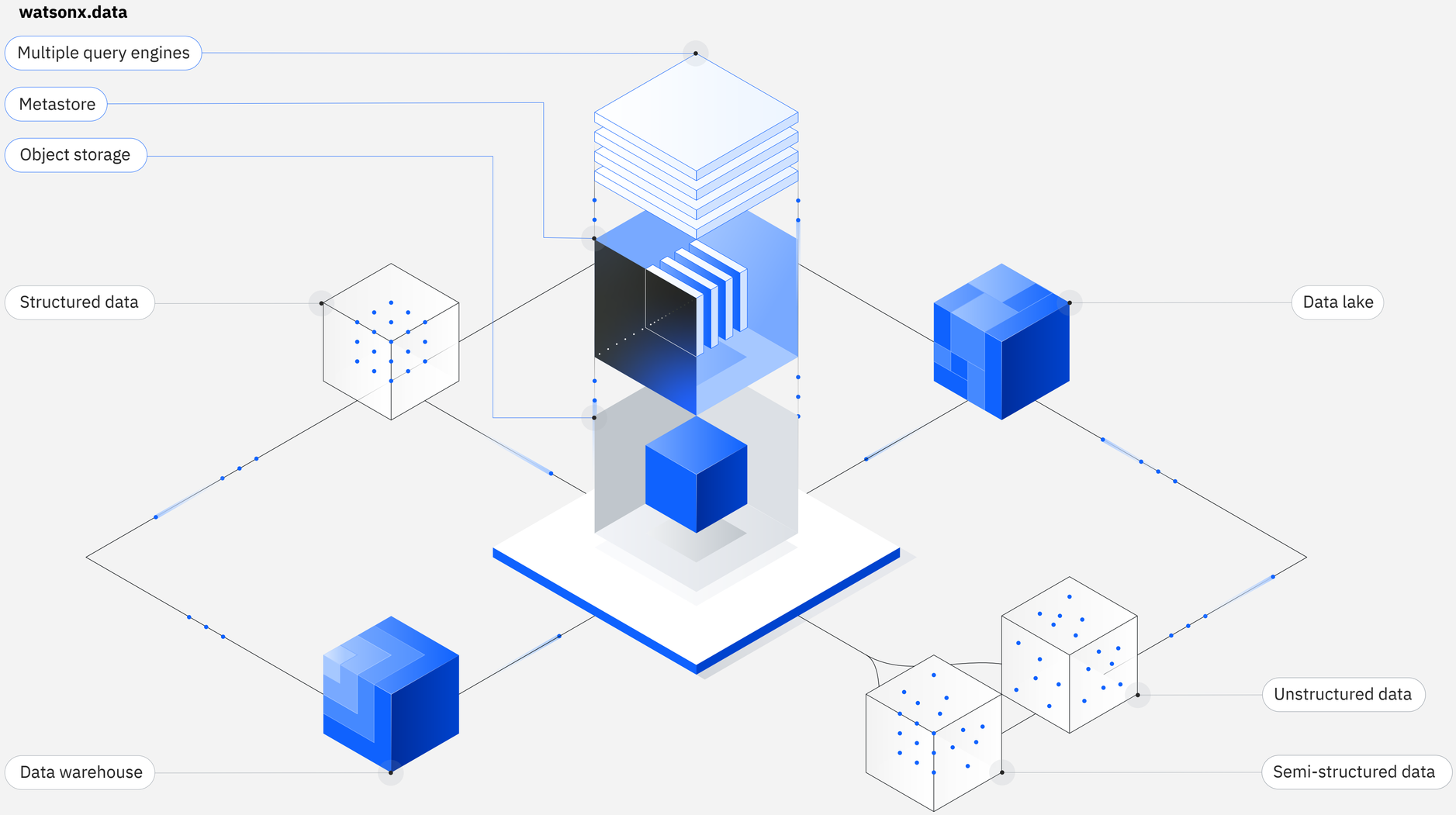
Seek out a lakehouse solution that provides a modern data foundation to scale AI.

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 ETL pipelines. Be aware that some first-generation lakehouses have key 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 or machine learning (ML) workloads could well be ineffective when it's used for another workload type.

The IBM data and AI team believes that every workload is unique and should be optimized with the best-suited environment that keeps cost at a minimum and performance at a maximum. Choose a lakehouse that delivers an optimal level of performance for better decision-making, along with the flexibility that's necessary to unlock value from all types of data.

Figure 1. How to best scale and accelerate the impact of AI



Components of the architecture

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

Storage
This layer is where the data is physically stored, which is stored as files and can be stored 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.

Open table formats
Open table formats, such as Apache Iceberg, help you 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. Share data 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.

Governance

Metadata is also stored with open table formats; it serves to define the file formats for any tool that can read or write open data formats.

Technical metadata service
This component is required to understand what data is available in the storage layer. The query engine requires the metadata for the data and tables to provide full lineage and know where it's located, what it looks like and how to read it.

Data catalogs
This component helps users find the correct data for the job and delivers semantic information for policies and rules. Expect to store business metadata such as business terminologies and tags to enable search and data protection.

