# **Analysis of the NVIDIA AI Data Platform**

The transformative potential of artificial intelligence (AI) continues to reshape enterprise operations across industries. As organizations increasingly integrate AI into their core processes, the demand for robust and efficient AI infrastructure becomes paramount. Data management and processing lie at the heart of effective AI applications, particularly in the burgeoning field of agentic AI, which involves the creation of intelligent agents capable of reasoning, planning, and acting autonomously. NVIDIA, a recognized leader in accelerated computing, has strategically positioned itself to provide comprehensive hardware and software solutions that address these evolving enterprise needs. The NVIDIA AI Data Platform represents a key offering in this landscape, specifically designed to tackle the demanding requirements of AI inference workloads and the complexities of agentic AI. This report provides a detailed analysis of the NVIDIA AI Data Platform, examining its branding, value proposition, technical architecture, integration within the NVIDIA ecosystem, commercial aspects, strengths, and potential weaknesses.

## **NVIDIA AI Data Platform Overview**

The NVIDIA AI Data Platform for Enterprise serves as a customizable reference design intended for the development of a new generation of AI infrastructure.1 This platform is built upon the foundation of integrating enterprise-grade data storage with NVIDIA's renowned accelerated computing capabilities.1 The central aim is to empower AI agents with the ability to access and process vast amounts of business data in near real time, thereby enabling the generation of timely and accurate insights.1 This is achieved through the use of specialized AI query agents that are designed to significantly accelerate AI reasoning workloads.2

The platform's core value proposition is multifaceted, offering high-performance computing infrastructure optimized for agentic AI, distributed inference capabilities tailored for enterprise-scale deployments, comprehensive support for both structured and unstructured data, and native integration for Retrieval-Augmented Generation (RAG) workflows.1 By leveraging NVIDIA AI Enterprise software, which includes NVIDIA NIM microservices and the innovative NVIDIA AI-Q Blueprint, the platform provides a robust foundation for building intelligent and responsive AI applications.2

The primary audience for the NVIDIA AI Data Platform comprises enterprises across diverse sectors that have substantial AI inference demands, particularly those venturing into or expanding their use of agentic AI.1 Notably, NVIDIA is collaborating closely with leading data platform and storage providers, such as DDN, Dell Technologies, HPE, Hitachi Vantara, IBM, NetApp, Nutanix, Pure Storage, VAST Data, and WEKA, to bring customized AI data platform solutions to the market.1 These platforms are engineered to empower businesses to effectively utilize their enterprise data for complex reasoning and response tasks through the deployment of AI query agents.2 The versatility of the platform extends to hybrid data center environments, offering flexibility in deployment options.2 Furthermore, its capability to handle a wide array of data formats, including text, PDF, images, and video, underscores its broad applicability across various industries that rely on sophisticated data analysis and AI-driven insights.1

## **Technical Architecture of the NVIDIA AI Data Platform**

The NVIDIA AI Data Platform is designed to seamlessly integrate with existing enterprise storage infrastructures.1 At its core, it leverages the power of NVIDIA's high-performance hardware components, including Blackwell GPUs, BlueField-3 Data Processing Units (DPUs), and Spectrum-X networking, to significantly accelerate the process of accessing data stored within these systems.1 The BlueField DPUs play a crucial role in enabling multiple GPUs to efficiently share the same data repositories, thereby optimizing data throughput and enhancing access speeds critical for AI reasoning tasks.1 Complementing this, the Spectrum-X networking technology is instrumental in accelerating overall storage performance and mitigating communication bottlenecks that can impede the smooth flow of data, especially in distributed AI environments.1 NVIDIA's commitment to ensuring robust and reliable AI infrastructure is further underscored by the expansion of its NVIDIA-Certified Systems program to encompass enterprise storage solutions, a move that validates storage systems meeting the stringent data requirements essential for enterprise-level AI deployments.1

The platform's architecture is also designed to facilitate efficient data ingestion and integration from a variety of sources and formats. It provides the capability to process and derive insights from both structured data, such as databases, and unstructured data, including text documents, PDF files, images, and video content.1 A key feature is its native support for Retrieval-Augmented Generation (RAG), which allows for the seamless integration of an organization's proprietary knowledge, such as internal documents, customer records, and domain-specific information, directly into AI workflows.1 This integration significantly enhances the accuracy and contextual relevance of the responses generated by AI agents. NVIDIA NeMo Retriever plays a vital role in this process by accelerating the extraction and retrieval of data required for RAG, ensuring that AI models can access and utilize relevant information quickly and efficiently.1 This capability is further exemplified by integrations with partner solutions like the Dell AI Data Platform, which offers accelerated data ingestion leveraging NVIDIA NeMo, and VAST InsightEngine, which provides instant automated data ingestion with NVIDIA DGX systems.8

