Unleash Data Science with Infrastructure Virtualization

Domino Data Lab Enterprise MLOps Platform and VMware Tanzu®

With Domino Data Lab and VMware Tanzu, code-first data science teams can accelerate research, increase collaboration, and deploy models across an optimized multi-cloud infrastructure.

Domino, in close collaboration with VMware Tanzu, supports open, collaborative, reproducible model development, training, and management free of DevOps constraints - powered by efficient, end-to-end compute.

VMware Tanzu® DOMINO

"We've implemented a multi-cloud strategy, along with Domino's Enterprise MLOps platform, to increase model velocity so we can address customer needs in a quarter of the time it used to take us."

Antione Ly
Head, Data Science
SCOR

Data Scientist Challenges

- **Inflexible Infrastructure:** Data scientists need specialized tools and scalable compute, and end up resorting to Shadow IT
- **Wasted Work:** Disparate tooling and lack of standardization limits collaboration, productivity and knowledge sharing among data scientists
- **Production Pitfalls:** Managing model development and production deployment is complex, delaying business value and increasing risk with inconsistent model monitoring

Data Science Brings Promise, But Not Without Challenges

From natural language processing for model-driven policy approvals to image classification in biotechnology to supply chain risk and anomaly detection in advanced manufacturing, data science brings tremendous promise—but not without challenges. In fact, **only 21 percent** of businesses are gaining a major competitive advantage through the use of data and analytics tools[1].

Why Companies Struggle to Scale Data Science

The primary challenge lies in finding the right tech stack to scale adoption of data and analytics tools, so that **IT can efficiently manage DevOps**, while **data science teams can focus on quickly delivering business impact** through fast iteration and speedy model deployment instead of on infrastructure challenges.

Data science requires specialized environments with access to data, tools, packages, and infrastructure to handle "bursty" workloads. Without Kubernetes to containerize applications, data scientists frequently have to:

- Work with IT to install and manage new packages, or provision new software or infrastructure
- Debug code issues caused by data scientists using different environments, or

[1] Althority article about DatalQ survey sponsored by Domino Data Lab, Feb. 2021

"The [Domino Data Lab] platform brings order to the chaos."

Matt Seaman
Director and Chief Data and Analytics Officer for Enterprise Operations
Lockheed Martin

Enterprise MLOps Benefits

- **Data scientists** spend less time on overcoming DevOps challenges, with self-serve access to their preferred languages (including Python, R, SAS, and MATLAB), IDEs, packages, and infrastructure.
- **Data science** leaders give their team an easy way to find, reproduce, and reuse past work so there's less "reinventing the wheel."
- **IT** securely provides the tools, data sources, and compute resources that data scientists need – now, and in the future, with increased governance over compute spend and costs.

- Wait unnecessarily long to onboard new team members due to environment, infrastructure, or data access issues
- Spend time updating old projects that have been rendered unusable due to environment changes, or a lack of collaboration and knowledge sharing
- Take manual steps across the data science lifecycle to operationalize and monitor models in production

Tailoring Kubernetes for Data Scientists

While Kubernetes is a great foundation for tool agility, faster iterations, and reproducibility within a data science platform, data scientists shouldn't have to become Kubernetes experts to realize its promise.

The Enterprise Machine Learning Operations (MLOps) software from Domino works together with VMware Tanzu portfolio products, such as VMware Tanzu® for Kubernetes Operations, to provide many benefits:

Usable interface for data scientists

Data scientists often come from different backgrounds than engineers and DevOps folks, and can require an intuitive interface and APIs that abstract some Kubernetes concepts.

Containerized environment management

Achieve reproducibility by allowing data scientists to create, update, and manage environment images that will be used for data science workspaces.

Integrated user management and permissions

Kubernetes provides a set of authorization primitives, but unique considerations arise when providing true user isolation for co-located workloads and managing sensitive information (e.g., access credentials for data connections).

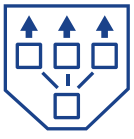
Data science-specific scheduling

Data science workloads often require complex, multi-pod workloads, which can be a challenge with Kubernetes alone.

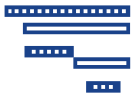
Resource controls

Kubernetes needs to be augmented so that administrators can balance user access to scalable compute resources, and to exercise sufficient controls to manage costs and prevent one user from monopolizing capacity.

