Enterprise Strategy Group™ by TechTarget

WHITE PAPER

Accelerate Enterprise AI With Cloudera and NVIDIA

Bring AI Models to Your Data to Build Powerful AI Solutions

By Stephen Catanzano, Senior Analyst
Enterprise Strategy Group

January 2025

Contents

Abstract

Modern enterprises are racing to unleash the potential of their enterprise data to drive business value. The convergence of Cloudera AI, NVIDIA's advanced GPU processing and microservices, and a true hybrid model offers a seamless path to scaling data, AI, and analytics while ensuring data readiness, governance, and inference. By centralizing data, AI, and analytics workloads in Cloudera's SDX-enabled data lake services and accelerating performance with NVIDIA NIMS (NVIDIA Inference Micro Services) and GPU-powered architectures, organizations can deliver faster insights, optimize costs, and maintain robust compliance within an integrated data security and governance framework. This paper explores how these technologies power a unified, scalable AI-driven enterprise framework to support the data-driven outcomes organizations are looking to achieve.

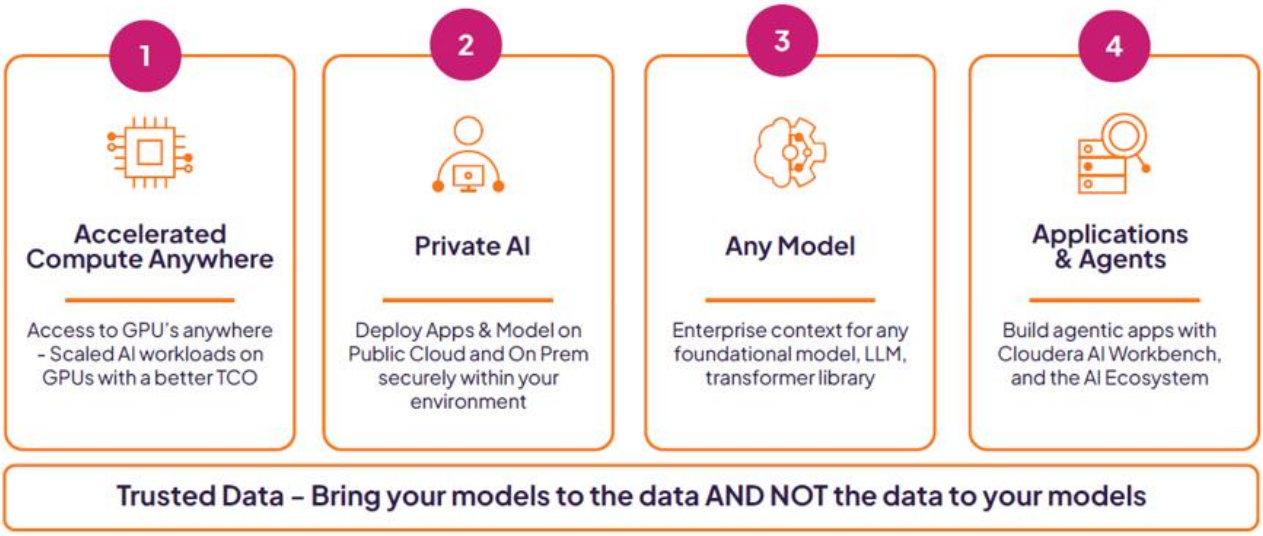
Introduction

As organizations increasingly adopt AI and analytics to drive innovation and decision-making, the need for trusted, high-quality data has never been more significant. Cloudera takes an active role in ensuring complete data accuracy and integrity for organizations. AI systems rely on vast amounts of diverse data to train models, deliver accurate insights, and enable intelligent automation. However, ensuring this data's reliability, security, and governance poses significant challenges, especially as data sources continue to expand for organizations and unstructured data becomes more prevalent. Organizations must address these issues to enable the full potential of AI, mitigate risks, and maintain user trust. A robust data foundation, combined with the ability to deliver insights at scale cost-effectively, is not just a prerequisite for AI and analytics success—it is a strategic imperative.

Organizations Need Agile AI Development

One of the core values of Cloudera is not just about building trusted data ready for AI but also delivering the flexibility organizations need in this ever-changing AI landscape, with data sources being both on premises and in the cloud. The four pillars of Cloudera support the agile development of AI solutions (see Figure 1).

