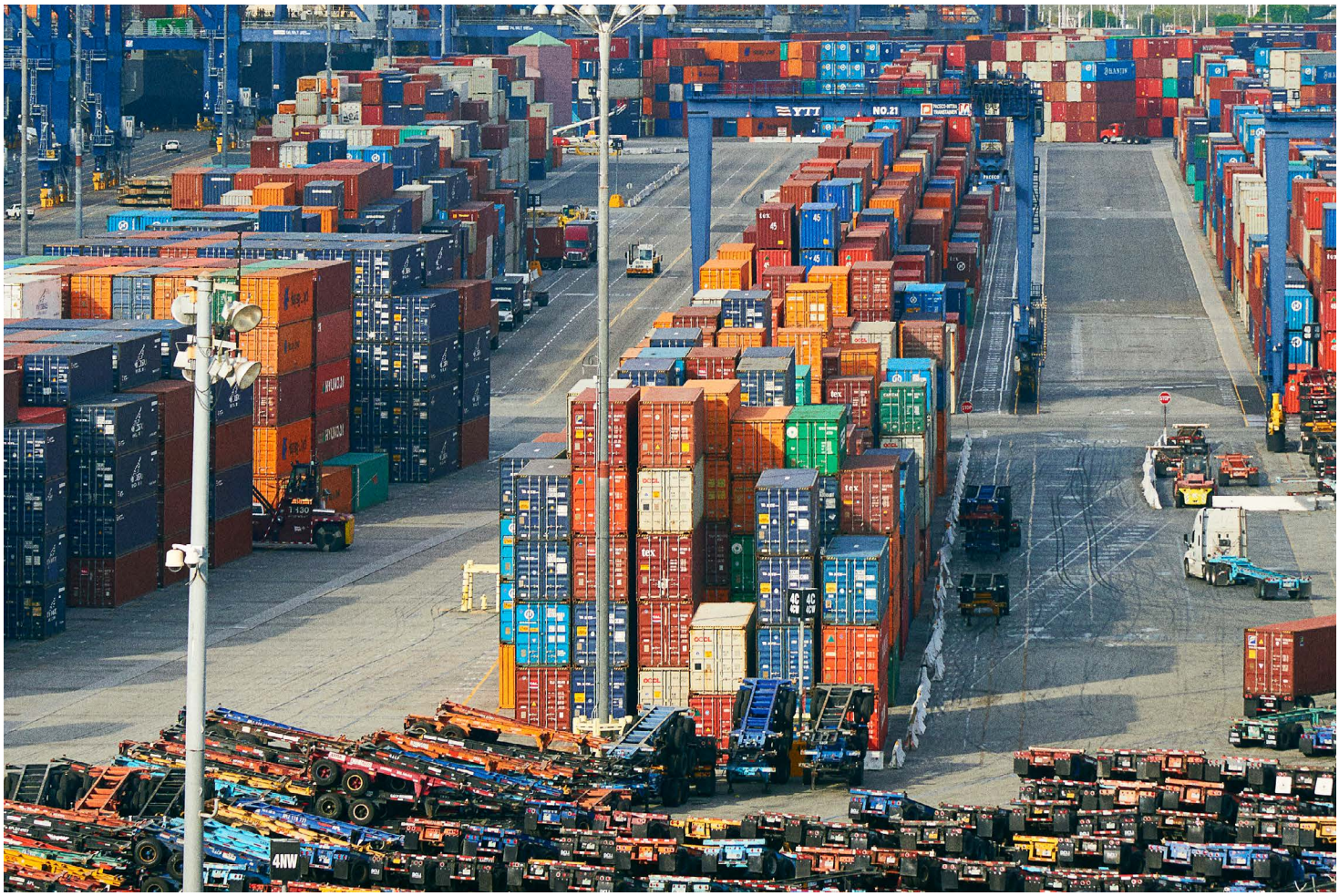
AI demands new ways of data management

The data leader’s guide for how to leverage the right databases for applications, analytics and generative AI



IBM

Table of contents

Introduction

The advent of generative AI has elevated the value of data, where companies are now racing to harness its potential faster than their competitors. The companies with substantial data wealth are gaining a competitive advantage. Enterprises that possess high-quality data and attest to the trustworthiness of their data among stakeholders have doubled the return on investment (ROI) from their AI capabilities.[1]

But it's not so simple. The effectiveness and trustworthiness of analytics and AI is inherently tied to the quality, availability, and management of the underlying data and many organizations are still faced with fundamental data challenges. In fact, 53%[1] of CEOs say that a lack of proprietary data will be a barrier to successful generative AI initiatives. Data is also exploding, both in volume and in variety. According to IDC, by 2025, stored data will grow 250% across on-prem and cloud storages.[2] With growth comes complexity—multiple data applications and formats that make it harder for organizations to access, manage and effectively use all their data.

For data and IT leaders, the challenge is formidable—they need to create more value from their data while improving resiliency, reducing costs and ensuring scalability for modern applications, analytics and AI use cases. And that's definitely not a task they can solely accomplish with traditional, on-premises databases, software and appliances. Applications now need to be able to store, manage and govern petabytes of data in a variety of formats to support new analytics and AI use cases across hybrid cloud deployments. By pairing the right workload type with the right database, data teams can help ensure that applications run with optimal performance and manage costs effectively.

Enabling an open and trusted data foundation for AI with IBM

Enterprises turning to AI today need access to a full technology stack that enables them to train, tune and deploy AI models—including foundation models and ML—all in one place and to run across any cloud environment.

