SOLUTION BRIEF

AI-ready Infrastructure for Quantitative Trading

Empower your quantitative trading strategies with the cutting-edge power of AIRI®.

In the race to generate alpha, time is money. Uncovering profitable relationships hidden in complex data sets requires the ability to generate new ideas, test strategies against massive quantities of data, and deploy models quickly and efficiently.

Enhancing Quant Trading with GenAI

The use of artificial intelligence (AI) is now table stakes in the hypercompetitive financial services industry. In fact, McKinsey estimates that AI can generate up to $1 trillion additional value annually for the global banking industry.[1] As one of the fastest growing areas of enterprise technology investment, AI supports successful quant trading by enabling teams to maximize the value of their data, including alternative and unstructured data, smaller files and metadata, and enormous volumes of historical data.

In the quant trading arena, credibility, accuracy, and timeliness are crucial. By leveraging retrieval-augmented generation (RAG) to incorporate specific proprietary vector databases, firms can enhance GenAI models and improve quant firms' ability to train their own LLMs. External data from financial statements, analyst views, compliance filings, or customer transactions can provide more specific and meaningful output, enabling AI systems to generate timely investment insights, enhance compliance, and mitigate risk.

To capture fleeting alpha, quant teams need high-performance infrastructure that can simplify, consolidate, and backtest all of that data at scale so they can optimize parameters and deploy proven algos with greater agility.

DIY Challenges and Roadblocks

The traditional IT infrastructure that supports legacy data science struggles to provide rapid access to the right data, limiting the advantage of GPU-powered AI-based data science. While cloud is an option, costs can become prohibitive as AI projects scale, strategies are spun up and down, and ingress and egress fees mount with each iterative test. Additionally, the performance simply can't meet the needs of latency sensitive strategies.

As a result, organizations seeking to enable AI with simplicity, speed, and cost-effectiveness often face roadblocks such as unexpected hardware costs, complex AI software and infrastructure, and lengthy and complicated deployment cycles, all of which delay the time it takes to turn data into insights and risks missing profitable market opportunities.

AIRI Technology Stack

- Pure Storage FlashBlade//S™
- Pure Storage Software: Purity//FB OE, Pure1® Management, Portworx® Kubernetes Data Platform, Pure RapidFile Toolkit.
- NVIDIA DGX H100, NVIDIA DGX A100
- NVIDIA Quantum and Spectrum Networking
- NVIDIA Software: NVIDIA AI Enterprise, NVIDIA Base Command, NVIDIA NGC, CUDA-X, Magnum IO, RAPIDS, PyTorch, NVIDIA TAO Toolkit, TensorFlow, NVIDIA TensorRT, NVIDIA Triton Inference Server, Merlin, Morpheus, Riva, and NeMo.

AIRI Simplifies AI-at-scale

- Reduces the complexity of integration with a proven and complete DGX BasePOD certified solution.
- Data scientists can focus on algorithms and outcomes, not infrastructure.

AI-at-scale Is an Advantage

- More compute = faster training
- More data = higher accuracy
- AIRI makes it simpler and faster to run multi-node training

Uncomplicate Your Data

AIRI®, a comprehensive AI infrastructure solution certified on the NVIDIA DGX BasePOD reference architecture, represents the latest evolution in complete AI-ready infrastructure. Developed by Pure Storage® and NVIDIA, AIRI incorporates the latest NVIDIA DGX systems, NVIDIA networking, and Pure Storage FlashBlade//S™ storage. It offers a complete hardware and software solution that maximizes AI results with the industry's most efficient scale-out storage platform.

DGX BasePOD is a prescriptive NVIDIA AI infrastructure architecture that eliminates the design challenges, lengthy deployment cycles, and management complexities associated with scaling AI infrastructure, enabling quant teams to train, test, optimize and deploy strategies before they lose their edge. DGX BasePOD certification ensures a validated, proven full-stack solution that is simple-to-use, allowing organizations to quickly benefit from AI with a fast and efficient infrastructure that meets enterprise-scale demands.

