insideBIGDATA

insideBIGDATA Guide to

Data Platforms for Artificial Intelligence and Deep Learning

By Daniel D. Gutierrez, Managing Editor and Resident Data Scientist, insideBIGDATA

BROUGHT TO YOU BY

DDN® STORAGE

Contents

Introduction

The stage is set for enterprise competitive success with respect to how fast valuable data assets can be consumed and analyzed to yield important business insights. Technologies such as artificial intelligence (AI) and deep learning (DL) are facilitating this strategy and the increased efficiency of these learning systems can define the extent of an organization's competitive advantage.

Many companies are strongly embracing AI. A March 2018 IDC spending guide on worldwide investments on cognitive and AI systems indicates the level will reach $19.1 billion for 2018, an increase of 54.2% over the amount spent in 2017. Further, spending will continue to grow to $52.2 billion by 2021. By all indications, this is an industry on an upward trajectory, but limiting factors such as data storage and networking bottlenecks must be addressed to assure the maximum benefit from AI and DL applications.

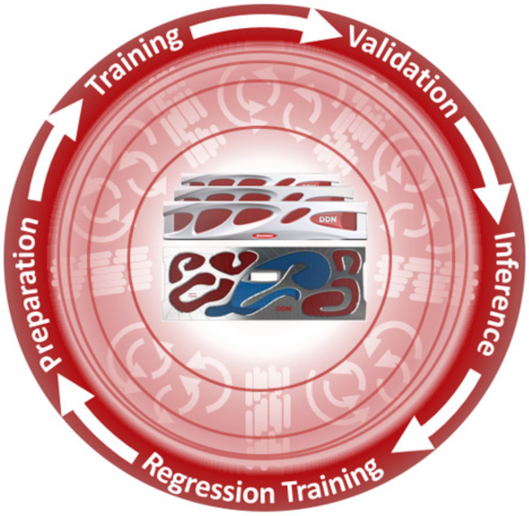
> With AI and DL, storage is cornerstone to handling the deluge of data constantly generated in today's hyper-connected world. It is a vehicle that captures and shares data to create business value.

Enterprise machine learning algorithms have historically been implemented using traditional compute architectures, where system throughput and data access latencies are measured by paring compute and storage resources through the same network interconnections that serve other business applications. With AI and DL, the increasing volume and velocity of arriving data are stressing these legacy architectures. Although compute has made great strides with GPUs, legacy file storage solutions commonly found in enterprise datacenters haven't kept pace.

With AI and DL, storage is cornerstone to handling the deluge of data constantly generated in today's hyper-connected world. It is a vehicle that captures and shares data to create business value. In this technology guide, we'll see how current implementations for AI and DL applications can be deployed using new storage architectures and protocols specifically designed to deliver data with high-throughput, low-latency and maximum concurrency. The target audience for the guide is enterprise thought leaders and decision makers who understand that enterprise information is being amassed like never before and that a data platform is both an enabler and accelerator for business innovation.

Data is the New Source Code

Data's role in the future of business cannot be overstated. DL is about growing autonomous capability by learning from very large amounts of data. In many ways, *data is the new source code*. An AI data platform must enable and streamline the entire workflow. AI and DL workflows are non-linear, i.e. not a process that starts and then ends, and then goes onto the next iteration. Instead, non-linear means the operations in the workflow happen concurrently and continuously (as depicted in the wheel graphic below). It's all about iterating, completing each step as fast as possible through the acceleration afforded by a parallel storage architecture. It's about getting the wheel going and allowing customers to grow their infrastructure seamlessly as the data sets grow, as the workflows evolve. Data is ingested then gets indexed and curated before being used for training, validation, and inference; all these different steps happen concurrently and continuously. Data continues to be collected as training occurs, as models are moving to production. The wheel gets bigger and more engaged as workflows evolve.



Unique Storage Demands for AI and DL Workloads

The primary components of AI and DL, artificial neural networks (ANNs), have extraordinary data consumption with limitless combinations of adjustments to hyperparameters and samples in data sets. These applications pose exceptional challenges and put significant strain on compute, storage and network resources. Legacy file storage technologies and protocols like NFS starve AI workloads of data, slowing down applications and deterring important insights. A true AI data platform must concurrently and efficiently service the entire spectrum of activities involved in DL workflows, including data ingest, data curation, training, inference, validation and simulation.