IBM® watsonx™ is a portfolio of AI products that accelerates generative AI into core workflows to drive productivity across your business. It includes the following products:

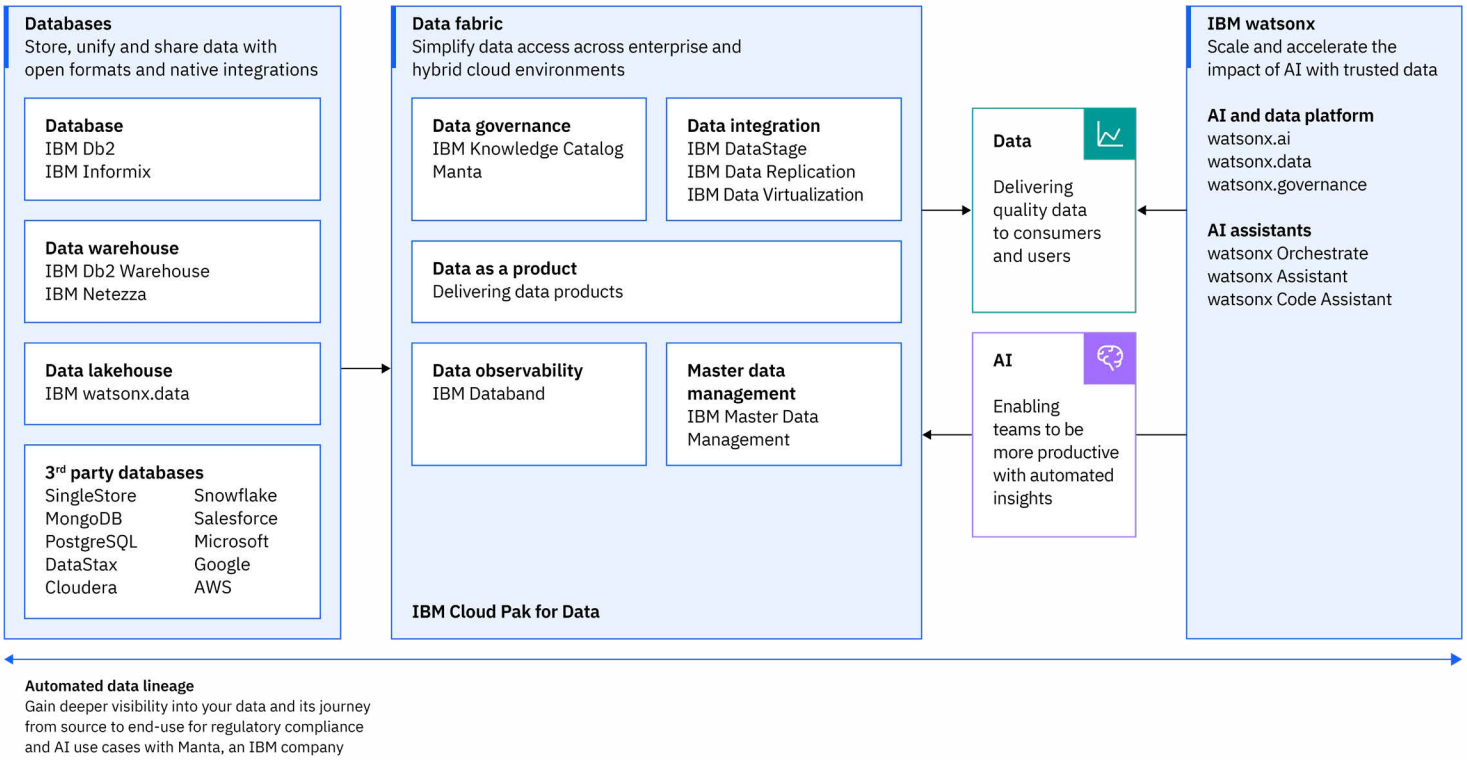
- IBM watsonx.ai™, a studio for AI builders
- IBM watsonx.data™, a fit-for-purpose data store built on an open data lakehouse architecture to scale AI workloads
- IBM watsonx.governance™, a toolkit for AI governance
- A set of AI assistants that can be deployed out of the box to help automate customer service, generate code, and automate key workflows in departments such as HR

IBM provides an open and trusted data foundation to help you simplify access and consumption of data for AI workflows. We provide native integrations across our databases portfolio and support open formats such as Iceberg, Parquet and ORC, enabling the ability to store, unify and share data for analytics and AI without additional ETL or duplication of data.

Additionally, training traditional ML or generative models on sensitive or regulated data, whether it be client-specific, customer-related or otherwise, introduces the risk of violating data privacy regulations. Your organization needs to ensure proper governance and lineage of this data, to avoid possible data leaks, financial penalties and reputational damage.

This is where a data fabric architecture comes in. With a data fabric, clients can automate discovery and enrichment of data for AI with centralized data governance capabilities, enable data observability across various pipelines, employ various data integration styles, and manage data quality to deliver reliable data for AI workflows. This feeds watsonx itself, where all your AI workloads can be executed with trusted and high-quality data.

Investments in an open and trusted data foundation will accelerate and scale your AI initiatives



