How to choose the right data warehouse for AI

Accelerate innovation and drive business outcomes by turning data into insights

IBM

Table of contents

Less than 1% of the global datasphere is currently used for AI.[2]

A modern data warehouse is the first step on the journey to AI

As centralized repositories that store and analyze organizational data from disparate sources, data warehouses have traditionally been essential to business intelligence. They have helped companies in every phase of the data maturity curve wrangle and make sense of massive volumes of data.

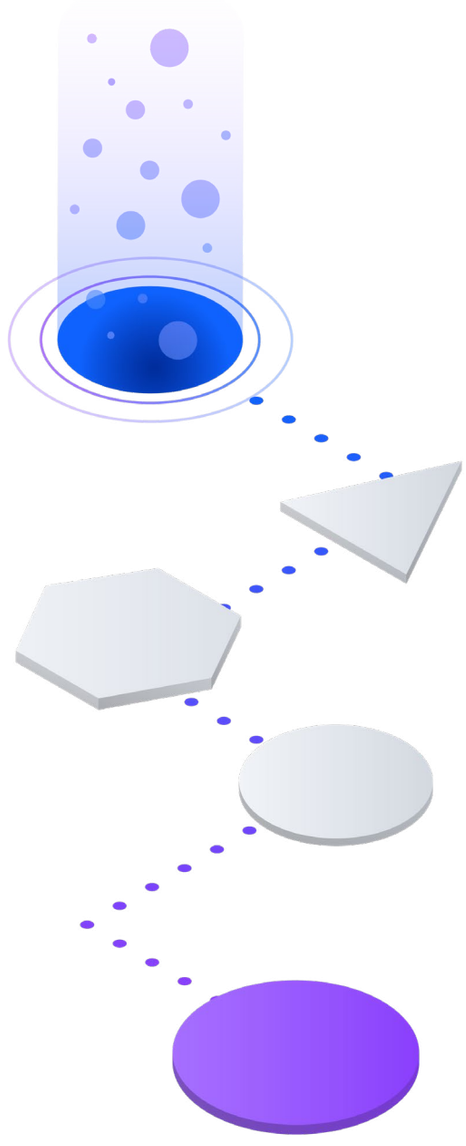
But today, artificial intelligence has changed the game. Beyond fulfilling conventional data management needs, the modern data warehouse has evolved into a catalyst for AI. It doesn't just provide reports and dashboards or simply overcome challenges in data volume and quality. Instead it is now the critical first step in helping companies digitally transform their business with AI innovations.

By automating data ingestion and analysis, the modern enterprise data warehouse (EDW) has become what Forrester calls a "system of insight"[1] that closes the loop connecting data, insight and action.

It is purpose-built to run complex queries that can be shared with various AI tools, enabling seamless machine learning and more accurate predictions. Companies can make better decisions faster because a modern data warehouse brings together all organizational data, at any scale, to deliver actionable insights.

This ebook will examine the role that data warehouses can play in realizing your company's AI aspirations. It will explore how a unified platform approach can advance this journey, why an EDW is the critical first step, and how you can choose the right EDW deployment to suit your unique business needs.

The journey to AI starts with collecting clean and complete data. Let's begin.

