

# Cloudera, Dell Technologies, and NVIDIA Deliver Private AI on Premises

Meeting Enterprise Demand for Secure,  
Cost-effective AI Where Data Lives

**Stephen Catanzano** | Senior Analyst

ENTERPRISE STRATEGY GROUP

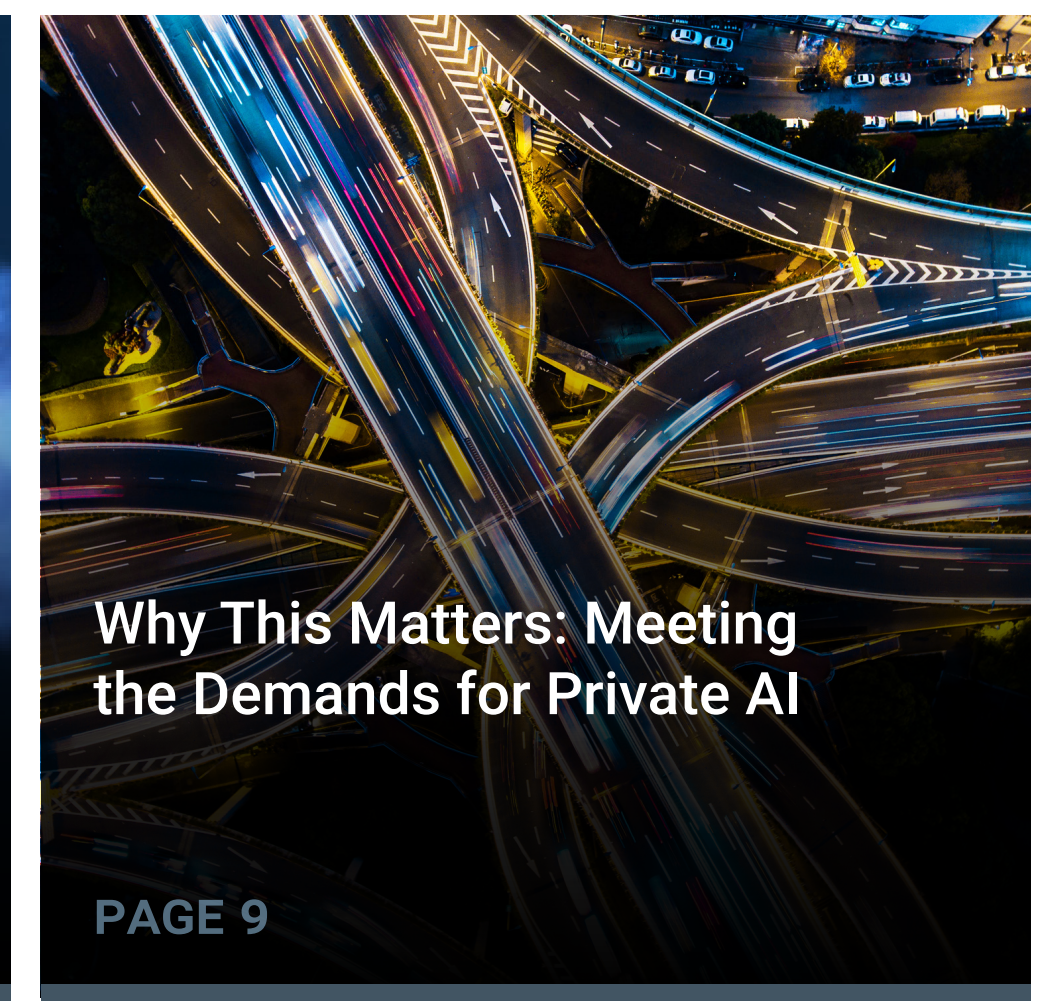
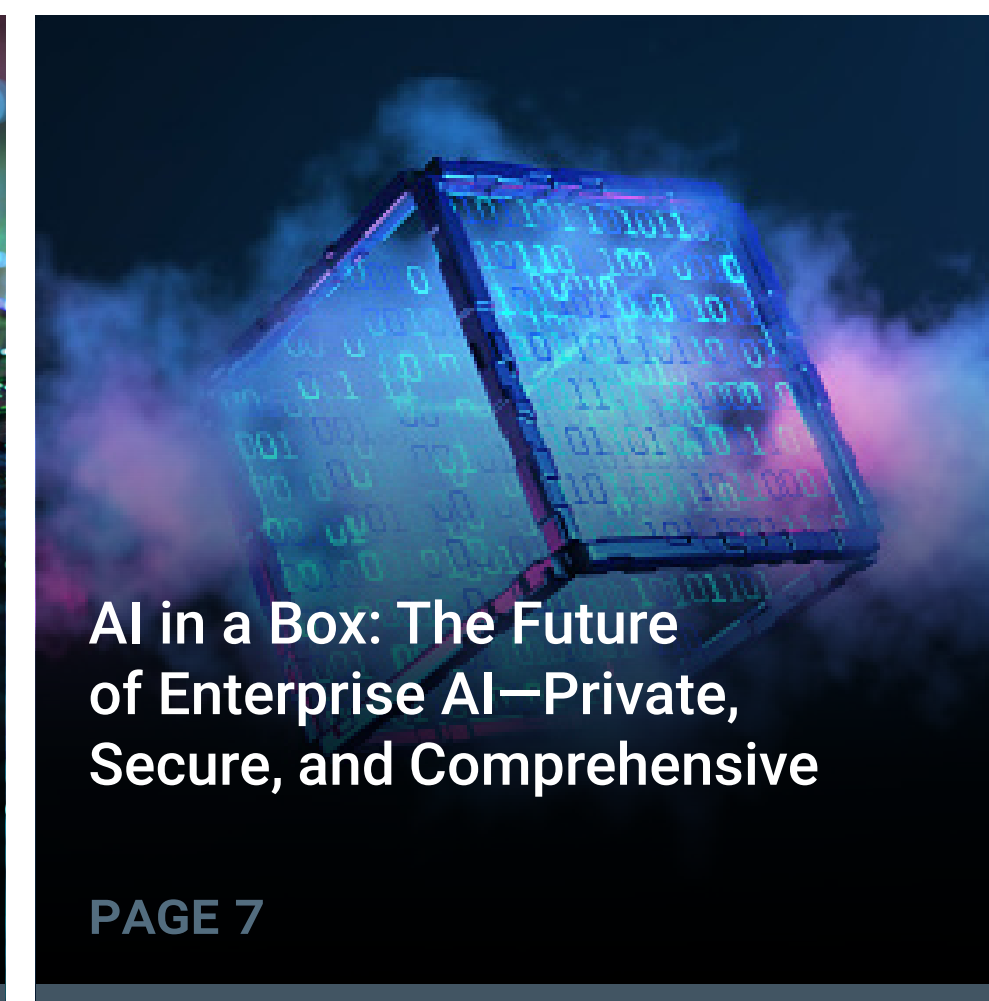
