


# Research Report:

## Solving AI Data Pipeline Inefficiencies

Steve McDowell, Chief Analyst

July 2024

This report was commissioned by **VAST Data**

# SOLVING AI DATA PIPELINE INEFFICIENCIES

The AI pipeline places unique demands on the underlying data infrastructure, introducing challenges that legacy enterprise data platforms, or even those originally designed for high-performance computing (HPC) applications, struggle to address.

Optimizing the data infrastructure to support the full AI lifecycle saves costs, leading to faster time-to-insight from AI workflows and maximizing an enterprise's return on investment for its AI infrastructure. Any organization implementing AI infrastructure, whether an enterprise or service provider, should explore how best to address these challenges early on.

## THE AI PIPELINE

The AI pipeline involves various stages, from data collection and preprocessing to model training, validation, and deployment. Each stage is data-intensive, necessitating robust storage solutions and efficient data handling mechanisms to facilitate rapid processing and insights.

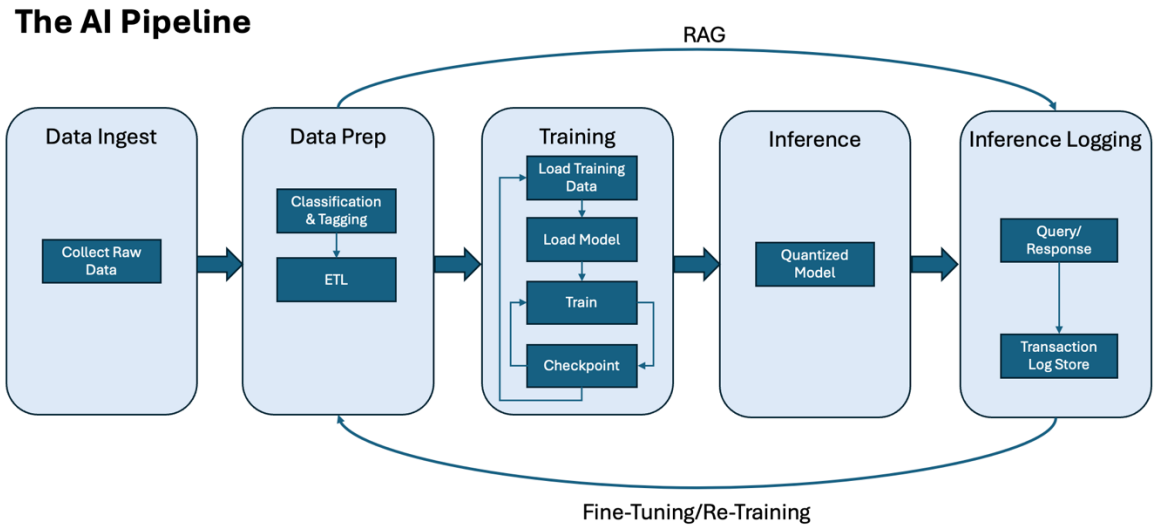



FIGURE 1: THE AI PIPELINE

Every stage in the AI pipeline involves specific tools, techniques, and methodologies. It often requires collaboration across multiple roles and disciplines, including data engineers, data scientists, AI researchers, and domain experts. The pipeline is not strictly linear and can involve

iterations and feedback loops, especially during the modeling, evaluation, and deployment phases.

## INEFFICIENCIES OF TRADITIONAL DATA INFRASTRUCTURE

While crucial for developing robust and accurate models, the iterative nature of the AI pipeline uncovers inefficiencies within traditional storage and data management approaches. These inefficiencies arise from how these systems manage, store, and process across the various stages of the pipeline. Failing to address them can hinder performance and increase costs.