In terms of data processing and transformation, the NVIDIA AI Data Platform is engineered to optimize the distribution of AI workloads across available GPU and compute nodes.1 It employs sophisticated techniques such as intelligent routing, load balancing, and advanced caching mechanisms to significantly accelerate data processing tasks.1 The NVIDIA Blackwell GPUs, with their advanced architecture, provide substantial acceleration for a wide range of AI workloads, specifically optimizing the data processing pipelines essential for agentic AI applications.1 Furthermore, the NVIDIA BlueField DPUs contribute to this acceleration by expediting data access and reducing the processing burden on the central processing units (CPUs).1 The NVIDIA Spectrum-X networking further enhances overall performance by boosting storage throughput and reducing communication bottlenecks within the AI infrastructure.1 The platform's design also facilitates integration with established data science tools and workflows. For instance, Dell's AI Data Platform incorporates NVIDIA NeMo Curator and RAPIDS Accelerator for Apache Spark, which enhance the efficiency and speed of data preparation and transformation processes.8 Similarly, the collaboration between Starburst, Dell, and NVIDIA offers a platform designed for ingesting, transforming, processing, and securing large datasets specifically for AI applications.12

While the initial documentation on the NVIDIA AI Data Platform does not delve deeply into specific data governance and security features, these aspects are clearly recognized as critical within the broader NVIDIA ecosystem. NVIDIA AI Enterprise, which is a foundational software component of the Data Platform, offers enterprise-grade security features and regular security updates.2 Furthermore, NVIDIA is actively collaborating with security-focused technology partners like Cisco to deliver secure AI infrastructure solutions such as the Cisco Secure AI Factory.15 Companies like Securiti have also integrated NVIDIA NIM microservices into their AI solutions to ensure data security and governance within enterprise AI environments.17 NVIDIA itself provides resources like a security white paper for AI Enterprise, underscoring its commitment to securing the AI software stack.18 Given that the AI Data Platform is a reference design implemented by storage partners, the specific data governance and security features in the final product offerings will likely be a combination of NVIDIA's core security technologies and the robust security capabilities inherent in the storage solutions provided by partners like Pure Storage, Nutanix, VAST Data, and Dell, each of whom emphasize security in their offerings.8

## **Integration within the NVIDIA Ecosystem**

The NVIDIA AI Data Platform is not a standalone offering but is deeply integrated within NVIDIA's comprehensive ecosystem of AI hardware and software. A key synergy exists with NVIDIA AI Enterprise, which serves as the essential software foundation for the Data Platform.2 NVIDIA AI Enterprise is a cloud-native software platform designed to accelerate data science pipelines and streamline the development and deployment of production-grade AI applications.13 It provides optimized model performance coupled with enterprise-level security, support, and stability, ensuring a smooth transition from AI prototypes to production deployments.13 This platform supports a wide array of NVIDIA AI software available through the NVIDIA NGC catalog, offering a rich set of tools and resources for AI practitioners.13 Furthermore, AI Enterprise includes pre-packaged AI solution workflows for various applications, such as intelligent virtual assistants and cybersecurity threat detection, facilitating the rapid development of AI-powered solutions.13 Notably, NVIDIA NIM (NVIDIA Inference Microservices), a crucial component for accelerating the deployment of generative AI models, is an integral part of the AI Enterprise suite.13

While not explicitly identified as a core component of the initial AI Data Platform announcement, NVIDIA RAPIDS, a powerful GPU-acceleration platform for large-scale data analytics and machine learning, is expected to play a significant role in the broader context of the platform.25 RAPIDS includes a suite of open-source libraries that seamlessly integrate with popular data science tools like Apache Arrow, Pandas, and scikit-learn, enabling significant speedups in data preprocessing and machine learning training tasks.25 The integration of RAPIDS Accelerator for Apache Spark within partner solutions like Dell's AI Data Platform highlights its importance in accelerating the data science workflows that are essential for leveraging the NVIDIA AI Data Platform effectively.8

NVIDIA Triton Inference Server is another fundamental technology within the NVIDIA ecosystem that is crucial for the AI Data Platform. Triton is an open-source inference serving software designed for deploying AI models at scale in production environments.31 It offers broad support for various AI frameworks and provides essential features such as dynamic model loading and unloading, efficient batching of inference requests, and seamless integration with Kubernetes for scalable AI application deployments.31 Given that the NVIDIA NIM microservices, which are part of the AI Enterprise software used by the Data Platform, are built upon inference engines like Triton Server 13, it is highly probable that Triton will serve as the underlying inference engine for the AI query agents that are central to the Data Platform's functionality.

