SOLUTION BRIEF

Qumulo for Artificial Intelligence

Qumulo Unified Data Platform for AI/ML – Faster Cycles, Lower Costs, Anywhere

One global namespace across edge, datacenter, and cloud. Low-latency access with NFS/SMB/S3, ML caching, and end-to-end visibility to keep the AI data flywheel spinning.

The AI Imperative: Data at Scale, Instantaneously, Everywhere

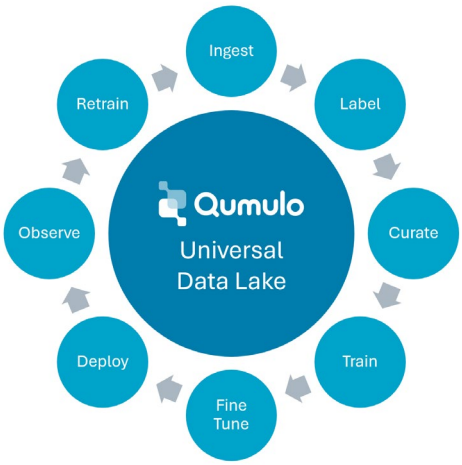
AI success depends on more than GPUs. The real bottleneck is unstructured data, comprising billions of files in diverse formats and workflows that span labs, data centers, and clouds. Copying data between storage silos stalls training, drives up egress costs, and leaves GPUs waiting.

Qumulo delivers a Unified Data Platform—a single, authoritative file/object foundation that spans edge, on-premises, and multi-cloud. With predictive caching, multiprotocol access, and global observability, Qumulo ensures the AI data flywheel—ingest → curate/label → train/fine-tune → deploy → observe → retrain—never slows down.

Value Pillars for the AI Data Flywheel

Global Data Access

- One namespace across sites and clouds.
- Geo-distributed caches warm hot shards near compute (edge ↔ core ↔ cloud).
- Keep on-premises footprints lean while enabling elastic bursts in cloud.
- Workloads move seamlessly to available GPU cycles, anywhere.



Serve Every Engine (Multiprotocol)

- Native NFS, SMB, and S3, simultaneously.
- The same dataset feeds Spark ETL, PyTorch/TensorFlow training, and RAG pipelines—no reshaping or costly copy hops.

See & Govern Everywhere (Nexus + GNS)

- Unified observability for performance, access, and spend.
- Policy, quotas, snapshots, and S3 versions tie embeddings and model checkpoints to exact dataset versions.
- Audit-ready lineage and compliance (HIPAA, GDPR, sovereign enclaves).

Perform & Scale Efficiently

- Elastic throughput over 1TB/s and IOPS over 4.7M with exabyte-scale headroom.
- WAN-aware caching trims egress and keeps GPUs fed.
- Up to 40% lower cloud file costs and >30% egress reduction with Cloud Data Fabric (CDF) caching.

A Modern File Data Experience

Qumulo Run Anywhere
Unlike legacy storage systems, Qumulo is 100% software-defined, meaning it runs consistently across on-prem, cloud, or hybrid environments—on any x86/ARM server. This lets modern enterprises avoid vendor lock-in, simplify infrastructure, and deploy anywhere while maintaining a single, unified file system.

Any Data, Any Location, Total Control
Qumulo's Cloud Data Fabric removes capacity and distance limits, making any data instantly available to any workload—anywhere—while ensuring sovereignty and security.

Intelligent Performance with NeuralCache
Qumulo NeuralCache uses machine learning to optimize reads and writes, improving performance and drastically reducing cloud I/O costs by up to 99%

Cloud-Native, But Not Cloud-Limited
Qumulo is cloud-native by design, using object storage with real-time file semantics to enable secure, predictable scaling. It eliminates the need for data replication and app rewrites, ensuring cost control and compliance with data sovereignty requirements.

