

The AD-ADAS Data Platform

Accelerate autonomous development and turn petabytes of sensor data into trained models faster.

Vehicles are shifting from software-defined to AI-defined, and nowhere is that shift more prominent than in autonomous driving (AD) and advanced driver assistance systems (ADAS). While automakers and suppliers are investing in GPU-powered AI factories in their race to deliver the most advanced systems, the constraint on progress is rarely compute capacity. The data platform is key to keeping those expensive GPU clusters saturated, turning petabytes of sensor data into training-ready datasets faster while governing and protecting safety-critical data across regions. This brief explains how AD-ADAS platform leaders can equip their engineering and infrastructure teams to remove the data bottleneck at each stage of the pipeline so GPUs stay productive, validation cycles shorten, and time to market for differentiated services improves.

The bottleneck is data, not compute

[McKinsey](#) projects that vehicles equipped with AD and ADAS could account for nearly 70% of vehicle sales by 2035, with the AI-addressable share of the AD-ADAS software market climbing from 79% to 95%. This makes AI-enabled features the fastest-growing, highest-value frontier in the vehicle.

That concentration creates a data problem. Modern Level 3+ vehicles stream camera, light detection and ranging (LiDAR), and other sensor data that can exceed 4TB per day,¹ and entire test fleets can generate hundreds of terabytes. This data must be prepared to extract the full value out of it. Preparation includes data ingestion, anonymization, cataloging, indexing, curating, and labeling, and all must happen before data is fed to GPU clusters for training. When the data platform can't keep pace with business requirements, GPUs sit idle waiting on I/O and project milestones slip. The organizations that will win the autonomous future are the ones that can turn the largest, best-organized data estates into validated, deployable models the fastest. Yet many AD-ADAS leaders struggle to pinpoint why programs underperform despite heavy compute investment. The answer usually lies in fragmented, siloed data estates that leave GPUs underutilized and slow engineering efforts in general.

As automakers and suppliers push toward L3+ autonomy, data pipeline speed directly gates how fast they can train, validate, and deploy new autonomous features. Unlike regulatory timelines or silicon supply, it is a lever they fully control through the underlying infrastructure decisions they make.



Faster time to validation

Move from raw sensor data to training-ready in hours, not weeks.



Higher GPU ROI

Keep expensive clusters saturated, not idle.



Resilient by design

Get immutable, rapidly recoverable, safety-critical data; grow from terabytes to exabytes with data in place.

Why programs stall despite compute investment

Sensor and simulation workloads generate billions of small files that overwhelm legacy scale-up storage and siloed on-premises or cloud object stores. Where GPU starvation is a throughput problem, this is a file-count problem: the platform runs out of metadata performance long before it runs out of capacity or bandwidth. The cost: metadata operations, not storage space, become the hidden ceiling on how fast you can train.

- **Slow raw ingest-to-training-ready cycles.** The data preparation pipeline often stretches to weeks when the data pipeline is stitched together using multiple disparate storage environments from multiple vendors. The end result is significant waste in the pipeline as measured by avoidable copy sprawl and non-value-added movement of the data, resulting in higher program costs and inefficient use of human and physical capital.
- **Management complexity.** Because the same datasets are duplicated across the data pipeline, significant complexity and costs are incurred to manage the underlying infrastructure from a provisioning and compliance perspective. Additionally, engineers often waste valuable time searching for the authoritative version of a given dataset due to the overly complex data management environment.
- **GPU underutilization.** Legacy scale-up storage and fragmented data silos can't sustain the throughput and metadata performance an AI factory demands, so multimillion-dollar clusters run well below capacity. The business impact is lower ROI on the capital investment and extended time to revenue for new end-customer-facing features and services.
- **Cyber and sovereignty exposure.** Automotive is one of the most targeted industries from a cybersecurity perspective, and AD-ADAS datasets are often irreplaceable. Regional regulations add a further constraint, dictating where that data can live and how long it should be retained. Fragmented and siloed data management infrastructure results in extreme difficulty and high cost as it relates to protecting enterprise data and maintaining legal compliance.

The AD-ADAS pipeline, and where Everpure removes the bottleneck

The AD-ADAS development loop runs from fleet data collection to a validated model redeployed to the fleet, with continuous analytics and insight and an AIOps and CI/CD feedback loop running across every stage. Everpure™ provides a common data management platform for all data types (file, object, and block) and all storage tiers, spanning each step of the AD-ADAS data pipeline. This architecture solves the challenges described in this brief by:

- Significantly reducing copy sprawl.
- Allowing multiple AD-ADAS workflows to operate against a common governed dataset.
- Reducing infrastructure management complexity through the [Everpure Intelligent Control Plane](#).
- [Keeping your GPUs fed](#) through Everpure DirectFlash® technology.

