August 2026

EBOOK

The AI Storage Cloud Playbook: Affordability at Scale

wasabi
hot cloud storage

FOREWARD

By James Montantes

Everyone's focused on AI compute. GPUs, training clusters, and inference infrastructure get the headlines and most of the budget. But between the orchestration layer and raw data sources, most AI architectures are missing a critical layer.

Organizations need a persistent, cost-predictable storage foundation built for the iterative nature of AI workloads.

Without this layer, every stage of the AI pipeline (ingestion, preparation, training, inference, and governance) runs on hyperscaler storage that charges fees for every data read, write, and transfer. Those fees compound with each training cycle, each fine-tuning pass, and each model update. Teams end up rationing data access, skipping training runs, and restricting inference. All of this degrades the AI models they invested in building.

I created this playbook to draw attention to the missing layer, define what it requires, and point you to tools that can fill in that missing layer while providing relief for your organization's budget.

James Montantes
Senior Industry Marketing Manager

© 2026 Wasabi Technologies

Is your AI stack incomplete?

When organizations plan AI investments, the budget flows toward compute. GPUs, TPUs, training clusters, inference infrastructure are the headline items. And they should be. But this focus can create a blind spot.

AI compute needs come in bursts by nature. Training runs last hours or days, then scale down or shut off. Storage plays a fundamentally different role. It's constant. Training data needs a home even when nobody's training. AI model weights need to be accessible whenever inference is performed. Logs accumulate around the clock. And unlike compute costs, storage costs compound over time. Especially when your provider charges you every time you touch your own data.

Picture your AI stack as a set of layers. Inference and serving sit at the top. Model training and fine-tuning sit below that. Orchestration and pipelines in the middle. Raw data sources at the bottom. Between orchestration and your raw sources, there's a gap. A layer that should be there, but isn't.

That's where the AI storage cloud belongs: the persistent storage foundation that connects every stage of your pipeline.

In most architectures, the AI storage cloud doesn't have a clear owner. And when it's absent or assigned to hyperscaler object storage with per-API fees, the cost compounds at every iteration. Data gets rationed. Training cycles get skipped. Model quality suffers.

AI data is inherently different from traditional enterprise data. Four properties set it apart:

01 Scope and storage

AI tools require massive amounts of data to build effective models. Since personalized AI requires access to real business data, this information is often decentralized and siloed across the organization.

02 Frequent movement

AI systems constantly access data to perform training, inference, and fine-tuning, making accessibility critical at every stage.

03 Constant growth

AI data starts large and only grows. Cleaning, chunking, indexing, archiving, and logging expand the data footprint as new fields, metadata, and intermediate versions are created for future use.

04 Data retention

Organizations must retain AI data for compliance with laws like the EU AI Act, as well as to support reproducibility, rollback, and model retraining.

© 2026 Wasabi Technologies

The pipeline loops

AI systems don't follow a straight line from data to deployment. They loop. AI models encode their training data, which grows out of date. Retraining, fine-tuning, and re-indexing are ongoing. Each pass through the cycle creates new durable artifacts and every one of them needs storage.

The cycle breaks down into 6 stages, and each one has specific storage demands.

Ingestion

What happens:

Raw data from documents, logs, APIs, and databases lands in a centralized repository organized by source. This provides centralized storage while maintaining attribution and allowing the training dataset to be efficiently updated. Efficient ingestion requires high-throughput, sequential write patterns.

Storage exposure:

Without a durable, high-throughput landing zone, this step becomes a bottleneck before the pipeline even starts. Organizations need cost-effective cloud storage to centralize raw data from every source for later processing.

Cleanse and prepare

What happens:

Multi-source data becomes clean, structured datasets: PII masked, data chunked for retrieval-augmented generation (RAG), formats normalized into canonical JSON/Parquet formats. This process dramatically expands the AI data footprint. Enterprises must store both the original raw data and the cleaned datasets to support regulatory compliance and retraining.

Storage exposure:

This step can more than double your storage footprint. Long-term storage requirements may more than double, and you need both the original and the cleaned copy accessible for compliance and retraining purposes.

Index and retrieve

What happens:

Indices are built for rapid, secure data retrieval with access controls and security filters embedded in indexing metadata. While indexing doesn't create a whole new copy of the dataset, it introduces additional metadata and requires a combination of fast local storage for operational vector databases and durable cloud storage for index snapshots.

Storage exposure:

Without durable storage for index snapshots, a failed index means a full rebuild from scratch. Organizations need to support recovery and portability across environments.

© 2026 Wasabi Technologies

AI Model training

What happens:

Structured datasets are read repeatedly as the AI model extracts patterns and compresses them into weights. Data is often copied from long-term storage into local scratch storage to minimize latency during the frequent reads and writes that training requires. The primary role of persistent cloud storage at this stage is storing final model artifacts and checkpoints to provide insight into how the AI model was trained and expediting future training cycles.

Storage exposure:

Every read from hyperscaler storage triggers an egress fee. Checkpoints need a cost-effective home after each run. Otherwise, teams may skip them and sacrifice the ability to resume training from a known good state.