Customer Success
Qumulo zero-latency support enables fast issue resolution through direct access to experienced Customer Success Managers or file system engineers—no tickets required. Qumulo support has earned an NPS score of 95.

Qumulo

 | info@qumulo.com

SOLUTION BRIEF

Qumulo for Artificial Intelligence

Before and After: The AI Data Flywheel

Before Qumulo:

- Multiple copy hops.
- Idle GPUs and unpredictable egress.
- Version drift across hybrid and multicloud.
- Inconsistent access across global sites.

With Qumulo CDF:

- One namespace spanning NFS/SMB/S3.
- Caches at edge, core, and cloud.
- Snapshots + versions tie datasets to lineage.
- Nexus insights provide real-time visibility.

Agility and Flexibility

With Qumulo, the flywheel never has to slow down because your lakehouse/lake can have Dataset A mainly supporting workload A on day one in Virginia, and then shift the workload to Virginia plus cloud, and then to cloud only, and finally to London, all without data copies or waiting for file transfers. We support a protocol-, site-, and workload-agnostic data lake, where each of these variables can change based on factors such as people, costs, resources, and availability targets.

Lakehouse and RAG Alignment

Most enterprises are building lakehouses on Delta or Iceberg for structured data. Qumulo complements them as the universal unstructured data lake for AI training, fine-tuning, and retrieval-augmented generation (RAG).

- Mount via NFS for high-throughput epochs.
- Ingest/checkpoints over S3—same path, no reshaping.
- Spark, PyTorch, and TensorFlow all consume the same files.
- Snapshots pin RAG context sets to exact versions.
- Datalake spanning multiple regions, clouds, and hybrid

The result: one foundation for structured and unstructured AI, without creating new silos.

Outcome-First Benefits

- **Faster time to data:** Hot shards pre-staged near GPUs cut queuing.
- **Copy-less pipelines:** Train, transform, and vectorize from one source that is available everywhere.
- **Right-sized spend:** WAN-smart caching trims cloud egress and allows for compute arbitrage
- **Keep GPUs fed:** Predictive warming minimizes idle accelerators.
- **Deterministic re-runs:** Snapshots and versions guarantee repeatable results.
- **Portability by default:** Same ACLs and paths across environments.
- **Flexible economics:** Buy capacity on-premises or pay for consumption in the cloud or both.

Proof Points and Benchmarks

Independent SPECstorage benchmarks validated Qumulo's AI_Image workload performance:

- **0.84 ms response time** scaling to 700 concurrent jobs.
- **Up to 40% faster epochs** with fewer stalls.
- **Lower idle GPU time** by as much as 40%.

These results mean steadier throughput, fewer retraining hiccups, and lower wall-clock times. For buyers, it's tangible proof that Qumulo turns storage from a bottleneck into a force multiplier.

Why Now?

AI adoption is surging, but infrastructure is lagging. GPUs are scarce and expensive, cloud egress can overwhelm budgets, and compliance requires strict lineage. Qumulo delivers measurable relief:

- Up to **40% lower file storage costs**.
- Over **30% egress reduction** via WAN-smart caching.
- Up to **4× fewer data-movement steps** across pipelines.

In an era of AI arms races, these advantages are not just efficiency wins—they are competitive edges.

Conclusion

The future of AI is universal, multiprotocol, and data-first. Qumulo's Unified Data Platform powers the AI flywheel with one namespace, predictive caching, and enterprise-grade governance. Whether you're training vision models at the edge, running simulations on superpods, or bursting RAG workloads into the cloud, Qumulo keeps data moving at GPU speed, without runaway costs.

About Qumulo

Qumulo is the leading provider of cloud file data platforms, offering unrivaled performance, scale, and data management solutions. Qumulo's platform is trusted by Fortune 500 companies and global enterprises to manage petabytes of data, enabling them to unlock the value of their data and drive innovation. For more information, visit www.qumulo.com

Qumulo

 | info@qumulo.com