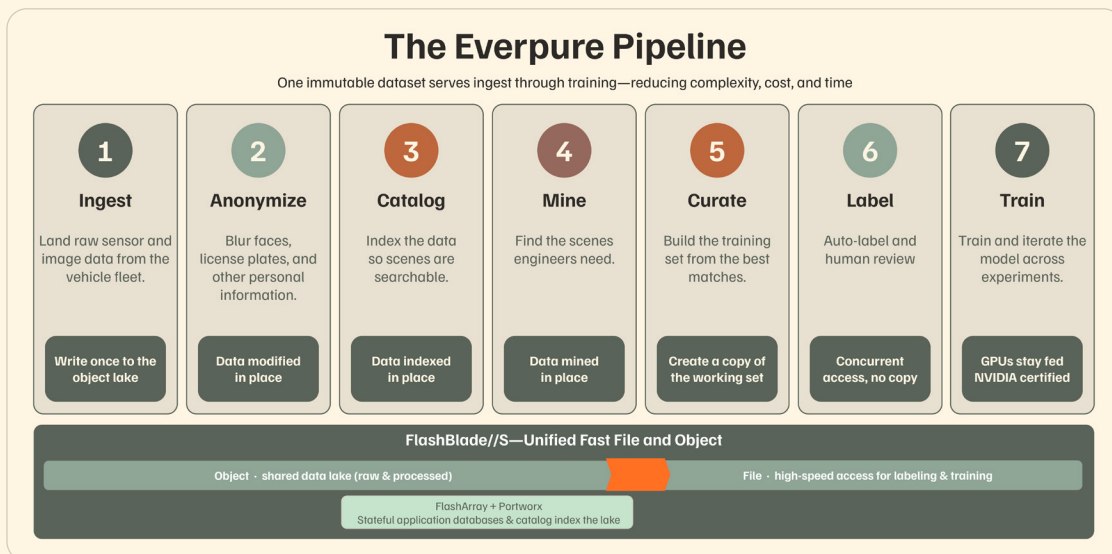


FIGURE 1 The AD-ADAS pipeline, where data meets GPU-scale AI

Ingestion and anonymization

The pain: Raw data coming off of test vehicles often requires a network-attached storage (NAS) landing tier because legacy hard disk drive (HDD) data lakes can't absorb fleet-rate offload, which is bursty. This results in an unnecessary copy to the landing NAS and then again to the HDD data lake. Once in the HDD data lake, the data must be anonymized before it can be further processed downstream. On an HDD data lake, the storage simply cannot supply frames fast enough to keep the GPUs busy, so the data corpus is staged yet again to a flash scratch tier, transformed there, and then staged back to the HDD data lake.

The Everpure Platform: [FlashBlade//S](#)™ absorbs sustained, multi-gigabyte-per-second writes straight into the lake, so there is no separate landing tier and no first copy. In addition, FlashBlade//S is able to read and anonymize the data corpus in place, creating a new object version using Object Lock to preserve the ground truth as it relates to the original data corpus.

Catalog and mine

The pain: Cataloging and indexing the data is metadata-compute-intensive. On spinning media, each operation is a physical seek pushing the limits of its performance. To compensate for this performance limitation, engineers often implement “shallower” indexing, which impacts their ability to find the scenes of interest downstream in the pipeline. The net result is extended time spent on these steps of the pipeline and suboptimal training datasets downstream, which directly impacts the end result of the final product.

The Everpure Platform: FlashBlade//S sustains the metadata performance of billions of files through a distributed metadata architecture, with tens of millions of file operations per second. FlashBlade//S provides the high-throughput data layer for the implemented cataloging and mining steps, allowing those tools to operate on large datasets. Due to the high performance of Everpure all-flash-based storage systems, engineers are not forced to compromise indexing quality for overall system performance. [Portworx](#)®, in combination with [FlashArray](#)™, provides persistent, high-performance storage for containerized data prep and MLOps pipelines.