Intuitive API



Containerized environments



Data science-specific scheduling

Domino Data Lab Enterprise MLOps Platform

Trusted by over 20 percent of Fortune 100 companies, Domino's platform helps customers overcome the biggest challenges to data science at scale—infrastructure friction, productionization challenges, and a lack of collaboration.

Figure 1: Domino Data Lab Platform Overview

Domino's secure, scalable enterprise MLOps platform gives data science teams a system of record to increase productivity through compounding knowledge and to make work reproducible and reusable. It's an integrated model factory that lets you develop, deploy, and monitor models in one place using your preferred tools and languages, and a self-service infrastructure portal providing one-click, governed access to the data, tools, and compute you need.

Ideal Use Cases – Domino Data Lab & VMware Tanzu

With Domino Data Lab and VMware Tanzu, code-first data science teams can accelerate research, increase collaboration, and deploy models across an optimized multi-cloud infrastructure – all aimed at building intelligent applications that truly drive enterprise value.

There are several key industry use cases where the combined portfolio could be well-applied.

Intelligent Edge & Streaming Applications

Domino Data Lab's MLOps Platform, combined with VMware Tanzu, can provide developers with an easy way to deploy models embedded in edge applications. Use cases can include predictive maintenance in energy/utilities and oil/gas, connected vehicles, smart cities, and more.

Risk Management & Regulatory Compliance

Domino's data science system-of-record capabilities paired with VMware Tanzu's consistency and resilience can streamline audit, governance, compliance, and reporting in regulated industries - from risk management and model monitoring in financial services, to clinical trials and regulatory approvals in pharmaceutical development.

GPU-Accelerated Models for R&D

Enterprises seeking to accelerate time-to-market with the latest machine learning and AI techniques can benefit from virtualized, GPU-accelerated data science workspaces with Domino Data Lab on VMware Tanzu. Use cases can include unsupervised anomaly detection in aerospace, and image

analysis for drug development, clinical trials, and diagnosis techniques in pharmaceuticals.

Edge

Risk modeling

GPU/Image Analysis

Deploying Domino Data Lab on VMware Tanzu® Kubernetes Grid™

VMware and Domino Data Lab have collaborated on a joint Reference Architecture (RA) that details how to deploy the Domino Data Lab platform on Tanzu Kubernetes Grid (TKG). The reference document covers details such as an architecture overview, bill of materials required, and cluster requirements.

The architecture included in the RA document should give users a path to creating a highly available, production-grade deployment of Domino Data Lab on TKG. However, users should not feel constrained by this exact path if specific use cases lead them to a different deployment architecture. Design decisions in the document reflect the main design issues and the rationale behind a chosen solution path - and if necessary can help provide rationale for any deviation.

The level of availability and redundancy required by the workloads being deployed will determine the topology of the clusters, from a simpler, single cluster running in just one availability zone, to a more complex deployment of multiple clusters distributed all over the world, either on a single cloud provider, or across multiple cloud providers.

The Reference Architecture document covers the Domino Platform and Domino Compute components. It also includes the Application services and Supporting services required for these components – for example, NGINX, Fluentd, and more.

Below, we represent a high-level Architectural View covering Domino Data Lab and VMware Tanzu.

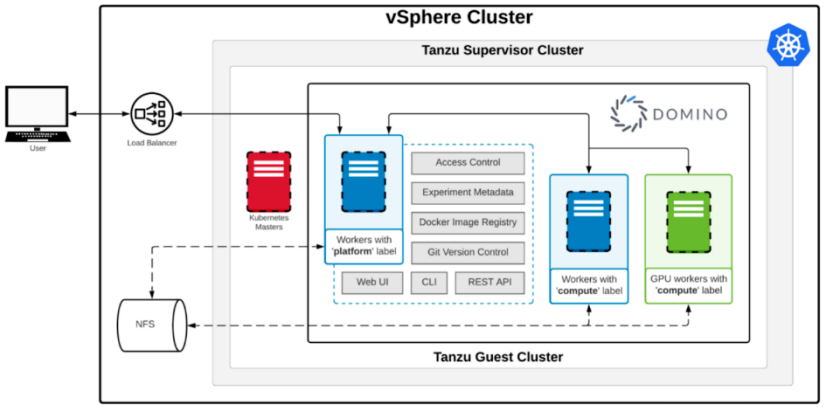


Figure 2: Architectural View - Domino Data Lab and VMware Tanzu

How to get started

For more information on using Domino Data Lab and VMware Tanzu, reach out to your Domino Data Lab sales rep or your VMware Tanzu sales rep, and we'll be in touch!