Key Components and Features

Powered by NVIDIA Base Command software, DGX BasePOD provides the essential foundation for AI development optimized for enterprise. NVIDIA Base Command is the operating system of the DGX data center, helping organizations speed the ROI of AI.

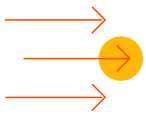
Base Command enables firms to tap into the full potential of their AI infrastructure with a proven platform that includes workflow management, enterprise-grade cluster management, libraries that accelerate compute, storage architected for AI, network infrastructure, and system software optimized for running AI workloads. It also includes the NVIDIA AI Enterprise software platform, which includes NVIDIA NIM, a set of easy to use microservices designed the accelerate the deployment of generative AI models, as well as proven and supported AI Frameworks, SDKs, and libraries relevant to the financial services industry, such NVIDIA RAPIDS, PyTorch, NVIDIA TAO Toolkit, TensorFlow, NVIDIA TensorRT, NVIDIA Triton Inference Server, Merlin, Morpheus, Riva, and NeMo.

FlashBlade//S, purpose-built from the ground up for modern, unstructured workloads, accelerates AI processes and simplifies scaling and operations. FlashBlade//S supports NVIDIA GPUDirect Storage. This means customers can maximize performance efficiency required for large AI workloads by eliminating CPU processing bottlenecks. A centralized data platform within a deep learning architecture allows data scientists to ingest, analyze, and backtest all the data, and provides an agile environment for the quant analyst.

Delivering the Next Generation AI Platform

Simplify Your AI Workflows

AIRI is a pre-validated reference architecture with full stack integration that allows for quick setup, deployment, and management as an end-to-end AI workflow solution. This means quant teams can focus more on delivering insights and less on infrastructure concerns.



Maximize Efficiency and Productivity

Pure Storage DirectFlash® technology maximizes storage efficiency and model training performance. Get higher capacity in a smaller physical and power footprint, while maximizing DGX and storage performance. Predictable high performance AI infrastructure shortens training, enables you to iterate faster, and minimizes time-to-insight.

Better Economics to Turn Your AI Data into Profits

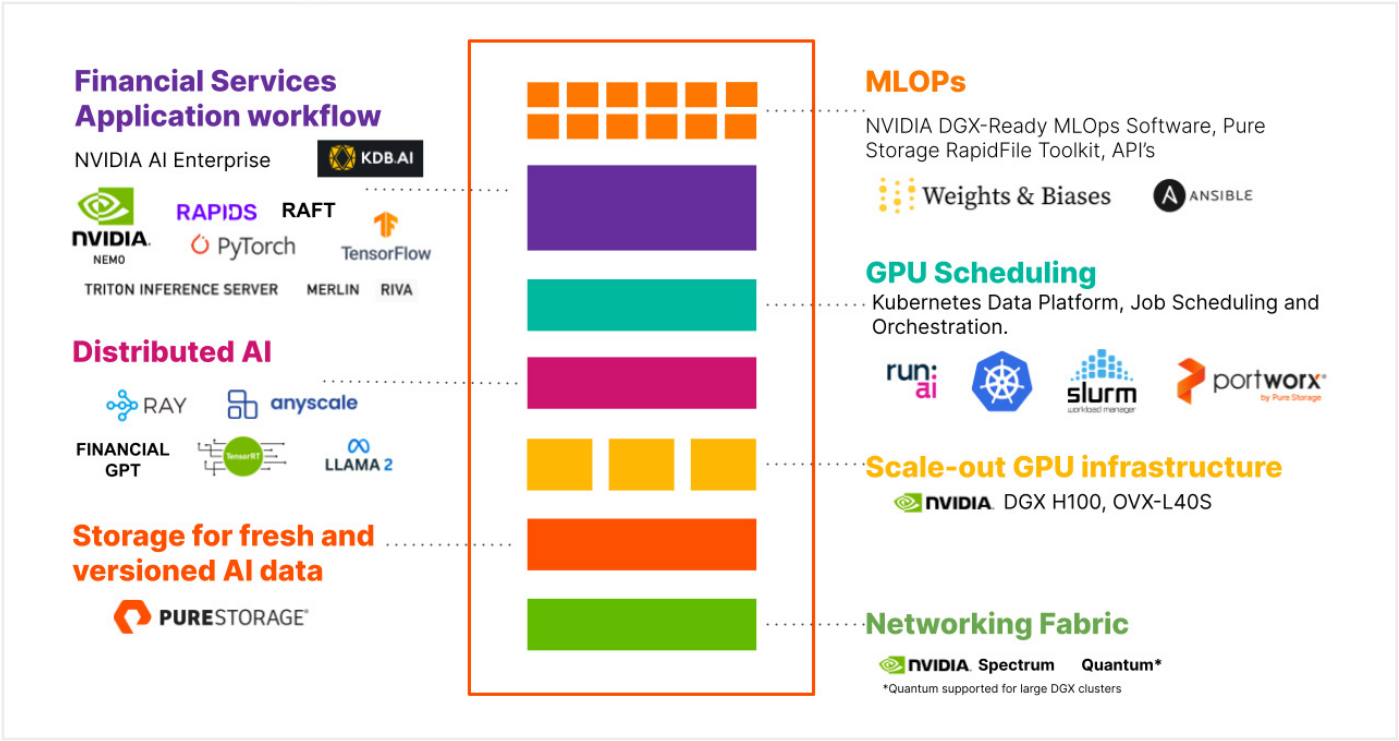
Lower TCO by consolidating data silos and analyzing diverse data sources at scale, providing performance as needed and offering all-flash features with disk economics for AI content repositories. A simple and consistent experience for all AI data reduces operational expenses on top of cost savings from efficiency and productivity benefits.

