run:ai case study

How a leading global bank scaled AI efficiently across regions and teams with Run:ai

A financial leader scaled its AI efficiently and improved time to market of strategic business initiatives. It did so by using Run:ai to better allocate GPU compute resources according to business priorities.

Customer Background

A global bank with trillions of dollars in assets under management. Has played a special role in global financial history and leading multiple financial innovations.

The bank has continuously invested in technology and innovation and views it as a competitive advantage. In 2021, it invested more than $3 billion in technology development, with technology employees making up a quarter of its workforce.

With access to one of the world's largest data sets and a strategic bank-wide AI/ML initiative, the pace of innovation has increased in recent years, with the goal of using data science to "uncover the hidden value behind every transaction."

Customer AI Infrastructure and Team

- On-premises environment with multiple clusters, each with tens of high end GPUs
- Hundreds researchers / data scientists across business units (e.g. HR, Real Estate, Risk)

After Implementing Run:ai's Platform

Enabled 10X data scientists to work on the same number of GPUs (using fractions)	Higher ROI faster time to market
Integration with all data science tools used across the bank	Happier more productive data scientists
Dozens of AI/ML projects being resourced according to business prioritization	More business impact, faster

Challenges

- **Improve GPU utilization.** Out of the bank's almost 400 data scientists, only a few tens were able to use the bank's GPUs concurrently. This was a tremendous waste, especially as many projects were in an initial phase of building, with small data sets, and no need for a whole GPU.

- **Allocate compute resources according to business priorities.** There were 15-25 active AI/ML projects vying for compute resources in the same GPU cluster. The bank had a riguous ROI estimation process taking into consideration investment, risks, and returns and prioritizing projects accordingly -- but GPU allocation was not done according to these priorities. Instead, researchers were using GPUs inefficiently, consuming whole GPUs instead of fractions, "hogging" GPUs and preventing other projects -- sometimes more important ones -- from using them.

- **Maintain tool flexibility - while ensuring coherence.** Different data science groups across regions and teams used different tools. The bank wanted to keep this flexibility and freedom, without creating a 'Wild Wild West' of siloed and inconsistent technologies sprinkled across the enterprise.

Solution

Run:ai's platform capabilities enabled the Company to achieve:

- **Fair scheduling and guaranteed resources.** Using Run:ai, admins now easily control GPU fraction allocation according to business priorities and other factors like seasonality.

- **Increased GPU utilization, leading to faster model development, training, and deployment.** Run:ai's unique ability to use fractional GPUs instead of a whole GPU enables the bank to use its GPU cluster in the most efficient way possible. Now, the entire data science / research force of ~400 employees can share the cluster and access only the GPU fractions they need, when they need it. This led to huge improvements in time efficiency and productivity for the data scientists, as well as faster model iterations and shorter time to production - getting business value from the models faster.

- **Regulatory compliance.** All projects using the Run:ai platform adhere to the bank's regulatory, risk, and compliance requirements. Data governance processes and robust access permissions are also an integral part of the environment.

> *With Run:ai, we were able to build out AI infrastructure from the ground up for scale. This ensures we get AI out of the lab and into production.*
>
> *- Principal Data Scientist, Head of Enterprise Data Science Platform at leading global bank*

About Run:ai

Run:ai's Atlas Platform brings cloud-like simplicity to AI resource management - providing researchers with on-demand access to pooled resources for any AI workload. An innovative cloud-native operating system - which includes a workload-aware scheduler and an abstraction layer - helps IT simplify AI implementation, increase team productivity, and gain full utilization of expensive GPUs. Using Run:ai, companies streamline development, management, and scaling of AI applications across any infrastructure, including on-premises, edge and cloud. Learn more at www.run.ai.