Developing your database management strategy to scale AI

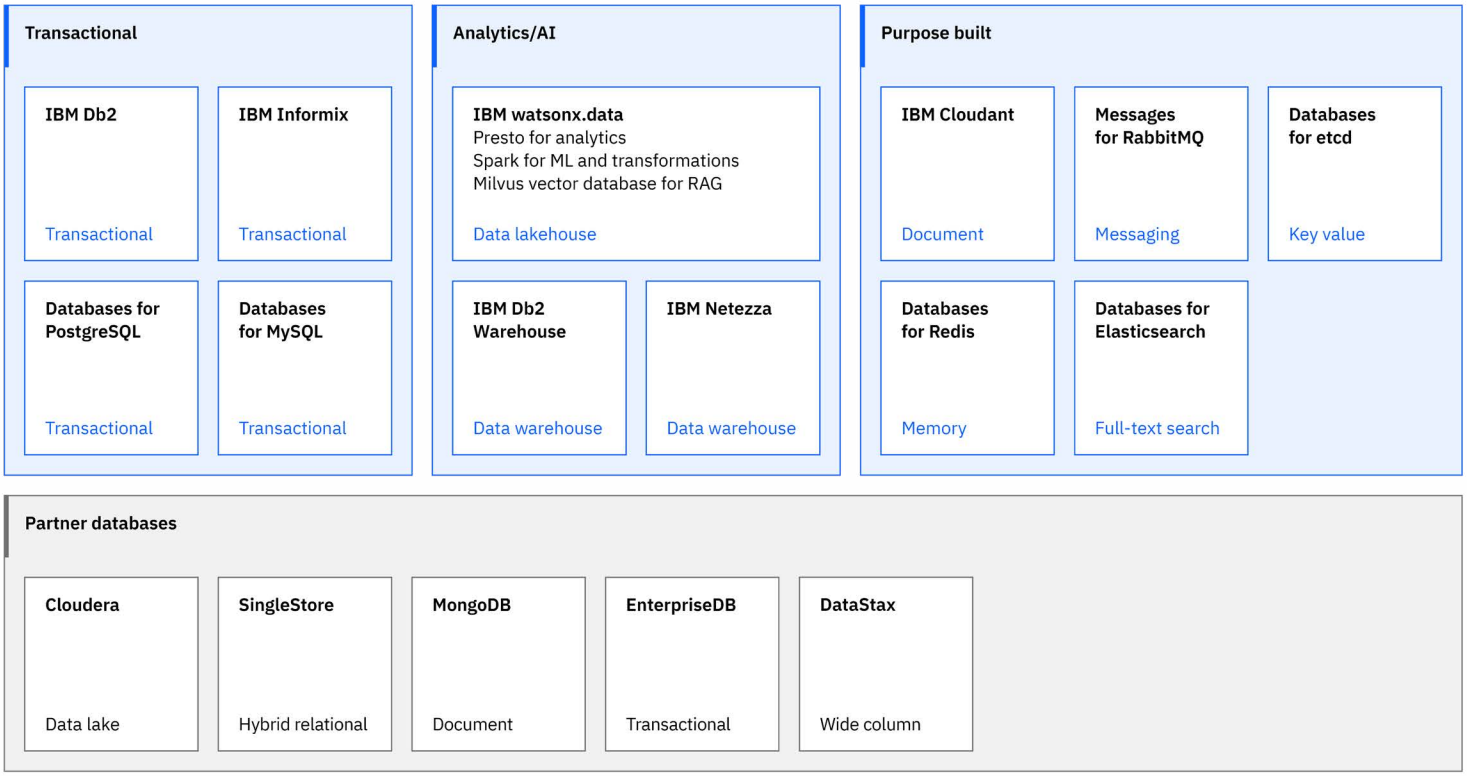
Think about the data needs of your organization. Do you manage data volumes in petabytes, if not exabytes? Do you need to scale to serve millions of users around the world? Is demand to support generative AI use cases such as retrieval-augmented generation (RAG) accelerating? Do you have workloads that need to run on premises and in the cloud?

At IBM, we believe no single database engine is built to manage all workloads, data types and use cases while maintaining performance and cost efficiency. Customers have different workloads and architectures and require purpose-built databases that can be accessed with a data fabric and integrated with AI and data products. As these demands are largely use case-specific, let's look at some of the common database types and use cases that best suit them, along with relevant database offerings from IBM.

Benefits of purpose-built databases:

- Right tool for the right job
- Better performance
- Cloud scale
- More functionality to support generative AI use cases
- Easier to debug and monitor
- Independence between teams
- Faster time to market
- Lower total cost of ownership (TCO)
- Reduced operations

A full portfolio of purpose-built databases



IBM offers an extensive range of purpose-built databases available across cloud, hybrid and on-premises deployments. These offerings are designed to support an open and trusted data foundation to run all your data, analytics and AI workloads.

There are 3 main types of databases you can use to support analytics and AI: open data lakehouses, data warehouses, and transactional databases.

Open data lakehouse

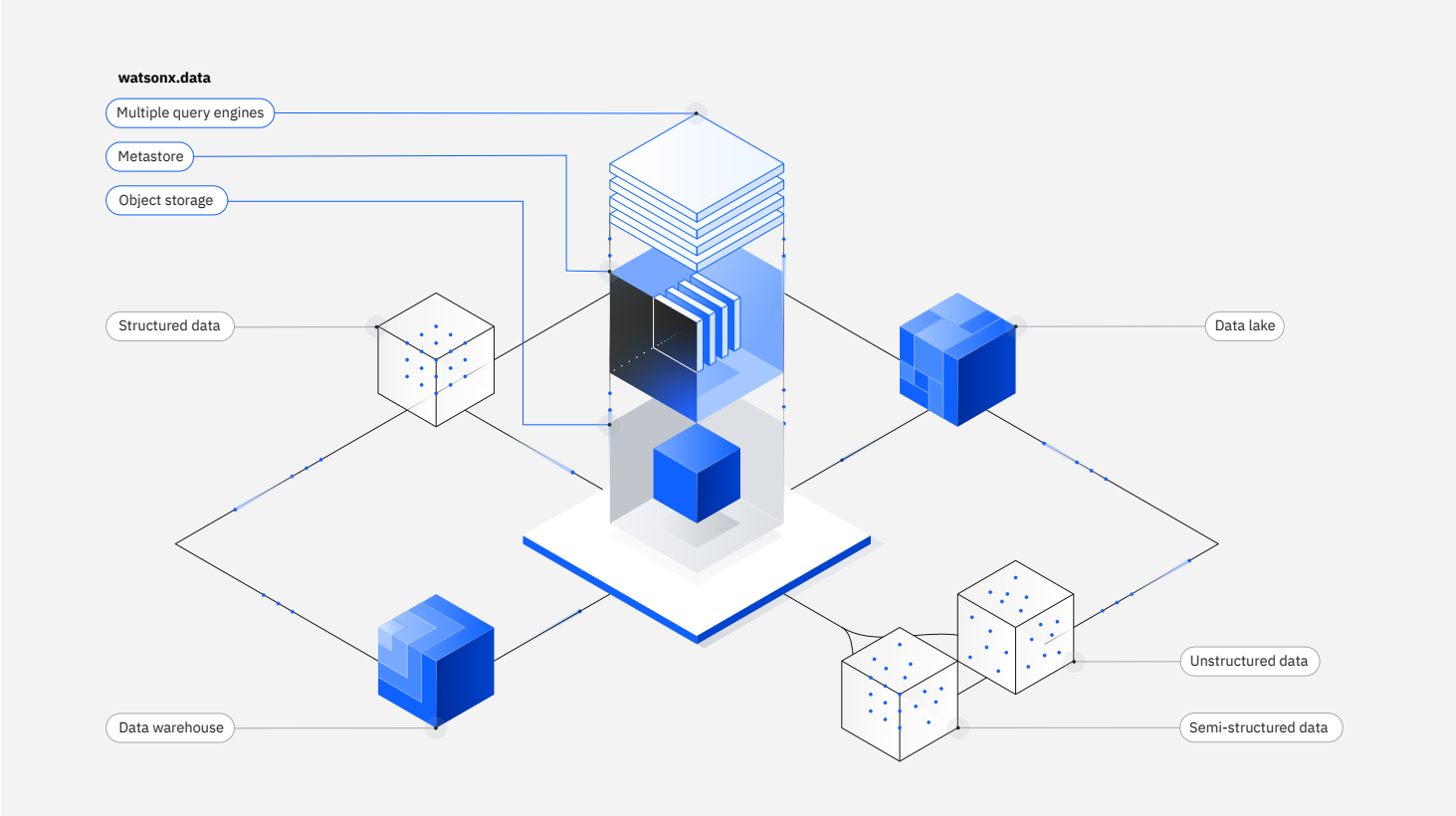
A data lakehouse combines the performance and governance of a data warehouse with the flexibility and cost effectiveness of a data lake. It provides the flexibility to support a variety of workloads such as heavy AI and ML applications, analytics and BI on data in your data lake, streamlined data engineering and transformations, secure data sharing, and self-service data exploration.

