IBM Spectrum Storage for AI with NVIDIA DGX Systems

The Engine to Power your AI Data Pipeline

Highlights

- Converged Solution: Ready to deploy for your most demanding Machine Learning / Deep Learning projects
- IBM Spectrum Scale v5: Software-defined to streamline data movement through the AI Data Pipeline
- NVIDIA DGX-1 and DGX-2 Servers: Purpose-built solution for AI and Machine Learning
- NVIDIA DGX software stack: Optimized for maximum GPU training performance
- Proven Data Performance: Over 120GB/s throughput to support up to 9 DGX-1 servers. and up to 3 DGX-2 servers in a single rack
- The foundation to build a shared data service for your containerized AI workloads
- Simplified support model via business partners with competency across the entire solution, backed by IBM and NVIDIA

The Engine to Power your AI Data Pipeline

IBM Spectrum Storage for AI with NVIDIA DGX is an integrated compute and storage solution to support the complete lifecycle of AI – from data preparation to training to inference – using the latest innovations of systems and software.

Challenges:

Industry predictions suggest almost all technology, services, and science will be enhanced with AI within a few years. Enabled by powerful GPUs and optimized ML/DL frameworks delivered as containers, AI is rapidly being adopted in every industry and discipline. Innovations in computer vision and object detection, human/computer interaction, data classification, and sophisticated pattern detection are available today and the applications of these techniques are expanding. The fidelity of Machine Learning, Deep Learning and Neural Network training is driven by the quality and quantity of data available. Matching data infrastructure to powerful servers can be a challenge because high-performance access to lots of shared data is critical.

The productivity of a data science team can be severely limited without access to the appropriate data. As widely reported, the preparation and ingest of data consumes the majority of data scientists' time. Every project requires training and testing data sets, properly organized and tagged, so they can be used for model development and on-going validation. Proper management of the AI data pipeline, including data governance, extensible metadata tagging, archiving and flexible shared storage provides organizations with a data repository against which to build and train multiple models.

According to an IDC Survey, the three major challenges to deploying AI workloads are Data Volume and Quality, Advanced Data Management, and Skills Gap.[1]

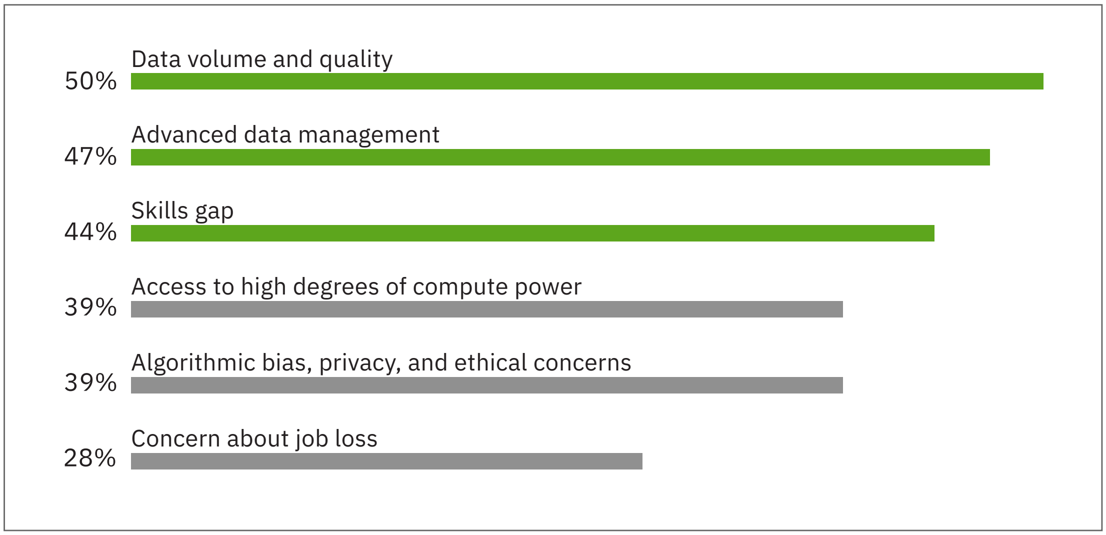

Figure 1: *Challenges deploying AI workloads*

[1] Source: Cognitive, ML, and AI Workloads Infrastructure Market Survey, IDC, January 2018; n=205, 1,000+ employees (U.S.); 500+ employees (Canada)

Solution Description:

At the core of data science productivity are the infrastructure and software used for building and training ML/DL workloads. As GPU density increases, the need for dense, fast storage increases as well. Therefore it is an absolute requirement to use high-performance shared storage with the latest NVIDIA DGX-1 and NVIDIA DGX-2 POD servers to run optimized container AI workloads.