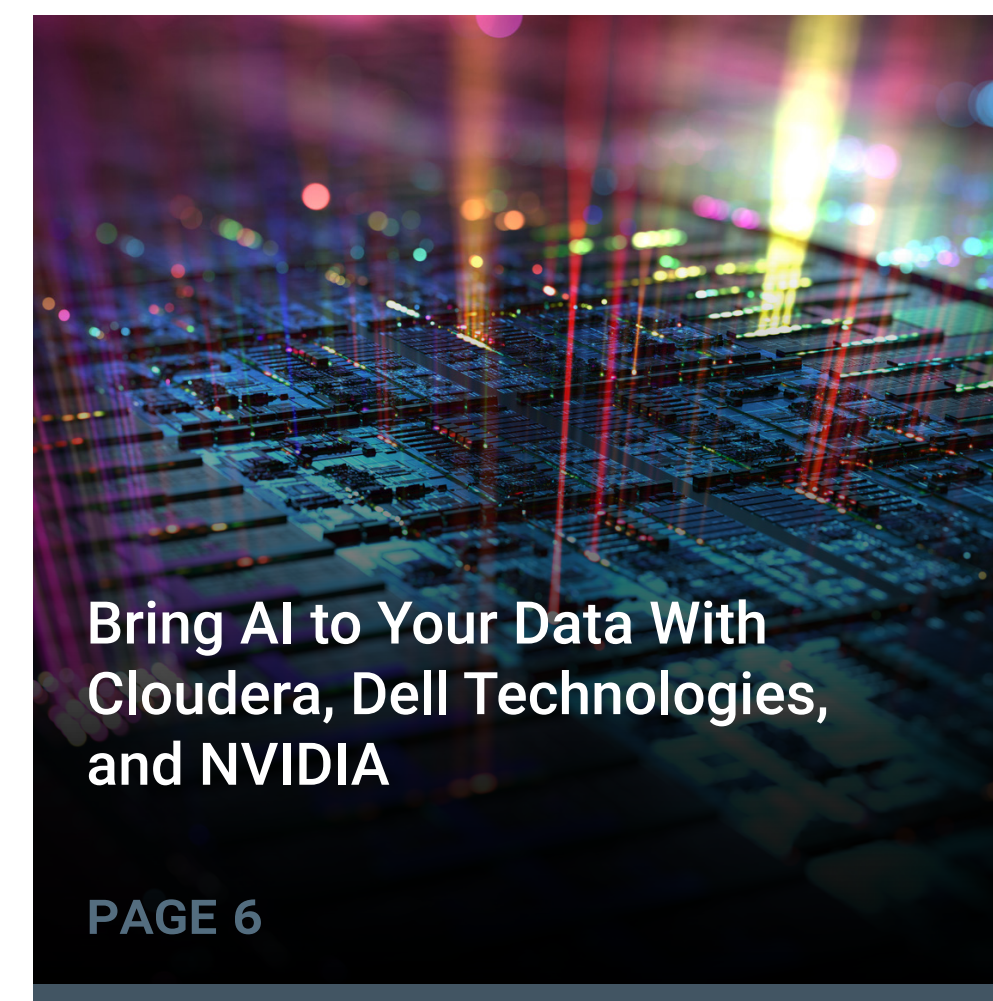
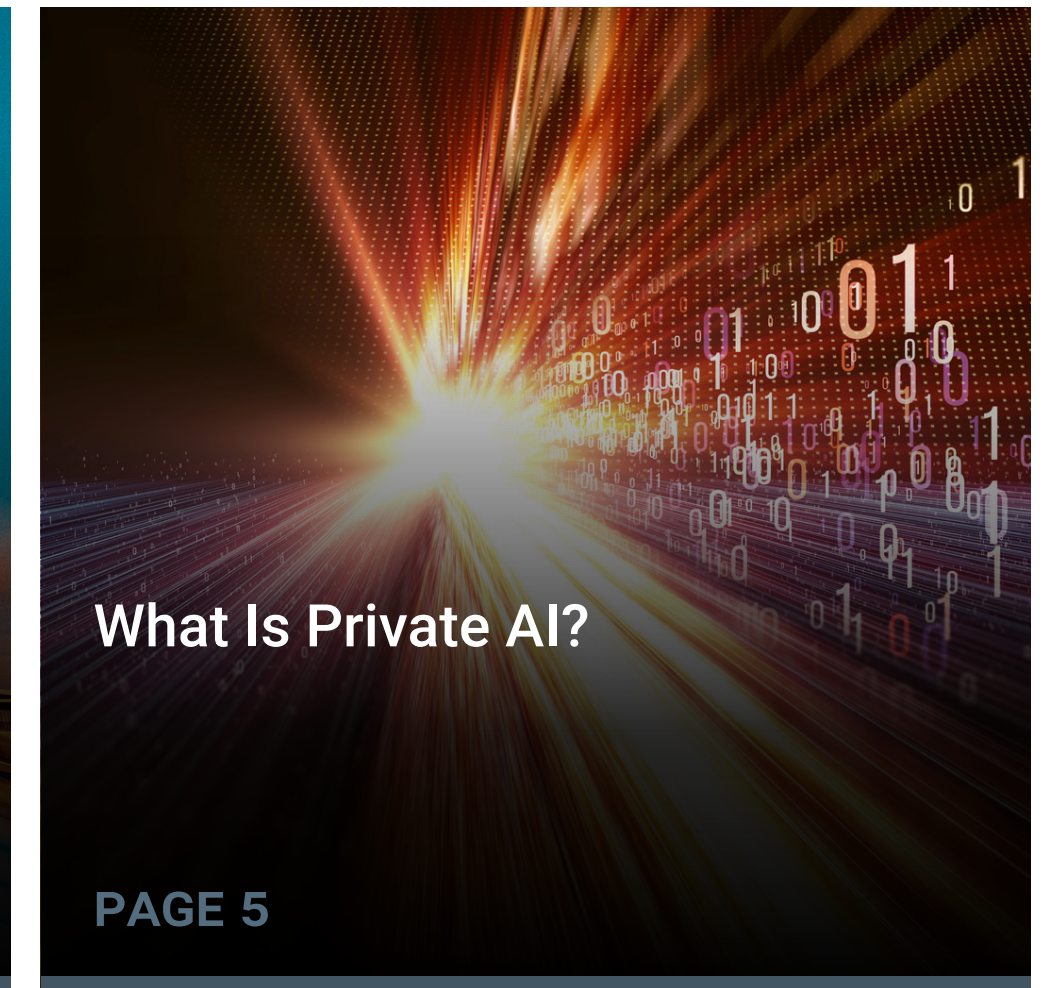
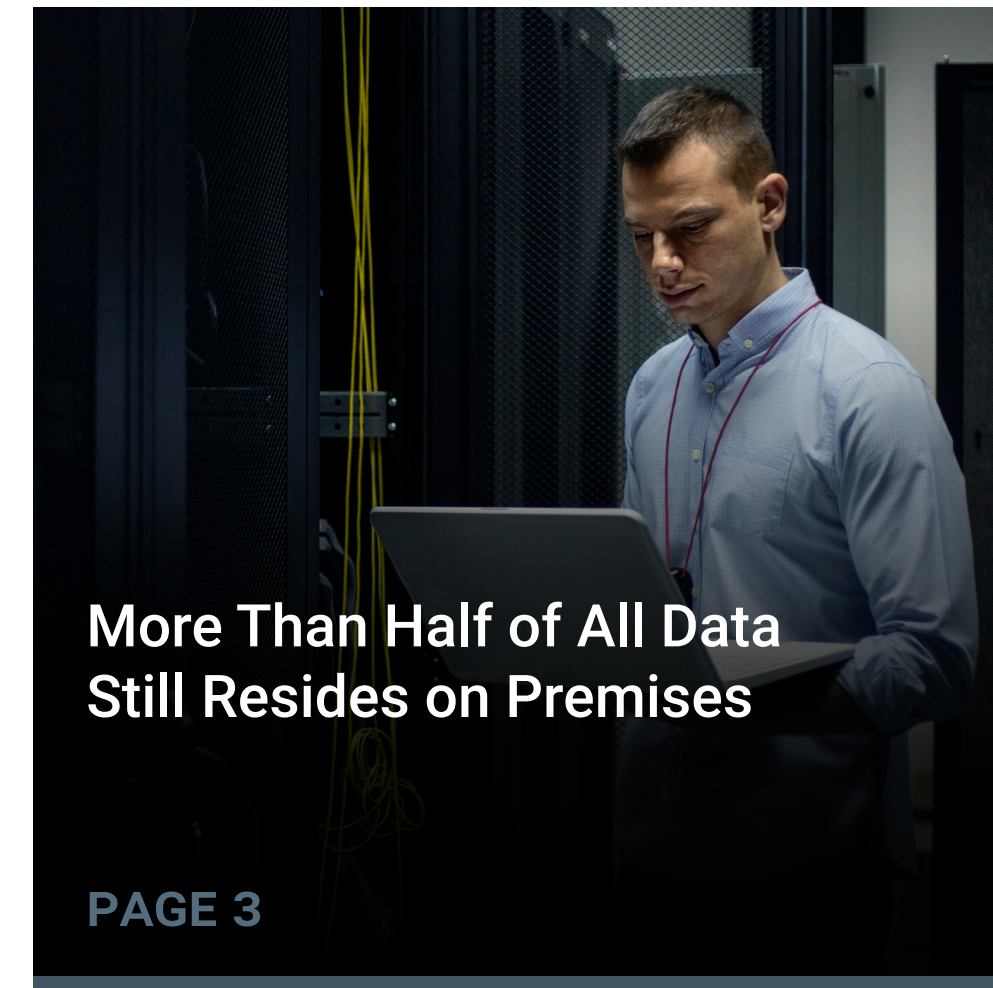
JUNE 2025

