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

Chapter 3. Strategic Vision: Security, Management, and the Future of AI

Simon Lightstone (EnterpriseDB): We've discussed performance. Now let's talk about security and the future of the industry.

From the software level, Postgres®, as you know, and EDB Postgres variants in particular, provide a lot more security than you're going to get out of the box with open source Postgres, which is the community edition. However, I want to say that security really starts at every single level. So you've got granular security with an actual username and password that you're logging into with Postgres. You've got the ability to synchronize with an LDAP server, with—you know, whether that's Active Directory—and make sure that you're pulling in security real time. If you have a thousand employees, you can imagine that that's not something you want to continually be rotating passwords on. And every single chunk of the stack, start to finish, has security in mind, because we know that that's going to be a very important thing for our customers.

Somik Behera (Supermicro): Just the way you guys have added security at every layer of the software and database, we have made sure this engineered system has zero-trust security architecture. So, in the hardware layer, we don't trust the software in that way. We want to make sure it's actually signed and certified software. And similarly, we don't trust any components of our hardware. We have a root of trust that secures the entire supply chain.

So every piece—even though these are hardware components—is signed and verified throughout the entire supply chain. So the customers, when they engage with the market leader, with the number-one OEM in the world, they can be assured that's what comes with it, right? That root of trust, secure supply chain that is attested at every step of the supply chain. So, even a customer can have a CA private key and talk to their public CA and validate every piece of this BOM, the "horror" BOM. And of course that, paired with your software security stack, provides something enterprises have probably never been able to accomplish themselves.

Nave Algarici (NVIDIA): Absolutely, and we at NVIDIA see that as absolutely fundamental on ensuring that, as part of that BOM of software that goes into an enterprise stack, NVIDIA also contributes to make sure that it is a no-brainer in putting in NVIDIA software into that. So when we produce NIMS, we ensure that all the critical vulnerabilities are addressed, all the APIs are completely secure, and work together with our partners to test that out rigorously in multiple environments.

And by doing that, we ensure high SLAs that are meeting our customer requirements in real-time enterprise production environments. So we're working together with the ecosystem. We've learned a lot, and we got to a place where NVIDIA NIM is the easiest way to build scalable enterprise deployments.

Behera: So, Simon, you know, there's hundreds of databases in a customer data center. How are customers going to manage this data estate?

Lightstone: So that's a great question, and with Postgres data in AI Factory, you're able to see hundreds of databases all in one single pane of glass. And not only that, but you can deep-dive very, very quickly. Traditionally, you need a whole bunch of little tools set up in order to do the sort of diagnostics that an administrator would need to do to determine what's slowing down your database. But for those pros who really want to dive very, very deep, you know, DBAs love EDB Postgres because of that level of granularity we give. It's just so simple. It's just a matter of clicks. You click into your database, you can click into your query, you can actually see the waits, you can get a pie chart of whether it's I/O or compute. What's slowing down your queries. And it allows you to basically diagnose these clusters that are mission critical for you, for your business, in just the easiest way you can imagine. We let you track over 200 metrics. So that way, you know, for those who do want to get that level of granularity and want to deep-dive into optimizing performance, it's right there for them, right? Without too much work and without some of the heavy lifting of having to ensure that your tools are installed everywhere.

Another thing to mention is that, again, there's no forced march for the enterprise, for the engineered system. Traditionally, you buy an expensive database and then you have to consolidate onto it if you buy a database appliance. With the engineered system, you can observe all the databases that exist in your data center today. So if you've already invested in awesome Supermicro hardware, you don't have to migrate that right away. You can observe it in place and get all the deep-dive statistics and metrics for your databases.

So that's one great thing. We imagine that every data center that has Postgres will probably want one engineered system to make sure that they can get the right sort of visual view of all their databases. And, of course, you can also manage cloud databases and observe cloud databases as well. So observability is a very important part of the story. So that way customers have a good sense of what their databases are doing, and it just eliminates all that complexity.

Algarici: And that also extends to the AI stack. So once you connect this data to the AI, you want to continuously know what's going on with it, who is retrieving it, how is it used, what insights does it generate. So continuing that, extending that stack into the AI that you deploy, is equally important, right? So all the names that—all the NVIDIA inference microservices that we deploy have the observability in mind. They connect to the EnterpriseDB ecosystem, and that single plane of glass is very important end to end.

And that is also part of the data flywheel. That we see how we observe the data, we learn from the data, and then we improve the systems continuously to make the enterprise deployments better.

