SOLUTION BRIEF

# Domino Data Lab Enterprise MLOps Platform

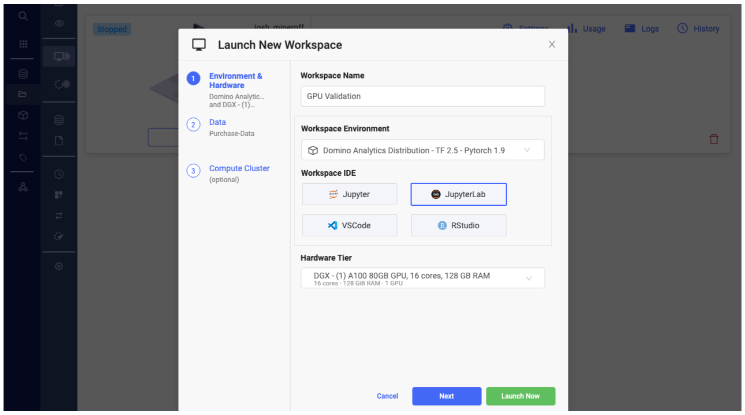
A single portal for all your data science infrastructure, tools, and assets

## Unleash Data Science

Be it lack of access to tools and infrastructure, silos that inhibit knowledge sharing, or complex processes to operationalize workflows, there are many barriers to becoming a model-driven enterprise. For companies with large teams of code-first data scientists, Domino Data Lab' Enterprise MLOps platform accelerates the process of developing and productionizing data science work by removing infrastructure friction in an open, collaborative, governed environment.  Trusted by 20% of Fortune 100 companies like Johnson & Johnson and Lockheed Martin,  Domino eliminates roadblocks to innovation.

## Integrated MLOps & GPU Solution

Under the hood, Domino automates the DevOps activities required to optimize utilization of the powerful NVIDIA DGX hardware on NetApp® ONTAP® AI systems, eliminating the low-value configuration tasks performed by valuable researchers.  Domino provides flexible, governed access to critical GPU resources, and it blends these workloads seamlessly with traditional infrastructure across a single system of record.  Data scientists have self-service access to custom GPU-enabled resources from their Domino workbench, with governance and access control that satisfy the strict requirements of Enterprise IT organizations.  All of the tools for experiment management, collaboration, reproducibility, governance, and operationalizing models are included.



Launch New Workspace

1 Environment & Hardware
Domino Analytic... and DGX - (1)...

2 Data
Purchase-Data

3 Compute Cluster
(optional)

Workspace Name: GPU Validation

Workspace Environment: Domino Analytics Distribution - TF 2.5 - Pytorch 1.9

Workspace IDE: Jupyter | JupyterLab (selected) | VSCode | RStudio

Hardware Tier: DGX - (1) A100 80GB GPU, 16 cores, 128 GB RAM
16 cores · 128 GiB RAM · 1 GPU

***Figure 1:*** *Choose the right hardware and software for any task. Use Domino Workspaces to run proprietary and open source tools side by side, governing scalable GPU Hardware Tier access with Domino's Compute Grid. Here, a JupyterLab IDE is configured to launch with access to a single NVIDIA A100 Tensor Core GPU from a DGX A100 with PyTorch 1.9 and TensorFlow 2.5.*

## Orchestrate resources from the workbench to production

The Domino platform, together with NetApp® ONTAP® AI proven architecture, supports open, collaborative, reproducible research free of DevOps constraints, and is powered by fast, efficient end-to-end compute. Streamlined use of accelerated DGX compute, accessible in the Domino platform, allows data scientists to focus on model work while IT supports from a single pane of glass.

### Get Started Faster with the Right Tools

NetApp ONTAP AI, powered by NVIDIA DGX servers and NetApp cloud-connected all-flash storage, is one of the first converged infrastructure stacks, built to help companies fully realize the promise of AI and deep learning. Streamline configuration and deployment of your data science stack with Domino's Enterprise MLOps Platform on ONTAP AI infrastructure, reducing risk and eliminating infrastructure silos with an optimized, flexible, validated solution.

### Provide Self-Serve Access to GPUs

NVIDIA DGX systems are easily accessible via Domino so data scientists can focus on critical work, and IT teams can eliminate infrastructure configuration and debugging.  DGX resources can be configured within the Domino Compute Grid, rather than depending on IT for one-off tools and deployments, reducing data scientist time spent on DevOps.

### Scale Across Multi-node GPU Hardware

Setting up a multi-node cluster for a deep learning or training job can be so hard that many teams opt to leave these resources in place. Domino enables the automatic creation and management of multi-node clusters, releasing them when training is done— eliminating dedicated resources and low utilization. Domino supports ephemeral clusters using Spark, Ray, and Dask.