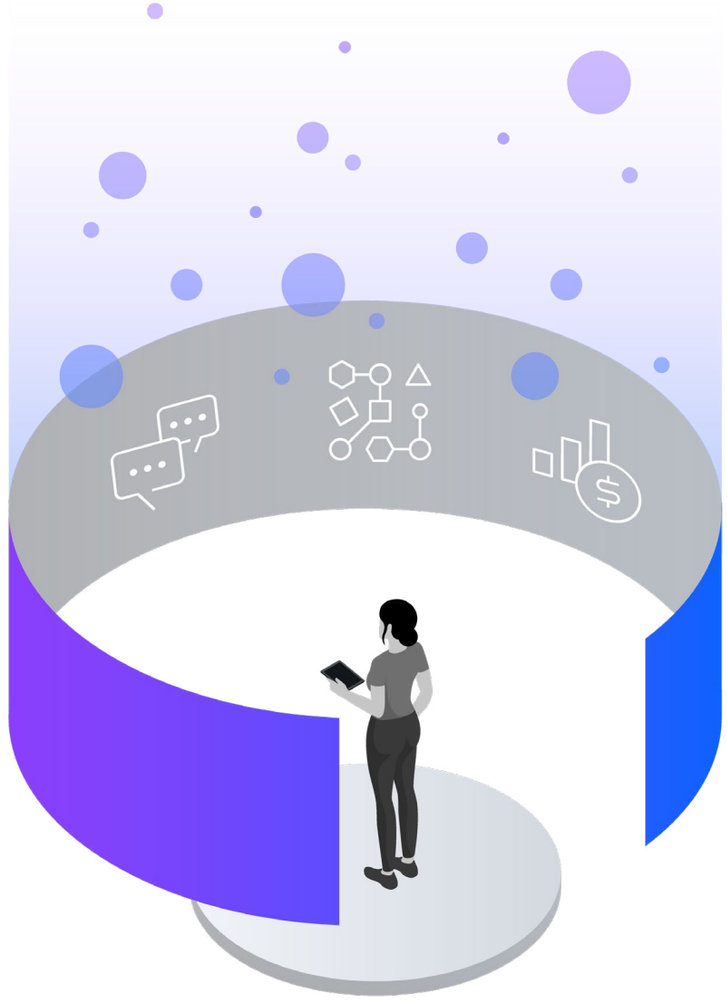


There’s nothing artificial about the impact of AI

In a saturated and constantly evolving market, AI can be a true differentiator, and hybrid, multicloud services that run from the edge to the core are becoming the new norm. In fact, IDC estimates AI investments will reach USD 97.9 billion by 2023.[3]

Automated customer service agents, IT automation and sales process recommendations are the current top uses cases of AI, while automated human resources, digital assistants for enterprise knowledge workers, regulatory intelligence and advanced digital simulation are not far behind.

According to McKinsey’s *The State of AI in 2020*,[2] 66% of businesses reported an increase in revenue and 40% saw a reduction in costs due to AI adoption. Because of proven business results, that adoption is growing. By the end of 2024, Gartner predicts, 75% of organizations will shift from piloting to operationalizing AI, driving a five-fold increase in streaming data and analytics infrastructures.[4]