This eBook from Enterprise Strategy Group was commissioned by Cloudera and is distributed under license by TechTarget, Inc.

## Introduction

Enterprises are increasingly turning to AI and generative AI to drive innovation, but they're recognizing that the greatest value lies in leveraging their most sensitive data—data that often resides on-premises. Industries with strict security, privacy, and governance requirements are facing growing concerns around data sovereignty, regulatory compliance, and the escalating costs of cloud-based AI solutions. This has led to a strategic shift. Enterprises are rethinking cloud-first approaches for their AI initiatives and investing in private AI on-premises infrastructures that align with their compliance needs and long-term cost models. Cloudera, Dell Technologies, and NVIDIA have come together to meet this need, delivering private AI solutions that enable enterprises to build, deploy, and scale AI securely within their own data centers, offering full control, compliance, and cost-efficiency without sacrifice.

### CONTENTS



## More Than Half of All Data Still Resides on Premises

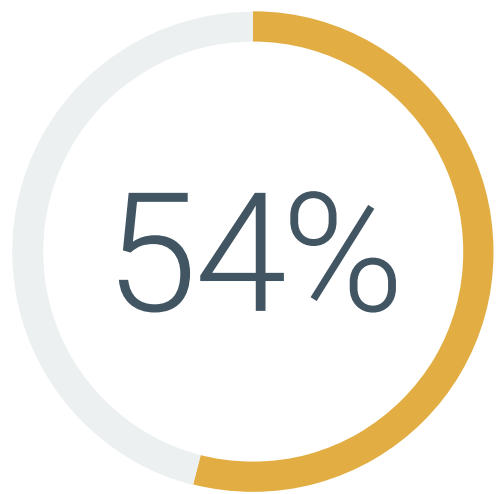
Despite the strong adoption of cloud technologies, a significant number of organizations continue to rely heavily on on-premises infrastructure for their data workloads. In fact, 51% of organizations still keep most of their data on premises. More notably, 54% of organizations have reversed course, moving some workloads back from the cloud to on-premises environments. This trend reflects a complex landscape where strategic, financial, and regulatory factors intersect—particularly in the age of AI.

