DIGITAL REALTY™

Case Study

Jina AI Cost-Efficiently and Swiftly Kicks Off AI Embedding Model Training with Digital Realty

Cutting edge AI company trains neural search models in HPC-ready data center and expects to reach global customer base with PlatformDIGITAL®

Challenge

Jina AI made a strategic decision to broaden its business model and better meet its customers' evolving high demands. This pivotal shift in focus necessitated a substantial increase in computing power, prompting a comprehensive realignment of Jina AI's digital infrastructure.

"We decided to shift our focus from providing a framework to training embedding models from scratch and offering them to our customers as a service. This approach required substantial computing power. Specifically, we needed multiple Graphics Processing Units (GPUs) capable of running at high capacity for extended periods, often months at a time. Given this requirement, it made sense for us to own these GPUs outright. To support efficient operations, we looked for a colocation partner that offered both reliability and a strong return on investment," said Maximilian Werk, Head of Engineering at Jina AI.

Expertise, high quality standards, and cost-efficiency were all crucial factors in choosing the right partner.

Jina AI decided to commission its own servers for the first time since its founding in 2020 and wanted an external data center partner. As this was Jina AI's first such deployment, finding an experienced and dependable partner was crucial. The data center partner needed to provide an appropriate environment and demonstrate experience in GPU operations.

Beyond providing services, the ideal partner should actively advise Jina AI, anticipate potential challenges, and support the entire process from planning to implementation and ongoing operation in a competent and straightforward manner.

About Jina AI

Jina AI is a leading search AI company for multilingual and multimodal data. Their open-source solutions enable companies and developers to index, search, and retrieve information of various types. Jina AI's B2B clients use their solutions to more efficiently search large volumes of documents, significantly reducing their support workload. Additionally, these solutions can categorize vast amounts of data, preparing it for use with generative AI models.

Industry
Service Providers

Headquarters
Berlin, Germany

Use cases
AI Training/Inference - Private Cloud

Jina AI's outcome from using Digital Realty's products and services:
Colocation and interconnection on PlatformDIGITAL®

50 %
Less energy required for active cooling 6 instead of 12 months - Cooling necessary

8 weeks
Time to market

99.999 %
Uptime

Solution

"During our procurement process, we consulted with NVIDIA, our GPU supplier. They provided us with a list of colocation partners, which included Digital Realty. We conducted a thorough evaluation of Digital Realty alongside several other colocation providers. After careful consideration, we determined that Digital Realty was the best fit to address both our current requirements and anticipated future needs," Werk said.

Global colocation on PlatformDIGITAL®, the data meeting place

As one of the largest and most experienced global providers of data center solutions, Digital Realty offers a robust infrastructure to power AI initiatives. With over 300 data centers strategically located across key metro markets worldwide, they provide the flexibility, scalability and advanced cooling methods Jina AI needed to accommodate growing AI workloads.

Digital Realty's data centers deliver the robust power and infrastructure required for High Performance Computing (HPC) and are fully AI-ready. Additionally, customers like Jina AI benefit from dedicated local experts and single points of contact who understand the unique needs of businesses in their country. These experts leverage Digital Realty's global reach to design and implement optimal solutions tailored to each customer's requirements.

"We had a local partner that helped us choose the best location for our GPUs from a number of options. We did not need to vet these options ourselves, but Digital Realty did this and explained it in a thorough way to us," Werk said.

Once Jina AI and Digital Realty agreed on the specifications—Jina AI chose a cabinet with two servers, each consuming up to 10 kWh located in Stockholm, Sweden—the process moved swiftly. Digital Realty prepared the data center space, including power supply, within just two weeks. This allowed Jina AI to seamlessly transition from testing at their supplier's site to full operation without any delays.

Jina AI benefits from Digital Realty's vast experience and customer focus

"Digital Realty's assistance after deployment was invaluable. When we experienced our first technical issue, we didn't need to send technicians from Berlin to Stockholm. Instead, we utilized Digital Realty's Remote Hands Services. This efficient approach resolved the outage within just a few hours," Werk said.

As far as Jina AI's future growth and customer acquisition are concerned, Digital Realty offers optimal conditions through its PlatformDIGITAL® and ServiceFabric® offerings. This comprehensive data center platform enables Digital Realty's customers to meet their data center requirements efficiently. It excels in facilitating robust and secure connections between partners, streamlining collaboration.