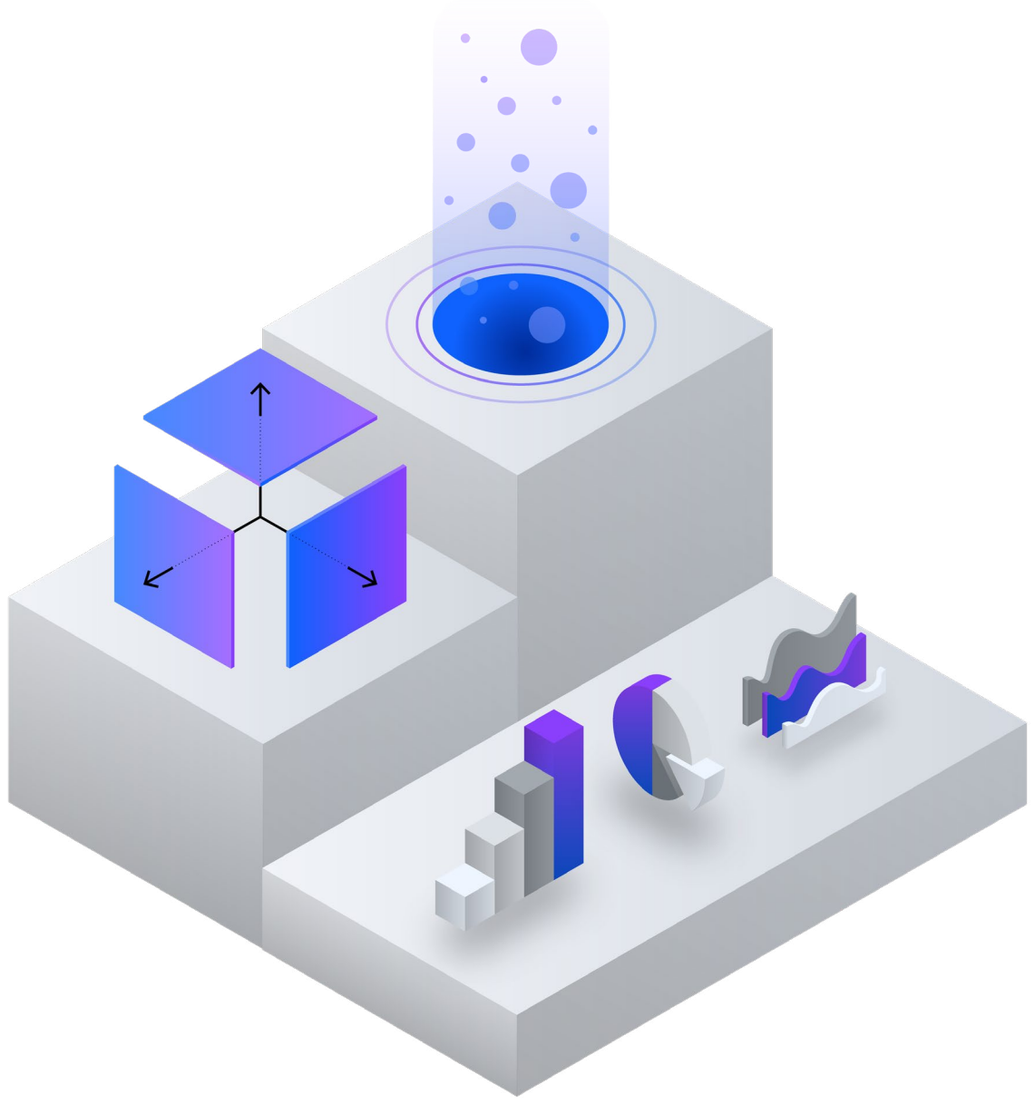
Challenges to adopting AI

Despite the growing case for AI, adoption isn't easy. What's stopping enterprises from fully embracing AI?

While there has been an explosion of data, only a tiny fraction is used to create insights and feed AI systems
IDC forecasts that by 2023, the Global DataSphere—all data created and consumed worldwide—will grow to 102.6ZB. However, less than 1% of the global datasphere is currently used for AI; the remainder is dormant or dark data.

Scaling AI is complicated
Data volume and veracity, intensive computing requirements, complex business processes and large numbers of users can hinder scaling efforts. In addition, expenses for highly skilled staff and project maintenance can add up. According to IDC, 58% of organizations cite cost as a major barrier.

A shortage of data science skills, challenges in algorithm explainability and a lack of data quality can lead to inaccurate machine learning models
Many organizations do not have dedicated data science talent, and without proper AI governance, models can drift and deliver biased results, yielding potentially inaccurate conclusions.



400

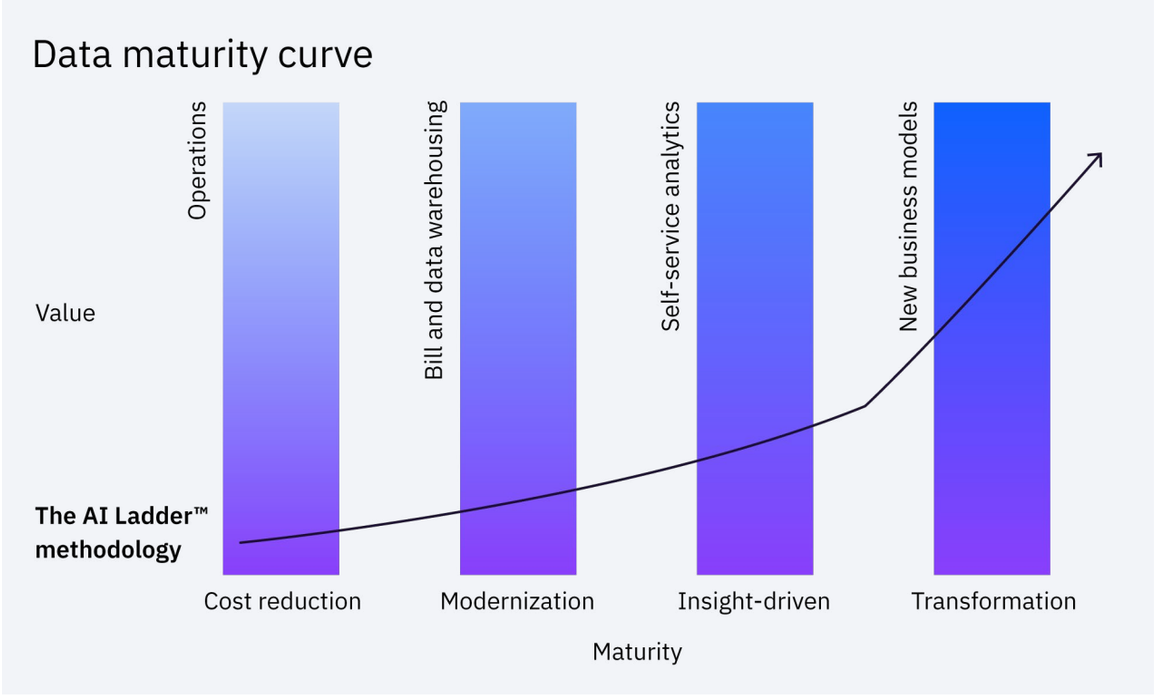
Average number of unique data sources that an organization uses for business intelligence and analytics.[5]

The need for a data and AI platform that integrates data warehouses with machine learning

In light of these challenges, a critical requirement of AI is a robust information architecture that implements an enterprise-wide data and AI strategy and helps organizations progress along the data maturity curve.