One of the primary drivers behind this shift is regulatory compliance, cited by 29% of organizations. Industries such as healthcare, finance, and government face strict regulations that often require sensitive data to be stored within specific geographic boundaries or under certain conditions that cloud environments can struggle to guarantee. Data security is another major concern (28%), as companies fear loss of control over sensitive information once it leaves their local systems.

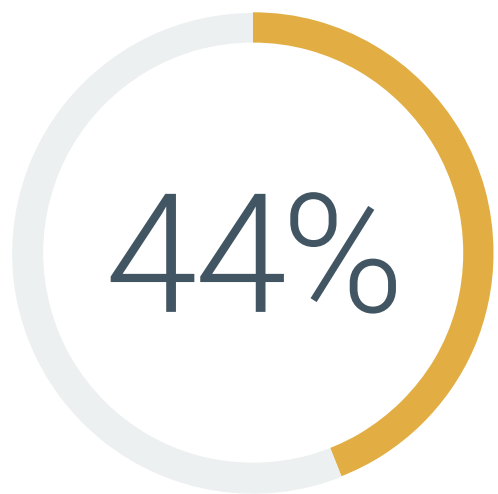
Financial factors are also at play. While cloud solutions promise flexibility, higher-than-expected costs have been reported by 22% of organizations. Additionally, 21% of organizations are deliberately relocating workloads as part of a broader strategy to lower cloud expenditures.

Crucially, the rise of AI is also influencing this shift. AI models require large, high-quality data sets—often proprietary or sensitive. Keeping this data on premises ensures tighter control, reduced risk, and better compliance, especially as data becomes a key differentiator in AI strategy. Organizations increasingly view data for AI as a strategic asset, and maintaining local oversight over it is becoming a competitive necessity.

## The Majority of Organizations Have Moved Workloads Back on Premises

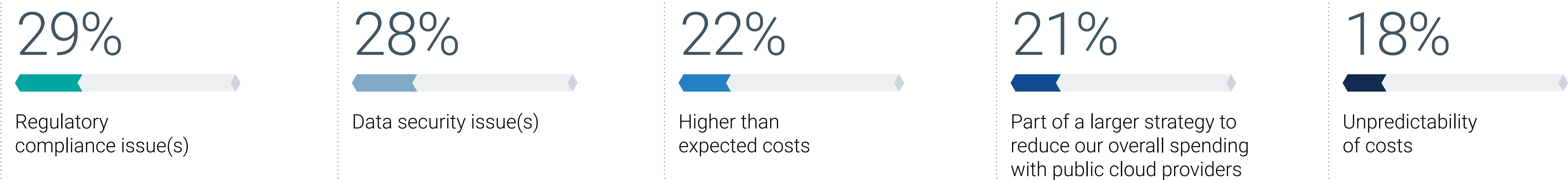


**Yes,** we have moved workloads back to on-premises data centers from public cloud infrastructure services



**No,** we have not moved any workloads back to on-premises data centers from public cloud infrastructure services

### Reasons Organizations Have Moved Workloads Back on Premises.



## Many AI Use Cases Require On-premises Enterprise Data

As organizations seek to enhance productivity and competitiveness, reduce operational costs, and discover new avenues for growth, AI is increasingly viewed as a core business enabler. However, realizing the full potential of AI often depends on the availability and accessibility of enterprise data—much of which remains on premises due to its sensitivity, scale, or regulatory requirements.

Enterprise Strategy Group research asked IT and data teams what their organization’s primary business drivers were for implementing AI. As shown, operational efficiency ranks as the top business driver and a key area where AI adds value by automating routine tasks and optimizing workflows. Many of these efficiency gains rely on direct access to internal systems and proprietary data sources, which are often better managed within on-premises environments to maintain performance and control.

In terms of customer experience, organizations are looking to AI to enable businesses to personalize engagement, anticipate needs, and deliver faster support. However, delivering such tailored experiences typically requires deep historical data on customer behavior and interactions—data that organizations might prefer to store and process locally to ensure compliance and data integrity.

AI is also fueling innovation by accelerating product development, enhancing R&D, and supporting real-time decision-making. This level of innovation demands robust, trustworthy data. On-premises data infrastructure gives enterprises the security, governance, and reliability needed to confidently train and deploy AI models on sensitive or business-critical data.

Risk management also plays a pivotal role as an AI business driver, especially across industries, from financial services to healthcare and manufacturing. By detecting anomalies, predicting failures, and identifying compliance risks in real time, AI enables organizations to proactively manage threats and reduce exposure. These capabilities, however, depend heavily on access to high-quality, historical, and often sensitive data sets, many of which are stored on premises due to regulatory or security considerations.