Advancements in all-flash storage technologies provide the basis for building these storage systems, connected with the latest high-speed networking to drive up to 72 GPUs with NVIDIA DGX-1 POD servers and 48 GPUs with NVIDIA DGX-2 POD servers in a single rack configuration. Expert design and configuration are required to eliminate data bottlenecks and keep the systems running at peak performance.

However, to really scale AI projects, data science and IT teams must collaborate to improve the efficacy of the overall AI data pipeline, from ingest to inference and archive. Data sources are typically varied. Optimization requires a common data store with sufficient power and flexibility to ingest data from multiple sources, support analytics and scripting tools, and transparently tier off little-used data.

Eliminating data movement without limiting the choice of tools will increase team productivity, reduce costs, and simplify data governance requirements. Software-defined storage provides the extensibility and management to create a single view of the data flow, scale-out, and tiered systems to meet business requirements. Using IBM Spectrum Scale removes barriers to data access and overhead of data copy, commonly found when using traditional storage arrays.

Productivity, Flexibility and Scalability

IBM Spectrum Storage for AI with NVIDIA DGX Systems is an AI data infrastructure workhorse on which companies can build their data science services. A converged offering from IBM and NVIDIA, the IBM Spectrum Storage for AI platform is built for the full lifecycle of data science including data preparation, training, and inference. A composable architecture that uses software-defined storage and the latest NVIDIA hardware and software stack innovations, IBM Spectrum Storage for AI provides the flexibility to fit into the extended data pipeline, and the performance to handle multiple users and multiple models. With IBM Spectrum Storage for AI data scientists have the power, data and flexibility they need.

Tested and tuned, IBM Spectrum Storage for AI with NVIDIA DGX is a converged, but separately scalable compute, storage and networking solution composed of NVIDIA DGX-1 or NVIDIA DGX-2 POD servers, IBM Spectrum Scale on all-flash storage, and Mellanox networking. Built so organizations can start small and grow, IBM Spectrum Storage for AI with NVIDIA DGX begins with three DGX-1 or DGX-2 POD systems with a single NVMe storage system or IBM Elastic Storage Server (ESS).

With three NVMe systems and nine DGX-1 POD systems, a full rack scales to 72 Tesla V100 Tensor Core GPUs supported by 120GB/s of data throughput. Or with three DGX-2 POD systems, a full rack scales to 48 Tesla V100 Tensor Core GPUs and 120GB/s of throughput. Because it is software-defined storage, IBM Spectrum Scale can span multiple racks or other types of storage with automatic tiering, data protection and archiving of exabytes of data.

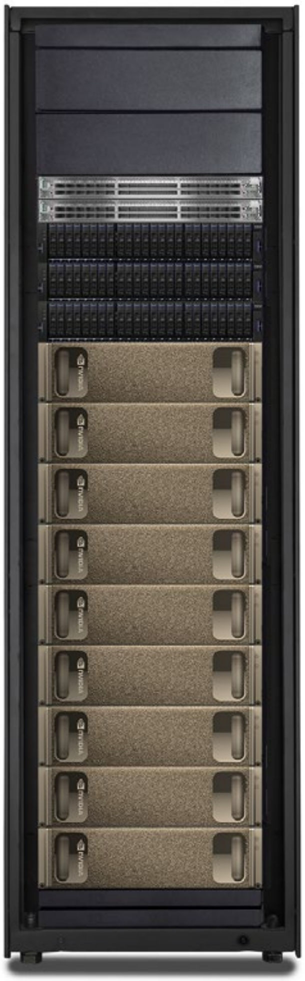

Figure 2: IBM Spectrum Storage for AI with DGX-1 POD, 9:3 configuration

The NVIDIA DGX-1 and DGX-2 POD systems are full-stack AI compute solutions, powered by the world's most advanced data center accelerators – the NVIDIA Tesla V100 Tensor Core GPU. Each DGX-1 provides 1 petaFLOP and each DGX-2 provides 2 petaFLOPS of mixed-precision training performance and is designed to be ready to use, with a rapid turn-up experience leveraging an optimized software stack that includes NVIDIA CUDA, deep learning libraries, and container management with Docker. The DGX container registry provides teams an extended catalog of NVIDIA-optimized AI and data science containers. NVIDIA DGX enables a data scientist to rapidly develop machine learning and deep learning models, effortlessly iterate and experiment, and deploy for training and inference at scale.