Organizations at the beginning of the curve apply data to operations, usually with an emphasis on cost reduction. As their data maturity advances, their use of information expands, shifting the focus to business intelligence and self-service analytics. At the top of the curve, organizations use data to develop newer business models and advance digital transformation.

Most organizations still have a long way to go in their data maturity curve before they can fully embrace AI. They face a series of decisions and challenges in building out a modernized information architecture that makes data ready for AI. From building a solid data foundation to delivering trusted insights to key decision-makers across the company, CIOs and other IT leaders must build a comprehensive data management strategy that supports their AI journey.



The need for a data and AI platform that integrates data warehouses with machine learning

An integrated, end-to-end platform for high performance analytics and AI provides the modernized information architecture needed to meet data maturity goals. This combination allows critical data to remain securely behind a private firewall and be accessible by cloud-based applications to generate new insights and machine learning models.

A unified data and AI platform such as IBM Cloud Pak® for Data is important because it can help companies:

1. Gain a complete view into their data.
2. Govern data to meet regulatory compliance.
3. Reduce bias and drift to produce trustworthy models.
4. Build AI applications that solve direct business needs.

The modern EDW is a critical component of a unified data and AI platform. It collects and analyzes data so that this data can be prepared for subsequent stages of the AI lifecycle.

Once an EDW has ingested data from various sources and processed it for insights, organizations can then activate data governance practices to make sure that data is secure and compliant. They can use governed historical and real-time data to build AI models, creating a “machine learning feedback loop”[5] that continuously processes new data to prevent model drift and bias. In these ways, the data collected by the modern EDW can be transformed into predictive analytics, paving the way for companies to build AI applications that are infused throughout the enterprise.

In the next section, let’s break down the modern data warehouse in more detail.

The anatomy of a modern enterprise data warehouse

How does a data warehouse turn data into insights? From the point of ingestion to a comprehensive business intelligence report, a modern data warehouse operates at various functional layers to collect, prepare and analyze data so that it can be used for AI.

Beyond the traditional data warehouse, a modern EDW supports key capabilities, such as a multi-model data store, data virtualization, mixed workloads, and deployment across hybrid clouds and other environments.

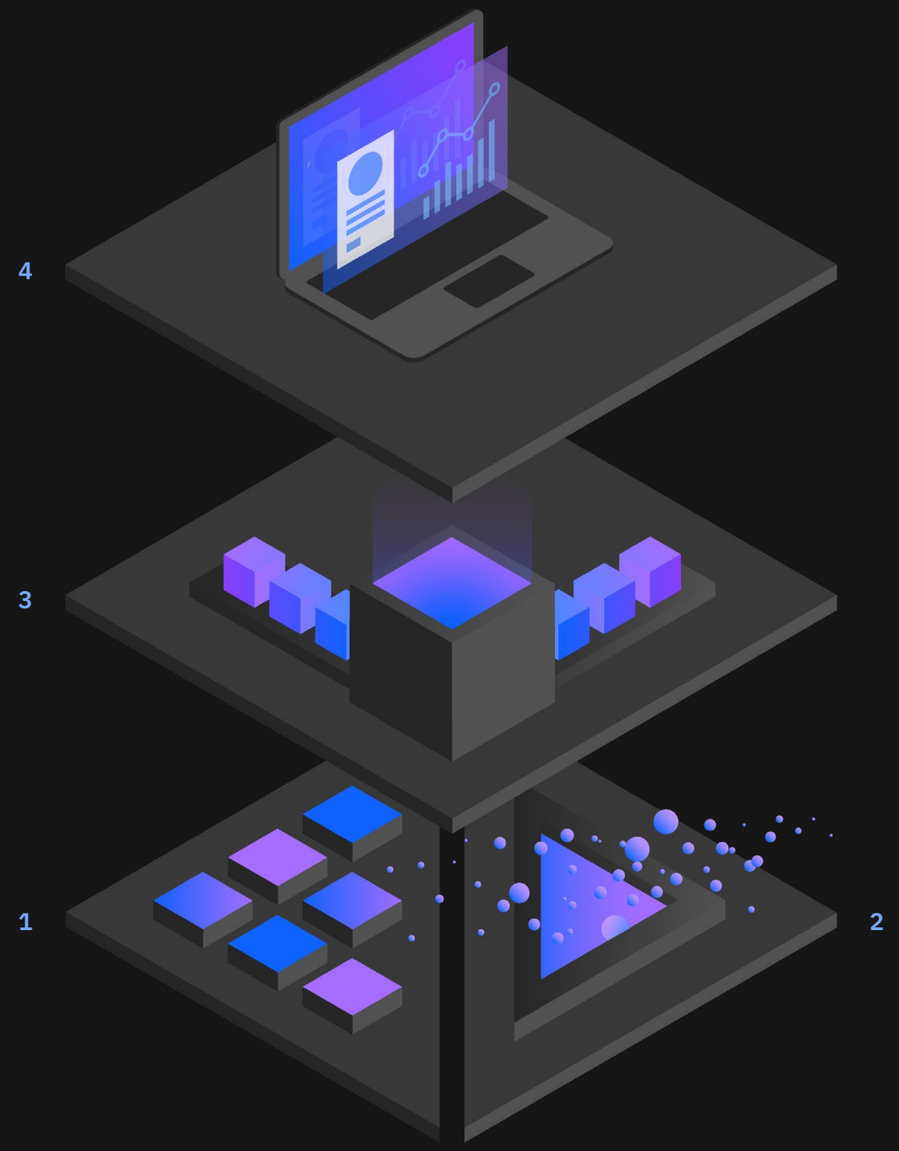
Key differentiators of a modern data warehouse