As the role of AI continues to grow, ensuring that enterprise data is both secure and accessible becomes a strategic imperative, driving many organizations to retain or repatriate data workloads on premises. This local control becomes particularly vital in data-intensive AI use cases that fuel operational improvement, customer engagement, and innovation.

### Primary Business Drivers for Implementing AI



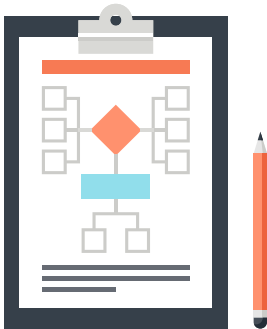
# What Is Private AI?

As part of a broader enterprise AI strategy, private AI plays a crucial role in enabling the secure, compliant, and scalable use of generative AI and AI agents with sensitive internal data through four key capabilities:



## Secure on-prem and hybrid AI deployments

Private AI refers to deploying AI models and systems in environments where data control, privacy, and governance are prioritized, such as on-premises infrastructure, private clouds, or secure hybrid environments. Unlike public AI models that rely on external, often cloud-based data processing, private AI ensures that sensitive enterprise data never leaves the organization’s trusted boundary. This is critical for industries like healthcare, finance, defense, and manufacturing, where regulatory compliance and intellectual property protection are non-negotiable. With private AI, organizations can harness generative AI and AI agents without exposing proprietary or regulated data to third-party models or cloud providers.



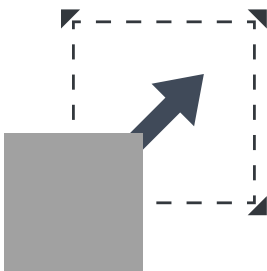
## Confidential AI workflows with full data sovereignty

A core attribute of Private AI is data sovereignty, which ensures that data stays within the legal and physical control of the organization. This enables confidential computing, localized model training, and inference directly on sensitive data sets. This is essential for use cases involving a customer’s personally identifiable information, financial transactions, employee records, or trade secrets. By bringing AI to the data rather than moving the data to the AI, companies reduce security risks and eliminate major legal and compliance roadblocks that can stall enterprise AI initiatives.



## Customizability and control over AI models and outputs

Private AI platforms enable organizations to fine-tune models, audit decision-making, and enforce strict access controls on both the data and the model outputs. This level of control is critical for aligning AI behavior with enterprise standards, brand voice, and ethical frameworks. It also reduces dependency on black-box third-party APIs, enabling companies to adapt AI agents to unique workflows and business contexts without compromising security or compliance.



## Trustworthy, scalable enterprise AI adoption

Trust is foundational to AI adoption, and Private AI directly addresses key concerns around reliability, explainability, and governance. It supports the secure deployment of AI agents that operate on sensitive internal data (e.g., HR assistants, contract analyzers, or executive briefing bots), without leaking information. Moreover, private AI enables horizontal scaling of AI capabilities across departments and regions while maintaining centralized governance and auditability, forming the backbone of a secure, responsible AI strategy.

# Bring AI to Your Data With Cloudera, Dell Technologies, and NVIDIA

With a tremendous amount of structured and unstructured data on premises, organizations need to bring AI to their data to unlock its full potential to create innovative and impactful AI solutions. Cloudera, in collaboration with Dell Technologies and NVIDIA, delivers an integrated AI solution designed to bring compute to data—wherever it lives. The solution offers the following:

## Accelerated compute anywhere

Organizations can leverage GPU-accelerated infrastructure across environments to scale AI workloads efficiently. Whether on premises or in the cloud, access to GPU resources ensures faster performance and optimized total cost of ownership.

## An any model, any framework approach

This approach offers support for any foundational model, large language model, or transformer library with full enterprise context. Whether an organization is building custom solutions or integrating open source tools, the platform is designed for maximum adaptability and innovation.