Curate and label

The pain: A manifest is the output of the mining step. This subset of the data must be copied from object to file, which is a requirement for the subsequent labeling and training steps. For the labeling step, an autolabeler is typically used for the first pass of the labeling, which is then verified by dozens of human annotators. Multiple concurrent reads and writes to the data corpus often results in annotators waiting for frames to load. This latency converts directly into higher labor costs and processing time, which gets exacerbated at scale.

The Everpure Platform: FlashBlade//S provides a strong storage foundation for dozens of concurrent annotators by combining distributed metadata processing, high-performance flash, and parallel file access. In [published testing](#), FlashBlade//S500 R2 delivered 7.2 million aggregate metadata IOPS and more than 200GiB/s of aggregate read throughput in a multi-client IO500 configuration.

Train

The pain: Curated frames and labels are packed into training shards, then fed to the GPUs. Each training experiment needs a frozen dataset that can be reproduced in the future in accordance with regulatory requirements. Without an automatable, space-efficient clone, that means a full physical copy per run, which results in excess time and inefficient use of storage infrastructure. Once actual training commences, GPU clusters often wait on data during the training cycle because the storage infrastructure simply cannot keep up. Lastly, some storage vendors require proprietary drivers to be installed on the GPU clusters, which makes maintaining and patching the infrastructure more complicated than it needs to be.

The Everpure Platform: FlashBlade//S delivers the massively parallel throughput to keep GPU clusters saturated—lifting utilization to as high as [85–98%](#)—and sustains the high-concurrency reads that replay and simulation demand, all from one governed data plane shared by training and validation. FlashArray supports the structured and transactional workloads in the loop, and Portworx orchestrates persistent storage for Kubernetes-based training and validation. In addition, the Everpure Platform supports cloning that eliminates the excess time and infrastructure capacity typically required in siloed solutions. Lastly, no third-party drivers from Everpure are required to unlock performance in the data pipeline, significantly reducing complexity for AI factory infrastructure owners.

Why Everpure for AD-ADAS

Accelerate AI and engineering velocity. Keep GPUs fed for faster time to insight and model iteration; go from raw drive data to training-ready in hours, not weeks. FlashBlade//S delivers [up to 20–25% better multidimensional performance](#) than competitive platforms for AI training, inference, and simulation workloads.

Unify and govern data. Everpure provides [one governed data plane and intelligent control plane](#) across product engineering, AD-ADAS, manufacturing plants, and enterprise applications, with the lineage, auditability, and residency that autonomy and regulation demand. Capacity and performance scale independently, with nondisruptive, data-in-place upgrades.

Cyber resilience and recovery. Immutable, tamper-resistant [SafeMode™](#) Snapshots that survive credential theft protect irreplaceable training datasets with rapid, tested recovery in one of the most attacked industries.

Sustainable scale and efficiency. High-density all-flash and Evergreen® upgrades scale from terabytes to exabytes with [up to 85% less energy](#) versus competitive all-flash storage. Run the right workloads in the right place—on-premises hot data for sustained training and validation and cloud for burst and aggregation—avoiding the egress and capacity costs that inflate cloud bills at petabyte scale.

Built with NVIDIA, validated at scale

Everpure is an active [NVIDIA partner](#). For teams starting an AI practice, the AIRI//S™ solution—a converged system built to the NVIDIA DGX BasePOD reference architecture—is a validated reference architecture for AI, analytics, and insight workloads. For AD-ADAS training and validation at scale, FlashBlade//S is a certified storage design for NVIDIA DGX SuperPOD, scaling to thousands of GPUs and multiple petabytes, with data and compute scaling independently rather than in a fixed ratio. [FlashStack®](#) goes a step further into a full-stack converged infrastructure, integrating Everpure storage with Cisco UCS compute and Nexus networking under Cisco Validated Designs for teams that want a single, turnkey stack to stand up their AI factory faster.

The bottom line

If model iteration is slow, GPUs are idle, or critical business data is hard to govern and protect, the constraint isn't the GPUs or the models; it's the data platform beneath them. Everpure turns that foundation from a bottleneck into a competitive advantage: higher GPU ROI, faster time to market, lower data management complexity, stronger cyber resilience and sovereignty, and lower total cost across on-premises and cloud.

Additional resources

- Explore [Everpure data solutions for the automotive industry](#).
- Read about [high-performance AI for automakers with AIRI//S](#).
- Discover [Everpure solutions for automakers](#).

¹ | Mordor Intelligence, Automotive Data Management Market, 2026

Visit Our Website

800.379.PURE