Figure 1. Cloudera AI



Source: Cloudera

At the core of every AI or analytic project is trusted data, which has the data quality, accuracy, and governance needed for decision-making and building generative AI or predictive AI solutions with context derived from proprietary data. Unless an organization builds its own large language model (LLM), every organization will have access to the same AI capabilities. The big differentiator will always be an organization's data. With Cloudera, organizations bring their models to their data. Organizations can leverage embedded GPUs such as NVIDIA in the platform and even build private AI solutions that can operate on premises or in the cloud using Cloudera true hybrid capabilities, making using data in both locations seamless. Another key advantage is Cloudera AI workbench, which provides access to generative AI testing and development tools and AI models to create generative AI applications and agents. Models are rapidly advancing, and each use case might have different model criteria, including using multimodal (i.e., text, images, graphics) in some cases. Model performance and pricing are important factors to be considered. The flexibility offered on Cloudera is just one core value-add that organizations should consider as they bring enterprise data-powered AI solutions to life.

Business Strategy Meets AI Strategy

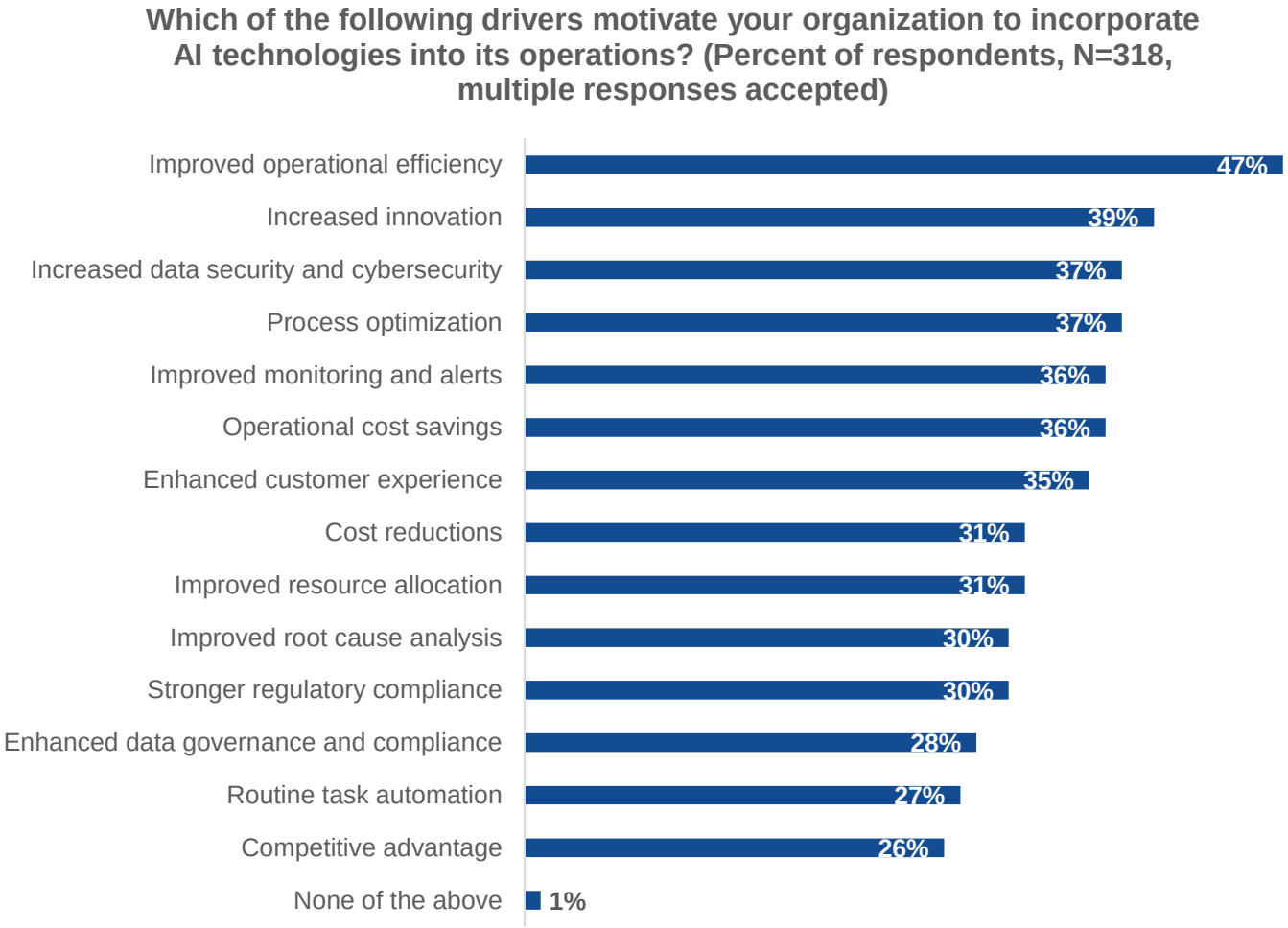
There has become a clear intersection of business strategy and AI strategy as organizations aim to achieve core organizational goals while driving innovation and competitiveness. When businesses focus on efficiency, savings, and customer experience, AI serves as a key enabler, automating workflows, accelerating decisions, and providing personalized solutions on a mass scale. A well-integrated AI approach supports business goals by discovering AI's value, including better use of resources, security, and regulatory requirements.