Why a data lakehouse is the data store for AI
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 extracting insights directly in a performant manner. An open data lakehouse addresses these limitations by handling multi-modal data formats—both proprietary and open data formats—supporting object storage, and combining data from a multitude of sources including existing repositories to ultimately enable analytics and AI at scale.

A lakehouse is built with shared metadata and governance to catalog your data and provide lineage, which enables trusted AI outputs. This creates a self-service, collaborative environment for your data scientists, AI builders, data engineers, analysts, and even non-technical users working with data.

Every data warehouse needs a lakehouse for AI
The decision is not whether to use a warehouse or a lakehouse. The best approach is to use a warehouse and a lakehouse—ideally a multi-engine lakehouse, to optimize the price-performance of all your workloads in a single, integrated solution. Add to that the ability to optimize deployment models across hybrid-cloud environments, and you have a foundational data management architecture for years to come.

Modernizing Hadoop data lakes with a lakehouse
Data lakes have proven successful where companies have been able to narrow the focus on specific usage scenarios. But what has been clear is that there is an urgent need to modernize these deployments and protect the investment in infrastructure, skills and data held in those systems. With an open data lakehouse architecture, you can now modernize your Hadoop data lakes with warehouse-like performance, security and governance.



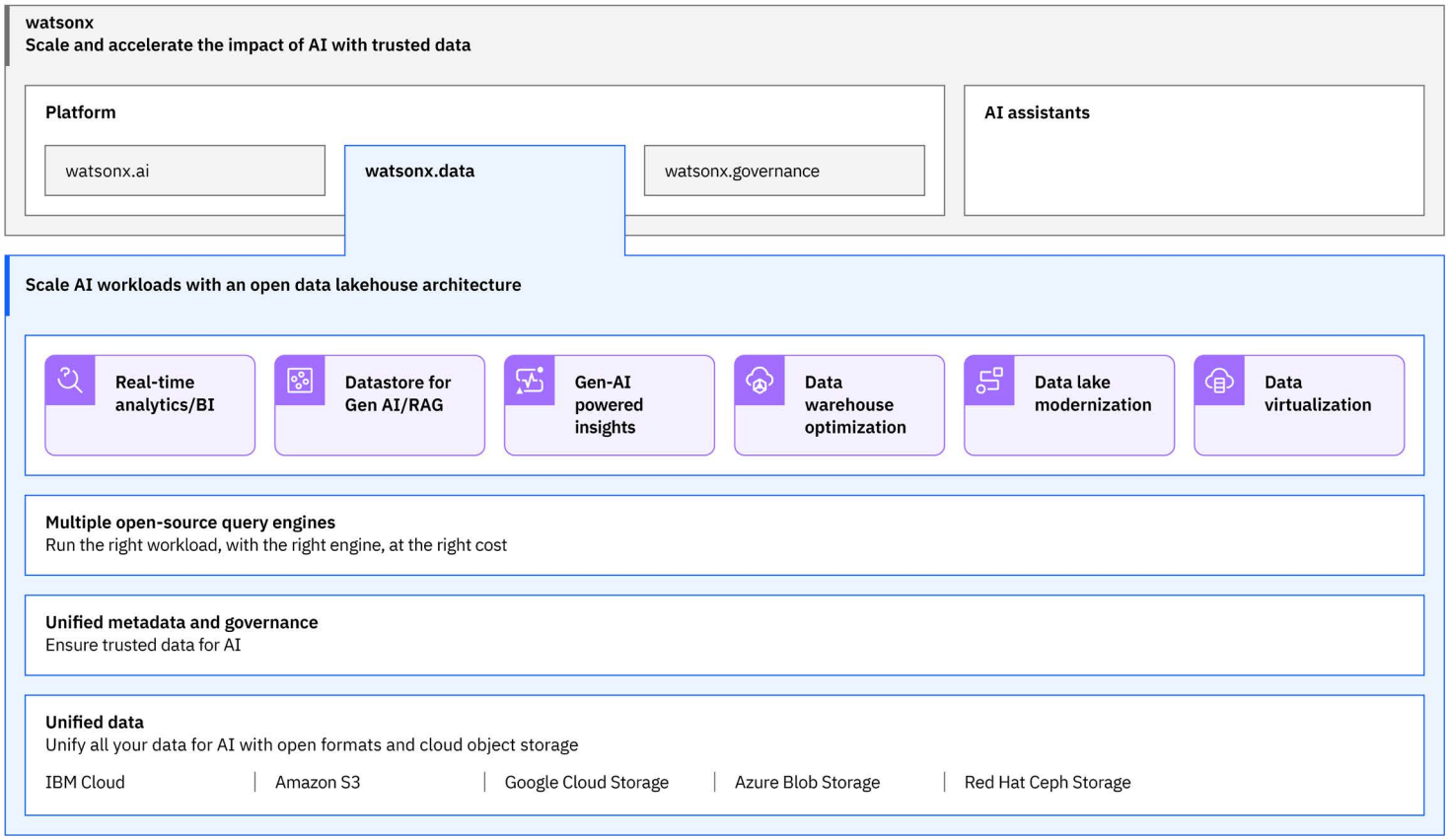
Data lakehouse use cases:

- Data warehouse cost optimization
- Data lake/Hadoop modernization
- Ad hoc and self-service analytics
- Real-time analytics and BI
- Data engineering and data transformation
- Machine learning and AI model building
- Data sharing

IBM data lakehouse solution

IBM watsonx.data is an open, hybrid and governed data store, built on an open lakehouse architecture and supported by querying, governance and open data formats. Watsonx.data enables you to scale AI and analytics workloads for all your data, anywhere across the hybrid cloud. You can connect to data in minutes, quickly get trusted insights and reduce your data warehouse costs by up to 50%.[3]

Bring AI to your data and trusted data to your AI with built-in generative AI and vector database capabilities. Discover and enrich data and metadata in watsonx.data using natural language—no SQL required. Store, query and search vector embeddings in watsonx.data with integrated vector database capabilities. Vector databases can store unstructured data and feed it to AI models. These capabilities are essential for generative AI because they enable retrieval-augmented generation (RAG), which is core to our ability to anchor large language models in trusted data to reduce model hallucinations. Watsonx.data is available as a SaaS on AWS and IBM Cloud—or you can deploy as software on premises.



Data warehouse

Data warehouse architectures are meant to centralize data to support business reporting, dashboards and other analytics projects. This architecture comprises the enterprise's disparate data sources, pipelines that ingest and process data, and the warehouse solution.

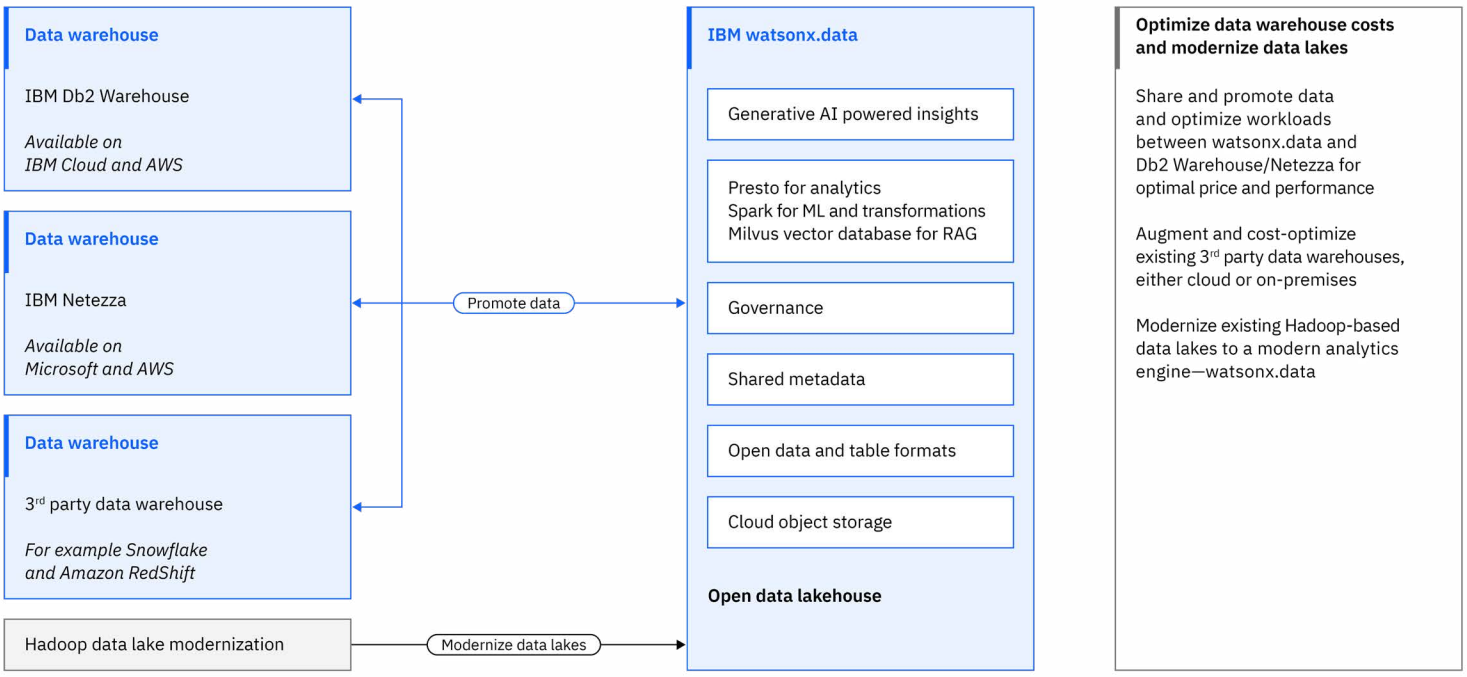
Customers turned to cloud data warehouses as a way to drive costs down with pay-as-you-go pricing, allowing you to start small, scale up as needed, and scale down your analytics environment when it's not in use. However, it's important for global operations to be running 24x7x365 and data warehouses are mission-critical assets for enterprises to make quick, real-time business decisions. Ensure your data warehouse provides cost flexibility and control of resources to avoid overprovisioning and prevents vendor lock-in by supporting open source technologies that are required to share data for data science and AI use cases.

