

Data Makes Moves with Generative AI

Data Management Considerations For Generative AI

(It's All About The Workflow)

It's the next big thing, newly embedded in every business model – from high-tech to home-grown. With the ability to generate human-like text, images, and even code, generative AI opens up a wide range of applications from content generation to natural language processing.

69% **Businesses using AI/ML to generate revenue streams.***

Generative AI became a household name thanks to widely available cutting-edge AI models like GPT-3 and its successors. Now every business must consider the upside of deploying GAI for insights, optimized efficiency, and new revenue streams. The success of any generative AI hinges not only on the quality and quantity of available data, but even more so on how easily that data can move into, and out of the model. For many projects, the agility with which a proprietary model handles data will be the make-or-break factor.

PROBLEM:

Most Data Architectures Are Not Prepared for Generative AI

The data that serves as the backbone of model training takes many forms. A given company's training data can consist of millions or billions of differently sized files, from structured or unstructured text to images and audio. Files may also be spread across a wide range of locations, including cloud, on-prem, or object storage, not to mention held in different data centers around the world.

*More businesses cite DATA MANAGEMENT as their primary technological inhibitor to AI/ML deployment, saying it's more challenging than both security and compute performance.

AI Development is ALL About the Workflow

To assemble and ingest such a wide diversity of data from multiple sources on a continuous basis, one can embark on a painstakingly manual process, or lean on automated data management. Data management shepherds data as it passes to/from storage and through the pipeline, ensuring a consistent flow of data from one end to the other and minimizing the expensive risk of idle GPUs. Speed of storage is important here but, it's not everything. Storage utilization will determine how effectively data moves into, though, and out of the pipeline. A 530B parameter LLM only requires .95MBps of performance for training. To checkpoint that same model, however, requires 103GB/s of write performance. A slow checkpoint generates 7,420GB over just 2 hours while GPUs sit idle. Running continuously, this process generates almost 90T of new checkpoint data that may never be used. That means more storage spend, more time invested, and more manual data movement. Meanwhile, your project ROI just got harder to achieve.

The Old Way: Throwing Money at the Problem

Historically, engineers had two options for effectively managing their AI workflow without letting their GPUs sit idle. Hire more people to handle the manual data movement. More data wranglers can help with diverse I/O profiles and move data to the to correct storage tier by hand. Add more capacity so the pipeline could move more data before getting jammed up. Both options are highly feasible for billion-dollar budgets. For everyone else, and for the efficiency-minded, there's a better way.

The New Way: End-to-End Data Management

Because your AI workflow must run concurrently with your data workflow, data management is a must-have to orchestrate data movement and optimize usage for GPUs and Tier 0 storage. AI development isn't about just fast or performant storage, it's about effective movement through the pipeline, and that means every step.

Ingest	Preparation	Training	Checkpoint	Fine Tuning	Inference	Archive
Write intensive	Read/write intensive	Read intensive	Write intensive (sequential)	Iterates on many of the same I/O profiles of Preparation and Training steps	Read/write intensive	Read intensive (sequential)
Huge, scattered datasets	Millions / Billions of files	Long running: reliability is key	Large output		Results must be stored; adds capacity pressure	Adds performance pressure to get data off archive fast enough
Laborious if performed manually	Metadata intensive		Pauses training execution while running			Laborious if performed manually
			Adds huge capacity pressure			



Ingest: Data ingestion is the foundation of data management in Generative AI and is like gathering the pieces into a puzzle before assembling your masterpiece. This is the time to identify your data sources and ensure they're clean, structured, and ready for what's next. If your data management is working, you can feed substantial amounts of data from multiple sources continuously into your hungry GPUs to keep the pipeline going.



Preparation: Data preparation involves transforming and structuring the ingested data in a way that is suitable for training the GAI model. This includes tasks like data normalization, encoding, and feature engineering. The quality of data preparation directly impacts the model's ability to learn and generate meaningful content.



Training: During training, the GAI model learns to generate content based on the input data. Accurate and coherent results require high-quality, well-structured training data delivered to the right place at the right time. Data management practices here ensure data availability, version control, and the ability to scale training processes efficiently.



Checkpoint: Checkpoints save intermediate states of the model during training to make sure training can be resumed or fine-tuned without starting from scratch if necessary. Proper checkpoint management is essential for optimizing model performance and reducing training time. If your data management is working, the massive amount of checkpoint data is processed quickly and automatically moved away from Tier 0 to a less expensive storage tier, where it can be pulled back immediately as needed. Without seamless data management here, training comes to a halt and expensive GPUs sit tragically idle.



Fine Tuning: Fine-tuning is an iterative process where the model is adjusted to improve its performance on specific tasks or datasets. Effective data management at this stage allows for controlled experimentation and the ability to adapt the model to changing requirements.



Inference: In the inference phase, the model generates content based on input queries or prompts, which must be stored for the model to continue learning. Effective data management ensures an efficient, reliable, and scalable process in which the data is accessible when and where needed.



Archive: Data management is not limited to active processes; it also involves archiving historical data and model versions for compliance, auditing, and retraining purposes. Archiving is critical for maintaining a historical record of data used for model training and inference, but it must be done quickly for the generative AI pipeline to run.

Without a data management layer, this becomes a manual process of offloading data from Tier 0 storage that raises risks of human error and delay.