NVIDIA NeMo, a scalable generative AI framework, is also deeply integrated with the NVIDIA AI Data Platform, particularly in enabling its advanced AI capabilities.36 NVIDIA NIM microservices for the new NVIDIA Llama Nemotron family of reasoning models, which provide on-demand AI reasoning capabilities, are a key component of the platform.2 Furthermore, the NVIDIA AI-Q Blueprint, which facilitates the development of agentic systems, utilizes NVIDIA NeMo Retriever microservices to accelerate data extraction and retrieval processes.2 The integration of NVIDIA NeMo in partner solutions like Dell's AI Data Platform further underscores its importance in enabling the generative AI and RAG capabilities that enhance the overall value proposition of the NVIDIA AI Data Platform.8

NVIDIA AI Workbench is likely to be a valuable tool for developers who are building applications and AI agents that will operate on the NVIDIA AI Data Platform.38 This developer toolkit simplifies the AI development process by allowing developers to begin their work on local machines and then seamlessly transition their workloads to scalable GPU resources available in data centers or cloud environments.38 AI Workbench provides access to NVIDIA's pre-built container environments and supports collaborative development through integration with Git repositories.38 The use of AI Workbench for building RAG chatbots, as highlighted by Dell, further suggests its role in the development ecosystem surrounding the NVIDIA AI Data Platform.42

While not explicitly detailed in the initial announcements regarding the AI Data Platform, NVIDIA Base Command Platform, a software service designed for enterprise-class AI training and providing centralized control over AI training projects, may play a role in managing the underlying infrastructure, especially in large-scale deployments.43 NVIDIA Base Command Manager streamlines critical tasks such as cluster provisioning, workload management, and infrastructure monitoring.45 Given that Base Command Manager Essentials is certified for use with NVIDIA AI Enterprise 45, it is plausible that it will be utilized, either directly or through integrations within partner solutions like the Dell AI Factory, to manage the resources associated with the AI Data Platform. The validation of NVIDIA Run:ai, which offers similar AI orchestration capabilities, by the Dell AI Factory further supports the need for such management tools.10

The NVIDIA NGC Catalog serves as a central hub for developers utilizing the AI Data Platform to access a wide range of GPU-optimized software resources.48 This catalog includes containers, pre-trained AI models, software development kits (SDKs), and other essential components that can significantly accelerate the development and deployment of AI applications on the platform.48 Since NVIDIA AI Enterprise, which is integral to the Data Platform, supports all NVIDIA AI software available on NGC 13, developers can readily leverage the resources within the catalog to jumpstart their AI projects and optimize them for the underlying NVIDIA hardware.

## **Ecosystem and Marketplace Presence**

The NVIDIA AI Data Platform benefits from a strong ecosystem of technology partners, most notably the leading data platform and storage providers who are collaborating with NVIDIA to bring the platform to market.1 These partners, including DDN, Dell Technologies, HPE, Hitachi Vantara, IBM, NetApp, Nutanix, Pure Storage, VAST Data, and WEKA, are leveraging the platform's reference design to build customized AI data platform solutions that are tailored to meet the specific needs of enterprises seeking to harness their data for advanced AI reasoning.1 Beyond these core storage partners, the broader NVIDIA AI ecosystem encompasses a diverse array of companies across various layers, including solution integrators, data and enterprise platform providers, hybrid and multi-cloud vendors, AIOps and framework developers, and application specialists.53

A network of consulting and implementation partners plays a crucial role in facilitating the adoption of the NVIDIA AI Data Platform by enterprises.53 These partners, such as Accenture, Deloitte, EY, Infosys, and others, possess deep expertise in various industries and can provide invaluable assistance to organizations in navigating the complexities of implementing the platform and leveraging it for their specific AI use cases.53 Certified NVIDIA Business Partners like Nextrealm.ai further contribute to this ecosystem by offering specialized expertise in enterprise AI integration and consulting services.59

The NVIDIA AI Data Platform also has a presence in various marketplaces, primarily through the availability of NVIDIA AI Enterprise, which is a key software component. NVIDIA AI Enterprise is offered on major cloud marketplaces, including AWS, Azure, and Google Cloud.53 These marketplaces provide optimized and standardized environments, often in the form of Amazon Machine Images (AMIs) or Virtual Machine Images (VMIs), that simplify the process for enterprises to access and deploy the NVIDIA AI software suite within their preferred cloud infrastructures.53 Additionally, the NVIDIA Marketplace itself serves as a comprehensive platform listing a wide range of NVIDIA hardware and software solutions, including offerings relevant to enterprise AI deployments.1 NVIDIA AI Enterprise is also available on the Oracle Cloud Marketplace, further expanding its accessibility across different cloud platforms.63