Use cases:
- Operational reporting
- Business intelligence
- Deep analytics
- Data sharing with open formats
- Traditional machine learning (ML) models

IBM data warehouse solutions

IBM Db2 Warehouse: The next generation cloud-native Db2 Warehouse meets your price and performance objectives for always-on workloads and scales operational analytics, BI and mixed workload needs now 4x faster with 34x lower storage costs.[4] It enables governed access to data in open formats and natively integrates with watsonx.data open data lakehouse to create a singular view of your analytics and AI estate. Available as SaaS on AWS and IBM Cloud, or deploy as software.

IBM Netezza: is IBM's cloud-native enterprise data warehouse optimized to run deep analytics, BI, and ML workloads at petabyte scale. Netezza can store and analyze governed data in open formats, control costs with AI-driven elastic scaling, and natively integrates with watsonx.data open data lakehouse to create a singular view of your analytics and AI estate. Available as SaaS on AWS and Azure, or deploy as software.



Transactional (OLTP) databases

Transactional databases are relational databases ideal for powering ERP, CRM, web and mobile apps, microservices, heavy transaction-based applications such as banking and finance. They store structured data in tabular format with columns and rows and data is queried with SQL.

The value of transactional data for AI

When it comes to new analytics and AI use cases, you may have petabytes if not exabytes of valuable data stored in your transactional databases that can be leveraged for new insights and ML/AI models. It's important to ensure your transactional database can run on a flexible, hybrid cloud infrastructure required for AI applications and can seamlessly integrate with existing analytics and AI repositories such as your data warehouse or data lakehouse to avoid additional ETL or duplication of data.

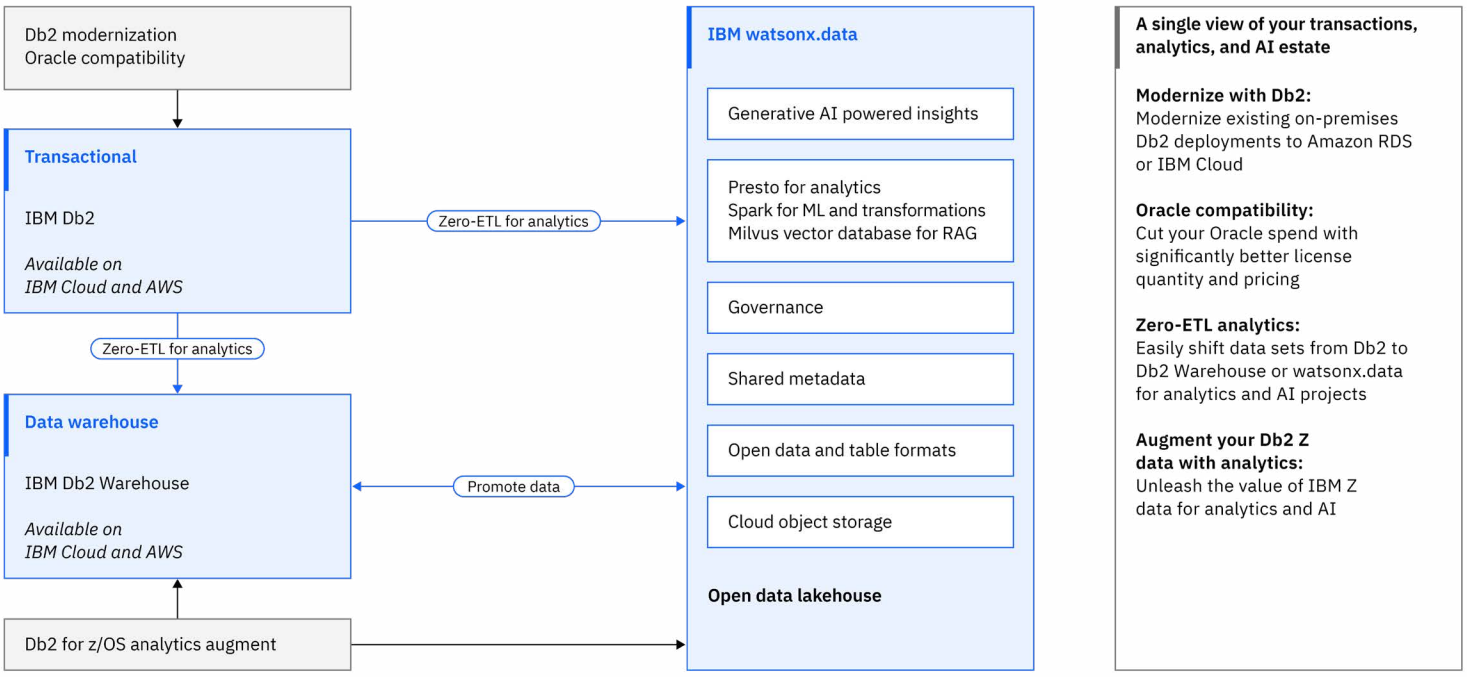
Use cases:

- ERP and CRM systems
- Microservices
- Cloud web and mobile applications
- Modernize legacy applications

