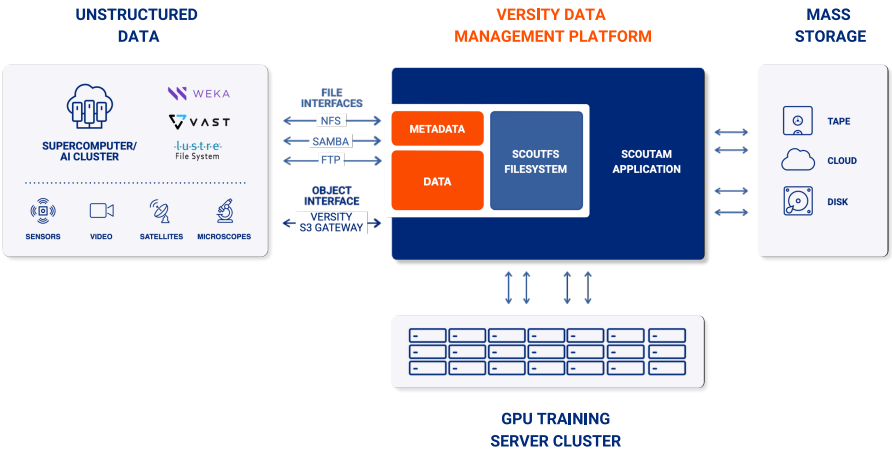
Optimizing Deep Learning AI for Science with Versity

Deep-learning AI workflows generate massive datasets that demand fast, affordable, and accessible storage. Versity's integrated stack, featuring ScoutAM, ScoutFS, and the Versity S3 Gateway, streamlines the entire AI lifecycle with exabyte-scale management, hybrid cloud compatibility, and intelligent tiering that lowers costs without disrupting existing workflows.

Comprehensive Data Management for Deep Learning AI

Scientific deep-learning AI workflows generate vast amounts of data, requiring robust storage and archival solutions that balance performance, cost, and accessibility. Versity's integrated data management stack provides a comprehensive solution to this challenge, combining the power of ScoutAM for exabyte-scale data management, ScoutFS for high-performance file operations, and the Versity S3 Gateway for seamless S3 protocol integration. This ecosystem enables organizations to efficiently manage the entire deep learning lifecycle—from initial data acquisition and preprocessing to model training, evaluation, and deployment—while cutting storage costs through intelligent tiering and maintaining compatibility with existing workflows through hybrid cloud capabilities.



Data Ingestion & Preprocessing

ScoutAM is designed to handle exabyte-sized collections of scientific data at scale with efficiency and reliability. At its core, ScoutAM integrates tightly with **ScoutFS**, a high-performance file system that accelerates data capture with:

- High-throughput parallel I/O for ingesting massive datasets quickly
- POSIX-compliant interface for seamless integration with existing applications
- Built-in checksumming to ensure data integrity from the start
- Automated tiering and lifecycle management through direct integration with ScoutAM

The **Versity S3 Gateway** extends ScoutAM's reach by bridging file-based and S3 object storage systems, enabling hybrid workflows and seamless ingestion of S3 data. With this, ScoutAM delivers true hybrid cloud capability, spanning on-premises S3 object storage and public cloud providers such as AWS, Azure, and Google Cloud. For long-term retention, ScoutAM automatically tiers both object and file data to tape, providing cost-efficient archival and long-term durability without requiring specialized tape expertise.

Model Training

Training deep learning models requires fast, reliable access to massive datasets across compute clusters. ScoutFS and ScoutAM together ensure performance and resiliency during this critical stage:

- **High-speed access to hot datasets** through ScoutFS eliminates bottlenecks
- **Automatic staging of active data into performance tiers** ensures efficient GPU utilization
- Efficient **checkpointing** protects progress, reducing downtime in case of failures or interruptions
- **Version control integration** tracks both code and datasets for reproducibility and collaboration across teams
- **Parallel transfer capabilities** in ScoutAM deliver high-throughput movement of training data across GPU nodes, ensuring models aren't slowed down by storage

Training Data Management

Once datasets are ingested, effective management ensures that training clusters always have access to the right data without unnecessary overhead:

- **Active datasets** remain on ScoutFS, ensuring maximum performance for compute clusters
- ScoutAM's **policy-based tiering** moves less frequently accessed data to lower-cost media, such as tape, freeing premium storage for current workloads
- Metadata is retained within ScoutFS, making searches and queries across the entire dataset fast and efficient
- Versity S3 Gateway provides **S3-compatible access to datasets**, enabling distributed GPU clusters to train directly on **unified namespaces** that combine both object and file data

Results Management

Once training and evaluation are complete, results must be preserved and made accessible for future use. Versity simplifies this process through automated archival and intelligent lifecycle management:

- **Training logs, validation results, trained models, and metadata** are automatically archived for long-term retention

- ScoutAM's tiered storage policies optimize cost efficiency by moving inactive results to lower-cost storage without impacting accessibility
- **Built-in verification and checksumming** maintain data integrity over time
- **Multiple copy support** ensures redundancy and reliability for decades of research outcomes
- Versity S3 Gateway allows production applications to interact with historical or real-time data through familiar S3 commands, simplifying deployment and re-use of trained models

This ensures that results remain both durable and accessible, supporting iterative research cycles and enabling new discoveries built on prior work.

Seamless Deployment and Scalable Archival

Versity unifies high-performance data ingest, scalable training data management, efficient model training, and secure results archival into one platform. At scale, ScoutAM provides exabyte-level management with hybrid cloud flexibility, intelligent tiering, and protocol compatibility to accelerate discovery and reduce infrastructure costs while ensuring long-term control of critical data.

Deploying AI models demands reliable access to both current and historical datasets. ScoutAM delivers this with intelligent, policy-driven automation that tiers data by usage, lowering costs while preserving high performance. With the Versity Gateway extending seamless S3 access across storage tiers, ScoutAM enables scalable, future-ready deployments that keep infrastructure efficient and workflows uninterrupted.

About Us

Versity is a leading independent technology company focused on delivering innovative mass storage solutions. With a commitment to performance, scalability, and open-source collaboration, Versity empowers organizations to efficiently manage their data in a rapidly evolving digital landscape.

Worldwide consumer base managing over **4 exabytes** of data across 95 sites

Accelerated innovation driven by customer-focused feature requests and a strong, next-generation engineering team

Unrestricted access to expert guidance and **personalized support** from our highly-skilled engineering team

Key Benefits

Scalability & Performance

- Modular architecture allows easy expansion of storage capacity as data volumes increase.
- **Parallel data transfers** and **enhanced caching** keep workflows fast
- **Policy-driven tiering** ensures the right data is always on the right tier
- **Versity S3 Gateway** bridges S3 object protocols with file-based systems for seamless integration

Cost Efficiency

- Intelligent tiering moves inactive data to low-cost tape storage
- Up to **10x lower TCO compared to Amazon Glacier**
- **Eliminates unpredictable egress and transaction fees**
- Cost-effective archival of completed training runs and historical datasets
- **Optimized resource allocation** ensures high-performance tiers are only used when needed, lowering TCO

Sustainability & Efficiency

- **Archival tape** uses almost no power when idle, ideal for large AI datasets
- **Low-power object storage tiers** reduce energy demand compared to Tier 0 disk
- **Intelligent caching** keeps hot data on high-performance media, moves cold data to efficient tiers
- **Lower carbon footprint** through adoption of tape and emerging sustainable storage technologies

Trusted by Industry Leaders

Versity

CONTACT
sales@versity.com
1-844-726-8826