1. **Multi-model data store**
   All data stored in the data warehouse. Provides the best performance and integration for selected business data.

2. **Data virtualization**
   Data from outside the data warehouse, accessed and analyzed at the source.

3. **Mixed workloads**
   Real-time data captured daily and continuously.

4. **Hybrid cloud deployment**
   Business insights from analytics, including NoSQL and data in motion.



1

Multi-model data store

More and more, business data is being stored in data models other than traditional relational databases. Given the business value of these data models, there's an increasing desire and requirement to easily integrate them into single analytic queries. A modern EDW natively stores these various data models, supports SQL-based functions that distinguish them, indexes this data in a meaningful way and secures this data in a consistent manner with the rest of the data in the data warehouse.

2

Data virtualization

In many cases, data is not stored inside an EDW but accessed from disparate, remote sources. Organizations may have structured, unstructured and semi-structured data from a variety of on-premise and cloud systems. Data virtualization can overcome the complexity, cost, time and risk of error when it comes to analyzing this data. It helps to speed time to market and eliminates added hardware costs, data inconsistences and data governance issues by processing queries on the server where the data source exists. Companies can remove the risks of data movement because queries are no longer performed on data that has been copied and stored in a centralized location.

3

Mixed workloads

Supporting a mix of analytic workloads is a key feature of a modern EDW. If a data warehouse can handle different workloads—such as reporting, analytics (bulk scan), operational analytics (single record look-ups) and operational data stores (normalized snap-shots of source systems)—it can drastically reduce the cost of the overall EDW ecosystem.

Mixed workloads also allow for a shift to real-time warehousing. Traditionally, warehouses would support batch windows to separate the analytic workloads from data ingestion. A modern EDW needs to have continual data ingestion occurring at the same time while the mixed analytic workloads are running. Further, with the growth in data science, data volumes and data sources, high concurrency is required. These three trends—mixed analytic workloads, ongoing data ingestion and high concurrency—put a new demand on modern data warehouses.

4

Hybrid cloud deployment

Hybrid clouds are increasingly the infrastructure of choice because they allow an enterprise to move workloads seamlessly between both private and public clouds to optimize performance, security, compliance and cost-effectiveness.

In a hybrid cloud model, organizations can run sensitive, highly regulated and mission-critical applications or workloads with reasonably constant performance and capacity on a private cloud. At the same time, they can run less-sensitive, more-dynamic, or even temporary workloads on a public cloud.

Hybrid clouds help a modern data warehouse operate as a single EDW in parallel with data models, remote data sources, and mixed workloads. In addition, organizations can enable development and test environments for new applications and support disaster recovery.

Advantages of various cloud deployments:

Public cloud

- **Elastic and scalable**, flexibly adjusting to meet changing workload demands
- **Greater efficiency** since customers pay only for what they use
- **Reduced spending** on hardware and on-premises infrastructures