AIRI Technology Stack

AIRI combines groundbreaking solutions from two industry leaders. All-flash storage from Pure Storage, architected to meet the exacting demands of quant trading and backtesting is paired with the latest AI infrastructure and software from NVIDIA. Optimized out-of-the-box, AIRI seamlessly scales storage and AI compute performance, allowing firms to embark on their AI journey at any scale and adapt quickly as their models evolve.

NVIDIA Base Command software and NVIDIA AI Enterprise software coupled with the Purity//FB operating environment and Pure RapidFile Toolkit provide a robust software foundation to jump-start and accelerate your quant trading journey.

Pure Storage Validated Financial Services AI Stack



Pure Storage: Enabling AI Success with NVIDIA

Pure Storage has worked with NVIDIA to help organizations exceed expectations in their AI initiatives across a variety of industries.

The following are just a few of the customers Pure Storage and NVIDIA have enabled AI success:

- Accelerating data science and faster data analysis at Options Technology, a leading provider of IT infrastructure to global capital markets firms
- Increasing data processing speeds by 2x and expanding GPU usage from 30% to 80%, maximized efficiency and management of data used to develop and train Chunkbuk Technopark's AI models.
- Accelerated genetic research and medical imaging 7x at the Center for AI in Medicine at Chang Gung Memorial Hospital.
- Accelerated medical image processing which increased customer satisfaction and market competitiveness at Olympus, a leading MedTech provider.
- Advanced AI analysis of large data sets at NavInfo to produce real-time maps and navigation models used by car manufacturers worldwide.
- Advancing the national security mission of the United States and other nations' intelligence agencies (private references).
- Accelerated EDA to support the pace of chip development where for each nanometer a die is reduced in size, the volume of data required to be processed in the development almost doubles from the prior size (private references).

Learn More

AIRI from Pure Storage powered by NVIDIA accelerates and simplifies the process of deploying and running complex AI infrastructure, allowing quant firms to focus on their core mission and enabling models to react faster to any market conditions.

- Explore the AIRI solution.
- Learn how Pure Storage solutions accelerate your quant trading.

1 Building the AI bank of the future

PS2597-01-en 0524