Lightstone: Yeah. And in fact, you can even customize your dashboards with the engineered system. So you can actually create dashboards for your business that look at the metrics that you want to look at in terms of orders or in terms of AI analytics, anything like that. You can build your own dashboards right in the engineered system itself.

OK, so let's talk a bit about sovereignty and what it means for us, and what it's going to mean for our customers. All of our customers are asking for it.

Algarici: The enterprise landscape has changed so fast, especially in the models. If you just look over the last two or three months, you see the state-of-the-art model of LLM generation has changed probably every other week. So it's really amazing to see the ecosystem come to life and how the live competition really pushes the envelope forward.

At NVIDIA, what we do is we really want to make sure that we work with the open source community, take whatever is best, and even make it better. So through the Llama Nemotron and the Nemotron effort, we've taken multiple enterprise-grade open models and we've made them better through unique data, through synthetic data generation, through our optimization techniques, through post training, and released that to the open source community as well to contribute. So anybody can leverage both the models and the data sets and the scripts to really improve even further. So we love to contribute. We love the open source community and help that out.

Lightstone: So Postgres, of course, is an open source database, and everything inside the EDB system is going to be based on open technology. So, why don't we talk a little bit more about how we address open technologies? Why don't we start? I have a lot to say about that, but Somik, why don't you start and give a little bit of an update in terms of your approach from the Supermicro standpoint?

Behera: Of course, at Supermicro, and me personally, right, we are a big believer in open source and open standards. Before open source, there was open standards. And increasingly, open source *is* the standard, or the open standard, right? So at Supermicro, as part of this engineered system, you're going to have a few components. For example, we were one of the first adopters of Redfish API, an open standard—how you can program and manage in servers. In fact, before we were the market leaders, before we were ahead of Dell and HP—maybe that's why we got ahead of these vendors, is because we embraced it. We integrated it into SuperCloud Composer and SuperCloud suite of software products so we can manage in an open way.

More recently, we have leveraged OpenBMC to manage and control that server as well, using open source technology while working with partners like NVIDIA on GB200 projects, right, with industry-leading supercomputers that Jensen [Huang] released to the market.

And then, going forward, one of the exciting things working with Postgres and the EDB team is that it is not only the number-one database, it's the number-one open source database. Because we believe the future of AI is in the hands of the end users and customers. And they can only be in control of that to leverage open source and open standards and open technologies.

Lightstone: Yeah, absolutely. And in the end, the truth is that the customers decide, and that's why it's so important to have a product that's based on open standards. If customers ever want to choose, you know, another option of running Postgres, they have that option, right? Because we're going to be open about it, and we can actually monitor those and observe them from the engineered system. Of course, most people are going to want to just use the engineered system directly.

So by giving open standards, not only do you ensure that you have a more customer-directed product choice but you also ensure a sort of level of trust with the customer. Because customers know that Postgres is open, which creates transparency, creates trust, and also means that they're influenced by the greater community—and inspected by the greater community in terms of what we deliver.

Behera: Yeah, and you know open source and software is now an established practice. And this is kind of a really great example of bringing all of that together in this engineering appliance.

Lightstone: Right, so there's a lot of reasons why databases have become so important in the world of AI. Databases have been optimized over, really, decades to be the fastest low-latency way to capture data, store it, and also to be extremely highly available, right? So with Postgres, you can be available across three or more nodes. It really depends on what you need.

And you want to make sure that if something were to happen to one of your systems, something that's beyond your control—could be a fire, who knows what, could be an entire city going out. Of course you have redundant power, but it's impossible to know. You want to make sure that you're just always working. And that's why when people look at some of the big companies out there that use Postgres, they're just always working. Even if there's a power outage, that banking system is always working. And often, many of those banking systems are powered by Postgres, powered by EDB. So that's why, as a foundation for data that's going to be used by AI, a lot of people are moving towards Postgres.

And also, because now, with the engineered system, a great step forward is really getting the AI workloads and the database workloads together. And they go back and forth. The simple way of thinking about it is that the AI is sort of the brain, the one that's making logical conclusions with the data. But that brain needs an actual data repository with the data. So you have the very powerful AI brain being able to very quickly go and query or ask questions to the database to get data, right? So you might ask, "Hey, somebody's trying to get information about what car they purchased in 2014, some very long time ago. Can you, database, please give me that data?" And then it might make recommendations based on what it gets back.

