ENTERPRISEAI
ADVANCED COMPUTING IN THE AGE OF AI

Nvidia CEO Huang: Get Ready for Software 3.0

June 28, 2023 by Agam Shah

Nvidia's CEO Jensen Huang said that artificial intelligence is ushering in the era of Software 3.0, where creating and running applications will be as simple as writing queries into a universal AI translator, running a few lines of Python code, and selecting an AI model of your choice.

"That's the reinvention of the whole stack -- the processor is different, the operating system is different, the large language model is different. The way you write AI applications is different... Software 3.0, you do not have to write it at all," Huang said at a fireside chat during the Snowflake Summit this week.

Huang talked about the emerging software landscape as the company switches gears to a software-sells-hardware strategy, a complete flip of its past hardware-sells-software strategy. Nvidia hopes to sell more software that runs only on its GPUs.

Software 3.0 applications will change the way users interact with computers, Huang said, adding that the interface will be a universal query engine "that's super intelligent and you can get it to ... respond to you."

Users can type in prompts and context at the query engine, which goes through large language models, which may be connected to corporate databases or other data sources. ChatGPT is an early iteration of how this system will work, but Huang said this will impact every facet of computing.

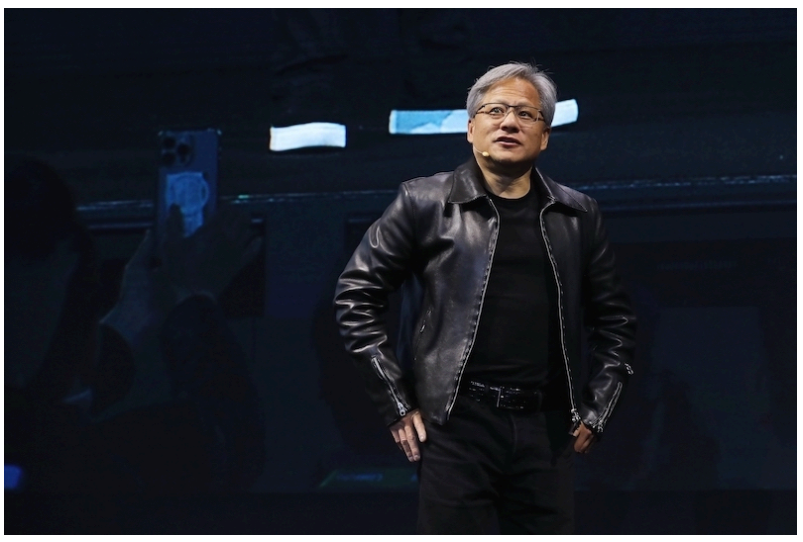
The Software 3.0 concept relies on a new structure of data, algorithms, and compute engines, Huang said, adding that instead of command lines, users will be able to talk databases and "ask it all kinds of questions about what, how, when and why."

He gave one example of ChatPDF, which analyzes and summarizes giant PDF documents. Large language models could also generate programming code if needed.

"We'll develop our own applications, everybody's going to be an application developer," Huang said, adding that conventional programs in companies will be replaced by hundreds of thousands of AI applications.

It is the early days of this new type of computing, which is a departure from the old style of computing that relied on bringing data through computers and processing it via CPUs. The entire structure of computing is untenable with the inability to scale performance.

The Software 3.0 approach will merge data from multimodal sources that include images, text, and voice. Huang said, added that "for the very first time you could develop a large language model, stick it in front of your data and you talk to your data... like you talk to a person."



Nvidia CEO Jensen Huang

Startups like Glean and Neeva (which was acquired by Snowflake) are investing in technologies that connect AI search within enterprises to large language models. On a consumer front, Microsoft and Google are sending queries from search to supercomputers with AI chips that process the queries and return a response.

Nvidia's strategy is to provide the hardware and software on both ends – the consumers and enterprises – to run artificial intelligence applications. Nvidia's involvement right now is mostly covert, but ChatGPT relies heavily on Nvidia GPUs to process queries.

Applications developed using LangChain, and intermediate agents and data sources can be added in betwe

en AI processing to provide more fine-tuned responses.

One such intermediary is Nvidia's NeMo Guardrails, which eliminates chatbot hallucinations so large-language models stay on track and provide relevant answers to queries. Huang also bet on large-language models with billions of parameters to make AI relevant, liking it to a college grad that was pre-trained to be super smart. The large models will be surrounded by smaller models augmented by specialized knowledge, which could support enterprises.

Microsoft and Google are already blending their old computing models by plugging their own large-language AI models into applications. Microsoft has a fork of GPT-4 powering Bing, while Google has the PaLM-2 transformer, and is also developing Gemini, which is still being trained.