IBM transactional database solutions

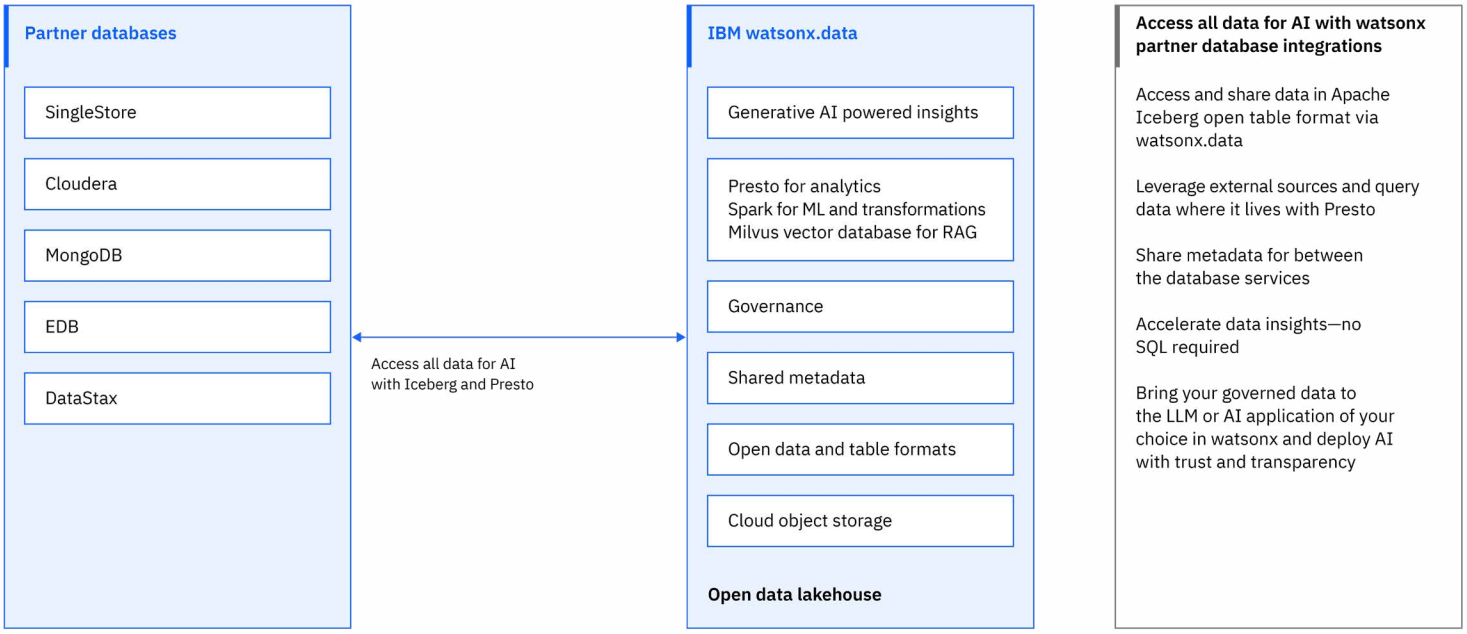
IBM® Db2® database: An always on, cloud-native database that offers data governance and security, low-latency transactions and continuous availability to your mission-critical data, analytics and AI-driven applications. It is built to run the world's mission critical workloads, including the largest banks, automotive manufactures, and retailers powering trillions of transactions per day and millions of users around the world. It natively integrates with Db2 Warehouse and watsonx.data open data lakehouse, to easily share and unify data for analytics and AI without ETL. Available as SaaS on Amazon RDS and IBM Cloud to reduce costs and automate database management, or deploy as software.

IBM Db2 for z/OS: Db2 for z/OS and its ecosystem of products offers agile, efficient, secure enterprise data serving for the world's most demanding hybrid cloud, and transactional and analytics mainframe applications. Db2 for z/OS Data Gate makes Db2 Z data easily available for analytics and AI via Db2 Warehouse and watsonx.data.



IBM partner databases

IBM partners with 3rd party databases such as SingleStore, Cloudera, MongoDB, EDB, and DataStax and integrates with watsonx, a portfolio of AI products, to scale AI with all your data.