## **Commercial Aspects of the NVIDIA AI Data Platform**

The pricing for the NVIDIA AI Data Platform itself, as a reference design, is not directly available. However, the pricing for NVIDIA AI Enterprise, which is integral to the platform, follows a multi-faceted model.66 It is available through subscription licenses, consumption-based pricing on cloud marketplaces, and perpetual licenses that require the purchase of support services.66 Subscription pricing for NVIDIA AI Enterprise Essentials has been observed starting at $3,750 for a three-year license covering a single GPU.67 Historically, perpetual licenses were offered at a price point of $3,595 per CPU socket, with additional annual costs for support services.68 On cloud marketplaces like AWS and Azure, NVIDIA AI Enterprise Essentials is typically priced on an hourly basis per GPU, following a pay-as-you-go model.61 For specific and up-to-date pricing details, it is recommended to contact authorized NVIDIA partners or the NVIDIA sales team directly.1

While the documentation does not explicitly detail cost management tools specifically for the NVIDIA AI Data Platform, NVIDIA does offer resources like an Energy Efficiency Calculator as part of its broader initiatives in sustainable computing.1 Partner solutions, such as the Dell AI Data Platform, emphasize the delivery of cost-effective AI systems through their integration with NVIDIA technologies.8 Furthermore, customer testimonials, such as the one from Amdocs regarding their use of NVIDIA NIM, highlight the potential for significant reductions in token consumption, which can translate to lower operational costs for deployed AI applications.69 The platform's inherent focus on hardware acceleration through DPUs and specialized networking, coupled with software optimizations provided by NIM and RAPIDS, suggests an underlying strategy to enhance efficiency and thereby reduce the operational expenditures associated with AI workloads.

The total cost of ownership (TCO) for the NVIDIA AI Data Platform is a significant consideration for enterprises. NVIDIA's overall aim is to provide a TCO advantage by delivering superior performance and maximizing the utilization of computing resources.70 Studies have indicated that leveraging NVIDIA data center GPUs for large-scale batch processing can lead to notable TCO benefits due to factors such as increased energy efficiency, reduced cooling requirements, and enhanced scalability.71 Understanding the TCO of AI infrastructure involves a comprehensive assessment that includes not only the initial purchase costs but also ongoing maintenance and operational expenses.72 Factors like infrastructure silos can negatively impact TCO, highlighting the importance of integrated platforms.73 The performance improvements offered by newer NVIDIA GPU architectures, such as Blackwell, are also expected to contribute to better TCO for AI inference systems.74 Ultimately, the TCO for the NVIDIA AI Data Platform will be influenced by the specific storage partner solution adopted, the scale of the deployment, and the efficiency gains achieved through the platform's hardware and software optimizations. A detailed TCO analysis would necessitate a thorough evaluation of specific partner offerings and the intended AI use cases.

## **Strengths and Weaknesses of the NVIDIA AI Data Platform**

The NVIDIA AI Data Platform offers several compelling strengths. Its ability to deliver accelerated AI reasoning and near real-time insights is a key advantage, particularly for demanding inference workloads.1 The platform's foundation on high-performance NVIDIA hardware, including Blackwell GPUs, BlueField DPUs, and Spectrum-X networking, provides the necessary computational power and efficient data pathways for complex AI tasks.1 The integration of NVIDIA AI Enterprise software and NIM microservices ensures a robust and scalable software environment.1 Native support for Retrieval-Augmented Generation (RAG) with NVIDIA NeMo Retriever enhances the accuracy and contextuality of AI agent responses.1 Furthermore, the strategic collaborations with leading data platform and storage providers ensure a strong ecosystem and a variety of tailored solutions for enterprises.1 The platform is specifically optimized for demanding AI inference workloads and the rapidly growing field of agentic AI.1 Its ability to process both structured and unstructured data broadens its applicability across diverse enterprise data landscapes.1 The potential for improved performance and efficiency compared to traditional CPU-based storage solutions offers a compelling reason for adoption.1 The platform's scalability across cloud, core, and edge environments provides deployment flexibility 75, and its integration with the broader NVIDIA AI ecosystem offers access to a wide range of tools and technologies.1