Research from Informa TechTarget's Enterprise Strategy Group shows organizations' motivating factors for AI implementation, specifically the promise of tangible operational and strategic benefits. The leading motivator, improved operational efficiency (47%), emphasizes increased process efficiency and productivity. Aside from productivity, companies increasingly see AI as an innovation agent (39%) and a process optimizer (37%), showing its ability to be creative and improve processes. Safety is also a top concern, with data security and cybersecurity (37%) ranked among the top three responses, suggesting the demand for AI-based solutions to protect essential assets. Cost savings, both operational savings (36%) and cost reductions (31%), are also significant, suggesting a dual focus on performance improvement and overall cost savings. In addition, incentives such as enhanced customer experience (35%), improved resource allocation (31%), and stronger regulatory compliance (30%) reflect the strategic imperatives of improving customer satisfaction and enforcing legal regulations (see Figure 2).[1]

Ultimately, this research highlights how businesses adopt AI for a cost-effective benefit and to stay competitive, creative, and safe in an increasingly data-driven and highly competitive business environment.

[1] Enterprise Strategy Group Research Report, *Navigating Data Governance in the Age of AI*, September 2024.

Figure 2. Organizations' Motivating Factors for AI Implementation



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

The AI Opportunity in Enterprise Data

It's an open space for organizations to use their existing, well-defined enterprise data to create genuinely unique generative AI solutions tailored to address many different use cases. It offers a way to differentiate and innovate. Whether it is customer data, operational processes, accounting data, or company-specific information, enterprise data represents an advanced, proprietary asset that can be used to generate company-specific generative AI models. Organizations can use their trusted data to build custom solutions that solve internal challenges, drive decision-making, and enhance customer experiences. For example, some organizations could use their historical customer data to produce adaptive AI-driven applications to create hyper-targeted marketing campaigns, product recommendations, or chatbots offering contextual customer support. In manufacturing, generative AI can harness operational and supply chain data to help improve designs, identify maintenance, or automate manufacturing. In the realm of finance, businesses can build models to identify a pattern of fraud or simulate a market situation in unprecedented detail.

Enterprise data offers the advantage of being highly specific and trusted so that organizations can build generative AI models that correlate with their business realities and strategies. These models can be fully developed internally, or techniques such as retrieval-augmented generation (RAG) can be utilized to work with leading models such as OpenAI, Gemini, or Claude and utilize the specific context of enterprise data without compromising. Agentic AI can also provide responses and use data and reasoning to take action.

Market Insight

Organizations Seeking Competitive Market Advantage Turn to AI

94% of organizations expected AI to have a moderate to significant improvement in leveraging analytics and business intelligence to gain a competitive advantage in the market.[2]

Some of the many use cases organizations are looking to build on top of their trusted data sources include:

- **Enhanced customer experiences:** AI systems can accurately understand and respond to customer needs and preferences. High-quality data enables AI models like chatbots and recommendation systems to provide personalized interactions and suggestions, significantly enhancing the customer experience. Effective data governance also ensures customer data is used responsibly and ethically, fostering trust and long-term customer relationships.
- **More effective support:** AI-driven support systems can access a comprehensive and accurate data pool to resolve customer issues more efficiently. AI models can use historical interaction data to anticipate problems and offer solutions proactively, leading to quicker resolutions and higher customer satisfaction.
- **Content creation and personalization:** AI can generate relevant content tailored to specific audience preferences and present it as personalized appearances or solutions to audiences. Good data practices ensure that the content creation process respects user privacy and adheres to regulatory standards, preventing issues like data bias and providing the context's appropriateness and relevance.
- **Data trust to empower intelligent decision-making:** Quality data and stringent data governance build a foundation of trust, which is critical for decision-making processes within an organization. Leaders rely on accurate, up-to-date data provided by AI systems to make informed strategic decisions. This trust is bolstered by data governance frameworks that ensure data integrity and compliance.
- **AI agents to execute tasks:** Working with agents to interact autonomously with business operations represents a pivotal advancement in organizational efficiency and innovation. These AI agents can execute tasks, make decisions, and even learn from their interactions. By leveraging agents, businesses can streamline repetitive tasks, improve accuracy, and free up human resources to focus on more strategic and creative endeavors. Additionally, these autonomous agents can adapt to changing circumstances, optimize processes over time, and provide valuable insights to drive further improvements in business operations.
- **Data democratization in the organization:** A well-managed data foundation facilitates data democratization, enabling employees across different levels of an organization to access and utilize data confidently for various tasks. Data quality ensures that the data employees access is accurate and relevant, enhancing productivity and innovation, while effective governance frameworks ensure that this data is accessed securely and in compliance with internal policies and external regulations, empowering more employees to leverage data-driven insights responsibly.

Cloudera is designed to manage data, AI, and analytics with the tools to test and deliver innovative generative AI solutions and accelerate the path to AI success for organizations.

[2] Source: Enterprise Strategy Group Complete Survey Results, *The State of Analytics and Business Intelligence Platforms*, April 2024.

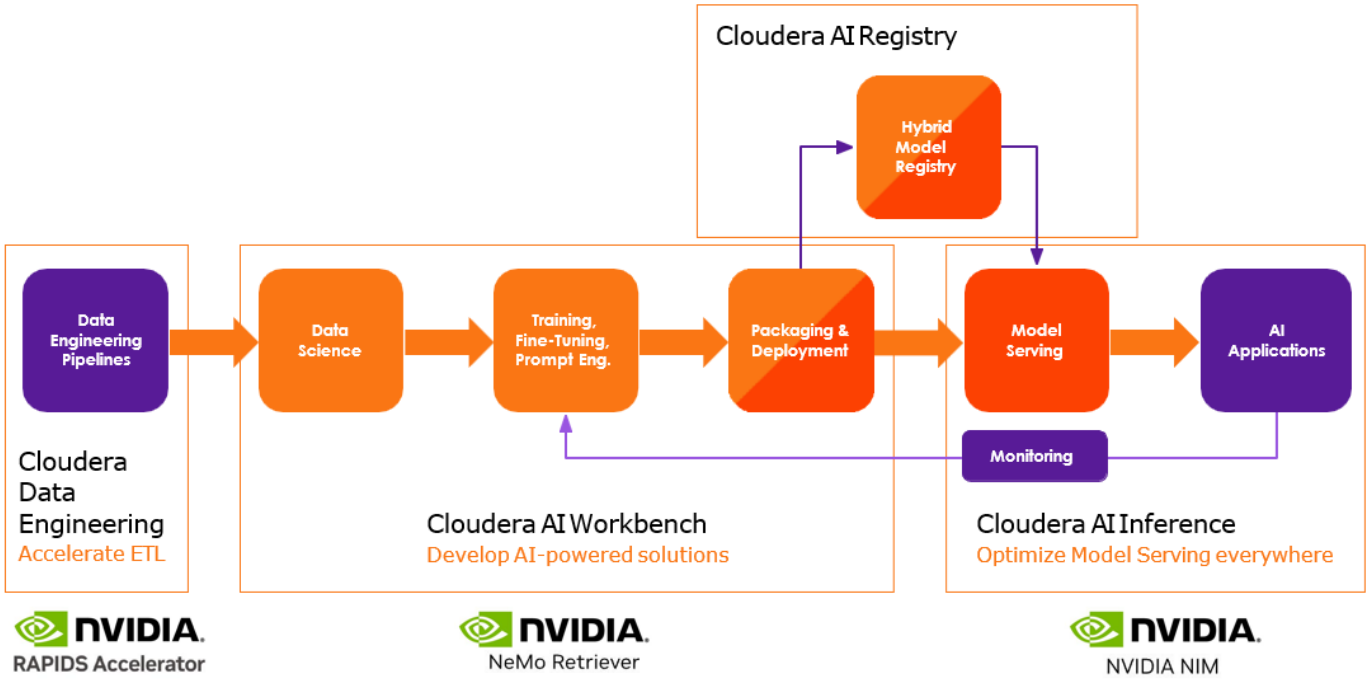
How Cloudera and NVIDIA Can Help Organizations Achieve Their AI Goals