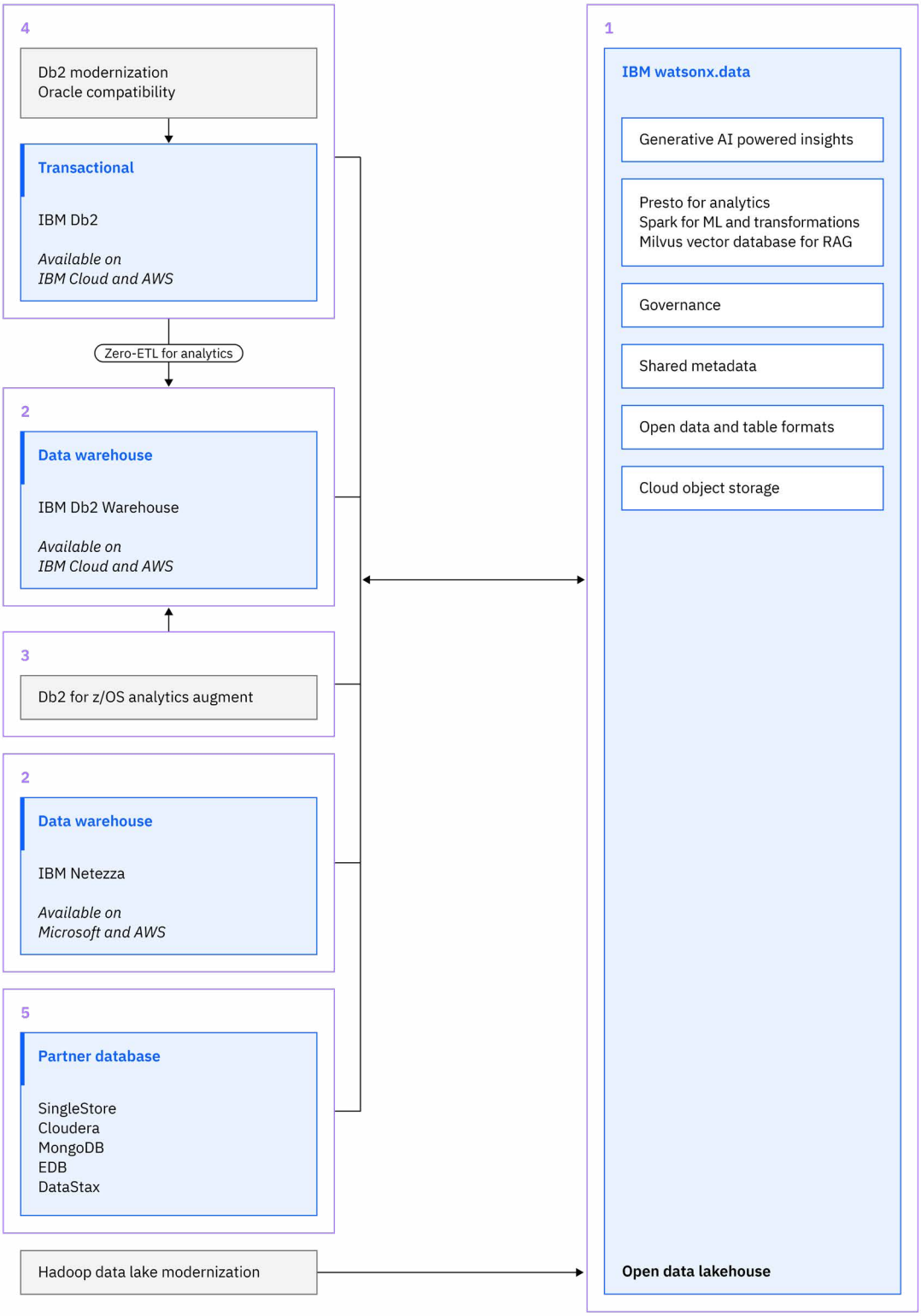


Bringing it all together

Take advantage of the integration across the entire IBM database portfolio with watsonx.data and simplify your data foundation for applications, analytics and AI.

1. Take an open data lakehouse approach with watsonx.data to unify, prepare, and scale AI workloads with all of your data across the hybrid cloud. Use built-in AI capabilities such as generative AI powered insights to discover and enrich data with natural language as well as vector database capabilities to support RAG pattern use cases.
2. Optimize your data warehouse costs and modernize your data lakes across multiple, fit for purpose query engines. Using open-source components, native integrations, and shared metadata, you can unify data across watsonx.data, Db2 Warehouse, Netezza, 3rd party data warehouses, and data lakes without ETL.
3. Unleash the value of Db2 for z/OS mainframe data for analytics and AI with IBM Db2 Warehouse and watsonx.data integrations.
4. Modernize your on-premises Db2 or Oracle transactional database instances to fully-managed SaaS with Db2 on Amazon RDS or IBM Cloud. Leverage native integrations with Db2 Warehouse and watsonx.data to unify data for analytics and AI without ETL.
5. Our portfolio of partner databases provide native integrations with watsonx to scale AI across all of your data.

Simplify data for applications, analytics, and AI



Simplify data for applications, analytics and AI

1. Leverage built-in AI capabilities with watsonx.data
2. Optimize data warehouse costs and modernize data lakes
3. Unleash value of IBM Z data for analytics and AI
4. Modernize Db2 and Oracle on-premises with Amazon RDS for Db2
5. Access all data for AI with watsonx partner database integrations

Next steps

Many leading companies trust IBM to help them efficiently manage their most mission-critical data and applications. Our innovations in enterprise data technologies include market-making database solutions, open-source technology, and enterprise-ready AI. We enable our clients to run solutions in any cloud or on-prem environment and believe that our clients' data solely belongs to them and not IBM.

Find out how you can meet your application needs with the right database solution.

Explore our database solutions →

Appendix

Other purpose-built databases—and when to use them for AI
Purpose-built databases like in-memory, key-value, and message queues can also connect to watsonx.data through the open-source Presto query engine. Query data in place with data virtualization using Presto, which has over 35 connectors to various external database vendors including Redis, etcd, and RabbitMQ. Watsonx.data also provides a vector database capability powered by Milvus, for generative AI use cases. Let's see how each of these databases can be used for AI.

Vector databases
Vector databases are fundamental to generative AI: Retrieval augmented generation (RAG) is core to our ability to anchor large language models in trusted data to reduce model hallucinations. This approach relies on leveraging vector databases store vector embeddings and enrich prompts with semantically relevant information for in-context learning by foundation models.

IBM vector database offering:
Watsonx.data: Store, query and search vector embeddings in watsonx.data with integrated vector database capabilities powered by Milvus, the highly scalable and blazing fast open-source vector database.

In-memory databases
One of the main features of memory databases for AI is its support for various data structures like strings, sorted sets, bitmaps, geospatial indexes, and more. It's low latency is particularly advantageous in AI scenarios that require real-time or near-real-time processing, such as streaming analytics or certain machine learning tasks.

IBM memory database offering
Redis: An open-source, in-memory database fully managed on IBM Cloud that delivers exceptional speed, reliability, availability and performance. Used as an application cache or quick-response database, Redis enables data to be placed physically closer to the user for the lowest latency.

Key-value databases
Each key stored in a key value database is unique, and the associated value of keys can be a simple data element or a more complex data structure. The simplicity and efficiency of key-value databases make them suitable for certain AI scenarios such as fast data retrieval, real time processing, and caching frequently accessed data.

IBM key-value database offering
etcd: An open-source, distributed key-value store fully managed on IBM Cloud that provides a reliable way to store data for large-scale distributed systems. It's used for configuration management, service discovery and coordination of distributed systems or clusters of machines.

Message queues
Message broker databases can be useful in AI applications, especially in scenarios where there is a need for asynchronous communication, event-driven architectures, and distributed systems.

IBM messages database offering
IBM Messages for RabbitMQ: IBM Messages for RabbitMQ on IBM Cloud is an enterprise-ready offering that supports multiple messaging protocols as a broker. It's fully managed, scalable and highly available and is used build web and mobile applications and in IoT.

1. IBM Institute for Business Value (IBV) 2023
2. IDC Global DataSphere Forecast 2022-2026: https://www.idc.com/getdoc.jsp?containerId=US49018922
3. When comparing published 2023 list prices normalized for VPC hours of watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendor.
4. Results derived from running the IBM Big Data Insights concurrent query benchmark on two equivalent Db2 Warehouse environments with 24 database partitions on 2 EC2 nodes, each with 48 cores, 768 GB memory and a 25 Gbps network interface. One environment used the new cloud object storage and advanced caching capability, and the other environment did not use the caching capability and was used as a baseline. The test revealed a 4x increase in query speed (213 seconds versus 51 seconds) using the new capability. Reduction of storage costs is derived from price for cloud object storage, which is priced 34x cheaper than SSD-based block storage.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America