Despite its strengths, the NVIDIA AI Data Platform also presents potential weaknesses. As a reference design, the final product offerings and their specific capabilities will be dependent on the implementation by NVIDIA's storage partners.1 This reliance could lead to variability in features, performance characteristics, and the overall quality of solutions available from different vendors. The initial announcements provide limited specific details regarding the data governance and security features of the platform itself, suggesting that enterprises might need to rely on the security capabilities provided by the chosen storage partner's solution. Pricing for the end-user is not direct from NVIDIA but rather tied to the pricing models established by these storage partners, which could make it challenging to readily assess the cost-effectiveness of the platform without evaluating specific partner offerings. The platform's strong focus on demanding AI inference workloads might mean it is not the optimal choice for organizations with less intensive AI requirements or those primarily focused on AI training activities. The potential need to also adopt other components of the NVIDIA AI ecosystem, such as AI Enterprise, could introduce additional licensing costs and complexities for some enterprises. Given the newness of the platform, real-world case studies and comprehensive performance benchmarks from end-users are likely to be limited in the immediate future.

## **Conclusion and Recommendations**

The NVIDIA AI Data Platform holds significant promise as a transformative infrastructure solution for enterprises seeking to enhance their capabilities in AI inference and agentic AI. Its core strengths lie in the integration of accelerated computing with enterprise storage, enabling near real-time insights and efficient processing of diverse data types. The platform's tight integration with NVIDIA's broader AI ecosystem, including AI Enterprise, RAPIDS, Triton Inference Server, and NeMo, provides a comprehensive toolkit for building and deploying advanced AI applications. The strategic collaborations with leading storage vendors are crucial for translating the reference design into tangible and readily available product offerings.

For Chief Technology Officers (CTOs) and Vice Presidents of AI/Data Science who are evaluating the NVIDIA AI Data Platform, several key considerations and recommendations emerge. It is essential to engage directly with NVIDIA's certified storage partners to gain a thorough understanding of their specific solutions built upon the AI Data Platform and how these solutions align with your organization's unique AI infrastructure needs. A careful evaluation of the potential performance gains and the total cost of ownership benefits, in comparison to your existing infrastructure and alternative solutions, is crucial, particularly considering the specific demands of your AI inference workloads. Assess the level of integration required with your current data science workflows and the potential implications of adopting other components within the NVIDIA AI ecosystem. For organizations that are heavily invested in or planning to adopt agentic AI, the NVIDIA AI Data Platform presents a compelling architectural approach that warrants serious consideration due to its focus on reasoning and real-time data utilization. Staying informed about the evolving ecosystem, including new partner solutions and emerging real-world case studies, will provide valuable insights into the platform's practical effectiveness. Finally, it is recommended to consider a pilot program with a selected storage partner to thoroughly evaluate the platform's performance and integration within your specific operational environment before committing to a full-scale deployment.