## AI inference on premises and in the cloud

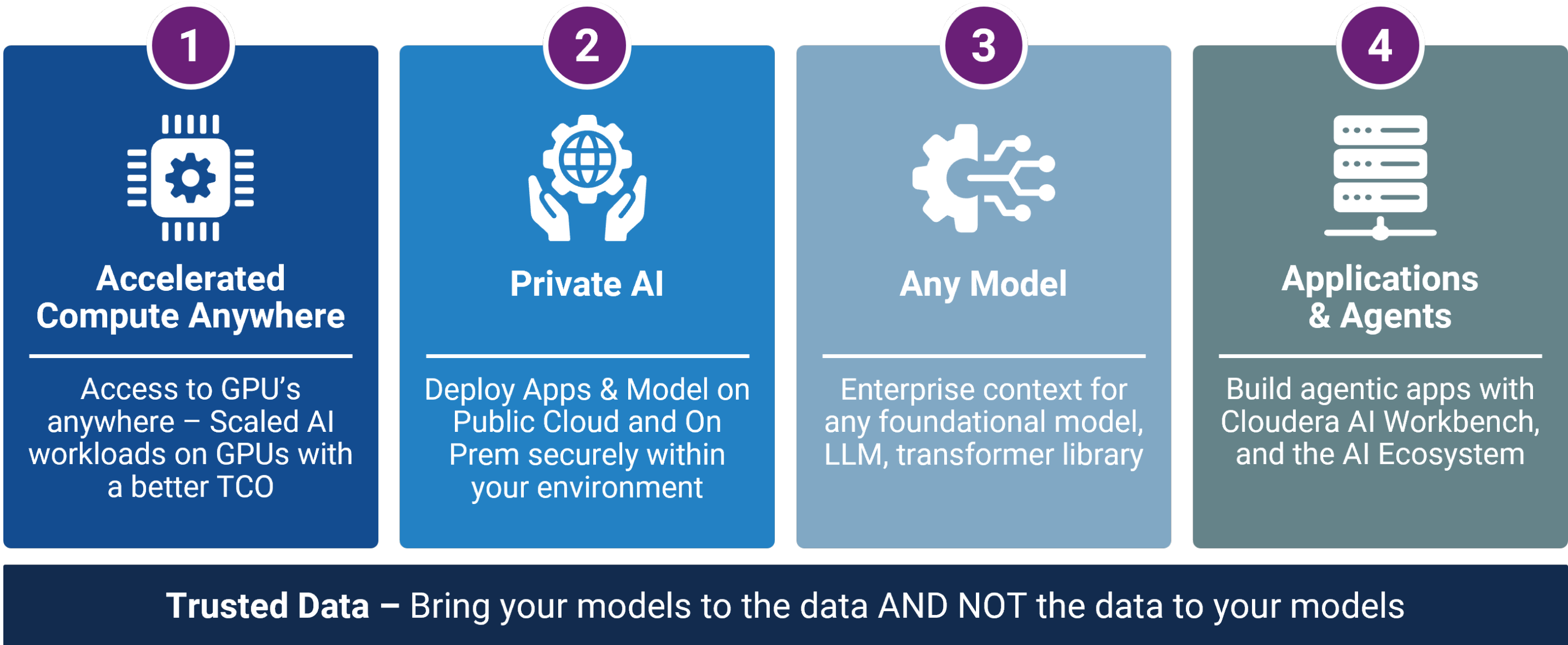
The solution enables organizations to run AI workloads on optimized software where it makes the most sense for their business. This hybrid flexibility enables organizations to deploy applications and models securely in their environment—on premises or in public cloud—ensuring data governance and compliance remain intact.

## Applications and agents

Users can create and deploy intelligent, agentic applications with Cloudera AI Workbench and a robust AI ecosystem. This empowers data teams to develop real-world AI solutions faster, with tools designed for scale and enterprise readiness.

At the core of this collaboration is trusted “AI-ready” data that is secure, scalable, and enterprise-grade. Instead of moving sensitive data to where models live, this solution flips the paradigm so that users can bring their models to the data, preserving control, performance, and privacy. Cloudera, Dell Technologies, and NVIDIA are reshaping enterprise AI with infrastructure, tools, and workflows that meet organizations’ data where it is, making AI transformation practical, secure, and impactful.


























### The Cloudera, Dell Technologies, and NVIDIA Joint Solution



# AI in a Box: The Future of Enterprise AI—Private, Secure, and Comprehensive

Cloudera’s AI in a Box, powered by Dell Technologies and NVIDIA, is a pre-validated, end-to-end AI solution. It’s designed to drive rapid AI innovation securely, efficiently, and cost-effectively in on-premises enterprise environments. This proven framework is designed to simplify and accelerate AI deployment on premises. Built for enterprise scale, AI in a Box offers a pre-validated, end-to-end stack that enables organizations to develop, deploy, and operate AI workloads securely and efficiently, leveraging best-of-breed infrastructure.

## AI in a Box

Services	<div><div> Professional Services</div><div> Strategy</div><div> Infra.</div></div>	<div><div></div><div> Use Case</div><div> Operate</div></div>	
AI Platform	<div><div> Frameworks</div><div> </div></div> <div><div> Inferencing</div></div> <div><div> Models</div><div> ANTHROPIC</div></div> <div><div> Vector Search</div><div></div></div>		
Data Management	<div><div> Open Data Lakehouse</div><div></div></div>	<div><div> NIFI INGEST</div><div> OBJECT STORAGE</div><div> OPEN TABLES</div></div>	
Acceleration	<div><div></div></div>	<div><div> GPU'S</div><div> Networking</div></div>	
Hardware	<div><div></div></div>	<div><div> PowerEdge (Servers)</div><div> PowerScale (Storage)</div></div>	

- **Dell Technologies: Trusted Enterprise Infrastructure**

Dell Technologies provides the hardware foundation for AI in a Box with its PowerEdge servers and PowerScale storage. Dell Technologies’ services span strategy, data, use case alignment, and operations, ensuring smooth end-to-end AI execution.