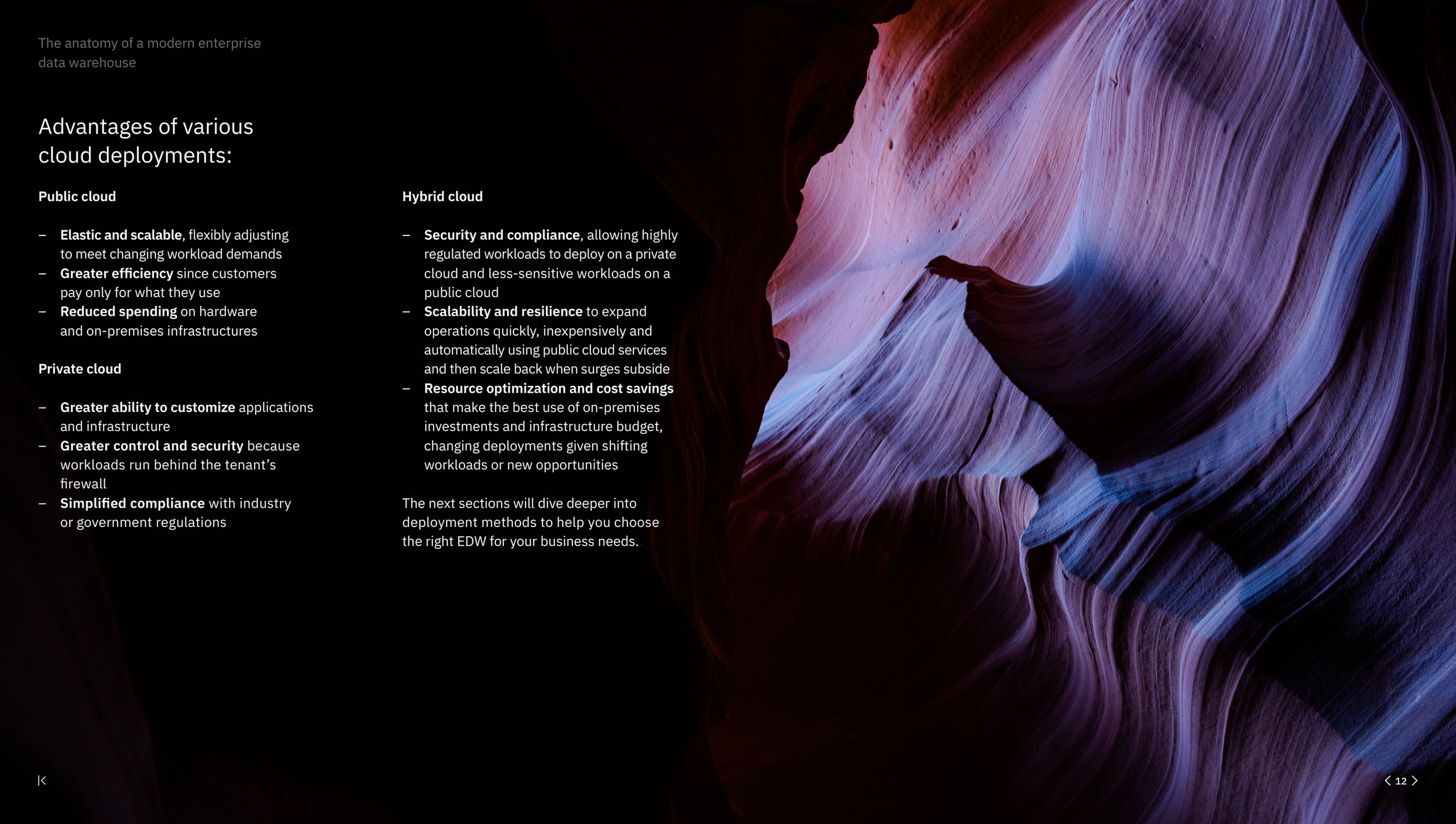
Private cloud

- **Greater ability to customize** applications and infrastructure
- **Greater control and security** because workloads run behind the tenant's firewall
- **Simplified compliance** with industry or government regulations

Hybrid cloud

- **Security and compliance**, allowing highly regulated workloads to deploy on a private cloud and less-sensitive workloads on a public cloud
- **Scalability and resilience** to expand operations quickly, inexpensively and automatically using public cloud services and then scale back when surges subside
- **Resource optimization and cost savings** that make the best use of on-premises investments and infrastructure budget, changing deployments given shifting workloads or new opportunities

The next sections will dive deeper into deployment methods to help you choose the right EDW for your business needs.



According to McKinsey's *The State of AI in 2020*,[2] 66% of businesses reported an increase in revenue and 40% saw a reduction in costs due to AI adoption.