Inference

What happens:

Trained AI models run in production: generating responses, executing workflows, and supporting applications at scale. Production AI systems must maintain secure, comprehensive operational records. Interaction logs, system events, and configuration data are needed for compliance, monitoring, troubleshooting, and providing data that informs future optimization.

Storage exposure:

Every inference call generates logs: your audit trail, compliance record, and future fine-tuning dataset. They accumulate fast, and they need to be immutable. Production AI runs around the clock, and so does the data accumulation.

Govern and archive

What happens:

Enterprises must comply with regulations specifying how AI can be used (EU AI Act), how data must be secured (GDPR, CCPA), and how systems must be protected (NIST AI RMF). Maintaining a compliant, auditable record requires long-term, scalable, cost-effective storage with built-in immutability and access management.

Storage exposure:

Years of sensitive data at scale, requiring provable immutability. On most hyperscalers, Object Lock costs extra. The storage layer must operate as the AI system of record across all stages.

© 2026 Wasabi Technologies

The hidden cost of AI

Cloud storage costs come in two main categories.

Base storage fees are calculated based on the volume of data stored and its level of accessibility. Faster access often means higher costs, with potential discounts for reserved capacity.

API and egress fees are charged on top for data reads, writes, transfers, and operations like enabling immutability. These vary month to month based on usage and are notoriously difficult to predict.

~50% *of the average cloud storage bill is comprised of fees, not actual storage capacity.*

73% *of companies run hybrid cloud estates, with rising adoption of multicloud, triggering fees at every step across environment boundaries.*

For most companies, these additional fees are where budgets blow up. Nearly half (~50%) of hyperscaler storage spend goes to fees rather than actual capacity. And most organizations now run hybrid or multi-cloud estates, crossing boundaries that trigger fees at every step.

For traditional workloads, those fees are painful. For AI workloads, they're structural, because AI touches storage at every stage of an iterative pipeline.

Where fees hit hardest

Training

Data is read from storage repeatedly. Each read triggers an egress fee. Add fine-tuning cycles and the bill compounds fast. Companies often need to transfer entire datasets from long-term storage to local scratch storage for compute-heavy tasks, with each transfer carrying its own egress charge.

Inference

Every inference call writes logs. Every log read for compliance or analysis triggers an API fee. Production AI runs around the clock, and so does the billing. At scale, inference logging alone can generate millions of API operations per month.

Governance

Object Lock costs extra on most hyperscalers. Compliance isn't optional, so teams either pay the fee or accept the risk. Backup applications that frequently utilize Object Lock can generate significant numbers of related API operations, adding up to a price tag that can rival storage capacity costs.

© 2026 Wasabi Technologies

When every iteration of your AI model carries a storage fee, teams start rationing. Fewer training cycles. Less data retained. Tighter restrictions on inference access. That degrades model quality with every cycle that gets skipped. AI teams need to focus on AI, not the storage bill.

Organizations that remove this tax don't just spend less. They iterate more, retain more data, and build better AI models.

The hidden costs of hybrid and multicloud AI

Most organizations run more than one cloud. Flexera's 2026 State of the Cloud Report puts hybrid cloud adoption at 73%, with multi-cloud adoption continuing to rise, embracing hybrid cloud or multiple public and private clouds. As AI initiatives grow, they follow similar trajectories. AI systems co-located with related workflows reduce latency, with storage and compute scattered across platforms based on pricing, features, and other factors.

While this architecture has advantages, it contributes significantly to budget challenges. With hyperscaler API and egress fees, additional charges are levied each time you write data to storage, read data for preparation, training, inference, or auditing, export data to GPU-attached local storage for training, transfer data over the network, or enable data immutability and similar protections.

For traditional cloud applications, fees account for nearly half of cloud spend (Wasabi 2026 Global Cloud Storage Index, Vanson Bourne). With AI (and especially hybrid and multicloud AI), frequent data access can cause fees to overtake storage capacity as the primary source of cloud spending.

How the costs compare

	Hyperscaler (S3 Standard)	Wasabi Hot Cloud Storage
Storage (per TB/month)*	$23.55	$7.99
PUT fees	Yes (per 1,000 ops)	None
Retrieval fees	Yes (tiered)	None (always hot)
Immutability request charges	Additional API fees	Included
Egress fees	Yes ($0.09/GB+)	None, subject to fair use policy
Predictable monthly bill	No	Yes

**Hyperscaler figures are AWS S3 Standard list price, US East (N. Virginia), first 50 TB, as of August 2026. Storage is $0.023/GB-month, billed at 1,024 GB per TB. AWS storage rates step down above 50 TB and 500 TB, and egress rates step down above 10 TB per month. Wasabi pricing is list Pay-As-You-Go and is subject to a 1 TB monthly minimum and a 90-day minimum storage duration. Free egress and free API requests are subject to Wasabi's fair-use policies.*

© 2026 Wasabi Technologies

The AI storage cloud

The AI storage cloud is the persistent storage layer that connects every stage of your AI pipeline, from raw ingestion through governance. This lets your tools repeatedly read, write, version, and move data across environments without API or egress fees, cost penalties, or lock-in.