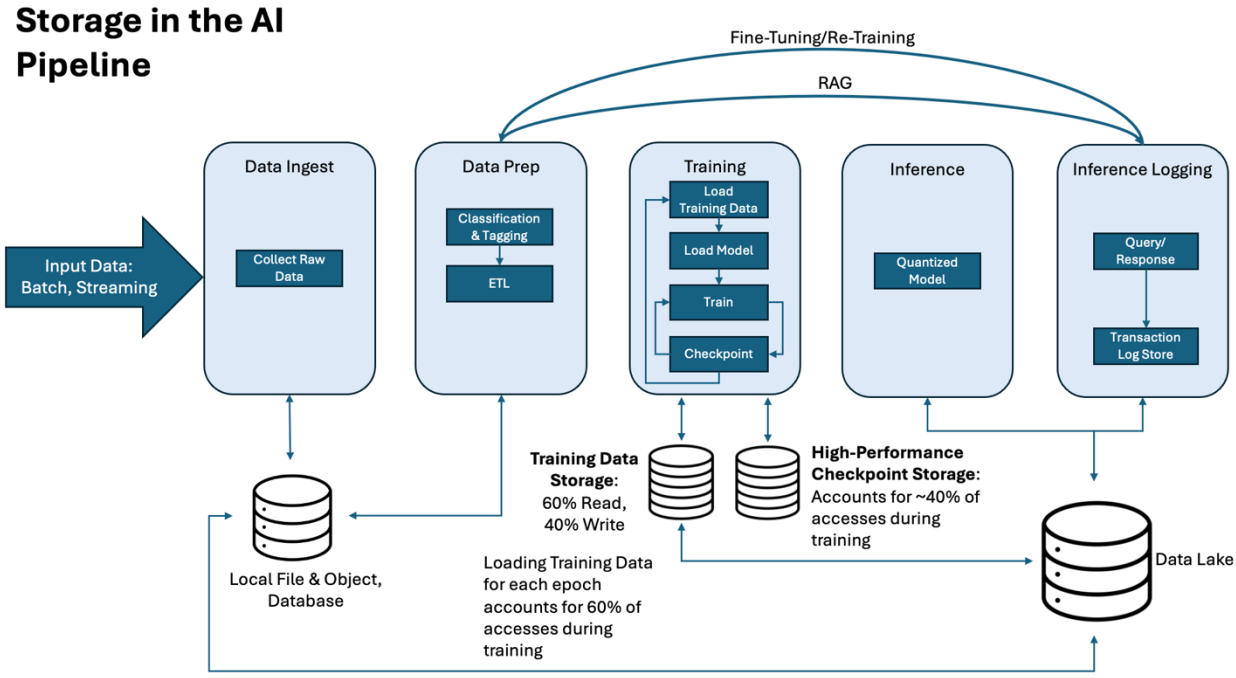



FIGURE 2: TRADITIONAL APPROACH IS COMPLEX AND INEFFICIENT

There are a set of issues common to multiple phases of the AI pipeline. These are described in Table 1 as follows:

TABLE 1: INEFFICIENCIES IN TRADITIONAL DATA INFRASTRUCTURE

| | Inefficiencies |
|---|---|
| Data Duplication and Movement | **Unnecessary Copying**: Each iteration in the training pipeline often requires copying data between different environments or tools (e.g., from storage to preprocessing servers to training environments). This movement increases infrastructure costs and can slow down the entire process due to data transfer latency.<br><br>**Multiple Versions**: Iterative improvements and experiments often lead to systems storing multiple versions of datasets, which consumes significant storage resources. |
| Latency | **High Latency**: The need to repeatedly access large datasets for training and validation introduces latency. If the infrastructure is not optimized for fast retrieval, this can significantly slow down the training iterations.<br><br>**Network Bottlenecks**: Moving large datasets over networks, especially in or between cloud environments, can create bottlenecks, further exacerbating latency issues.<br><br>**Data Retrieval Latency**: High latency in retrieving data from storage can slow down the training and inference phases, making iterative development slower and less efficient. |
| Infrastructure Utilization | **Idle Resources**: The iterative nature of AI training means that computational resources (e.g., GPUs, CPUs) often wait for data movement or processing, leading to periods of underutilization and inefficiency.<br><br>**Non-Optimized Storage Utilization**: High-value infrastructure, such as fast SSD storage, might be under-utilized due to inefficient data handling and unnecessary data movement, leading to suboptimal use of available resources. |
| Scalability | **Scalability Limitations**: Legacy infrastructure systems often struggle to scale efficiently with the iterative demands of AI workflows, leading to performance bottlenecks as data volumes and computational demands grow. |
| Complex Data Management | **Managing Multiple Systems**: Each element of the AI pipeline involves moving data between different systems for storage, processing, and analysis, which creates complexity and increases the risk of errors. |