At the core of AI and DL, the training process involves scale and complexity. Training is essential to reach the desired accuracy for these algorithms, requiring immense I/O, data storage and computational resources. Parallelizing the training process serves to accelerate model refinement, with faster transition to production.

Reliable and rapid inference requires an iterative training process to achieve validation of accuracy — models with hyperparameter variations are run through multiple epochs (complete passes of the data sets).

AI and DL workflows happen concurrently, continuously, and benefit from distributed computing. A shared storage architecture provides simultaneous access to data from multiple systems, enabling multiple operations to happen at the same time. An AI data platform must provide collection and access of large amounts of heterogeneous data from a wide variety of sources. To be useful for the DL application, the ingested data sets must be indexed and curated. From a user perspective, it's important to enable easy data discovery, which means making the information available everywhere, anytime, and through an interface easily accessible by data scientists and the applications.

Characteristics of Storage Solutions Optimized for AI & DL

Achieving successful AI and DL deployments requires ingest, processing and continuous engagement of large and diverse data sets, often at scale. Further, the combination of different concurrent, mixed workloads requires a storage system which can cope with a wide range of workloads. As a result, the characteristics of a data storage solution in support of these types of workflows must be specifically optimized for AI and DL. What's needed is an AI data platform.

An AI data platform must enable rapid iteration for model training and refreshing. Rapid prototyping, experimentation and refinement of models are critical for achieving best possible yield from neural networks. The AI data platform must enable easy transition of models from training into validation and inference. Seamless transition from development to production enables organizations to quickly leverage and monetize innovation.

The AI data platform must be capable of ingesting, handling and delivering heterogeneous data types and mixed workloads simultaneously and without compromise.

Where data is the new source code, the process begins by learning from a very large amount of data. The idea of success for a DL development environment is having access to very large data sets that are meaningful to the problem domain. The size, the quality, the diversity of the data sets are all critical. The AI data platform must be capable of ingesting, handling and delivering heterogeneous data types and mixed workloads simultaneously and without compromise. The AI data platform must scale seamlessly in multiple dimensions — capacity, capability, performance — to match evolving workflow needs. It should provide flexible configuration options to achieve optimal technical and economic benefits for organizations.

AI thinking, making the AI think fast, is accomplished by using large amounts of information and making it accessible faster. The more reliable the access, the better the AI results will be.

There is a balance between technical and economic realities, i.e. you want to be efficient in the way you store your data, you want to be smart about the way you store your data, and you want to spend a lot of time thinking about these realities when you're defining your infrastructure.

Data protection also is very important. Organizations engaging AI and DL programs incur significant expense to collect data sets. Enterprises need to start thinking about how to accumulate data from multiple sources, heterogeneous types of data, images, text, etc., and hold onto it for an infinite period of time. How can this data be protected? When you're talking about 10s or 100s of PB over decades, this can raise overwhelming and complex considerations. How can you make sure your data is there when you need it? And maybe even more importantly, how can you organize this data in a way that you can find it again? A thorough data governance plan is a cornerstone for enterprises engaging in AI and DL.

"Data plasticity" for AI and DL is an important notion where data is the lifeblood of the march toward insights (similar to neuroplasticity where the brain is able to reorganize itself). When it comes to AI pipelines, data must be easily molded into forms needed by the application. Once captured, you have a giant pool of very valuable information, and you want to be able to consume it easily, rapidly and in very different forms. AI thinking, making the AI think fast, is accomplished by using large amounts of information and making it accessible faster. The more reliable the access, the better the AI results will be. These are all virtues of a robust AI data platform.