Cloud-native platform-based data warehouse

Should you deploy your data warehouse on a cloud-native platform? The agility, scalability and elasticity of cloud solutions are spurring many organizations to consider cloud IT initiatives. Here are several benefits:

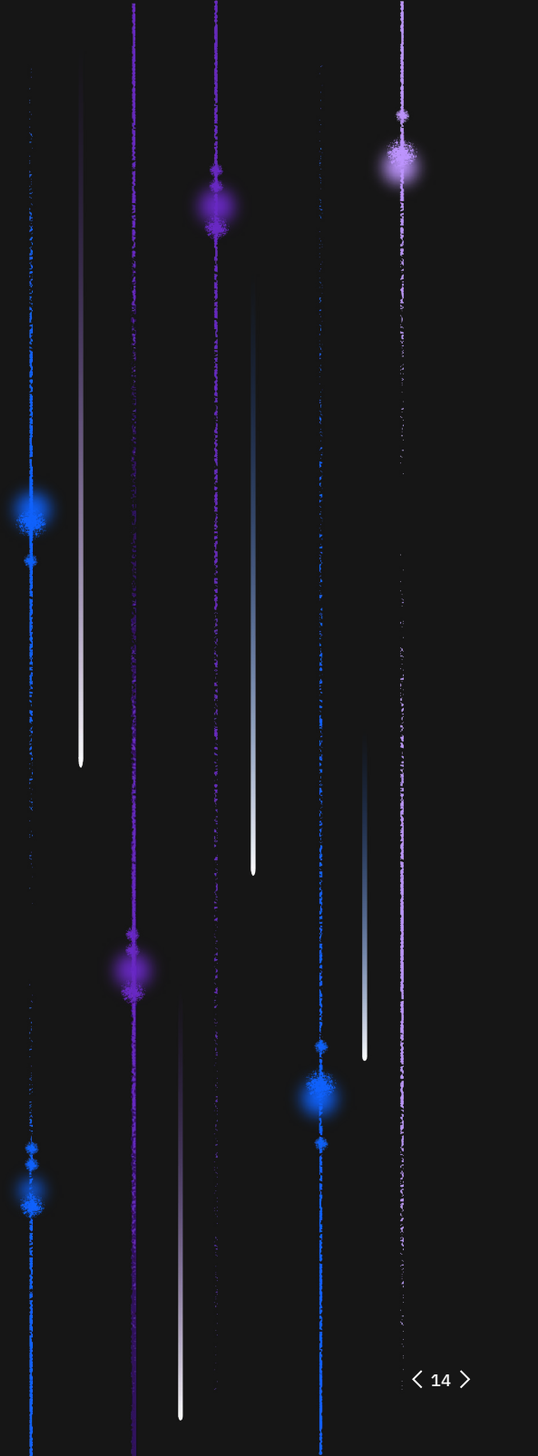
Integrate with other clouds and best-of-breed AI services
Deploy data warehouses on any cloud without moving data and save costs by consolidating all tools under a single infrastructure. On cloud-native platforms like IBM Cloud Pak for Data, you can connect a modern data warehouse with a full spectrum of AI tools or services to turn data insights into machine learning and AI:

- **Protect trust and compliance in data and AI:** Ensure that your data is governed, secure and compliant with regulations through IBM Watson® Knowledge Catalog.
- **Automate AI lifecycles:** Build and scale trusted AI with advanced data science and machine learning capabilities on Watson Studio.
- **Infuse AI across your organization:** Boost operational efficiencies and reimagine customer engagement with intelligent applications like Watson Assistant.

Analyze born-on-the-cloud data
Is your data "born on the cloud?" If you collect IoT data from sensors or mobile devices, you might decide to analyze that data in the cloud as well. Avoid the risks of transferring tremendous volumes of cloud-generated data on premises.

Streamline budgeting and speed deployment
How fast do you need it? The budgeting and planning processes for a new on-premises data warehouse can be time-consuming. You might need to pull together people and information from multiple departments. Installing, configuring, testing and upgrading the new data warehouse can be streamlined through a containerized, cloud-native platform with less upfront costs and effort.

Scale rapidly
Provision a data warehouse on demand with a few clicks, whether it's 1 terabyte or 1 petabyte. Automate all administrative functions, including backup and recovery, tuning and optimization, and patching and upgrading.



Hyper-converged data warehouse

For some organizations, a hyper-converged data warehouse that integrates optimized hardware and software in a single solution is the best choice. By combining storage, compute, networking and software, they can speed deployment and time to value with preconfigured, governed and security-rich high-performance. Since capabilities are tightly integrated, you can quickly get started with instant pre-assembled provisioning and save costs typical with public clouds.

Hyper-converged data warehouses, like Netezza® Performance Server, also provide greater scalability with pay-as-you-go models for resource expansion. According to analysts at Cabot Partners, a hyper-converged EDW can increase total value of ownership.

Keep sensitive data in-house
Protect sensitive data by keeping it in-house, where it's easier to comply with rigorous privacy regulations.