| | |
|---|---|
| | **Fragmented Data**: Data silos and fragmentation across multiple storage solutions can complicate data management, making it difficult to ensure data consistency and integrity.<br><br>**Security Concerns**: Moving data between multiple storage systems often introduces gaps or inaccuracies in audit logs that track how data is handled. |
| Data Processing | **Inadequate In-Database Processing**: Traditional databases often lack the capabilities to perform the comprehensive data processing required for AI directly within the database. This necessitates moving data to separate processing environments, adding overhead and complexity.<br><br>**Redundant Processing Steps**: Iterative processes might involve repeated data processing steps (e.g., normalization, feature engineering) for each cycle, leading to inefficiencies if the system does not support incremental updates or caching. |
| Multi-Tenancy | **Resource Contention**: In environments where multiple teams or projects share resources, iterative AI processes can lead to resource contention, causing delays and inefficiencies.<br><br>**Bolted-on Multi-Tenancy:** Legacy infrastructure with bolted-on multi-tenancy support often fails to provide the necessary isolation and performance for iterative AI processes, leading to inefficiencies and potential data security issues. |

## OPTIMIZING THE AI PIPELINE WITH VAST DATA

Addressing these inefficiencies requires a unified and optimized data infrastructure that can seamlessly handle the iterative nature of AI development, minimize data movement, reduce latency, and ensure efficient utilization of computational and storage resources.

It's important to provide a quick overview of the VAST Data Platform before discussing how it overcomes the challenges inherent in managing the data path for AI.

## WHAT IS THE VAST DATA PLATFORM?

The VAST Data Platform[1] is a revolutionary, next generation data storage, database, and management solution from VAST Data that caters to the needs of modern data-intensive applications.

The platform combines a multi-protocol (file and object) unstructured data store that delivers high performance, scalability, and efficiency with an integrated database and an event-driven data engine to provide data processing functionality extending far beyond legacy data platforms or storage systems.