To support better data classification, tracking and governance, IBM Spectrum Discover is modern metadata management software providing data insight for unstructured data across IBM and third party file and object storage both on-premises and in the cloud. It can rapidly ingest, consolidate and index metadata for billions of files and objects, providing a rich metadata layer for data scientists, storage administrators, and data stewards to efficiently manage, classify and gain insight from massive amounts of unstructured data.

IBM Spectrum Storage for AI network connectivity is provided by Mellanox. Choosing the Infiniband option provides reliable 100Gbps networking that can leverage advanced technologies such as RDMA. Proper sizing and configuration provide throughput to storage and for inter-node communication between NVIDIA DGX-1 and DGX-2 POD systems.

Supercharge your AI Data Pipeline

IBM Spectrum Storage for AI with NVIDIA DGX Systems delivered ready-to-deploy by IBM and NVIDIA expert channel partners, it enables data scientists to quickly ramp up their work – as well as be ready for the future. Backed by NVIDIA expertise and support, data scientists will have access to the latest in NVIDIA optimized GPU accelerated tools with the confidence in the storage performance to use them.

As data and business requirements expand, IBM Storage provides the choice and innovation to meet the performance, economics, and data governance needs of any AI data pipeline. IBM Spectrum Scale provides storage services across multiple media, including AWS public cloud. It can share data and metadata with IBM Cloud Object Storage and tape to provide geo-dispersed or local storage flexibility.

By considering an end-to-end AI data pipeline, forward thinking organizations will benefit from continued productivity, lower costs, and simplified data governance. Data science teams realize rapid time to insight with a minimum number of data copies while, infrastructure teams benefit from simplified management, scalability and improved TCO.

IBM Spectrum Storage for AI with NVIDIA DGX Systems is the data science platform on which to develop and deliver new AI/ML/DL applications – from data preparation, to training, to inference and visualization. This proven enterprise solution delivers a rapid, plug-in / power-up deployment, with groundbreaking performance. IBM Spectrum Storage for AI with NVIDIA DGX Systems integrates the latest innovations in AI computing, high-performance storage and networking to deliver the fastest path to insights, supporting multiple development projects, backed by a simplified support model that keeps your AI workloads up and running.

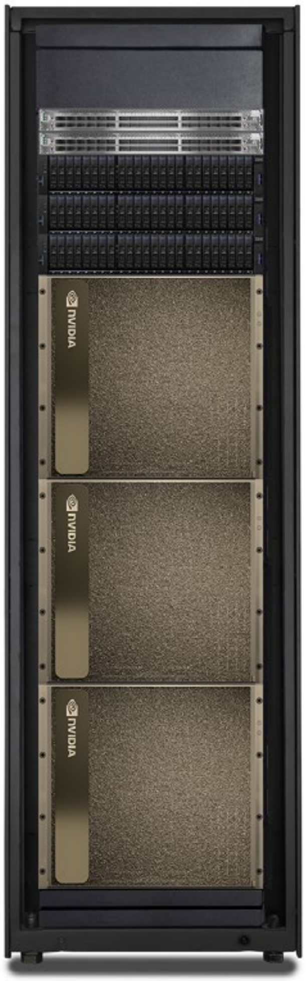

Figure 3: IBM Spectrum Storage for AI with DGX-2 POD, 3:3 configuration

IBM Spectrum Storage for AI with NVIDIA DGX Technical features

- DGX-1 and DGX-2 POD Systems – purpose-built solutions for AI and machine learning
- The NVIDIA DGX software stack optimized for maximized GPU-accelerated training performance
- IBM Spectrum Scale v5, the leading software-defined file storage, architected specifically for AI workloads with enhanced small file, metadata and random IO performance.
- NVMe all-flash storage for extremely low latency power efficiency and data density. Using IBM Spectrum Scale distributed data protection it delivers over 300TB in every 2U building block and up to 120GB/s of data throughput in a rack.
- Seamless data pipeline connectivity across multiple racks, other IBM Spectrum Storage for AI with NVIDIA DGX Systems configurations, and workstations to provide data scientists with a unified view of their IBM Spectrum Storage for AI data environment.
- Integrates with IBM Spectrum Discover for extensible data governance and metadata tagging across IBM and third-party file and object storage on-premises and in the cloud.