Policy engine
This component 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
This component is at the heart of the open data lakehouse. A query engine, which can be open source or proprietary, accesses data in open table format and is often known as the compute component. Query engines typically come in two types: an SQL-based query engine, such as the open-source Presto, or an open-source Apache Spark engine or its equivalent.

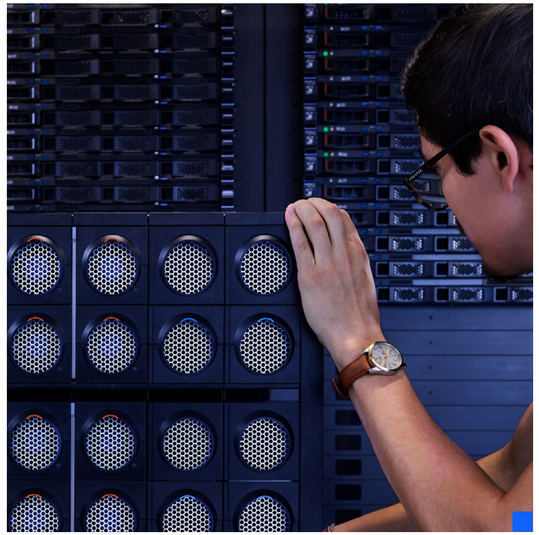
In an open lakehouse architecture, the query engine is fully modular, which means that the engine can be dynamically scaled to meet workload demands and concurrency. Query engines can also attach to any catalog and storage.

↓50%

Now it’s possible to achieve faster, trusted insights while you cut data warehouse costs in half.[4]

05

Cost optimization opportunities



If your organization has existing on premises big data implementations, a lakehouse offers a less-expensive alternative for storing data in open formats on object storage. You'll lower the cost of analytics, decrease complexity and improve time to value.

If you have an existing warehouse implementation, a lakehouse approach can represent a massively scalable, lower-cost alternative for your large analytics workloads that are less sensitive to service-level agreements (SLAs). Warehouses are often expensive and proprietary—but with a lakehouse, you can dramatically reduce storage and compute costs. You can optimize warehouse workloads using fit-for-purpose engines that are based on your workload requirements. The open nature of a lakehouse frees you from proprietary warehouse technology, which means less vendor lock-in and a reduction in IT infrastructure overhead costs.

IBM watsonx.data is an open, hybrid, and governed data store optimized for all data, analytics, and AI workloads.

Analytics and data science enhancements

> “We are moving in the direction where the data lakehouse becomes a best practice.”[3]
>
> **Adam Ronthal**
> Vice President
> Gartner

Proprietary data formats and high storage costs limit AI and ML model collaboration and deployments within a data warehouse environment; data lakes are challenged with low-performing data science workloads. The isolation of these technologies has led to downstream infrastructure challenges, along with the security and governance implications that come with the duplication and movement of data for development of AI and ML models.

A data lakehouse is a great way to help colleagues who are hungry for the insights that lie waiting in your organization’s data. If you’re serious about extracting business value from the firehose of data that’s coming at you, do consider the lakehouse strategy.

Adam Ronthal, vice president and analyst at Gartner, says that “We are moving in the direction where the data lakehouse becomes a best practice.”[2] The best approach will offer an open, collaborative and governed environment for the end-to-end management of data science workloads.

Let’s examine IBM® watsonx.data™—the open, hybrid, and governed data store that’s optimized for all data, analytics, and AI workloads.

07

IBM watsonx.data

Scale AI workloads, for all your data, anywhere. Watsonx.data is an open, hybrid, governed data store optimized for all data, analytics, and AI workloads, built on a data lakehouse architecture (see figure 1).

Access all of your data and maximize workload coverage across all your hybrid-cloud environments. Expect seamless deployment of a fully managed service across any cloud or on-premises environment. Access any data source, wherever it resides, through a single point of entry and combine it using open data formats. Integrate into your existing environment with open source and open standards, and interoperability with IBM and third-party services.

Accelerate time to trusted insights. Start fast with built-in governance and automation; strengthen enterprise compliance and security with unified governance across your entire ecosystem. A clear UX and click-and-go console helps your teams ingest, access and transform data and run workloads. Watch how quickly they'll embrace a dashboard that makes it easier for them to save money and deliver fresh, trusted insights.

Reduce the cost of your data warehouse by up to 50%[4] through workload optimization across multiple query engines and storage tiers. Optimize costly warehouse workloads with fit-for-purpose engines that scale up and scale down automatically. Reduce costs by eliminating duplication of data when you use low-cost object storage; extract more value from the data in ineffective data lakes.

Next steps

Take advantage of the IBM team’s data management and optimization knowledge honed by decades of handling the world’s most demanding data workloads. See how quickly you can gain value from watsonx.data.

1. Why Unstructured Data is the Future of Data Management, Venturebeat, July 2021.

2. Worldwide IDC Global DataSphere Forecast,2022-2026, IDC, May 2022.

3. The rise of the data lakehouse: A new era of data value, CIO Magazine, 18 August 2022

4. When comparing published 2023 list prices normalized for VPC hours of IBM watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendors.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
May 2023

It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is compliant with any law or regulation.