Accelerated, Any-scale AI Solutions

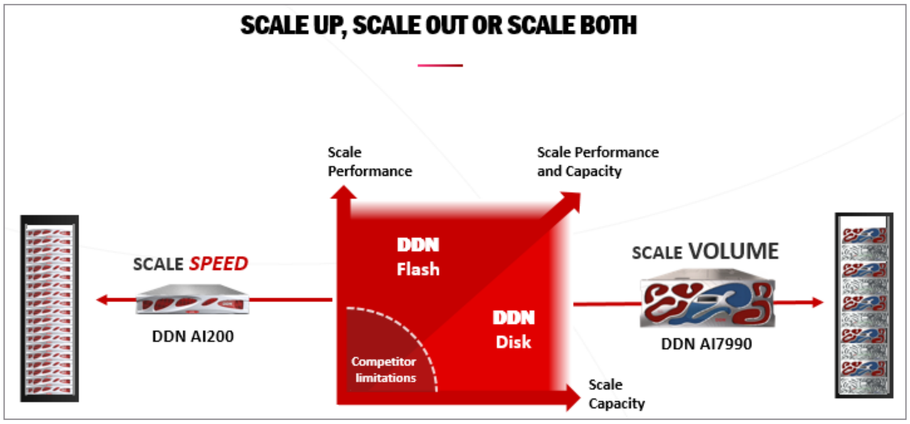
DataDirect Networks (DDN) is a global technology leader, with two decades of experience in designing, implementing, and optimizing storage solutions that enable enterprises to generate value and accelerate time-to-insight from their data — both on premise and in the cloud. Engineered as an AI data platform, DDN's A³I (Accelerated, Any-Scale AI) can truly maximize business value and optimize the AI environment — applications, compute, containers, and networks. It enables and accelerates AI applications and streamlines DL workflows using a scalable shared parallel architecture and protocol.

One of the foundational elements for success with AI and DL is establishing the centrality of data in the AI workflow. Many existing projects in the enterprise might depend on a collection of smaller data sets at this point in time, but as AI methodologies mature, and as more enterprises engage AI, the amount of data that can be applied to a problem, as well as the performance that can used to process it, is becoming more and more important. The DDN shared parallel storage architecture embraces this notion.

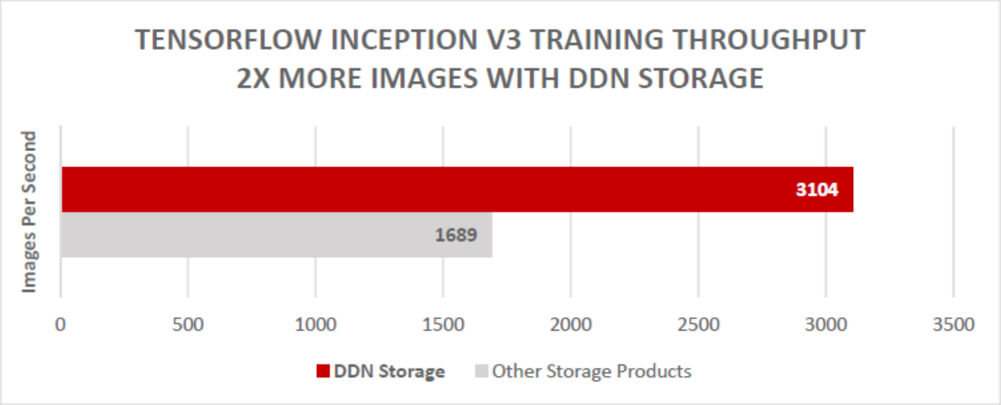
DDN's large data ingest capability can run concurrently with training and inference processes. It easily handles large and diverse data streams from multiple sources. The DDN shared parallel storage architecture incorporates unique features that accelerate, streamline and secure end-to-end AI and DL workflows.

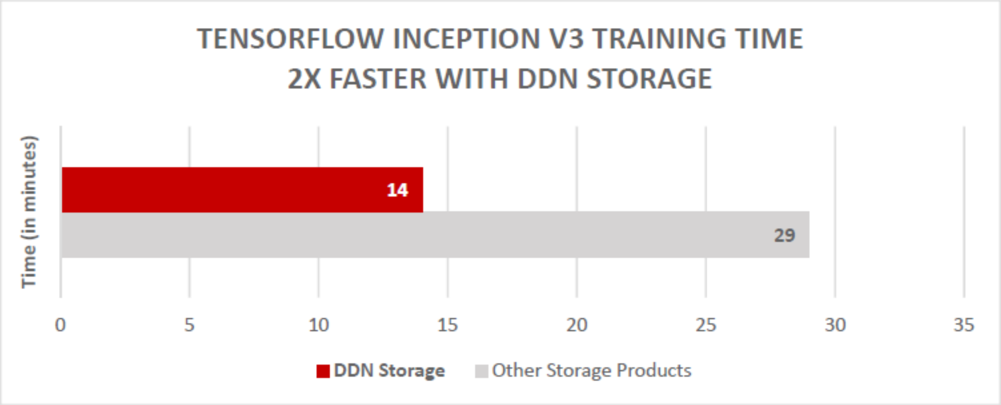
The following is a short list of benefits of A³I for AI and DL deployments:

- DDN A3I is a turnkey AI data platform, fully-integrated and optimized for AI and DL applications. It has been thoroughly tested with widely used CPU and GPU computing platforms, AI and DL applications, and can be easily integrated into any IT environment.
- The parallel architecture and protocol extends the performance of NVME from the disk all the way to the application for maximum acceleration.
- Data is delivered with high-throughput, low-latency and massive concurrency to achieve full GPU saturation. This ensures that all compute cycles are put towards productive use.
- The shared architecture allows multiple systems to access data simultaneously, enabling multiple phases of the workflow to happen at concurrently and continuously.
- The platform provides flexible and seamless scaling of performance, capacity and capability to match evolving workflow needs.
- The platform can ingest, process and deliver heterogeneous data from a wide variety of sources, and supports mixed workloads.
- DDN's shared parallel file systems use metadata to enable file system level tagging of assets, and then making it easy for applications to find the data they're looking for based on these metadata tags.
- The platform includes robust data protection and integrity capabilities, and can be architected for maximum availability.
- These solutions are designed, deployed and supported by DDN's global R&D and field engineering organizations.



DDN storage solutions are unique in that they are continuously tested and optimized with commonly used AI and DL applications and various networking and computing platforms. DDN technology is tightly integrated with GPUs, providing optimal data fulfillment from storage-to-GPU and GPU-to-GPU, for fastest and most efficient use of computing resources. DDN has equipped its laboratories with leading GPU compute platforms to provide unique benchmarking and testing capabilities for AI and DL workloads.





Engineered from the ground up for the AI data platform, DDN A³I solutions are fully-optimized to accelerate AI applications and streamline DL workflows for greatest productivity. DDN A³I solutions make AI-powered innovation easy, with faster performance, effortless scale, and simplified operations — all backed by the data at scale experts. The DDN shared parallel architecture fully saturates GPUs and ensures all efforts go towards productive AI use.

- The DDN AI200 is an all-NVME flash appliance that is an efficient, reliable and easy to use data storage system AI and DL applications. The AI200 reference architectures are designed in collaboration with NVIDIA® to provide highest performance, optimal efficiency, and flexible growth for NVIDIA® DGX-1TM servers. With AI200, Caffe applications running on a DGX-1 server demonstrate 2.4x increased image throughput and 2x shorter completion times. TensorFlow training applications demonstrate double image throughput and complete twice as fast on a DGX-1 server using a AI200 solution.
- The DDN AI7990 is a hybrid storage appliance for ultimate flexibility that allows intermix of performance flash and large capacity disk in a high-density system. The AI7990 keeps DGX-1 servers saturated with data ensuring absolute maximum utilization whilst also managing tough data operations from bursty ingest to large scale data transformations.

Used for image, voice and sound recognition, TensorFlow DL applications depend on large and diverse data sets with rich media content. DDN storage solutions provide the capacity needed to store and deliver massive heterogeneous data sets. They sustain the performance required to ensure data saturation of multiple GPUs engaged in distributed, accelerated training of node-based, multi-layered deep neural networks.

Data Storage for AI/DL Case Studies

In this section we'll consider some compelling use case examples of how DDN storage systems have enabled customers to maximize the value of their data and easily and reliably accelerate time to insight using AI and DL. DDN enables thousands of customers all around the world, in a wide cast of industries, to accelerate their businesses using AI and DL. DDN A³I solutions are fully-optimized to deliver massive performance acceleration to these enterprise applications.

AUTONOMOUS VEHICLES

Autonomous vehicles engage some of the toughest workloads in AI at unprecedented scale. They require the handling, ingest and delivery of a broad range of data set types and sizes, generated from many different sources such as video cameras, lidar, radar and other sensors. Very large data sets captured over millions of miles undergo many cycles of processing, labeling, sub-sampling and categorization, before being presented to the DL applications.