Optimize Every HPC Workflow with NGenea

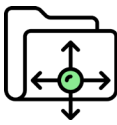
HPC and AI-ready data management must be many things – highly performant, location agnostic, agile, automated, and rich in capability. NGenea delivers a suite of solutions that solve for the challenges of a busy AI pipeline by performing several duties at once:



Orchestration



Unification



Data Movement

Optimized Tier 0 Capacity	Unified namespace offers global accessibility with NO data lock-in	Fully automated and protocol-based movement
Why it matters:		
Data is always in the right place at the right time. Drives performance, reduces required capacity, and optimizes TCO	Data is searchable and retrievable from anywhere by anyone	Whether On-prem, cloud, hybrid, data moves quickly through the pipeline

High Performance at the Edge – A Case Study

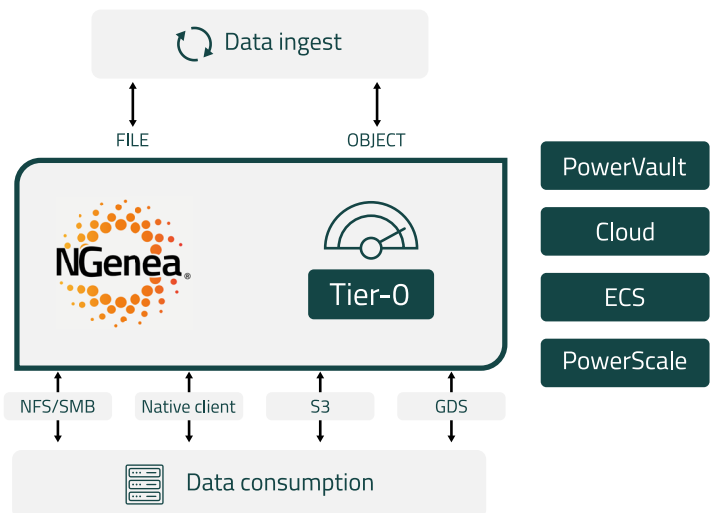
Duos Technologies leverages automated data management to keep the freight business on track. Originally built to detect illegal smuggling, the Railcar Inspection Portal (RIP) used 190 cameras to capture machine vision images from trains traveling at speeds up to 125mph. It was quickly discovered that the same technology could be deployed to perform early detection of maintenance issues on passenger trains – but that required significantly more processing power to capture, compress, store, and analyze raw images, then feed that data into an AI model to detect anomalies and alert proper authorities.

The RIP cameras generate 60GB of data every second and, as an edge application, Duos did not have the luxury of placing server warehouses on the side of the tracks. So, they worked with Kalray to engineer a system to house the AI pipeline including Dell servers, NVIDIA GPUs, 2 NGenea NG-Stor Management nodes, and 4 NG-Box NVMe nodes that offered half a petabyte of usable capacity with an ingest rate of over 80GBps.

Thanks to fast and secure processing by Dell servers and hyper-efficient Kalray storage, Duos helped their customer strategically place these monitoring portals to catch maintenance issues before they snowball into track-stopping disasters.

“We realized it was going to work when we turned it on. When we hooked up the 190 cameras and pushed it, and it started filling that drive up with images, we didn’t miss a bit.”

Derrick Schmenk, Duos Technologies



Q&A - Data Management Considerations for Generative AI

If you plan to join the ranks of businesses deploying generative AI to build revenue streams, improve customer experience, or streamline operations, consider data management a crucial building block. When selecting your data management partner, consider these questions.

Q. Who owns my data?

A. Many companies are happy to manage your data for you - and even do a great job while they're at it. But when you need to retrieve your data, they charge you to access your own information. NGenea refuses to play this "legitimate ransomware" game. Instead, we deliver transparent pricing and highly flexible data management solutions with no data lock-in, ever.

Q. Where do I need to store my data before feeding it to the model?

A. Most generative AI training data sets include a diverse array of file sizes, types, and locations. NGenea data management works with data whether it's on-prem, in the cloud, or both.

Q. How hands-on do I need to be?

A. The level of orchestration delivered by your data management solution will impact how quickly you are able to get to market. Look for data management solutions that offer automation to streamline data processes, reduce manual intervention, and improve efficiency.

Q. How well do you play with my existing storage?

A. NGenea's data management platform sits on top of any storage solution, whether it's our own NG-Stor, Dell PowerScale, AWS (Amazon Web Services) cloud storage, or anything else.

Q. How do I collaborate with my global / distributed team?

A. A unified global namespace means data is visible and accessible from a single interface by anyone on your team. For that reason, Kalray NGenea is the data management choice for modern distributed and asynchronous teams from Los Angeles to Lagos, Paris to Singapore.

Conclusion: Meet the Unsung Hero of Generative AI

How data is stored and managed data is the key to ensuring smooth operations, resource allocation, and optimized price performance for any generative AI undertaking. Data storage plays the bookend, holding data pre-ingestion and post-archive, as well as storing the extremely large files generated by data checkpoints midway through the process. But data management's role is critical to saving time in an already labor-intensive process.

For example, Kalray NGenea consistently reaches 60GBps while maximizing usage of Tier 0 storage to keep costs low and projects moving forward. And the system scales linearly to keep up with growth as more GPUs are added for greater performance. Selecting a unified storage and data management solution is a step toward smoother operations, better resource allocation, and optimized price performance.