#### Works cited

1. AI Data Platform for Enterprise Agentic AI - NVIDIA, accessed on April 2, 2025, <https://www.nvidia.com/en-us/data-center/ai-data-platform/>
2. NVIDIA and Storage Industry Leaders Unveil New Class of Enterprise Infrastructure for the Age of AI, accessed on April 2, 2025, <https://nvidianews.nvidia.com/news/nvidia-and-storage-industry-leaders-unveil-new-class-of-enterprise-infrastructure-for-the-age-of-ai>
3. NVIDIA Unveils AI Data Platform for Accelerated AI Query Workloads in Enterprise Storage, accessed on April 2, 2025, <https://www.hpcwire.com/off-the-wire/nvidia-unveils-ai-data-platform-for-accelerated-ai-query-workloads-in-enterprise-storage/>
4. Nvidia Launches AI Data Platform at GTC 2025 - Technology Magazine, accessed on April 2, 2025, <https://technologymagazine.com/articles/gtc-nvidia-launches-ai-data-platform-and-reasoning-models>
5. NVIDIA Launches AI Data Platform with Major Storage Leaders | NVDA Stock News, accessed on April 2, 2025, <https://www.stocktitan.net/news/NVDA/nvidia-and-storage-industry-leaders-unveil-new-class-of-enterprise-zoorq5d20tgz.html>
6. NVIDIA unveils AI Data Platform for enterprise AI infrastructure - Engineering.com, accessed on April 2, 2025, <https://www.engineering.com/nvidia-unveils-ai-data-platform-for-enterprise-ai-infrastructure/>
7. nvidianews.nvidia.com, accessed on April 2, 2025, <https://nvidianews.nvidia.com/news/nvidia-and-storage-industry-leaders-unveil-new-class-of-enterprise-infrastructure-for-the-age-of-ai#:~:text=Leading%20data%20platform%20and%20storage,reason%20and%20respond%20to%20complex>
8. Transforming AI Data Pipelines with Dell and NVIDIA, accessed on April 2, 2025, <https://www.dell.com/en-us/blog/transforming-ai-data-pipelines-with-dell-and-nvidia/>
9. Transforming AI Data Pipelines with Dell and NVIDIA, accessed on April 2, 2025, <https://www.dell.com/en-hk/blog/transforming-ai-data-pipelines-with-dell-and-nvidia/>
10. Dell Technologies Accelerates Enterprise AI Innovation from PC to Data Center with NVIDIA, accessed on April 2, 2025, <https://investors.delltechnologies.com/news-releases/news-release-details/dell-technologies-accelerates-enterprise-ai-innovation-pc-data>
11. VAST Data Announces Enterprise-Ready AI Stack via VAST InsightEngine with NVIDIA DGX, accessed on April 2, 2025, <https://www.vastdata.com/press-releases/vast-data-ai-stack-insightengine-with-nvidia-dgx>
12. Powered by Starburst: Dell AI Data Platform with NVIDIA, accessed on April 2, 2025, <https://www.starburst.io/blog/starburst-dell-nvidia/>
13. Enterprise-Grade AI Software Platform - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/ai-enterprise/reference-architecture/latest/platform-overview.html>
14. The Forrester Wave™: AI Infrastructure Solutions, Q1 2024 - Alibaba Cloud, accessed on April 2, 2025, <https://www.alibabacloud.com/en/about/forrester-wave-ai-infrastructure-solutions-2024?_p_lc=1>
15. Cisco to Deliver Secure AI Infrastructure with NVIDIA, accessed on April 2, 2025, <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m03/cisco-and-nvidia-secure-AI-factory.html>
16. Cisco Expands Partnership with NVIDIA to Accelerate AI Adoption in the Enterprise, accessed on April 2, 2025, <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m02/cisco-expands-partnership-with-nvidia-to-accelerate-ai-adoption-in-the-enterprise.html>
17. Securiti to Empower Organizations to Build Safe, High-Performance Enterprise AI Systems with NVIDIA NIM Microservices, accessed on April 2, 2025, <https://securiti.ai/press-release/build-safe-enterprise-ai-systems-with-nvidia-nim-microservices/>
18. NVIDIA AI Enterprise Security White Paper, accessed on April 2, 2025, <https://docs.nvidia.com/ai-enterprise/planning-resource/ai-enterprise-security-white-paper/latest/index.html>
19. Pure Storage Integrates NVIDIA AI Data Platform into FlashBlade to Fuel Enterprise AI Innovation - PR Newswire, accessed on April 2, 2025, <https://www.prnewswire.com/news-releases/pure-storage-integrates-nvidia-ai-data-platform-into-flashblade-to-fuel-enterprise-ai-innovation-302403778.html>
20. Pure Storage Integrates NVIDIA AI Data Platform into FlashBlade to Fuel Enterprise AI Innovation, accessed on April 2, 2025, <https://investor.purestorage.com/news-and-events/press-releases/press-release-details/2025/Pure-Storage-Integrates-NVIDIA-AI-Data-Platform-into-FlashBlade-to-Fuel-Enterprise-AI-Innovation/default.aspx>
21. Nutanix Cloud Platform Empowers Enterprise Agentic AI Built with the NVIDIA AI Data Platform, accessed on April 2, 2025, <https://www.nutanix.com/blog/nutanix-cloud-platform-empowers-enterprise-agentic-ai-built-with-the-nvidia-ai-data-platform>