ServiceFabric® specifically empowers AI-focused customers by providing seamless, low-latency interconnection between applications, clouds, and ecosystems, which is essential for high-performance AI and HPC workloads. The platform lets customers automate workflow orchestration and reduce manual configuration, simplifying the management of complex AI environments. By supporting both private and hybrid AI architectures, ServiceFabric® facilitates robust security, scalability, and optimal uptime for sensitive, data-intensive applications, allowing businesses to rapidly adapt to evolving AI demands and drive innovation with data-driven decisions.

> "We decided to shift our focus from providing a framework to training embedding models from scratch and offering them to our customers as a service. This approach required substantial computing power. Specifically, we needed multiple GPUs capable of running at high capacity for extended periods, often months at a time. Given this requirement, it made sense for us to own these GPUs outright. To operate them efficiently, we needed a cost-effective and reliable colocation partner."
>
> **Maximilian Werk**
> Head of Engineering at Jina AI

Rapid response, on-site technicians

Digital Realty's 'Remote Hands' service has been extremely valuable to Jina AI: Instead of sending their own technicians to the data center, Jina AI can deploy Digital Realty's local staff. By avoiding long journeys and associated travel times, GPU downtime has been reduced to a minimum. This not only underlines the efficiency of the service, but also contributes to a significant saving of resources, as it reduces travel costs. The ability to solve technical problems quickly and flexibly on site is a decisive factor for the smooth operation of AI infrastructures.

"This was our first colocation deployment, but Digital Realty approached it like a true partner—not just a vendor. They guided us through every stage of the process and proactively advised us on potential issues to watch out for, which helped us avoid delays and unnecessary costs," Werk said.

Outcome

100 % green energy in Stockholm

Jina AI swiftly operationalized its GPUs in Stockholm with Digital Realty's assistance. The entire process—from decision-making through consultation to commissioning—took only eight weeks. Currently, sixteen GPUs are running on two servers in a Digital Realty cabinet, with a combined power requirement of up to 20 kWh and with 99.999 % uptime.

This setup is powered entirely by renewable energy. Thanks to its strategic location, the data center in Stockholm benefits from an innovative cooling system: half of the year, the environment operates without cooling, while the other half utilizes sustainable seawater cooling, further lowering the overall energy need. By choosing Digital Realty's Stockholm campus, Jina AI requires 50 % less energy for active cooling.

"Digital Realty's innovative cooling infrastructure in Stockholm has lowered our energy spend, and their pricing model provides the cost predictability that's critical for long-term planning," Werk said.

Jina AI is well-positioned for future growth with Digital Realty. "We're witnessing a large and rapidly expanding market for our solutions. Our clients possess extensive, well-organized document repositories that form the basis of their customer support," Werk said.

> "Digital Realty was always straightforward with us, which I greatly appreciated, given that I had never deployed any server at a colocation data center before. They guided us through every stage of the process and proactively advised us on potential issues to watch out for, which proved to be both helpful and efficient."
>
> **Maximilian Werk**
> Head of Engineering at Jina AI

Key Insights

Context
Jina AI, an innovator in multilingual and multimodal AI search, shifted its focus to training embedding models.

Challenge
Jina AI needed an AI-ready data center environment for high-performance GPUs to enable customers to train embedding models.

Solution
Digital Realty's global data center PlatformDIGITAL® enabled AI-ready colocation and global interconnection, expert consulting, and rapid deployment at the sustainable Stockholm campus.

Outcome
Jina AI's AI embedding model training environment was ready in just eight weeks, accelerating their business for future expansion. They can operate their GPU infrastructure at scale, with 99.999 % uptime, control energy costs and a clear path for future expansion.

About Digital Realty

Digital Realty brings companies and data together by delivering the full spectrum of data center, colocation, and interconnection solutions. PlatformDIGITAL®, the company's global data center platform, provides customers with a secure data meeting place and a proven Pervasive Datacenter Architecture (PDx®) solution methodology for powering innovation, from cloud and digital transformation to emerging technologies like artificial intelligence (AI), and efficiently managing Data Gravity challenges. Digital Realty gives its customers access to the connected data communities that matter to them with a global data center footprint of 300+ facilities in 50+ metros across 25+ countries on six continents. To learn more about Digital Realty, please visit **digitalrealty.com** or follow us on **LinkedIn** and **X**.