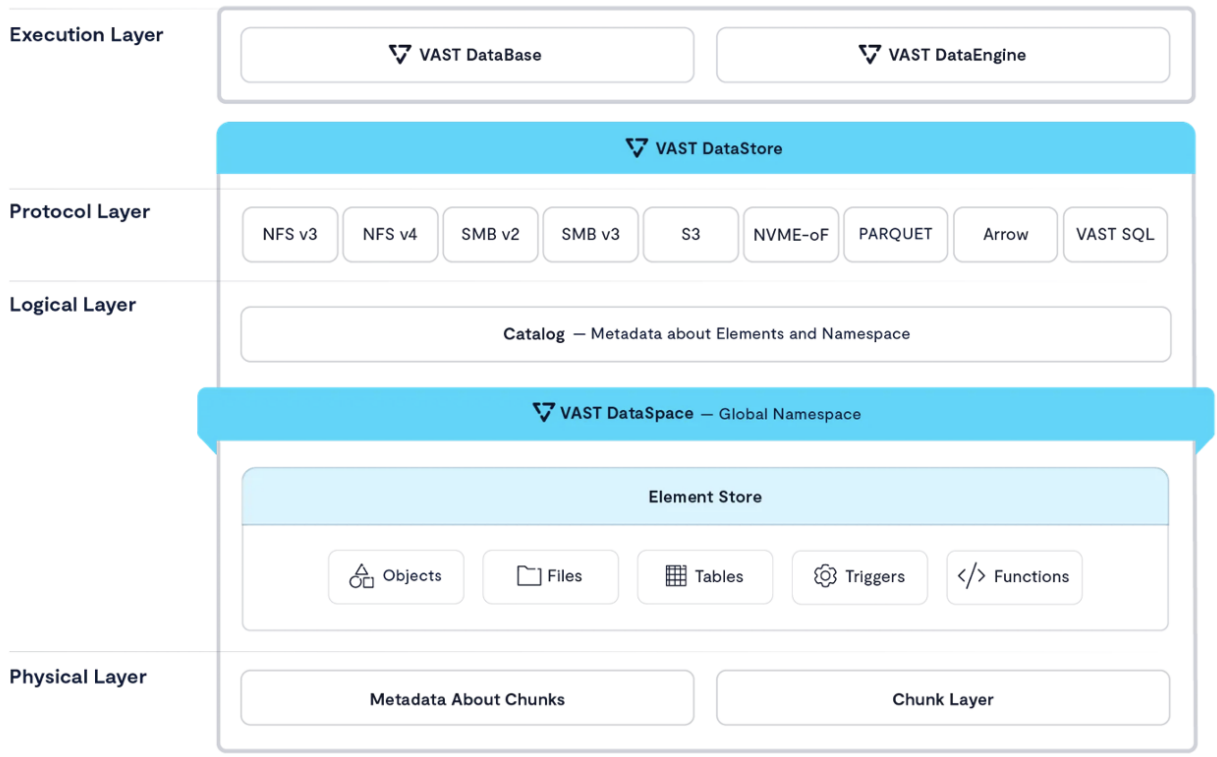



FIGURE 3: THE VAST DATA PLATFORM (SOURCE: VAST DATA)

[1] VAST Data, *The VAST Data Platform*, 2024, https://www.vastdata.com/vast-data-platform-explained

The key features of the VAST Data Platform include:

1. **Universal Storage**: The VAST Data Platform offers a multi-protocol file and object content repository that combines the attributes of all-flash performance with cost-efficiency. This makes it capable of supporting diverse workloads such as AI, big data analytics, and cloud-native applications. It does this while scaling linearly up to exabytes of capacity across billions of files and objects. VAST Data ensures consistent performance, bringing structure to unstructured data, even as data volumes grow with its disaggregated, shared-everything architecture.
2. **Disaggregated, Shared-Everything (DASE) Architecture**: The platform's architecture separates compute and storage resources, allowing each to scale independently. This optimizes resource utilization, provides flexibility in managing storage and compute needs, and provides direct access from every compute node to every SSD, delivering the ultimate parallel IO performance while eliminating the performance penalty of metadata crosstalk on the network as the system scales out.
3. **Global Namespace**: The VAST Data Platform provides a unified view of data across geographically distributed clusters, simplifying data management and access to facilitate handling large-scale data environments.
4. **Advanced Data Reduction Technologies**: The platform employs sophisticated data deduplication and compression techniques and a new data reduction method, similarity reduction, to minimize storage footprint and reduce overall storage costs.
5. **High Availability and Resilience**: VAST Data ensures data integrity and availability through robust fault-tolerance mechanisms and the DASE architecture, which provides seamless failover and disaster recovery capabilities.
6. **Advanced Database Capabilities**: The VAST Data Platform has a built-in advanced database that allows data processing within the platform. The database's capabilities include:
    - **Columnar Data Storage**: The VAST DataStore stores tables into reduced data chunks with a columnar structure, similar to how Parquet files are organized but at a much finer granularity. Because data is stored in its natural form as a table, users are relieved of the tasks of partitioning their tables across files of the right size to maximize large sequential reads. This relieves data scientists from planning table layouts to accommodate storage limitations.
    - **High-Performance SQL Queries**: The platform supports high-performance SQL queries, enabling fast and efficient data retrieval and analysis.

- **ACID-Compliant Transactions**: The VAST Data Platform provides robust transactional support, ensuring full ACID (atomicity, consistency, isolation, and durability) capabilities that provide data consistency and integrity for critical business applications.
- **Multi-Model Database Capabilities**: The platform handles various data models, including relational, key-value, document, and graph databases, providing flexibility in managing different data types.
- **Scalable and Performant Metadata Handling**: The platform includes a data catalog with a built-in metadata index. The catalog enables high-performance, scalable indexing and searches across billions of files and objects, eliminating the need to perform inefficient operations that walk filesystems. The catalog can be managed and queried through a web user interface, API, or an external query tool such as Apache Spark.
- **In-situ Processing**: Data-intensive processing occurs directly on the platform without the burden of external compute resources, removing the inefficiencies of moving data for processing.
- **Functions**: The VAST Data Platform supports programmable procedural functions. These functions are stateful and integrated directly into the platform, allowing seamless data processing.
- **Triggers**: Triggers in the VAST Data Platform allow for real-time event-driven processing. When data is added, modified, or removed, triggers can initiate functions that act on this data immediately. This capability is essential for perpetual AI training, real-time analytics, and other dynamic data processing needs.
- **Event Streaming and Real-Time Data Processing**: The platform includes a Spark-compatible, event-driven data engine that supports real-time analytics and processing. This allows users to stream multiple data types directly to the platform and store them in the most appropriate format.