Self-driving vehicles require the maximization of the number of testing scenarios to improve vehicle perception accuracy and operational autonomy. This requires a reliable data storage framework that scales to TB/sec of throughput and hundreds of PB of capacity is essential.

For this customer, a massive data set for training neural networks was developed, data from experimental vehicles and ridesharing engagements was collected, and an extensive and complex DL framework was trained, tested and refined for the autonomous driving capability. The resulting software was loaded onto experimental vehicles for evaluation in the field, and operational data from the ride fed back into the loop to further enhance the DL process.

The customer's requirement called for the creation of a very large scale parallelized data storage system to feed an extremely large scale GPU-based computing platform. The storage solution had to ingest, keep and deliver massive amounts of data rapidly and reliably, scaling linearly to extreme levels in performance and capacity. With original increments set at nearly one hundred petabytes of capacity, highest data center density and efficiency with low management and support overhead were additional must haves.

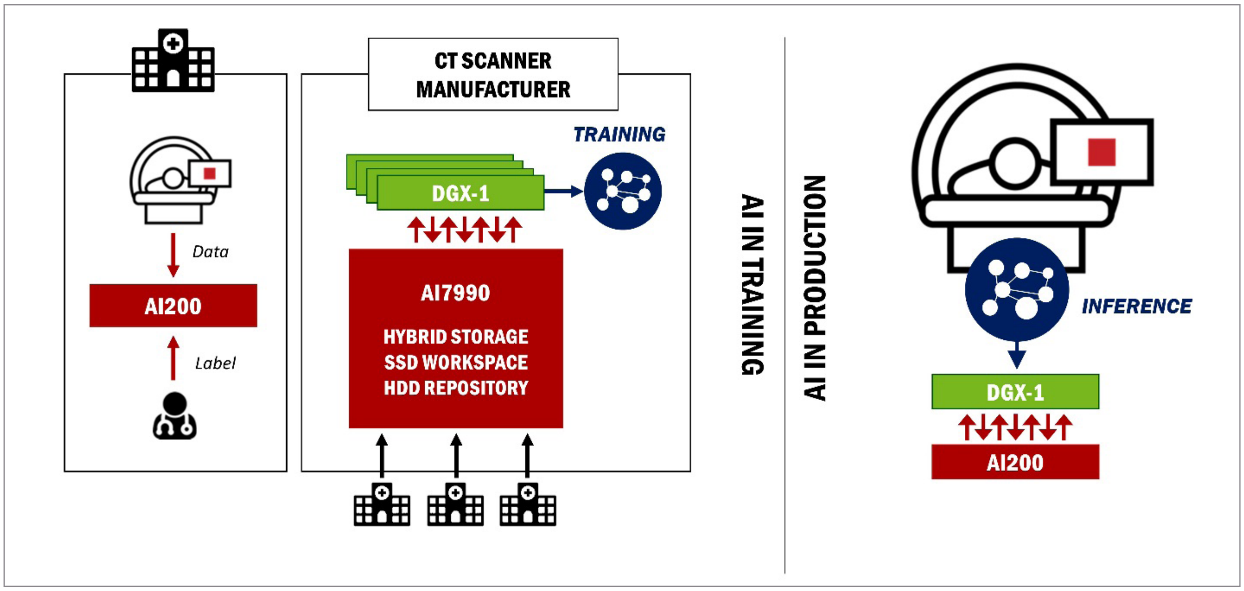
> Autonomous vehicles require the handling, ingest and delivery of a broad range of data set types and sizes, generated from many different sources such as video cameras, lidar, radar and other sensors.

The DDN storage platform effortlessly handles the concurrent ingest of these massive data streams, organizing and structuring the underlying data sets.

Millions of GPU cores continuously access the DDN storage system, executing extensive and complex training processes, continuously refining the self-driving capabilities of the fleet of vehicles. DDN storage has enabled this customer to harness data at immense scale, successfully and reliably building an advanced AI framework that is revolutionizing the transportation industry.