22. IBM Taps NVIDIA AI Data Platform Technologies to Accelerate AI at Scale - IBM Newsroom, accessed on April 2, 2025, <https://newsroom.ibm.com/2025-03-18-ibm-taps-nvidia-ai-data-platform-technologies-to-accelerate-ai-at-scale>
23. Enterprise-Grade AI Software Platform - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/ai-enterprise/deployment/spark-rapids-accelerator/latest/platform-overview.html>
24. Enterprise-Grade AI Software Platform - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/ai-enterprise/overview/latest/platform-overview.html>
25. NVIDIA RAPIDS: Running ML and Big Data Faster with GPUs, accessed on April 2, 2025, <https://www.run.ai/guides/ai-open-source-projects/nvidia-rapids>
26. RAPIDS | GPU Accelerated Data Science, accessed on April 2, 2025, <https://rapids.ai/>
27. RAPIDS Suite of AI Libraries | NVIDIA Developer, accessed on April 2, 2025, <https://developer.nvidia.com/rapids>
28. Ecosystem | RAPIDS | RAPIDS | GPU Accelerated Data Science, accessed on April 2, 2025, <https://rapids.ai/ecosystem/>
29. Accelerate Data Science and Analytics using NVIDIA GPU | by Ramakrishna Sanikommu, accessed on April 2, 2025, <https://medium.com/@ramakrishna.sanikommu/accelerate-data-science-and-analytics-using-nvidia-gpu-048ee3d27df1>
30. GPU Accelerated Data Science with RAPIDS | NVIDIA, accessed on April 2, 2025, <https://www.nvidia.com/en-au/deep-learning-ai/software/rapids/>
31. Triton Inference Server with Ultralytics YOLO11, accessed on April 2, 2025, <https://docs.ultralytics.com/guides/triton-inference-server/>
32. Triton Inference Server - NVIDIA Developer, accessed on April 2, 2025, <https://developer.nvidia.com/triton-inference-server>
33. NVIDIA Triton Inference Server, accessed on April 2, 2025, <https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/index.html>
34. Triton Inference Server - GitHub, accessed on April 2, 2025, <https://github.com/triton-inference-server>
35. The Triton Inference Server provides an optimized cloud and edge inferencing solution. - GitHub, accessed on April 2, 2025, <https://github.com/triton-inference-server/server>
36. Build Custom Generative AI | NVIDIA NeMo, accessed on April 2, 2025, <https://www.nvidia.com/en-us/ai-data-science/products/nemo/>
37. NVIDIA/NeMo: A scalable generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (Automatic Speech Recognition and Text-to-Speech) - GitHub, accessed on April 2, 2025, <https://github.com/NVIDIA/NeMo>
38. NVIDIA AI Workbench User Guide, accessed on April 2, 2025, <https://docs.nvidia.com/ai-workbench/user-guide/latest/overview/introduction.html>
39. NVIDIA AI Workbench — RAPIDS Deployment Documentation documentation, accessed on April 2, 2025, <https://docs.rapids.ai/deployment/stable/platforms/nvidia-ai-workbench/>
40. NVIDIA AI Workbench, accessed on April 2, 2025, <https://docs.nvidia.com/ai-workbench/index.html>
41. An NVIDIA AI Workbench example project for Retrieval Augmented Generation (RAG) - GitHub, accessed on April 2, 2025, <https://github.com/NVIDIA/workbench-example-hybrid-rag>
42. Creating a Chatbot using Precision and NVIDIA AI Workbench - Dell Technologies, accessed on April 2, 2025, <https://www.dell.com/en-us/blog/creating-a-chatbot-using-precision-and-nvidia-ai-workbench/>
43. NVIDIA Base Command Platform - Advanced HPC, accessed on April 2, 2025, <https://www.advancedhpc.com/pages/nvidia-base-command-platform>
44. NVIDIA Base Command - Advanced HPC, accessed on April 2, 2025, <https://www.advancedhpc.com/pages/nvidia-base-command>
45. NVIDIA DGX Platform, accessed on April 2, 2025, <https://docs.nvidia.com/dgx/index.html>
46. NVIDIA Base Command Platform - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/base-command-platform/index.html>
47. NVIDIA Base Command Manager - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/base-command-manager/index.html>
48. NVIDIA NGC catalog - D2iQ Docs, accessed on April 2, 2025, <https://archive-docs-old.d2iq.com/dkp/kaptain/2.2.0/tutorials/ngc>
49. NGC Catalog User Guide - NVIDIA Docs, accessed on April 2, 2025, <https://docs.nvidia.com/ngc/gpu-cloud/ngc-catalog-user-guide/index.html>
50. NVIDIA® GPU Cloud (NGC) Catalog CLI, accessed on April 2, 2025, <https://docs.e2enetworks.com/docs/myaccount/gpu/ngc_catalog_cli/>
51. NVIDIA NGC containers - Purdue's RCAC, accessed on April 2, 2025, <https://www.rcac.purdue.edu/knowledge/ngc>