Cloudera focuses on creating trusted data solutions for organizations, enabling the seamless democratization of data across the business. This empowers informed decision-making and supports efficient business operations. Key features of the Cloudera platform include:

- **True hybrid:** Cloudera's true hybrid provides a unified, consistent, and flexible data platform that seamlessly operates across on premises, public, and private cloud environments. This capability empowers organizations to build and manage advanced data architectures, including enabling data, AI, and analytics across hybrid environments, which is a unique and powerful capability.
- **Open data lakehouse:** Cloudera open data lakehouse design bridges the gap between data lakes and warehouses and provides scalable and secure data storage for structured and unstructured data. It enables businesses to easily collect and visualize data in hybrid environments without compromising scalability and vendor lock-in.
- **Self-service analytics:** Cloudera offers self-service analytics, enabling business users and data scientists to view, visualize, and interpret data in their own ways. This data democratization drives faster insights and gives teams the power to make informed decisions without heavily investing in IT.
- **Data transformation and pipelines:** In addition to powerful data transformation and pipeline tools, Cloudera automates and simplifies data preparation for organizations' analysis. From ingest to analysis, the platform facilitates many different data flow scenarios, delivering high-quality data fit for big data and AI use cases.
- **Scale and performance:** Cloudera aims to scale naturally, meeting the needs of even the largest businesses with vast amounts of data. Its hyper-performance design makes it possible for organizations to analyze and consume data in real time to aid mission-critical processes and rapid decision-making.
- **AWS relationship:** In collaboration with Amazon Web Services (AWS), Cloudera leverages AWS to offer its customers cutting-edge analytics and AI tools. This integration provides high performance, savings, and the flexibility to use the entire suite of AWS services and Cloudera's hybrid data infrastructure. NVIDIA CPU-based instances are also featured in Amazon Elastic Compute Cloud offerings and provide a seamless, cost-effective, and highly performant option for deploying Cloudera solutions. They include a wide range of NVIDIA-based AWS instances that are available globally.

Cloudera with NVIDIA provides a fast route to achieving trusted and secure generative and predictive AI and facilitates effortless integration with any language or foundation model for efficient AI development. When an organization has trusted data running on a high-performance platform, it can accelerate its path to using enterprise data to create unique AI-powered experiences. Figure 3 shows how Cloudera has integrated NVIDIA into the platform to deliver the high performance and efficiencies organizations require to deliver exceptional generative AI solutions.

Figure 3. Cloudera AI Unified Platform



Source: Cloudera

The platform consists of the following core elements:

Cloudera Data Engineering – Accelerate ETL

Cloudera uses the NVIDIA RAPIDS framework to power data science and machine learning operations with GPU-based processing. By bringing together Cloudera's scalability and the performance of RAPIDS, organizations can process massive data sets faster to get insights in real time and train models more effectively.

Cloudera AI Workbench – Develop AI-powered Solutions

Cloudera AI Workbench provides all the tools to test and develop AI solutions using an organization's trusted data on Cloudera, which is integrated with NVIDIA NeMo. This enables organizations to leverage LLMs within their secure enterprise data ecosystems. This integration ensures scalability, compliance, and high performance. Key features of the workbench include:

- **RAG Studio:** RAG Studio facilitates the creation of RAG applications by integrating knowledge graphs. It enhances the performance of RAG systems by capturing relationships and contexts that are not easily accessible by vector stores alone.
- **Fine Tuning Studio:** An all-encompassing application for managing, fine-tuning, and evaluating LLMs, this feature enables users to organize data, design training prompts, train adapters, and assess model performance within Cloudera's AI ecosystem.
- **Agent Studio:** This tool assists in building and deploying AI agents and agentic applications tailored to specific tasks, streamlining the process of creating intelligent agents that can interact with users or systems effectively.
- **Notebooks:** Cloudera provides interactive notebooks as a collaborative environment for data scientists and developers to write, execute, and share code, facilitating exploratory data analysis and model development.