So by having those together, you can just scale as fast as you need to go to get the answers to your customers or to make business decisions from these AI models, which are sort of the brains behind the operation.

Algarici: Absolutely. One important observation that we've seen is the scale of agents compared to human–AI interactions. So AI–AI interactions are going to scale deployments quite significantly. Imagine you are sitting in front of your computer, you ask questions. There is only a limit to how fast you can type in your questions. With AI and AI agents, there is no limit. We're seeing deployments where AI agents ask 30 questions at a time, multiple iterations, to really learn as much as they can from the enterprise data and generate the most relevant insights.

So these iterative agents and reasoning agents that really want to learn from that data need that data highly available and keeping that data fresh, keeping that data up to date, highly accurate, so the AI can make the best decisions and create the most relevant insights. So that's where we see the industry going, in using more AI to create even more data and more insights.

Lightstone: Yeah, these agentic workflows, right, where you have basically an AI that's going and asking multiple questions, going back and forth in the data, doing its own independent discovery, so to speak: This is where we see things going, and this is where we have to be ready. And, of course, we have to be highly available and we have to be performant, which is why we all got together and made it happen for our customers.

Behera: Yeah, we have actually, for a number of years, seen data being the core of enterprises. It's always been on top of some kind of database, right? Be it a 3-tier architecture or monolith, or even microservices, they always needed that data as the core beating heartbeat. And we're excited that this becomes now the open source–based core beating heartbeat to bring any kind of application faster to enterprises.

Lightstone: Yeah. And it's not just about databases, but I think another thing is, we work very hard to make it simple, right? So yes, there's that high tech there, but often the simplicity is the barrier, and the reason why customers really struggle to take the latest technology and then actually apply this to their business. And I think that's a place where we worked so hard. You know, when you can deploy high availability in three clicks and you can deploy 100 databases for all your microservices without having to worry about it too much, that's really key. When you can use the latest NVIDIA technology and the latest AI models and you can do that easily, without having to be a deep AI expert.

It's also important for it to be simple for another reason, which is that when you come to any of us and you say, "Hey, I want the engineered system," you want something that you can validate is easy. When we show some of the demos that I think we're all very excited to share, I think customers will say, "You know what? Wow, thank goodness I saw that demo. We thought this was going to be a multiyear project." But because everything is combined, because we've got the speed—like the speed of the actual components in terms of performance—but we've also got the speed in terms of being able to deliver a demo and deliver some of the fundamentals, and just the simplicity. I think that's what we really hope customers are going to give us a big nod on, right?

Behera: Yeah, absolutely. As you know, we're jointly going to market. And I see three industries and three different use cases, which are very prominent. The first one, of course, we have been talking about a lot is enterprise. Enterprise is adopting AI, adopting AI on top of existing sovereign, private, structured SQL data, right?

But we haven't talked about two more use cases, which you just reminded me of. First, the service providers, neoclouds, GPU clouds, AI factories, right? All of these guys need to help their customers build an application, which needs a database, which needs that AI workflow engine like AWS Bedrock or what have you—but on their sovereign estate.

And then the second—the third one is edge, edge environments. Factory floors, right? These robots are going to need some control applications that need to run somewhere. You know, you drop in this rack at every edge location. We have customers which have, like, hundreds of factories who are automating that, leveraging NVIDIA technology. And that would be a great use case for this enterprise engineered system—in addition to, you know, service providers as well as enterprise environments.

Lightstone: Yeah, absolutely, and this is really the next generation of AI, right? And so we know AI is going to be everywhere. We know it has to be reliable, we know it has to be fast, and we also know that it's just very expensive when you don't have all the components working together, right? So I think definitely this will allow us to help our customers take things to the next level.

Algarici: And I think it's only the beginning, right? This is the first evolution, and we'll have many more to come, with new capabilities, advanced models, advanced acceleration techniques. And that will continue to get value more and more from the platform as it grows.

Lightstone: Yeah, absolutely. I mean, we've been working with developers and teams across Supermicro and EDB and NVIDIA, and we're collaborating on quite a lot of different projects in terms of just making things even better and better in terms of where we want to go from here. We know that this is the beginning of a bright future, on top of the really exciting news we have today about the engineered system being available.

All right guys, so it's been great catching up and sharing what we've got in store for the future for our customers. So that's a wrap. And we hope that our customers give it a try. They can order a POC of the engineered system or contact EDB or Supermicro or NVIDIA about the next steps to get started.