NVIDIA AI Factory & NIM Inference Microservices

STEVE MCDOWELL, CHIEF ANALYST
7/24/24

CONTEXT

NVIDIA announced its new NVIDIA AI Foundry, a service designed to supercharge generative AI capabilities for enterprises using Meta's just-released Llama 3.1 models, along with its new NIM Inference microservices.

The new offerings significantly advance the ability to customize and deploy AI models for domain-specific applications.

NEW: AI FOUNDRY

NVIDIA AI Foundry is a newly launched service designed to enhance enterprises' generative AI capabilities by leveraging Meta's Llama 3.1 models. The platform allows organizations to create domain-specific AI models tailored to their unique industry needs using NVIDIA's advanced software, computing resources, and expertise.

KEY FEATURES

1. **Custom Supermodels**:
    - **Tailored AI Models**: Enterprises and nations can create custom "supermodels" for their use cases. These models are built using Meta's Llama 3.1 and can be trained with proprietary and synthetic data generated by Llama 3.1 405B and the NVIDIA Nemotron™ Reward model.
2. **Scalable Infrastructure**:
    - **NVIDIA DGX™ Cloud AI Platform**: The AI Foundry is powered by the NVIDIA DGX Cloud AI platform, which was co-engineered with

leading public clouds. This platform provides significant compute resources that can scale quickly as AI demands evolve, ensuring enterprises have the necessary power to handle their AI workloads.

3. **NVIDIA NIM™ Inference Microservices**:
   - **High Performance**: NIM inference microservices are available for Llama 3.1 models, offering up to 2.5x higher throughput than traditional inference methods. These microservices enable efficient deployment of AI models in production environments.
   - **Enhanced AI Pipelines**: NIM microservices can be paired with NVIDIA NeMo Retriever NIM microservices to create advanced retrieval pipelines for AI applications like copilots, assistants, and digital human avatars.

NEW: NIM INFERENCE MICROSERVICES

NVIDIA's NIM (NeMo Inference Microservices) are designed to optimize the deployment and performance of AI models. The new microservices provide a robust, scalable solution for running AI models in production environments, enhancing throughput, accuracy, and efficiency.

As background, NVIDIA NeMo is NVIDIA's existing framework for building, training, and fine-tuning state-of-the-art conversational AI models, including models for natural language processing (NLP), speech recognition, and text-to-speech synthesis. NeMo offers a modular approach, where users can combine different neural network components (such as encoders, decoders, and classifiers) to create customized models.

The new NeMo Inference Microservices extends NeMo's capabilities in several significant ways.

KEY FEATURES

1. High Throughput and Performance:
   - Enhanced Speed: NIM inference microservices offer up to 2.5x higher throughput than running inference without NIM. This significant boost in speed enables enterprises to handle larger volumes of data and more complex AI tasks efficiently.

- Optimized for Llama 3.1: Specifically designed to support the Llama 3.1 models, NIM microservices ensure that these models perform at their peak potential in production environments.

2. Seamless Integration:
   - Deployment Flexibility: NIM microservices can be easily integrated into existing MLOps (Machine Learning Operations) and AIOps (Artificial Intelligence Operations) platforms. This flexibility allows organizations to deploy and manage AI models using their preferred tools and infrastructure.
   - Compatibility: Supports deployment on various cloud platforms and NVIDIA-Certified Systems™ from global server manufacturers, providing broad compatibility and ease of use.
3. Advanced AI Pipelines:
   - NeMo Retriever NIM Microservices: In addition to the core NIM inference microservices, NVIDIA offers NeMo Retriever NIM microservices. These are designed to enhance retrieval-augmented generation (RAG) pipelines, improving response accuracy for applications such as AI copilots, assistants, and digital avatars.
   - Improved Accuracy: These microservices deliver the highest open and commercial text Q&A retrieval accuracy for RAG pipelines, ensuring precise and reliable AI interactions.
4. Customizable and Scalable:
   - Domain-Specific Models: Enterprises can use NIM microservices to deploy custom AI models tailored to their specific industry or use case, enabling precise and effective AI applications.
   - Scalable Infrastructure: The microservices are built to scale, allowing organizations to expand their AI capabilities as their data and processing needs grow.

ANALYSIS

By enabling the creation of highly customized, domain-specific AI models using Meta's Llama 3.1, NVIDIA empowers enterprises to harness the full potential of generative AI. The combination of scalable infrastructure, advanced software, and extensive partner support positions NVIDIA AI Foundry as a critical tool for driving the next wave of AI innovation and adoption across industries.

NVIDIA AI Foundry faces competition from well-established AI platforms, each offering unique strengths and capabilities. IBM Watsonx, Google Cloud AI Platform, AWS SageMaker, Microsoft Azure Machine Learning, Hugging Face Transformers, and OpenAI API all provide powerful tools for developing, deploying, and managing AI models.

The choice between these platforms depends on specific needs such as performance, compliance, integration, and scalability. Each platform has its strengths, making it suitable for different types of AI projects and organizational requirements.

For example, NVIDIA AI Foundry and IBM Watsonx offer powerful platforms for AI development and deployment, but they cater to different needs and strengths. NVIDIA AI Foundry is ideal for enterprises requiring high-performance, customized AI models focusing on scalability and flexibility.

In contrast, IBM Watsonx excels in providing a comprehensive, governance-focused AI platform suitable for industries with stringent regulatory requirements and a need for explainable AI. Choosing between them depends on an organization's specific requirements, including performance, customization, governance, and deployment flexibility.

NVIDIA’s AI Foundry service and NIM inference microservices provides enterprises with the tools to build and deploy powerful, custom AI models with unprecedented efficiency and scalability. As AI continues transforming industries, NVIDIA's latest offerings are set to lead the charge, driving the next wave of AI innovation and adoption across the global business landscape.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.