While the VAST Data Platform shares many characteristics expected from a scalable high-performance storage platform, it goes much further. It integrates advanced database capabilities and real-time analytics functions to provide a solution that excels at addressing the challenges inherent in traditional storage and data solutions for AI.

## OVERCOMING INEFFICIENCIES IN THE AI PIPELINE

Unlike traditional storage solutions or even parallel file systems, the VAST Data Platform takes an integrated approach to data storage and management. This allows the platform to overcome many of the inefficiencies inherent in legacy approaches.

Here's a look at how the VAST Data approach mitigates these inefficiencies:

TABLE 2: HOW THE VAST DATA PLATFORM OVERCOMES TRADITIONAL STORAGE INEFFICIENCIES

| | |
|---|---|
| Data Duplication and Movement | **Unified Storage Architecture**: VAST Data's platform consolidates data storage, eliminating the need for multiple copies of datasets across different environments. This reduces redundancy and storage costs while enhancing data accessibility.<br><br>**Zero Data Movement**: By enabling operations directly on the data where it resides, VAST Data minimizes the need for data transfers between different pipeline stages, significantly reduces latency, and speeds up the iterative process. This is especially critical when transitioning between the data prep and model training phases. At the same time, traditional approaches require moving the prepared, tokenized data set used for training from enterprise systems to single-purpose parallel training systems; the unique VAST Data design allows you to support both phases. This allows users to immediately run training jobs in place against the prepped data, eliminating the cost (CAPEX and OPEX), complexity (M&O), and significant data transfer time associated with using discreet storage systems for each phase. |
| Latency | **High-Performance Storage:** VAST Data leverages NVMe-based storage, which provides high throughput and low latency for data access, minimizes data retrieval times, and accelerates the training and validation phases.<br><br>**GPUDirect Technology:** Integrating with NVIDIA GPUDirect®, VAST Data allows GPUs to access data directly from storage without intermediate CPU involvement, reducing data access latency and enhancing training efficiency.<br><br>**Low-Latency Access**: VAST Data's platform's NVMe-based storage and optimized data paths ensure low-latency data access, which is critical for maintaining high performance in iterative AI processes.<br><br>**Consistent Performance**: VAST Data's architecture delivers consistent performance even as data volumes grow and iterations increase, preventing performance degradation over time and ensuring reliable operation. |
| Infrastructure Utilization | **Efficient Resource Utilization**: By reducing idle times and enabling faster data access, VAST Data allows data-intensive workloads to utilize computational |

| | |
|---|---|
| | resources such as GPUs and CPUs more effectively, improving overall system performance.<br><br>**Adaptive Scalability**: VAST Data's architecture scales seamlessly with growing data volumes and computational demands, ensuring optimal utilization of infrastructure resources. |
| Scalability | **Global Namespace:** VAST Data provides a global namespace that simplifies data management across a geographically distributed environment, enabling seamless scalability and efficient data access across multiple clusters.<br><br>**Distributed File and Object Storage**: VAST Data's platform supports high-performance distributed file and object storage, which can handle the scalability requirements of large-scale AI workloads without performance bottlenecks. |
| Complex Data Management | **Integrated Data Management**: VAST Data's unified platform integrates storage, data processing, and management capabilities, reducing the complexity of managing multiple systems. This streamlines operations and reduces the risk of errors.<br><br>**Data Consolidation:** By eliminating data silos and providing a single source of truth, VAST Data simplifies data management, ensuring consistency and integrity across the AI pipeline. |
| Data Processing | **In-Database Processing**: VAST Data supports in-database processing capabilities, allowing data preprocessing and transformation operations to be performed directly within the storage system. This reduces the need to move data to separate processing environments and speeds up data preparation steps.<br><br>**Real-time Data Lakehouse**: VAST Data offers a real-time environment that supports batch and streaming data processing, enhancing the efficiency and flexibility of data preprocessing and ETL (Extract, Transform, Load) operations. |
| Multi-Tenancy | **Purpose-Built Multi-tenancy**: VAST Data's platform is designed with multi-tenancy in mind. It provides robust isolation and performance guarantees for teams or projects sharing the same infrastructure, as well as efficient resource utilization and data security for multi-tenant environments.<br><br>**Resource Allocation and Management**: VAST Data enables fine-grained resource allocation and management, ensuring that each tenant can access the necessary resources without interference, optimizing the overall system performance. |