- **NVIDIA: Accelerated AI Performance**

At the heart of AI acceleration are NVIDIA’s accelerated computing and optimized AI software in Dell Technologies hardware. These deliver the performance needed for high-scale AI inference and training, with support for vector search and generative AI.

- **Cloudera: AI and Optimized Models**

Cloudera brings the AI platform and data management layers. From MLflow and LangChain to Llama and Cohere, it supports a wide AI model ecosystem. It’s an Open Data Lakehouse with SDX that ensures secure, governed, and scalable AI across hybrid cloud environments.

## Cloudera, Dell Technologies, and NVIDIA

<div></div> <div>Secure Hardware</div> <div>Enterprise-grade processing power</div> <div></div> <div>Up to 75% lower TCO than public cloud</div>	<div></div> <div>Accelerated Compute &amp; AI Software</div> <div>Optimizing GPU &amp; model performance</div> <div></div> <div>End-to-end security &amp; governance</div>	<div></div> <div>Scalable Software</div> <div>Enabling enterprise context</div> <div></div> <div>Extensible &amp; interoperable</div>	<div></div> <div>Optimized Model &amp; Agents</div> <div>Model of choice, depth &amp; scale</div> <div>  </div> <div>Pre-validated &amp; optimized AI solution</div>
--	--	---	--

Technical advantages of the AI in a Box solution include the following:

- **Integrated AI workflows:** Manage the entire AI lifecycle with Cloudera AI Workbench and integrated MLFlow with Kubernetes orchestration.
- **Optimized AI models and customization:** Access both tailored open source and commercial AI models, with the capability to fine-tune proprietary models using internal data sets and advanced hyperparameter tuning.
- **Ultra-low latency and high performance:** Deliver sub-millisecond response times using Dell Technologies’ high-performance infrastructure combined with NVIDIA accelerated compute and software optimization with NVIDIA NIM microservices.
- **Predictable, budget-friendly costs:** Eliminate the unpredictability of cloud egress and scaling expenses with fixed-resource allocation.
- **Robust security and governance:** Ensure data remains on premises with seamless embedding of compliance, data security, and governance into every stage of the AI workflow.

AI in a Box enables organizations to unlock the full potential of enterprise AI.



### Generative AI

Content and Code Generation, Document Summarization, Virtual Assistants and Agents



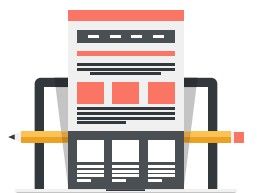
### Predictive Analytics

Fraud Detection, Customer Personalization, Predictive Maintenance



### Computer Vision

Quality Control, Visual Inspection, Safety Monitoring



### Digital Twin Applications

Prototyping, Simulation and Optimization, Logistics Management



### Data-driven Insights

Semantic Search, Entity Recognition and Classification



## Why This Matters: Meeting the Demands for Private AI

As organizations race to implement AI, many are faced with serious challenges: how to run AI securely, keep sensitive data on premises, and meet compliance without compromising speed or performance. AI in a Box addresses all of these needs.

This solution brings AI to an organization's data, not the other way around, enabling enterprises to develop and deploy advanced AI models directly within their own secure environments. For industries that handle regulated or proprietary data, this is critical. By keeping data local and governed through Cloudera's SDX framework, AI in a Box ensures full visibility, control, and compliance.

Security is built in at every level from Dell Technologies' trusted enterprise infrastructure to NVIDIA's AI-optimized performance and Cloudera's end-to-end platform governance. The result is a scalable, production-ready foundation for private AI that doesn't require sacrificing security, performance, or flexibility.

In short, AI in a Box empowers enterprises to innovate with AI confidently—securely, privately, and efficiently—while staying fully aligned with data sovereignty and enterprise governance requirements.

# CLOUDERA

## ABOUT

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world’s largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to use their data to solve what was once impossible—today and in the future.

LEARN MORE



©2025 TechTarget, Inc. All rights reserved. The Informa TechTarget name and logo are subject to license. All other logos are trademarks of their respective owners. Informa TechTarget reserves the right to make changes in specifications and other information contained in this document without prior notice.

Information contained in this publication has been obtained by sources Informa TechTarget considers to be reliable but is not warranted by Informa TechTarget. This publication may contain opinions of Informa TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent Informa TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, Informa TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of Informa TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at [cr@esg-global.com](mailto:cr@esg-global.com).



Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

© 2025 TechTarget, Inc. All Rights Reserved.