AI Data Pipeline with Storage Requirements

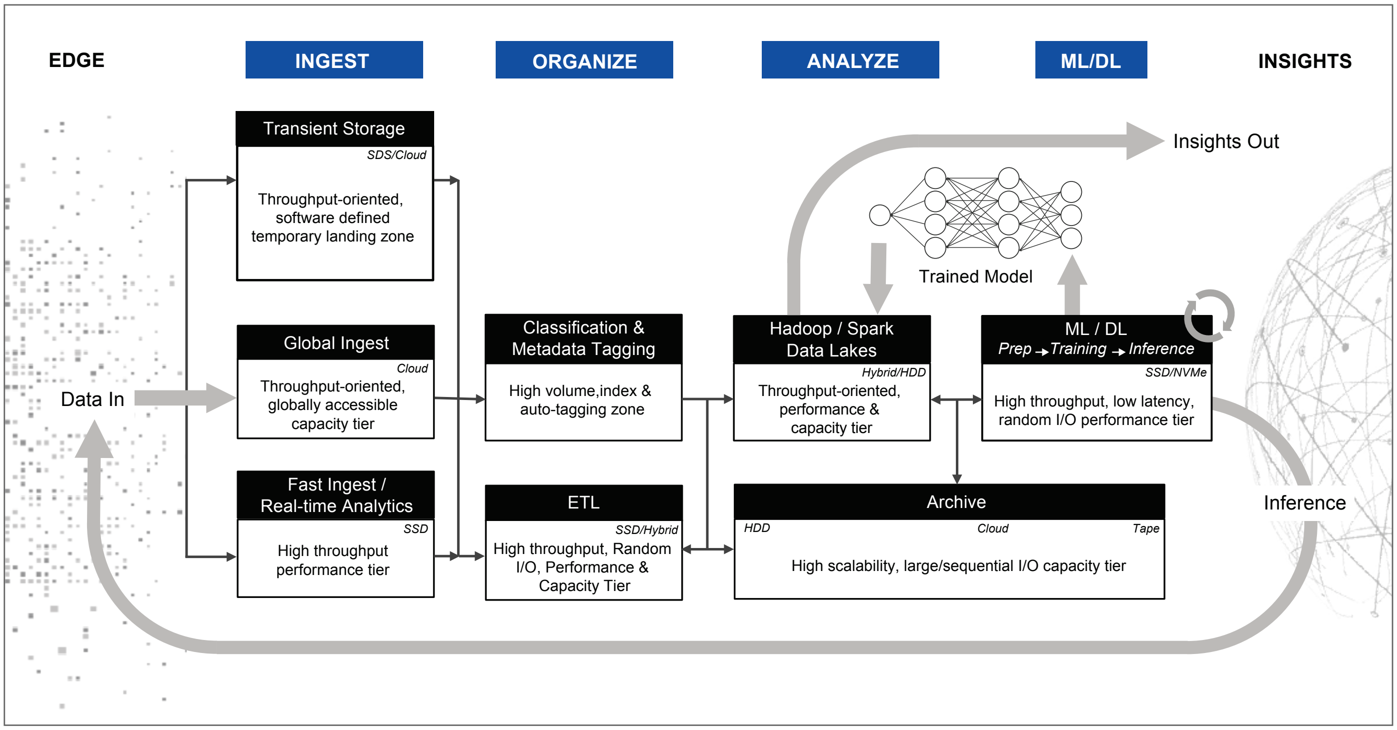
EDGE
INGEST
ORGANIZE
ANALYZE
ML/DL
INSIGHTS
Transient Storage
SDS/Cloud
Throughput-oriented, software defined temporary landing zone
Insights Out
Trained Model
Global Ingest
Cloud
Throughput-oriented, globally accessible capacity tier
Data In
Classification & Metadata Tagging
High volume,index & auto-tagging zone
Hadoop / Spark Data Lakes
Hybrid/HDD
Throughput-oriented, performance & capacity tier
ML / DL
Prep → Training → Inference
SSD/NVMe
High throughput, low latency, random I/O performance tier
Fast Ingest / Real-time Analytics
SSD
High throughput performance tier
ETL
SSD/Hybrid
High throughput, Random I/O, Performance & Capacity Tier
Archive
HDD
Cloud
Tape
High scalability, large/sequential I/O capacity tier
Inference

Figure 23*: An enterprise data pipeline with storage requirements*

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

References in the publication to IBM products or services do not imply that IBM intends to make them available in all countries in the IBM operates.

Please recycle