It's not GPU scratch space. It's not your primary database. It's the always-on backbone that makes iteration possible.

AI introduces new requirements for data storage. While traditional applications may rarely access the majority of their data, AI workflows require frequent, low-latency access to massive datasets. The AI storage cloud serves as the persistent system of record for AI datasets and artifacts across the full lifecycle, enabling teams and tools to repeatedly read, write, version, move, and retain AI data across environments without incurring unpredictable penalties, but still maintaining security, compliance, and cyber resilience.

It's defined by three properties:

01 Predictable economics

AI models are only as good as the data they can access. With PUT, retrieval, and egress fees, every time data is touched, it costs money—directly disincentivizing usage and innovation. Flat-rate pricing so every API call, every data read, every training cycle runs without triggering a surprise bill. A predictable pricing structure enables teams to budget confidently, defend proposed spending, and protect R&D funding against overruns. You know the cost before you run the pipeline.

02 Portability by default

AI data is constantly moving. At each pipeline stage, data is read for cleaning, training, or inference. At some stages, entire datasets are copied to local scratch storage for compute-heavy tasks. Accessible from on-prem, AWS, Azure, GCP, and SaaS tools so data moves freely to wherever your compute is, without paying an exit fee every time. Data portability eliminates the tradeoffs forced by vendor lock-in.

03 Trust by design

AI systems increasingly play a central role in critical business workflows, using sensitive customer and business data. Built-in immutability, AES-256 encryption, and granular identity and access controls. It's audit-ready from the first byte, not bolted on after a compliance review. AI data storage needs built-in durability and high availability, encryption and access measures, immutability to protect against unauthorized modification, and cyber resilience to manage risk and meet compliance requirements.

© 2026 Wasabi Technologies

How Wasabi sets you up to succeed

Wasabi was built around a belief that organizations would need to afford storing and accessing all of their data, at any time, without penalty. It's a belief that predates the AI era, but ironically is an approach that makes AI more viable and reliable for any organization that needs to iterate more, retain more data, and build better AI models.

Exactly what the AI storage cloud requires

Wasabi isn't the GPU scratch tier. Your compute environment handles that. Wasabi is the always-on storage layer providing a backbone for ingestion, data preparation, inference logging, and long-term governance.

Predictable, flat-rate pricing

No egress fees. No PUT, retrieval, or egress fees. No immutability surcharges. A single flat rate per TB per month. Up to 80% less than hyperscaler storage with no surprises when your pipeline runs more than you planned.

By eliminating fees that penalize data access, Wasabi enables companies to move and reuse data freely across the AI pipeline. Clear, up-front pricing enables teams to confidently budget for future iterations. Every training cycle, every data read, every compliance audit runs at the same predictable cost.

Hot, accessible storage. One tier, always on

AI systems require frequent, low-latency access to training data, model parameters, and configuration information. Everything lives in a single hot tier. No tiering decisions, no retrieval planning, no waiting for data to rehydrate. The same storage that holds your raw ingestion data holds your inference logs and compliance archives—at the same price and speed.

By avoiding cool and cold storage entirely, organizations can pursue AI initiatives without worrying about retrieval planning or slower data access.

Cyber-resilient design, built in

AI data, which includes training sets, model weights, inference logs, and configuration history, has become critical infrastructure. This makes it a prime target for cyber threat actors.

Wasabi Hot Cloud Storage is built around security by design. AES-256 encryption at rest, HTTPS in transit. Object Lock immutability included at no extra cost to protect against ransomware and unauthorized modification. Multi-User Authorization so no single compromised account can delete your AI artifacts. And Covert Copy™ technology gives you a protected copy that stays invisible to the console and the API by default.

© 2026 Wasabi Technologies

Where Wasabi fits at each stage

Stage	Wasabi's role
Ingestion	Hot, economical landing zone to centralize raw unstructured data from every source.
Cleanse and prepare	System of record for cleaned datasets and originals. Supports iteration without cost fear.
Index and retrieve	Durable storage for index snapshots and versions, enabling recovery and portability.
AI model training	Store AI training datasets, model registry artifacts, and checkpoints. Not the GPU scratch tier.
Inference	Store logs, configs, and outputs that drive evaluation, audits, and future fine-tuning.
Govern and archive	Immutable, auditable system of record for AI artifacts across time. Compliance-ready.

Next steps

Your AI pipeline is already running. The question is whether the storage layer underneath it is built for iteration or quietly taxing every cycle.

Wasabi Hot Cloud Storage is affordable storage for the Neocloud era, offering secure, predictably-priced cloud object storage for AI. Store more. Spend less. Move data freely with no fees, no lock-in, and no surprises at any scale

Calculate your savings

Use the Wasabi TCO calculator to see what you're paying in fees today and what changes with flat-rate storage.

Talk to a storage architect

Bring your current architecture. We'll show you where the AI storage cloud fits and what your pipeline looks like with the tax removed.

Get Wasabi Hot Cloud Storage starting at $7.99/TB/month

No egress fees. No API fees. No immutability surcharges. No tiering. Up to 80% less than hyperscaler storage.

© 2026 Wasabi Technologies

wasabi