- **Model Training and Fine Tuning:** Cloudera's platform supports the training and fine-tuning of machine learning models, enabling users to adapt pre-trained models to specific tasks or data sets.

Cloudera AI Inference – Bring AI Solutions to Life

After being trained, models are taken into production and deployed using a task known as inferencing. In contrast to data preparation and training, which involve taking large quantities of data into account, inferencing is focused on the speed with which a model will generate results. Models can be deployed to process thousands of requests per second from applications with special hardware, including NVIDIA GPUs, to ensure low latency. The inferencing infrastructure must be scaled out to prevent calls from becoming bottlenecks.

Cloudera AI Inference includes:

- **NVIDIA NIM Microservices:** NIM optimizes inference performance by leveraging NVIDIA GPUs for parallel processing, enabling efficient handling of large data sets and complex computations. Seamlessly integrated with Cloudera's AI platform, these microservices support deep learning and machine learning frameworks, ensuring models can be deployed and scaled efficiently in production.
- **Auto-scaling and high availability:** Auto-scaling dynamically adjusts resources based on workload demands, optimizing costs by scaling up during peaks and down during idle times. High availability ensures continuous operation through workload distribution, redundancy, and failover mechanisms, minimizing downtime and maintaining service continuity.
- **Monitoring:** Monitoring provides real-time visibility into system and model performance, tracking metrics like latency, throughput, and error rates. It enables proactive issue detection, timely alerts, and troubleshooting to maintain consistent workflows and meet service-level agreements.
- **Enterprise security and governance:** The platform enforces robust security measures, including role-based access control, encryption, and audit trails, to safeguard sensitive data and AI models. These features ensure compliance with governance policies and protect against unauthorized access or breaches.

The Cloudera AI Model Registry – Hybrid Model Registry

- This centralized repository is designed to give users access to AI models throughout their lifecycle. It provides version control to ensure traceability and rollback options across hybrid environments, enabling seamless deployment from the registry to the inference environment.

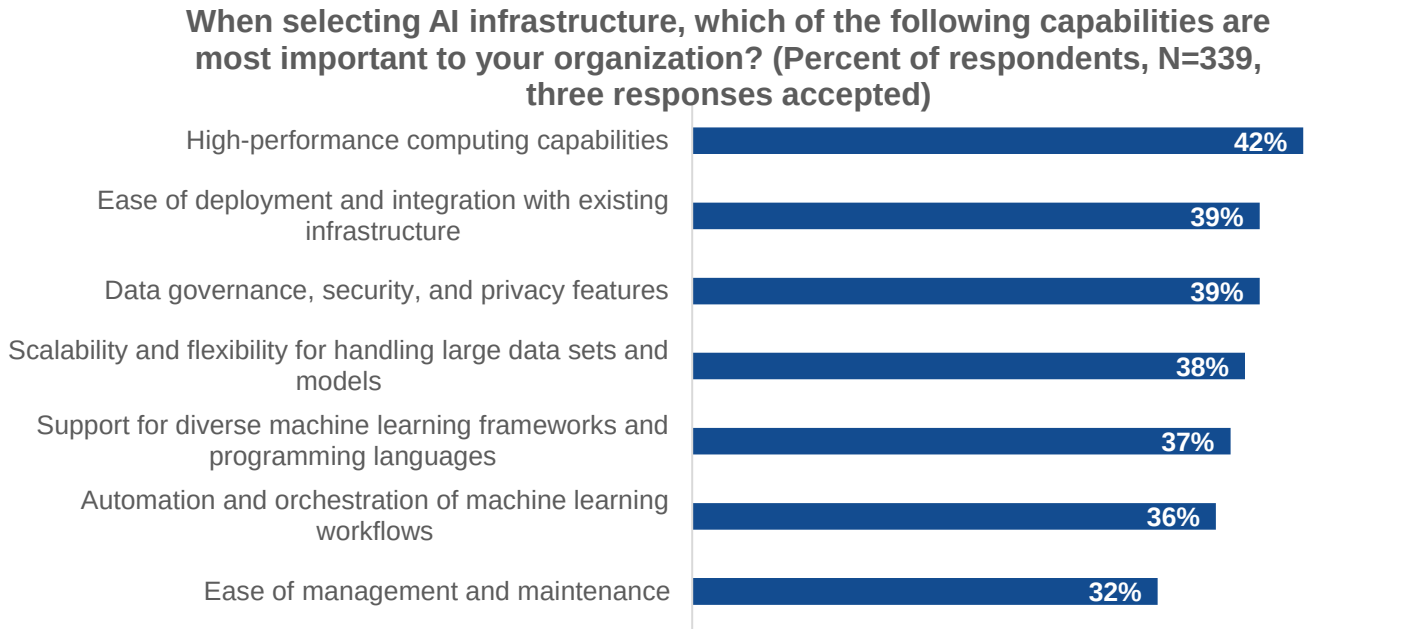
By combining Cloudera's robust data platform with NVIDIA's advanced AI acceleration, this workflow facilitates enterprise-grade AI adoption with speed, precision, and reliability, which is critical for organizations aiming to remain competitive and build impactful AI solutions.