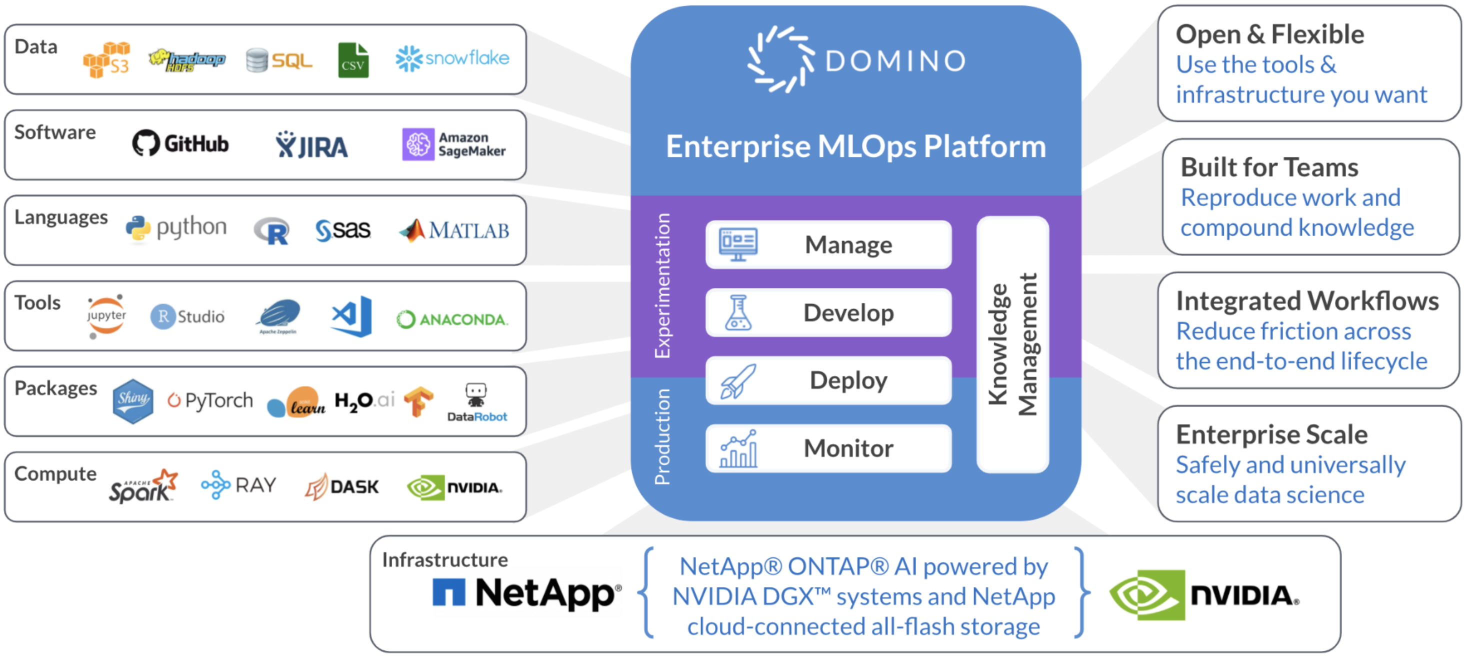
### Govern Utilization of GPUs by Role

Domino gives IT visibility into who is accessing GPU hardware and how it's being used.  Permissions can be set to ensure employees are not burning through valuable resources, while power users have full access to maximize the use of hardware. Assign access by role (i.e., according to development, validation, or production functions) with different corresponding hardware available.

For example, 1 or 2 GPUs can be allocated to users for basic research, while 4 or 8-GPU configurations can be used for training. With NVIDIA Multi-Instance GPU (MIG) technology in the NVIDIA  A100 Tensor Core GPU,  a single DGX A100 can support up to 56 concurrent notebooks or hosted models with independent GPU instances.

Dominodatalab.com

## Domino supports the broadest ecosystem of tools and infrastructure



**Open & Flexible**
Use the tools & infrastructure you want

**Built for Teams**
Reproduce work and compound knowledge

**Integrated Workflows**
Reduce friction across the end-to-end lifecycle

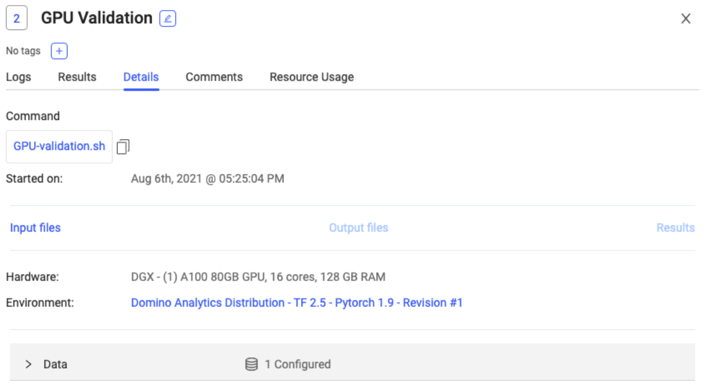
**Enterprise Scale**
Safely and universally scale data science

NetApp® ONTAP® AI powered by NVIDIA DGX™ systems and NetApp cloud-connected all-flash storage

## Accelerate value from your AI infrastructure

Domino accelerates data science initiatives—at velocity and scale—by complementing NetApp ONTAP AI with the best-in-class Enterprise MLOps platform. ONTAP AI simplifies, scales, and integrates your data pipeline for deep learning, consolidating a data center's worth of analytics, training, and inference compute into a single, 5-petaflop AI system.