52. NVIDIA NGC Catalog | Research Software Portal, accessed on April 2, 2025, <https://software.xsede.org/vendor-software-and-service-component/nvidia-ngc-catalog>
53. NVIDIA AI Enterprise | Cloud-native Software Platform, accessed on April 2, 2025, <https://www.nvidia.com/en-us/data-center/products/ai-enterprise/>
54. US Technology Leaders Tap NVIDIA AI Software to Transform World's Industries, accessed on April 2, 2025, <https://nvidianews.nvidia.com/news/us-technology-leaders-tap-nvidia-ai-software-to-transform-worlds-industries>
55. Tech companies team up with NVIDIA for AI innovation - ERP Today, accessed on April 2, 2025, <https://erp.today/tech-companies-team-up-with-nvidia-for-ai-innovation/>
56. Meet Nvidia's 14 Top Americas Partners Who 'Deeply Understand' Its Full-Stack AI Platform, accessed on April 2, 2025, <https://www.crn.com/news/ai/2025/meet-nvidia-s-14-top-americas-partners-who-deeply-understand-its-full-stack-ai-platform>
57. NVIDIA Honors Americas Partners Advancing Agentic and Physical AI, accessed on April 2, 2025, <https://blogs.nvidia.com/blog/partner-network-awards-2025/>
58. AI Computing Consulting | NVIDIA Partner Network, accessed on April 2, 2025, <https://www.nvidia.com/en-gb/about-nvidia/partners/ai-computing-consulting/>
59. NVIDIA Business Partner | Enterprise AI Integration & Consulting - Next Realm AI, accessed on April 2, 2025, <https://nextrealm.ai/nvidia-business-partner/>
60. NVIDIA and Global Partners Launch NIM Agent Blueprints for Enterprises to Make Their Own AI, accessed on April 2, 2025, <https://nvidianews.nvidia.com/news/nvidia-and-global-partners-launch-nim-agent-blueprints-for-enterprises-to-make-their-own-ai>
61. AWS Marketplace: NVIDIA AI Enterprise - Amazon.com, accessed on April 2, 2025, <https://aws.amazon.com/marketplace/pp/prodview-ozgjkov6vq3l6>
62. NVIDIA AI Enterprise - Microsoft Azure Marketplace, accessed on April 2, 2025, <https://azuremarketplace.microsoft.com/en-us/marketplace/apps/nvidia.nvidia-ai-enterprise?tab=overview>
63. NVIDIA AI Enterprise - Oracle Cloud Marketplace, accessed on April 2, 2025, <https://cloudmarketplace.oracle.com/marketplace/en_US/listing/155314141>
64. Enterprise Marketplace | NVIDIA, accessed on April 2, 2025, <https://marketplace.nvidia.com/en-us/enterprise/>
65. NVIDIA Marketplace US: Graphics Cards, AI, Gaming & Enterprise Solutions, accessed on April 2, 2025, <https://marketplace.nvidia.com/en-us/>
66. NVIDIA AI Enterprise Licensing Guide, accessed on April 2, 2025, <https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/licensing.html>
67. NVIDIA AI Enterprise Essentials - SHI, accessed on April 2, 2025, <https://www.shi.com/product/47017691/NVIDIA-AI-Enterprise-Essentials>
68. NVIDIA Unveils AI Enterprise Software Suite to Help Every Industry Unlock the Power of AI, accessed on April 2, 2025, <https://nvidianews.nvidia.com/news/nvidia-unveils-ai-enterprise-software-suite-to-help-every-industry-unlock-the-power-of-ai>
69. Fast, Low-Cost Inference Offers Key to Profitable AI | NVIDIA Blog, accessed on April 2, 2025, <https://blogs.nvidia.com/blog/ai-inference-platform/>
70. nVidia's GPU Price vs. Cost Perspective | Future of Selling - YouTube, accessed on April 2, 2025, <https://www.youtube.com/watch?v=f1YrfK5Zi7I>
71. What are the total cost of ownership (TCO) benefits of using NVIDIA data center GPUs for large-scale batch processing? - Massed Compute, accessed on April 2, 2025, [https://massedcompute.com/faq-answers/?question=What%20are%20the%20total%20cost%20of%20ownership%20(TCO)%20benefits%20of%20using%20NVIDIA%20data%20center%20GPUs%20for%20large-scale%20batch%20processing?](https://massedcompute.com/faq-answers/?question=What+are+the+total+cost+of+ownership+(TCO)+benefits+of+using+NVIDIA+data+center+GPUs+for+large-scale+batch+processing?)
72. Understanding the Total Cost of Ownership in HPC and AI Systems - Ansys, accessed on April 2, 2025, <https://www.ansys.com/blog/understanding-total-cost-ownership-hpc-ai-systems>
73. NVIDIA AI Enterprise - SHI, accessed on April 2, 2025, <https://www.content.shi.com/cms-content/accelerator/media/pdfs/nvidia/nvidia-030525-nvidia-ai-enter-solution-overview.pdf>
74. TCO of NVIDIA GPUs and falling barriers to entry | Continuum Labs, accessed on April 2, 2025, <https://training.continuumlabs.ai/infrastructure/the-modern-data-centre/tco-of-nvidia-gpus-and-falling-barriers-to-entry>
75. DDN Announces Support for NVIDIA Blackwell-Based System, accessed on April 2, 2025, <https://www.ddn.com/press-releases/ddn-expands-ai-data-infrastructure-for-enterprises-and-announces-support-for-nvidia-blackwell-based-system/>