Omdia
by informa techtarget

JANUARY 2026

Achieving Economical AI Value Securely at Scale

Cloudera Delivers Private AI Powered by AMD EPYC CPUs on Dell Technologies Infrastructure

Stephen Catanzano, Senior Analyst

Abstract: There is widespread interest in AI, but not every enterprise needs graphics processing units (GPUs) to deploy it. As organizations race to integrate generative and agentic AI into business operations, they are discovering that many of the highest-value use cases, such as chatbots, content generation, knowledge management, and process automation, can be powered efficiently and securely using central processing units (CPUs) and small language models (SLMs).

Cloudera, powered by AMD EPYC processors and optimized on Dell Technologies infrastructure, offers a practical, cost-effective path for organizations to realize AI's promise in their own data center, close to their data, and at a fundamentally superior total cost of ownership (TCO), challenging the economics of GPU-based deployments.

A New Chapter in Enterprise AI Economics

Research from Enterprise Strategy Group (now Omdia) found that an overwhelming 80% of organizations said AI agents are a top or high business priority,[1] underscoring the clear ROI potential that agentic AI can deliver through automation, productivity, and efficiency gains. Yet many enterprises still struggle to operationalize AI, constrained by the cost, complexity of GPU-centric approaches, and the operational burden of inefficient legacy infrastructure.

The next phase of enterprise AI will be defined not by model size, but by return on intelligence, the ability to turn data into actionable outcomes quickly, securely, and affordably.

80% of organizations said AI agents are a top or high priority for their organization compared to other AI initiatives.

Cloudera, in collaboration with AMD and Dell Technologies, is redefining what practical AI looks like for modern enterprises. By leveraging SLMs that run efficiently on AMD EPYC CPUs, organizations can deploy Private AI solutions in their data center on existing or slightly upgraded infrastructure. Private AI by Cloudera is the deployment and operation of AI systems entirely within an organization's own secure infrastructure, ensuring that all proprietary data, configurations, and AI models remain within the organization's control. This approach guarantees that no data or insights are shared outside the organization, whether the AI is deployed on premises, in a private cloud, or in a hybrid environment. This approach democratizes AI adoption, enabling

[1] Source: Enterprise Strategy Group (now Omdia) Research Report, *AI Agents: The Game-changing Generative AI Use Case*, August 2025.

businesses to achieve measurable productivity gains without the high capital expenditure and energy costs of GPU-based systems.

Right-sized AI: The Power of SLMs

Not all AI needs to be massive to be meaningful. SLMs are focused, efficient generative AI models designed to perform specific business tasks, such as summarization, chatbots, code generation, and analytics, without requiring the compute or energy intensity of large language models.

SLMs are smaller by design, typically with fewer than 13 billion parameters, enabling them to run effectively on CPUs rather than specialized GPUs. This makes them ideal for on-premises enterprise environments where latency, security, and cost control are as important as accuracy.

When deployed on AMD EPYC CPU-based (or -powered) Dell Technologies infrastructure, Inference on SLMs deliver:

- Low-latency inference for real-time AI agents and automation
- High performance-per-dollar, with benchmarks on foundational analytics workloads showing very strong performance-per-dollar
- Data sovereignty and compliance, with computation staying within enterprise boundaries
- Sustainability benefits through lower power consumption versus GPU clusters

This "right-sized AI" approach embodies Cloudera's Private AI vision, delivering high-impact results using the infrastructure and data that enterprises already own.

Use Cases Are Driving Value Across Industries

According to research from Enterprise Strategy Group (now Omdia), the top enterprise AI use cases included increasing productivity (39%), improving and automating processes and workflows (38%), improving decision-making and accuracy (33%), and enhancing customer experience and engagement (31%).[2] These workloads are ideally suited to Cloudera Private AI running SLMs on AMD EPYC CPUs within Dell Technologies proven infrastructure. Following are some examples that show that enterprises can achieve powerful AI outcomes without GPUs.

Many AI use cases, including data insights, chatbots, content creation, and more, can have a strong TCO when running on Cloudera with AMD EPYC CPUs on Dell Technologies servers.

Internal Knowledge Base Creation

Enterprises hold vast troves of unstructured information across emails, documents, and reports. Cloudera Private AI ingests and structures this data into intelligent, queryable knowledge bases powered by SLMs. Employees can ask natural-language questions and get instant, accurate answers all within a secure,

[2] Ibid.

on-prem environment. The business impact includes faster decisions, better insights, less manual search, and full compliance with data-sovereignty requirements.