Consolidated and centralized support with version control for data science workspaces ensures stable and consistent access to cutting-edge deep learning compute and frameworks such as Keras, TensorFlow, Torch, and TensorRT. That way, when data science teams deploy their on-demand notebooks, they select tailored DGX resources and software for their tasks with administrator controlled permissions. And, the compute environment is fully traceable. The Domino open platform streamlines workflows and scales, taking full advantage of the power of NVIDIA DGX systems and NVIDIA NGC™ optimized containers out-of-the-box.



2 GPU Validation

No tags +

Logs | Results | Details | Comments | Resource Usage

Command
GPU-validation.sh

Started on: Aug 6th, 2021 @ 05:25:04 PM

Input files | Output files | Results

Hardware: DGX - (1) A100 80GB GPU, 16 cores, 128 GB RAM
Environment: Domino Analytics Distribution - TF 2.5 - Pytorch 1.9 - Revision #1

> Data | 1 Configured

***Figure 2:** A Domino Workspace is used to validate in-notebook access to a single NVIDIA A100 Tensor Core GPU from a DGX A100 for development.*

> *"Domino makes it easy for our data scientists to rapidly access NVIDIA GPUs so we can support complex use cases like training deep neural networks."*
>
> Mike Johnson, Lead Data Scientist, Lockheed Martin

## A system-of-record for models

Unifying all data science work streams into one common platform makes collaboration effortless in Domino. Version control of all models, tools, and environments is automated. Collaborators see their team's work, then quickly access and fork prior works to progress research, building upon the efforts of their peers. Domino is the system of record for all data science efforts that helps align IT and data science teams and standardize on best practices—one system lets organizations manage and organize all data science work.

## Learn More

dominodatalab.com/partners/netapp

Domino *Website:* dominodatalab.com
*Blog:* blog.dominodatalab.com
*Try:* dominodatalab.com/trial

NETAPP *Website:* netapp.com/artificial-intelligence/
*Blog:* netapp.com/blog

Dominodatalab.com

## Domino, NVIDIA, & NetApp Integrated Solution

Preconfigured offering from Domino, NetApp, & NVIDIA provides optimal performance through field-proven architecture including the best-of-breed, purpose-built AI hardware validated with Domino's Enterprise MLOPs Platform. The integrated solution is available through Arrow's network of resellers, including Mark III Systems, Insight, WWT, CDW, and more.

**Single PO**

*All-in-one solution from NVIDIA & NetApp*

**No-risk Installation**

*Pre-configured, validated, and installed*

**Effortless Support**

*Simple and centralized support with a single phone number*

**Faster ROI**

*Enables accelerated time to insights - unleash you data scientists!*

## Matching Infrastructure To Workloads

*Our bundled solutions are a great fit for many different use cases, allowing customers to scale as needs grow! Key questions to consider are...*

1. How many data scientists do you have running AI/ML/Data Science Jobs?
2. How many GPU jobs are you running?
3. How are you planning on intending to grow?
4. What kind of storage capacity is required in the future?

| | **SMALL**<br>2xNVIDIA DGX A100,<br>1xNetApp A400 | **MEDIUM**<br>4xNVIDIA DGX A100,<br>1xNetApp A700 | **LARGE**<br>8xNVIDIA DGX A100,<br>1xNetApp A800 |
|---|---|---|---|
| **Use Cases** | Natural Language Processing, Imaging / Pathology, Healthcare Research, Risk Management, Fraud Detection, Preventive Maintenance, Supply Chain Optimization, Inspection, Autonomous Systems, ... | | |
| **Config best suited for** | AI, ML/DL, Training, Inference, Entry-level Training, Inference | AI, ML/DL, Training, Inference, Medium-level Training, Inference | AI, ML/DL, Training, Inference, Large-level Training, Inference |
| **Designed for a Data Science Team of** | Up to 15 named Domino users | Up to 30 named Domino users | Up to 60 named Domino users |
| **Concurrent GPU Jobs (max)** | 112 per configuration, scales to 224 per AFF A400 | 224 per configuration, scales to 448 per AFF A700 | 448 per configuration and per AFF A800 |
| **GPU performance** | 10 petaFLOPS | 20 petaFLOPS | 40 petaFLOPS |
| **I/O throughput** | 11 GB/s per AFF A400 | 17 GB/s per AFF A700 | 25 GB/s per AFF A800 |
| **Scalability** | Plans to scale by 1-2x DGX at a time; Effective storage capacity: 65TB, 131TB and expansion to 702.7PB | Plans to scale by 1-4x DGX at a time; Effective storage capacity: 131TB, 263TB and expansion to 702.7PB | Plans to scale by 1-8x DGX at a time; Effective storage capacity: 263TB, 526TB and expansion to 316.3PB |

* Dependent on # of GPUs per job and # of jobs per user

Dominodatalab.com

DOMINO NetApp® NVIDIA.