Nvidia's future is in Software 3.0 concept, with the main computing hardware being its GPUs. Nvidia saw the AI opportunity many years ago, and has invested heavily in developing a complete AI stack -- including software, services, and hardware -- to chase the opportunity, said Jim McGregor, principal analyst at Tirias Research.

The company's AI operating system is the AI Enterprise Suite, which includes large language models like NeMo, compilers, libraries, and development stacks. The software developed via AI Enterprise will need Nvidia's GPUs, which can be found on-premise or in the cloud.

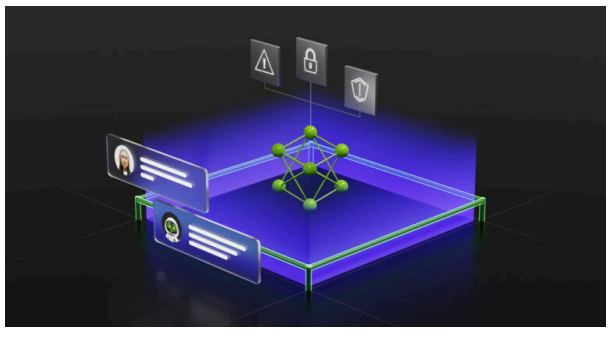
At this week's Snowflake Summit, Nvidia announced software partnerships that provided clarity on how it would lock customers into using its software and GPUs in the cloud.

Nvidia said it was bringing its NeMo large-language model to Snowflake Data Cloud, which is used by top organizations to deposit data. The NeMo LLM is a pre-trained model in which companies can feed their own data to create their own models. Enterprises can generate their own tokens and customize the models, and queries to the database will deliver more fine-tuned answers. For example, employees could generate an expense report for a specific quarter from a prompt.

Nvidia's NeMo transformer model is trained from a generic corpus of data, and companies will augment the models with their own data. The proprietary corporate data will remain locked in their model, and will not be sent back to the larger models, said Manuvir Das, vice president for enterprise computing at the company, during a press briefing.

Users of the Snowflake Data Cloud will be able to connect the software to hardware on cloud service providers, which have set up their own supercomputers with Nvidia's GPU. Google a few months ago announced the A3 supercomputer, which has 26,000 Nvidia GPUs. Microsoft has its own Azure supercomputer with thousands of Nvidia GPUs.

Nvidia is also providing the ability for third-party customers to use large language models and smaller custom models via a partnership with ServiceNow, which was announced earlier this year. In this partnership, ServiceNow is using NeMo to create models for their customers. But the software-as-a-service company also provides access to other AI models such as OpenAI's GPT-4, giving flexibility for customers to use GPT-4 instead of Nvidia's NeMo.



ServiceNow also provides connectors that provide customers access to many AI options. For example, Glean, which uses multiple LLMs, integrates with ServiceNow.

Nvidia is a top player in the AI market, and its main hardware competitors, Advanced Micro Devices and Intel, are far behind with no proven commercial success.

AMD this month introduced a new GPU, the Instinct MI300X, which is targeted at AI but has no clear software strategy with its focus squarely on hardware. Tirias Research's McGregor said that AMD was late to the game, and as a smaller company does not have the resources to pour into software.

Intel has many AI chips in its portfolio, including the Gaudi2 AI chip and Xeon GPU Max for training, but those chips are still not being sold in volume. Intel's contrasting software strategy revolves around an open approach so developers can write AI applications that can run on any hardware.

Categories: AI/ML/DL, Cloud, Software, Systems

Tags: AI software, Google, hardware, Intel, Jensen Huang, Microsoft, NVIDIA, Snowflake

RECENT NEWS

Faster Interconnects and Switches to Help Relieve Data Bottlenecks
March 22, 2024

Global IT Spending Forecast: Gartner Predicts 6.8 Percent Growth in 2024
January 23, 2024

AWS Unveils Major Bedrock Upgrade: More AI Models and Enhanced User Flexibility
December 8, 2023

The IBM-Meta AI Alliance Promotes Safe and Open AI Progress
December 5, 2023

Nvidia Showcases Domain-specific LLM for Chip Design at ICCAD
November 2, 2023

CONTRIBUTORS

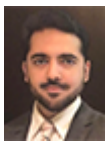

Tiffany Trader
Editorial Director


Kevin Jackson
Managing Editor


John Russell
Contributing Editor


Alex Woodie
Contributing Editor


Douglas Eadline
Contributing Editor


Ali Azhar
Contributing Editor


Drew Jolly
Assistant Editor

UPCOMING EVENTS

15th Annual HPC-AI Swiss Conference
April 15 - April 17

AI in Finance Summit New York
April 18 - April 19

AI X Summit 2024
April 23

World Summit AI Americas 2024
April 24 - April 25

Call & Contact Center Expo
April 24 - April 25

View All Events