EXPERIENCE THE FUTURE OF AI: THE SOVEREIGN DATA AND AI FACTORY

Chapter 1. The Sovereign Data and AI Factory: Your Hybrid AI Foundation

Simon Lightstone: Gravity. The unseen force pulling together the world around us. In today's enterprise, we're faced with a new notion that your data carries that same gravity. The more concentrated the data, the greater the pull. And within that pull lie both the greatest innovations and new vulnerabilities.

At EDB, we understand the importance of data sovereignty. To be sovereign is to protect, to control, to build with confidence. And we can't shoulder this endeavor alone. That's why we've partnered with the industry's leading innovators to create the first Postgres® AI sovereign data and AI factory, designed to power and protect the future of the enterprise.

Here at the Supermicro campus, I'll be joined by Somik Behera, GM of Software and AI Products at Supermicro; Nave Algarici, senior PM of GenAI at NVIDIA; and myself, Simon Lightstone, director of Product Management at EDB Postgres AI.

The future isn't forecast, it ships. Here's how. Let's dive right in.

Well, it's great to be here, all of us. We have NVIDIA, Supermicro, and EDB. NVIDIA with GPU is an accelerated software hardware from the Supermicro side. And here at EDB, you know, we're really excited about the future of AI and database software, our AI software. And we have a very exciting piece of technology to talk about, which is the engineered system. And of course, we also want to talk a little bit about the future of AI, the future of hardware, and how we're going to be moving towards that, as companies together with the engineered system—and, in general, where all these technologies are headed.

So, Somik, why don't you give a quick little bit about yourself?

Somik Behera: Cool, thank you. And welcome to the Supermicro campus. It's lovely to host you all. I'm Somik Behera. I'm the general manager of Software and AI Products over here, responsible for our product suite, like SuperCloud Composer, SuperCloud

Orchestrator, SuperCloud Director, and SuperCloud Developer Console, to provide you the total building-block software solution experience for joint solutions.

Nave Algarici: I'm Nave Algarici. I lead product management for NeMo Retriever at NVIDIA, and the focus really is connecting enterprise data into generative applications through accelerated computing. And that's why we're here working with EDB as well as Supermicro.

Lightstone: Yeah, and I'm Simon Lightstone. I am a director of Product Management, specifically for the engineered system, which we're going to talk about today—as well as some other componentry, which I love to talk about in terms of Postgres. So, very excited for us to all catch up and discuss what we've got in terms of the engineered system and, really, what's next for AI.

So, Somik, why don't you start off and give a quick overview about Supermicro's view on AI and where they see things going at the high level?

Behera: Cool. As you know, we have been the market leader in AI infrastructure for years, since the beginning, together with NVIDIA—in fact, 30 years ago, before the term "AI" was coined. And what we are increasingly seeing as we became the number-one OEM server vendor and manufacturer worldwide, due to the power of AI, is that AI is increasingly becoming distributed, hybrid, and available in every region, in every country, in every sovereign environment. And that's why we're excited about, what can we do to bring the power of AI applications? Not only the AI hardware, not only the power needed to supply that hardware, which we are again market leading in providing direct liquid cooling solutions, as well as power-efficient computing, green computing solutions—but build those applications on this hybrid infrastructure, on customer data, where they are, within their sovereign infrastructure.

Lightstone: Yeah, and I think that's where the engineered system comes in to help with part of that story. So, to Nave: So, one thing that we're seeing is an increase in the number of people choosing a hybrid deployment option on premises for AI workloads for solving AI problems. And it's an interesting thing, because by investing a little bit in their own data center, they have full sovereignty, full control.

A very big concern that we keep hearing is, "Well, where am I sending my data? I'm sending it to the public cloud." And really, you need that absolute control when it comes to your data. And that's sort of what we're delivering with the upcoming engineered system, where we're all working together and providing the GPUs, the hardware, and the software necessary.

Algarici: We at NVIDIA built the NVIDIA NIM, which is a containerized service that enables you to deploy AI models into production and into enterprise applications. With only a few lines of code, you can take a state-of-the-art model and scale it up on NVIDIA GPUs, and it leverages all the greatest technologies from TranspoRT and Triton, all packaged in a container that is deployable on prem or in cloud, in any way or matter.

The main pull is around the gravity of the data, right? So if, previously, the paradigm was to bring your data to where your AI sits, now it shifts to bringing your AI to where your data resides. And by combining the solutions from NVIDIA AI software and NVIDIA NIM to EnterpriseDB, we're enabling that turnkey solution, where your data is already sitting in EnterpriseDB. You turn the key—now it's AI powered, AI accessible, and connected to your AI applications.

Lightstone: Yeah, exactly. So now finally you're going to have such low latency between your database and your AI infrastructure that you're going to be able to get answers very quickly, right? We're talking about a situation where data could be updated in real time, and immediately your chatbots, your decision-making, any of your agentic workflows are going to incorporate those changes right away. We're beyond that era where you have to push the data from one point to another. We're talking about real-time decision-making and very, very fast, low-latency responses in terms of getting from the AI models to data as it exists today in your databases and Postgres, and then delivering results.

Behera: Absolutely, and we see that as well. As you know, we are the number-one OEM server vendor, server manufacturer now. But before that, we were the number-one AI infrastructure provider, NVIDIA GPU infrastructure provider. And time and time again, when we went to the customer environment, we saw that data is so critical to be sovereign, within the control of the enterprise, than it has ever been.

So, what does that mean? That means, because of generative AI, the amount of value and agentic workflow a copilot provides to the business, to the enterprise leveraging NVIDIA NIMS—leveraging, you know, EDB Postgres and EDB AI capabilities—is needed more than ever. That we need to bring these AI capabilities right to where enterprise data is. Because that data is truly sacred, right?

Lightstone: And not just that, but now we're giving the ability to manage your entire Postgres database estate, right? So, yes, it's in the data center. But one challenge was, all right, you have some items in your data center. You've got that. And then you might even also have items still in the cloud. Or you have some in cloud, some on premises,

depending on your requirements and also depending on the need to get it closer and closer to your actual AI hardware. But there was no easy way to just see it all from one single pane of glass, all of your Postgres databases.

So now, with the first Postgres AI sovereign data in AI factory, you're going to get that ability to basically control everything from one single pane of glass. Even those databases that aren't on the engineered system. We aren't forcing a march to an engineered system. You can take your equipment you have today, running Postgres, and monitor it all from one view, diagnose queries, deep-dive into exactly what's causing the delays, the weights inside your database. And that's going to allow you to just operate your entire database system very, very easily across as many nodes as you really want.

Behera: Most of the enterprise data is in SQL stores and SQL database. So are you guys saying that, you know, what our solution is doing is letting enterprises bring the value of AI—the same kind of productivity benefits of generative AI that you see with ChatGPT—on top of their own database and on their own data, running around accelerated, industry-leading kind of super-microservice in hardware? Anywhere in the world, any data center?

Lightstone: Yeah. Basically, yes. Of course, Supermicro makes it so easy for us to deploy this hardware and ship it out to exactly where it needs to be anywhere in the world. Then, honestly, once this is shipped and we set it up for the customer, we curate that. They're a few clicks away from either production Postgres across three nodes or even just building their own AI, full, end-to-end solutions. All the way up to the point where you can deploy a chatbot on WhatsApp or on Slack. We just make it easy to tie all those things together and get it done faster than ever before.

Behera: That is so cool, because I can't just wait for my PG&E bill to chat with it that next time. Or my AT&T customer support, you know—instead of waiting 10 minutes to disconnect a line. And it would be even more gratifying to know that it's built on our solution that we worked together.