Enterprise Chatbots and Agent Assist

Internal service teams face a constant flood of repetitive employee questions. Deploying a secure, on-prem chatbot powered by Private AI with Cloudera, and running entirely on CPUs provides immediate ROI. The chatbot can handle HR, IT, and policy queries 24/7, freeing human experts for higher-value work. The business impact includes improved employee satisfaction, reduced operational overhead, and faster response times using Dell Technologies servers.

Automated Content Generation

Marketing and operations teams spend hours producing repetitive content. Private AI with Cloudera enables secure, rapid content creation by running SLMs on energy-efficient CPUs. The business impact includes accelerated production cycles, consistent messaging, and measurable Opex savings.

Code Generation and Documentation

For software teams, SLM-powered assistants hosted on Cloudera can automatically generate code snippets, tests, and documentation, all within the enterprise firewall. The business impact is faster development without GPU dependency.

Predictive Maintenance and Optimization

In manufacturing, SLMs running on AMD EPYC-powered Dell Technologies servers analyze sensor data to predict equipment failures and optimize processes. The operational impact includes reduced downtime, lower costs, and higher energy efficiency.

Data Sovereignty and On-prem Readiness

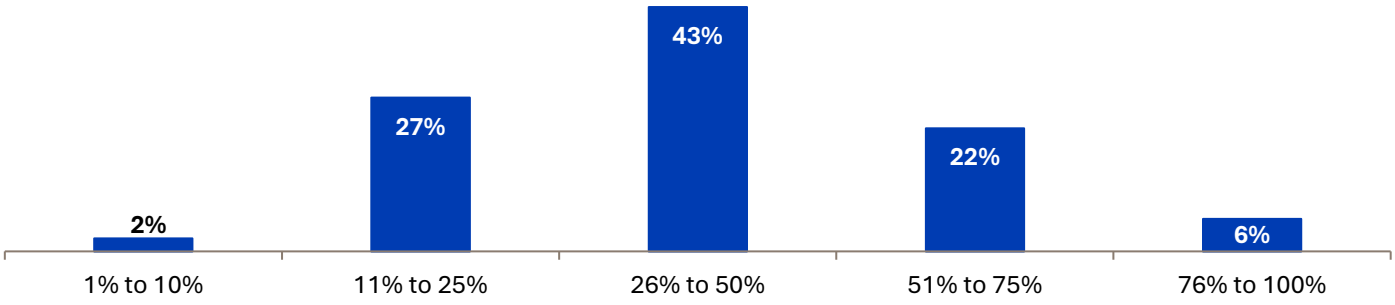
Research from Enterprise Strategy Group (now Omdia) shows that 65% of organizations keep between 26% and 75% of their data on-premises (see Figure 1),[3] highlighting a critical reality that data gravity remains in the enterprise data center. Moving sensitive or regulated data to the cloud for AI processing introduces risk, latency, and cost.

Private AI with Cloudera turns this challenge into an advantage. Its underlying Cloudera open data lakehouse architecture brings compute to the data, enabling enterprises to train and deploy AI models directly within their existing environments. AMD EPYC CPUs, with built-in security features such as AMD Infinity Guard and Secure Encrypted Virtualization, help safeguard data in use, while Dell Technologies infrastructure provides trusted, high-performance systems optimized for these workloads.

[3] Source: Enterprise Strategy Group (now Omdia) Research Report, *The Critical Role of Storage in Building an Enterprise AI Infrastructure*, September 2025.

Figure 1. Significant Amounts of Data Remain on Premises

Approximately what percentage of your organization's primary/active data capacity is deployed in on-premises locations? Please provide your best estimate. For the purposes of this question, please regard all data stored in colocation and edge facilities as being "on premises." (Percent of respondents, N=350)



Source: Omdia

Together, this stack enables organizations to monetize the value of generative and agentic AI while maintaining full control over their data, meeting the strictest regulatory requirements, including GDPR, HIPAA, and financial-sector mandates.

Positive Outcomes From Pilot to Production

While many organizations remain in proof-of-concept mode, Cloudera customers are already deploying production-ready, CPU-powered AI solutions. By leveraging existing infrastructure, they avoid expensive GPU procurement cycles, reduce time to deployment, and simplify management. Key outcomes include:

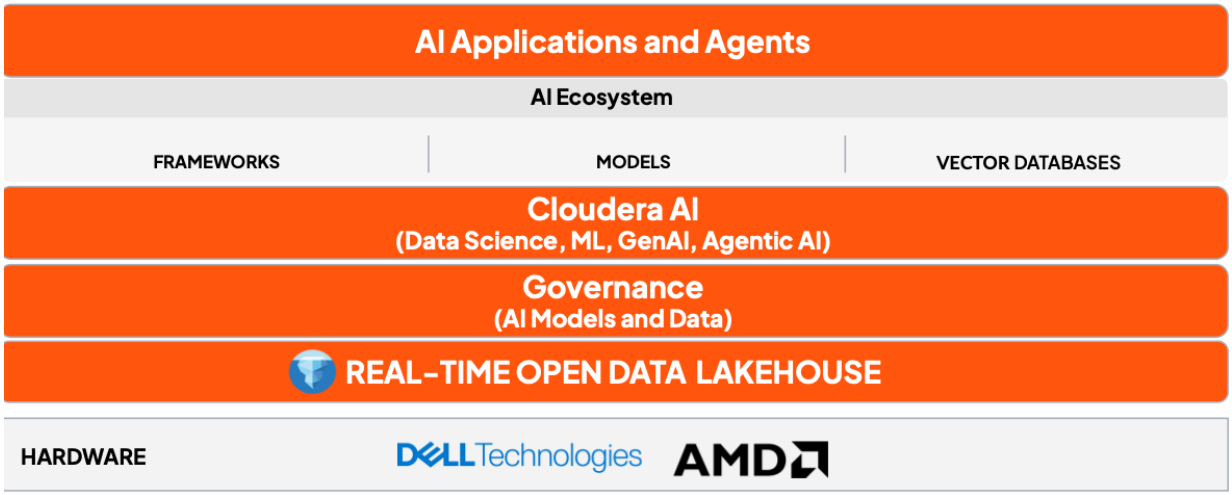
- Substantially lower TCO for common AI workloads
- Faster time to value from pilot to production
- Reduced energy consumption, supporting sustainability initiatives
- Higher ROI by matching compute to workload requirements

Modern CPUs are cost-effective compared to GPUs for many real-time inference tasks, especially when running SLMs optimized for single-query or low-latency operations, proving that right-sized AI delivers enterprise value faster and more efficiently.

Building the Private AI Foundation

Every AI success story starts with a trusted data foundation. Cloudera's unified data platform, shown in Figure 2, delivers an end-to-end lifecycle underpinned by its Shared Data Experience (SDX), providing a single control plane for unified security, governance, and metadata management and leveraging components like Apache Ranger and Atlas to enforce consistent policies for all data and AI workflows, management data ingestion, governance, MLOps, and model serving under a single control plane—all built on Dell Technologies infrastructure and powered by AMD EPYC CPUs.

Figure 2. Cloudera SDX



Source: Cloudera

Some advantages include:

- **Unified governance:** Cloudera's SDX ensures consistent policies, lineage tracking, and role-based access across AI workflows
- **Flexible architecture:** Support for containerized SLM deployment via Cloudera AI enables seamless scaling and monitoring across CPU clusters
- **Real-time open data lakehouse:** Cloudera's single, unified architecture, where organizations can store both raw and processed data, enables data scientists, analysts, and engineers to work from one platform
- **AI ecosystem:** This includes all of the AI frameworks, models, and vector databases needed to build and deploy AI solutions
- **Leadership:** AMD EPYC processors deliver market-leading performance-per-dollar and energy efficiency, resulting in superior ROI and sustainability outcomes

Scale With Cloudera, AMD, and Dell Technologies

The combined ecosystem of Cloudera, AMD, and Dell Technologies provides a proven platform for on-premises Private AI.

- **Cloudera** delivers the data and governance foundation
- **AMD** drives AI workflows with powerful, energy-efficient CPUs
- **Dell Technologies** provides scalable, enterprise-grade infrastructure and a Dell Technologies Validated Design for the Cloudera Data Platform

Together, they make Private AI practical and deployable today, on infrastructure organizations already own and aligned to their governance and sustainability goals

Conclusion

The future of enterprise AI isn't about chasing ever-larger models; it's about achieving better economics, stronger governance, and faster ROI by addressing the massive amount of data already existing in private data centers, achieving strong inference performance, and realizing lower TCO. With Cloudera Private AI powered by AMD EPYC CPUs on Dell Technologies infrastructure, organizations can realize the benefits of generative and agentic AI while maintaining data control and cost efficiency.

Whether the goal is to deploy intelligent chatbots, automate workflows, or extract insights from enterprise data, Cloudera provides a practical, private, and performant AI foundation. For organizations seeking strong TCO for AI solutions, Omdia recommends engaging with Cloudera to explore how CPU-based AI on Cloudera, AMD, and Dell Technologies could be the answer. To learn more visit http://www.cloudera.com/.

Get in touch: www.omdia.com askananalyst@omdia.com