Speed deployment
Avoid the time-consuming processes of procuring equipment, installing software and configuring the environment—the solution components are designed to work together right out of the box. Enjoy cloud agility in your own data center, where you can plug into your network and start loading data the same day.

A hyper-converged data warehouse can increase SQL performance 3x with 1/5 of the footprint.[7]

Reduce management complexity
Increase efficiency by minimizing or eliminating administrative tasks such as tuning, indexing and aggregating tables.

Support fast-growing data volumes
Easily scale up your hyper-converged solution to support fast-growing data volumes. You can accommodate petabytes of data in a single environment.

Capitalize on data science technologies
Support advanced analytics by leveraging the latest data science technologies to make better and faster decisions. You can improve stock recommendations, produce targeted advertising, enhance fraud detection and more.

See a client case study of a hyper-converged EDW in action.

Watch the video →

On-premises data warehouse

Does your data already exist on premises? Then analyzing the data where it already resides might be your most effective option.

On-premises data warehouses are deployed on your company's own infrastructure behind an internal firewall. This has been one of the most popular and traditional deployments because you can have complete control over the management, configuration, customization and security of the infrastructure and data. In short, you know exactly where your data is.

A data warehouse on premises, like IBM® Db2® Warehouse, can minimize analytics latency and cut down the costs of moving large amounts of data to another environment. Expenses for high-speed network lines can be especially steep in global regions that lack a robust network infrastructure.

Comply with regulations
Are you prohibited from moving data? In healthcare, financial services and other fields, regulations might require you to keep sensitive data on premises. Even if there are permissions to transfer data to the cloud, you might be in a country with rigorous restrictions on where the data can reside and how it can be transferred across state or country lines. You might decide to keep data on premises so you can retain better control of your data.

Maintain flexibility
An on-premises environment does not necessarily compromise flexibility. For example, you can choose a data warehouse that lets you use your preferred hardware in a private cloud or virtual private cloud configuration.

A virtualized environment can also help enhance agility. With the right solution, you can deploy a pre-configured data warehouse on a Docker container in minutes. Automated scalability helps accommodate new analytics demands easily.

Leverage your existing IT environment
Have you made significant investments in your on-premises data warehouse? If you have an advanced infrastructure and strong, established skills for managing your data warehouse, you have good incentives to continue using them.

Enhance performance
By choosing an on-premises data warehouse that combines in-memory processing with in-database analytics, you can enable faster processing of complex queries. Keep latency to a minimum and lower the complexity and risk of moving data to an analytics cluster.

Which mix is right for you?

Simply put, AI isn't possible without a modern data warehouse. As the first step on the journey to AI, a modern EDW helps companies gain a complete view into their data and produce actionable insights. When integrated on a unified data and AI platform, the modern EDW helps companies master data management and form the robust information architecture needed for AI.

IBM provides choice and flexibility so you can identify the best EDW deployment for your business.

Need help selecting the right solution?

Talk to an expert today →

Cloud-native platform

IBM Cloud Pak for Data

A unified data and AI platform that runs on the cloud of your choice and modernizes data management, data governance and machine learning to help companies accelerate their journey to AI.

Read the Forrester study →

Hyper-converged

Netezza Performance Server*

An advanced data warehouse and analytics platform available both on premises and on cloud.

See competitive benefits →

On-prem and on-cloud

Db2 Warehouse*

A client-managed data warehouse that features in-memory data processing and in-database analytics for fast and flexible deployment.

Visit the website →

**Integrates with IBM Cloud Pak for Data*

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
March 2021

01 Forrester. "The Anatomy of a System of Insight" https://www.forrester.com/report/The+Anatomy+Of+A+System+Of+Insight/-/E-RES120088

02 McKinsey. "The State of AI in 2020." https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

03 IDC. "Accelerate Your AI Journey with a Hyper-converged Data and Analytics platform." https://www.ibm.com/account/reg/us-en/signup?formid=urx-43121

04 Gartner. "Gartner Predicts the Future of AI Technology." https://www.gartner.com/smarterwithgartner/gartner-predicts-the-future-of-ai-technologies/

05 https://www.oreilly.com/library/view/data-warehousing-in/9781491997963/ch01.html

06 CIO. "Optimizing Business Analytics by Transforming Data in the Cloud." https://www.cio.com/resources/form?placement_id=149fd5e1-6995-42d7-91ca-774eecc9c2b4&brand_id=256&locale=1

07 https://www.ibm.com/blogs/journey-to-ai/2020/11/ibm-cloud-pak-for-data-system-updates-provide-a-modern-hyperconverged-solution-for-transformation/