Organizational Needs Align With Cloudera

Enterprise Strategy Group asked research participants about the most important capabilities to them when selecting AI infrastructure. According to the research, organizations' needs align with Cloudera's ability to manage their data, AI, and analytics to maximize their strategic value. High-performance computing capabilities were ranked most important among respondents (42%); followed by ease of deployment and integration with existing infrastructure (39%); data governance, security, and privacy features (39%); scalability and flexibility (38%), and more (see Figure 4).[3]

[3] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

Figure 4. Importance Considerations for AI Infrastructure



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Use Cases Across Industries

Across industries, organizations are looking to build unique, data-driven, generative AI applications to increase productivity, reduce costs, innovate, and remain competitive. Some of the primary industry use cases include the following:

- **Financial services:** Cloudera enables companies to leverage their enterprise data as a context for AI-based solutions, enabling breakthroughs in financial services. Graphene-accelerated AI detects transaction fraud in real time by analyzing historical context to detect fraud before it happens. Powered by enterprise data, predictive models deliver more reliable risk-management information to drive better decisions and help financial institutions mitigate risks in more informed ways.
- **Healthcare:** In medicine, Cloudera's use of enterprise data as context complements AI-based personalized medicine and drug discovery. By mining genomic and patient data—from history and clinical records—AI creates individualized treatment strategies that yield better outcomes. GPU-accelerated computing also leverages this data context to accelerate research and simulations, shortening the time to develop and launch new drugs.
- **Retail and e-commerce:** Cloudera's combination of enterprise data and AI results in robust retail and e-commerce solutions. AI-powered recommendation engines augment shopping with relevant customer information to maximize user engagement and sales. Integrated, real-time analytics utilizes business data to predict demand better, reduce inventory and waste, streamline supply chains, and drive efficiency.
- **Public sector:** Cloudera enables public sector agencies to bring their enterprise data into the context of AI solutions for real-world problems. Predictive maintenance models use historical infrastructure statistics to predict when something will fail—cutting downtime and costs. With smart cities, real-time analytics harness information from urban systems to better plan, manage resources, and provide services, resulting in smarter, more effective cities that enhance citizens' lives.

By bringing enterprise data to AI solutions, Cloudera enables businesses across industries to deliver more specific, context-aware insights on their unique use cases.

Conclusion

Cloudera provides the high-performance data platform organizations need to accelerate AI innovations across hybrid environments. Seamlessly extending capabilities from on premises to public clouds ensures enterprises retain complete control over their data, AI models, and applications. Cloudera's advanced data, AI, and analytics solutions are engineered to meet the diverse needs of modern enterprises, offering the scalability, reliability, and confidence required to power all AI use cases. Combined with the unparalleled performance and security of NVIDIA technology and the scalability of AWS, this solution lays a strong foundation for success in today's competitive landscape. Enterprise Strategy Group highly recommends Cloudera's all-in-one AI solution for organizations across industries.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group 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.

contact@esg-global.com
www.esg-global.com