## OPTIMIZING THE AI PIPELINE

VAST Data's integrated data approach optimizes every stage of the AI pipeline, from data ingestion and preprocessing to model training, evaluation, and deployment, ensuring faster and more reliable AI development cycles.

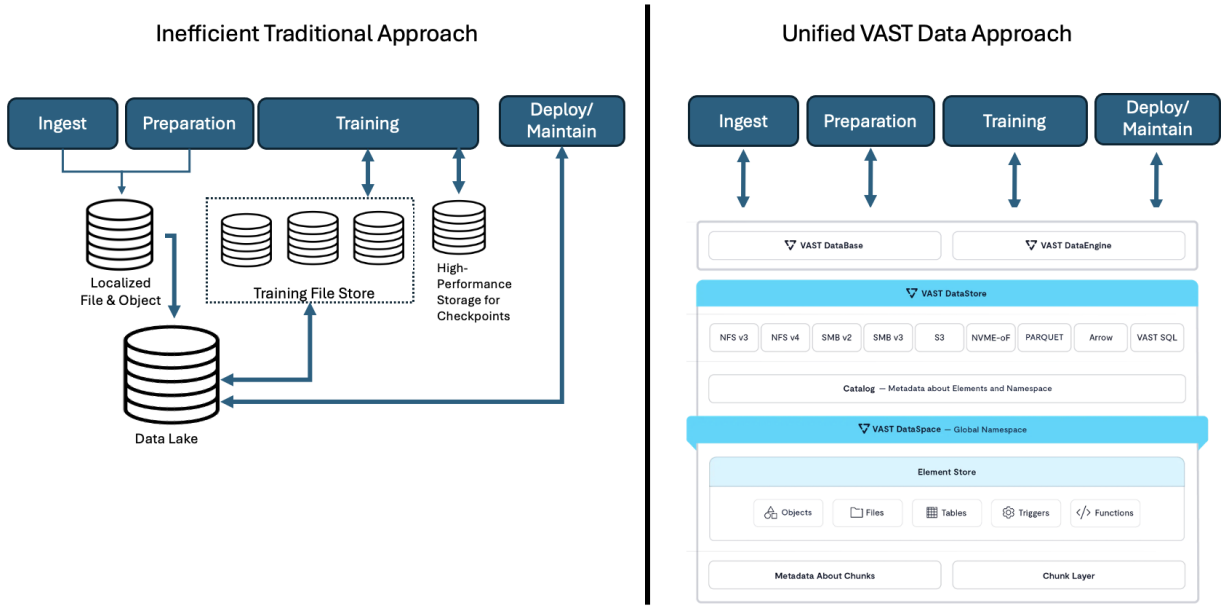



FIGURE 4: VAST DATA OPTIMIZES THE AI PIPELINE (SOURCE: VAST DATA)

Here's how VAST Data enhances each step within the AI pipeline:

