NVIDIA

Digital Realty recently announced a collaboration with NVIDIA and Core Scientific to deploy the industry's first Data Hub featuring NVIDIA DGX A100 systems at the Interxion Digital Docklands Campus in London. With access to a new AI-ready infrastructure solution, businesses can rapidly deploy AI models in close proximity to their data sets globally, opening up a new artificial intelligence platform-as-a-service (AI PaaS) solution developed specifically for data science teams.

In Conversation with Tony Paikeday, Senior Director, AI Systems, NVIDIA

How does data impact AI development?

Most enterprises make a huge investment in data science talent to build and deploy AI applications. But there is a real gap between hiring data science talent and actually building AI models that are deployable in a production environment. Many of these AI models never actually move into production. AI is fundamentally different in how it is conceived, prototyped, tested, trained at scale and gets deployed. You are not just building a single, monolithic application, like in a traditional enterprise. Even when you deploy an AI model in production, you need a human in the loop to continually evaluate if this model is performing well, and models drift and degrade over time, because they are feeding on real live data from your operation.

Data is basically the source code that builds great AI models, and data gravity—which explains the nature of large data sets to attract applications and resources towards it—is critical for AI development. Many enterprises do not realize that if there is a lot of time and distance separating critical data from the computing infrastructure that needs to work on that, then you are going to immediately suffer the impact of data gravity.

What is the importance of Data Gravity on AI projects?

Many organizations lean on the cloud as a great way to engage in early productive experimentation. The cloud is very good for making a fast start, and supporting what I would call temporal needs that power early prototyping and development, especially as your AI project is starting to get underway. Over time, your AI model inevitably starts to get more and more complex through ongoing iteration. So, as you iterate and build a more complex model, it is consuming more and more computing cycles. In parallel, the data sets that feed the model training get exponentially larger. And this is the point at which your costs can escalate.

This is a fundamental data gravity problem that many organizations face, and it presents a speed bump and kind of an escalation in the cost of building AI. What ultimately happens is that the rate at which data science teams can build a better, higher quality, more creative model starts to slow down, while the costs rise, because they're spending more time on it. When that happens, the quality of the AI model that you're trying to deliver is affected. That is the inflection point at which many organizations realize that there's a benefit to a fixed-cost infrastructure that supports rapid iteration at the lowest-cost-per-training run. But, how do you get there? You get there by moving your computing infrastructure to where your data lives. This is why we think the architecture and the offer put together by the combination of Digital Realty and NVIDIA is so valuable. You are eliminating time and distance between the data sets. You also are regaining control of your costs since you now have a highly deterministic platform that delivers incredibly fast performance, but in a predictable way.

ABOUT TONY PAIKEDAY

Tony Paikeday is Senior Director of AI systems at NVIDIA, responsible for the go-to-market for NVIDIA's DGX portfolio of AI supercomputers. Paikeday helps enterprise organizations infuse their business with the power of AI with infrastructure solutions that enable insights from data. Paikeday has also held key roles at VMware, where he was responsible for bringing desktop and application virtualization solutions to market, and at Cisco, where he built its data center solutions. Paikeday, who started his career as a manufacturing engineer at Ford Motor Company, holds an engineering degree from the University of Toronto.

We are jointly attacking the problem of helping enterprises industrialize their AI development pipeline and that's at the heart of the Data Gravity Index DGx™ and solving the Data Gravity problem. Digital Realty has the architecture to facilitate that. Our customers also need high performance AI computing, which obviously is why we built the DGX.

How can enterprises benefit from the Data Hub established by Core Scientific and Digital Realty, which features NVIDIA DGX A100?

All verticals have a similar architectural problem related to AI development and infrastructure requirements. When it comes to AI models, they need a large compute footprint that has the performance to train complex models. You also need easy and effortless access with very low latency and very high-speed interconnect to your data infrastructure. With the Data Hub, our solutions and our technology are coming together to solve the same problem and make it easier for organizations to build great AI. We are jointly attacking the problem of helping enterprises industrialize their AI development pipeline and that's at the heart of the Data Gravity Index DGx™ and solving the Data Gravity problem. Digital Realty has the architecture to facilitate that. Our customers also need high performance AI computing, which obviously is why we built the DGX.

NVIDIA DGX systems are the world's first portfolio of purpose-built AI supercomputers. What problems will it solve now and how will the platform evolve?

We are currently in our third generation of NVIDIA DGX systems; our latest generation is DGX A100. A typical enterprise data center is built on legacy computing, i.e. traditional CPU servers with three silos of server infrastructure. Each silo is designed and scaled to tackle only one kind of computational problem: analytics, AI training, and AI inference. This inflexibility has been driving up capital and operating costs in the enterprise data center. So we built DGX A100 to solve this challenge. We consider it to be a universal building block for the enterprise AI data center supporting analytics, training and inference in one agile, flexible platform. DGX A100 is fully optimized to run the entire lifecycle, from AI development and prototyping all the way to production deployment.

What key trends do you see in the future?

In 2020, we saw organizations deploying large-scale AI infrastructure at a rapid rate. If I look to the future, I see two simultaneous trajectories. I am continually impressed by the incredible scale at which organizations are actually building out infrastructure. The need to build state-of-the-art AI solutions and to tackle complex problems with large infrastructure is one trend. Many organizations are consolidating previously siloed AI teams and centralizing people, process and infrastructure to speed AI innovation, especially as they employ AI to thrive in turbulent times.

The other is that many organizations also see that their data scientists and developers can do better work if they have computational power available to them within arm's reach.

We see a rise in putting these powerful computing resources like DGX Station A100 in the hands of data science teams, closer to development and prototyping. More and more organizations now need supercomputing power in the hands of their developers and data science teams as they build and test complex models ahead of scaled training.

"With the vast majority of workloads running in colocation facilities, enabling best-in-class AI PaaS infrastructure in a near-cloud environment helps customers unlock the value of their data lakes. The ability to train models near your data lake, and then use the Data Hub to move around to global edge facilities via a single pane of glass is a game changer for data scientists."

- Ian Ferreira, Chief Product Officer, AI, Core Scientific