STREAMLINING AI DATA MANAGEMENT: FIVE RECOMMENDATIONS TO OPTIMIZE DATA PIPELINES

Produced by Tabor Custom Publishing
in conjunction with:

HPC wire

ddn

For supreme AI data management, the AI data lifecycle needs a clear, continuous flow, both for training of AI models, and for applying AI models to real time decision making.

Artificial Intelligence (AI) involves several different stages from data ingest to model training, inference, and more. Each phase can become a bottleneck in the process, slowing the time to build new models or deliver new insights. Increasingly disparate data sources and model complexity magnify pipeline bottlenecks, so it is not surprising that many AI projects take too long, go over budget, or fail to reach production.

AI data management across the data lifecycle and capable AI storage are crucial to improving the efficiency of AI systems. Certainly, a model-centric approach for AI can benefit from increasing the quantity of training data and enhancing the quality and diversity of the data used for learning, so building a data-centric pipeline to manage the end-to-end data lifecycle is critical.

This whitepaper examines common issues and considerations in AI data management, and the data journey in AI workloads, and explores how to address these issues using a comprehensive AI data management system.

The AI Data Lifecycle

AI systems are data-centric and each application may need a unique data management workflow to meet business requirements across multiple stages. Diverse data types, throughput needs, and the level of support for different compliance regimes vary across these phases.

The typical data journey in a single AI project might include the following:

- **Data Ingest** – The gathering of raw data, such as images or time series data, used to create and refine a model or as input to the decision process. There may also need to be some upstream filtering or buffering to manage the ingest process.
- **Data Preparation** – Raw data needs to be cleansed or prepared before processing. For example, learning data needs to be appropriately categorized and labeled, while real-time data needs to be filtered or normalized before input to the model.
- **Training and Reinforcement Learning** – A complex AI model may take a significant time to build (says, weeks or more), as the pre-prepared training data needs to be loaded and iterated many times and validated against known correct data.
- **Inference and Recommendation** – Real-time data drives recommendations or decisions once a model is validated. There may also be a feedback loop to tune the model against new data via reinforcement learning.
- **Data Archive** – Data audit and compliance regulations may require logging every AI decision, along with the input data and the model information. For example, an image recognition system would need to register the original image in addition to the AI recommendation to provide documentary evidence of how a decision was reached.

Each of these pipeline stages has a unique set of requirements for capacity, throughput, and latency – and could each represent a potential bottleneck in the AI pipeline, delaying data delivery for the next step in the process. A common mistake is to treat each stage as a separate silo – each optimized to meet the characteristics of that stage. However, this often overlooks the most common challenge shared by each stage – the need for continuous and seamless data flow through the pipeline. Without this end-to-end approach, the individual silos tend to work in batch mode, with no continuity and frequent data stalls, failing to deliver an efficient and frictionless data journey from one stage to the next and onward.

Support Systems for AI Workloads

While scalable, high-performance file services are essential for most AI workloads, support for other workloads is also required. As organizations mature their development processes and start to bring AI into more areas of the business, several supporting systems become increasingly important. In addition, the evolution of AI environments introduces additional storage requirements for AI data management, such as:

- **Home directories** — Data scientists and developers need home directories to store their working documents, including reports, specifications, and program scripts. These would typically be interactive workloads, supporting day-to-day operations and collaboration.

- **Containers** — Virtualized and container technologies are ideal for supporting today's dynamic environments, especially where a team of developers is launching new applications or analytics for prototype applications. Storage systems optimized for virtual machines and container technologies can deliver greater agility, performance, and security than traditional IT storage systems.

- **Databases** — Databases can support interactive and embedded systems across workloads, including the data labeling and preparation phase of AI model training. DB-optimized storage systems can also help with integrated monitoring and AIOps capabilities.

- **Backup** — AI systems often have such a high data volume, and traditional backup systems may have difficulty with peta-scale datasets where "data gravity" is a problem. Data snapshots are not an ideal way to manage backups, so AI systems may need data sync/replication tools to support backup operations. In either case, backup and recovery operations may be protracted, resulting in significantly increased I/O activity with a potential performance impact.

- **Archive** — While AI Backup is an essential process to capture or restore data for recovery from data loss or outage, an AI Archive centralizes copies of data to act as a record or audit for later evaluation. For example, we might archive a training data set or a model and subsequently archive copies of input data and the decision or recommendation along with the model reference for audit purposes. The workload may be a continuous stream of ingested and AI-generated data to a 24x7 archive platform, potentially impacting I/O and performance.

Planning each of these additional demands in the overall AI data lifecycle context is essential. While it may be possible to identify unique storage

architectures to optimize each style of workload discussed above, it is vital to carefully orchestrate the entire AI data journey to deliver continuous 24x7 data operations.

Five Recommendations for AI Data Storage

Business and Technical leadership are essential for success in AI projects. It is also crucial to manage data through the entire lifecycle to ensure a streamlined data pipeline and assist operations and governance, including compliance, audit trails, and process knowledge.

Here are five critical considerations for AI data management:

1. **Assign executive leadership for your AI program.** Set the overall business objectives for the program, key milestones, and agree on what success looks like for this project. An AI program can gain cross-functional engagement and much greater success with executive sponsorship.

2. **Verify data capacity and throughput needs.** Identify which steps in the process are likely to be bottlenecks in the data journey, and level up your architecture or design to meet the performance levels required to meet the business requirements and technical parameters needed by your AI model.

3. **Assess workload requirements from end to end.** AI systems handle a unique set of workloads, so traditional IT technologies may be a poor fit for AI applications, which need to operate at much larger scales. For example, suppose the aim is to drive an AI learning system with a petabyte or more of learning data; in that case, you must be able to run that data through the compute platform many times to train the model – which may take weeks or months using traditional IT technologies.

4. **Evaluate the backup and archive storage needs.** Backup and Archive processes can significantly impact data throughput and may need considerable additional investment to meet capacity and throughput requirements. An archive may be imperative when applying an AI model to an image where you must simultaneously store the modified and original for compliance reasons. Similarly, consider the impact of scale on backup, recovery, and continuity plans.

5. **Plan ahead for explainability.** Interpretability (also known as explainability) is becoming increasingly important as AI systems directly impact businesses and consumers. Some industries, such as the financial or security sectors,

need to be able to interrogate and justify decisions or provide supporting evidence for a recommendation, which can have implications for the governance and compliance of AI systems.

DDN, The AI Data Company

To deliver peak performance in production, an AI system needs to manage the entire data journey specific to each workload, encompassing both the ingest data path to the workload and the downstream workflows at any scale.

Particular pressure points are with Model Training, Backup, and Archive of AI systems – which may each significantly impact I/O congestion and performance degradation in AI systems. Systems that support parallel, shared access, such as DDN's A^{3}I AI Storage, are uniquely suited to continuous 24x7 AI storage systems because they eliminate bottlenecks in data throughput, supporting concurrent access to multiple stages of the AI pipeline.

DDN has significant experience designing, building, and delivering consultancy for at-scale systems to support global enterprises. To learn more about how to manage the end-to-end data pipeline purpose-built for AI data management, go to DDN.com.