TABLE 3: VAST DATA ENHANCES THE AI PIPELINE

| Phase | VAST Data Platform |
|---|---|
| Data Capture | **High-capacity, Scalable Storage**: VAST Data's platform efficiently handles large volumes of diverse data types (structured, unstructured, streaming), facilitating the initial collection and ongoing data management.<br><br>**Streaming Ingest**: The platform supports efficient ingestion from streaming data sources, ensuring that data is readily available for subsequent stages without delays.<br><br>**Fast Data Access**: The high-performance NVMe storage allows quick access to data, making preprocessing faster and more efficient. |

| | |
|---|---|
| Data Preparation | **Efficient Querying**: VAST Data supports rapid querying and retrieval, enabling data scientists to perform more effective exploratory data analysis (EDA).<br><br>**Centralized Data Store**: By consolidating data in a single, unified system, VAST Data eliminates data silos, making it easier to access and analyze data comprehensively.<br><br>**In-Database Processing**: VAST Data allows preprocessing operations like normalization, imputation, and encoding to be performed directly within the data platform, reducing the need to move data to separate environments and speeding up the process.<br><br>**Accelerated Feature Engineering**: The platform's integrated data management capabilities streamline the feature engineering process, allowing data scientists to iterate and test different features quickly. It accomplishes this by using its integrated database to enable advanced analytics and machine learning operations directly on the data platform. This reduces the time and complexity of feature engineering. |
| Model Training | **GPUDirect Storage Access**: VAST Data supports NVIDIA GPUDirect storage, which enables GPUs to access data directly from storage without intermediate CPU involvement, significantly speeding up model training times.<br><br>**Consistent Performance with High Throughput and Low Latency:** The platform's consistent and scalable high throughput enables rapid processing of large datasets while providing low-latency access to validation and test datasets. This ensures that model evaluation is conducted efficiently and without delays.<br><br>**Rapid Iterations**: VAST Data's high performance and fast data access capabilities enable rapid iterations during hyperparameter tuning, allowing data scientists to identify optimal configurations quickly.<br><br>**Job Mobility**: VAST DataSpace gives users the flexibility to move jobs between GPU clusters across different physical locations. This capability facilitates the preemption of training jobs for higher priority use cases, such as model serving, while allowing those jobs to be resumed in a different location.<br><br>**Efficient Resource Utilization**: VAST Data's efficient resource utilization ensures that computational resources are maximized during the tuning process. Its ability |

| | |
|---|---|
| | to generate many permutations rapidly and in a storage-efficient manner ensures that computational resources are maximized. |
| Model Serving | **Seamless Integration**: VAST Data's platform supports seamless integration with existing IT infrastructure, making deploying AI models smooth and efficient.<br><br>**High Availability**: The platform's robust storage provides high availability and reliability for deployed models, supporting continuous operation and real-time inference.<br><br>**Scalable Maintenance:** The platform's scalability allows for easy updates and maintenance of models, ensuring they remain effective as data patterns evolve.<br><br>**Efficient Data Handling**: VAST Data's ability to handle and process feedback data efficiently enables the rapid incorporation of insights and improvements into AI models.<br><br>**Iterative Improvements**: The platform supports the iterative nature of AI development, enabling continuous improvements and refinements to models based on real-world feedback.<br><br>**Generative AI Optimizations**: The VAST Data Platform includes targeted optimizations for generative AI workflows. An example is its enhancements for distributed RAG, where the platform can simultaneously load and retrieve data from a single vector database in numerous locations and rapidly replay and update vector embeddings as required by model changes. |

By addressing specific needs at each stage of the AI pipeline, VAST Data enables a seamless flow from data collection to deployment, minimizing delays and maximizing efficiency. This comprehensive support accelerates the AI development process and enhances AI systems' overall effectiveness and reliability.

## IN SUMMARY

AI is foundational to the next wave of digital transformation, impacting the competitiveness of enterprises across industries. Traditional infrastructure isn't designed for the demands required to support today's AI lifecycle, which places unprecedented demands on storage and data infrastructure.

Developing an effective data infrastructure for AI requires a holistic approach, considering data, database, processing, and storage as a unified entity.

The innovative approach taken by VAST Data with its VAST Data Platform is designed from the ground up to support the needs of AI at scale. Its approach significantly alleviates the typical bottlenecks faced in AI model development and deployment, offering a scalable, efficient solution that enhances the entire AI lifecycle.

© Copyright 2024 NAND Research. NAND Research is a registered trademark of NAND Research LLC, All Rights Reserved. All other trademarks mentioned herein are the property of their respective owners.

This document may not be reproduced, distributed, or modified, in physical or electronic form, without the express written consent of NAND Research. Questions about licensing or use of this document should be directed to info@nand-research.com.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.