LIFE SCIENCES AND HEALTHCARE

By using AI techniques such as machine learning and artificial neural networks, researchers are building systems to improve the detection, diagnosis, treatment and management of diseases. In addition, clinicians, researchers and industry players are working to co-develop and validate algorithms that can recognize patterns of disease and advance diagnostic capabilities. Corps of data scientists, developers, and fellows train and test models with the potential for commercialization. There is a focus on the pipeline of translation — from model conceptualization to clinical validation. AI platforms enabled by DDN storage greatly enhance the ability of researchers to identify and cure diseases.



DDN AI and DL in life sciences use case

A research facility selected DDN to implement a solution capable of covering all ingest, processing and management of the data sets, training and inference from the DL applications, and real-time visualization.

The storage system was required to hold a large repository of data sets for neural network training with rapid shared access to multiple GPUs that execute intense training, testing and inference. The DDN all flash system deployed reliably, handles complex data ingest while simultaneously supporting post-processing, inference, visualization, training and validation operations.

CONSUMER RETAIL

Another compelling use case involves a leader in next generation retailing technology that developed ground breaking software enabling consumers to shop without having to go through the cumbersome check out process. A series of high-definition cameras within each store are coupled with advanced computer vision and DL to identify shoppers and keep track of which items they collect in real time. Shoppers are billed automatically for the items as they leave the store. Live feeds are ingested from each store's video cameras during opening hours, while an intensive training activity is engaged in the limited window after closing time, leveraging the day's collected data sets.

The customer selected DDN for their requirement of an all flash component due to the limited training time window and in order to ensure saturation of the GPUs used by the DL application. DDN delivered a solution which ingests live feeds from cameras in real-time and provides built in scalability to handle the collection of additional daily data sets over time. The DDN solution combines an all flash layer, with integrated controls for automatic staging of the day's freshly acquired dataset, with a hard disk layer for longer term economical storage. GPUs get fastest and most efficient access possible to the daily capture data and achieve highest productivity.

Summary

With the help of storage solutions fully optimized for AI and DL training and inference, data scientists, data engineers as well as academic researchers are able to focus their complete attention on what really matters most — transforming valuable data assets into important insights with unparalleled velocity and accuracy.

In this technology guide, we've reviewed the unique storage demands for AL and DL workloads, along with the characteristics of storage solutions optimized for AL and DL. We also provided a description of products available from DDN and how they suit the requirements of storage solutions well-adapted for workflows involving AI and DL. Here are some important takeaways when considering next steps to take in choosing your storage solution:

1. **Performance** is a critical aspect of data storage for AI and DL workloads. Parallel data access is the key for keeping pace with the demands of these popular technologies.
2. **Flexibility** in the AI workflow is also vital in order to be able to deal with multiple data types, and engage multiple workflows.
3. **Scalability** enables the ability to think ahead. Your needs today may be of limited scale. You may have a small data set in 2018, but there is high likelihood that you'll be on a path of collecting more data because you have new sensors, new connectivity such as the new 5G coming out, and higher resolution data sets. The technologies that are enabling AI like GPUs have a very fast refresh cycle—every 8 months your GPUs are quadrupling in capability. Suddenly you're able to collect and process more information. In terms of scaling, enterprise applications are built on software and that iteration is in real- time as data scientists are able to come up with new algorithms for consumption. Benefit comes from maximum amounts of performance. This is the difference between break through innovation vs. incremental upgrade.

> By simultaneously expediting deployment and delivering acceleration in time to insight, DDN's groundbreaking approach enables you to manage the entire AI lifecycle in-place and simplify your data center.

Time is of the essence in making strategic decisions about storage solutions for managing accelerating demands put in place by AI and DL applications. Your competitors are making the same decisions to gain strategic advantage in the marketplace. To take important next steps for learning how you can facilitate breakthrough innovation by easily leveraging the power of new turnkey AI solutions for the data center visit DDN. By simultaneously expediting deployment and delivering acceleration in time to insight, DDN's groundbreaking approach enables you to manage the entire AI lifecycle in-place and simplify your data center. DDN can show how their storage solutions have the following advantages:

- Easy to deploy AI solutions that immediately transform your AI concepts into business innovation
- Possess long-term advantages that enable you to achieve high-performance AI at every stage of your growth
- Show you how to realize the greatest technical and economic benefits through leveraging deep AI-expertise