How Domino and AWS accelerate research
for world-class data science teams
Cloud computing has revolutionized enterprise infrastructure, enabling organizations of every size and stage to access elastic and scalable compute, storage, and networking resources. This movement to the cloud for both new and existing workloads has impacted every industry: finance, insurance, pharmaceuticals, retail, even government.

Amazon Web Services has emerged as a leader in public cloud computing services, providing data science teams with the resources they need, like bursts of elastic computing power and specialized resources such as GPUs, to build and train models. However, not every organization can provide their data science teams with these resources and benefits in an efficient, transparent way.

This white paper describes several workflows we have found critical to enabling cloud-based data science in an enterprise. These are challenges that any organization must solve to realize the promise of the cloud. To address them, you will need a solution — either something custom built, or an integration of several best-of-breed point tools. We also describe how Domino addresses these challenges by serving as an elegant orchestration layer on top of AWS, offering functionality tailored to data science workflows.
Introduction to Domino

Domino resides within your AWS virtual private cloud (VPC) and provides a workbench that data scientists can use to run experiments, share their work, reproduce work done by others, and deploy models. Behind the scenes, Domino automates compute elasticity, container orchestration, version control, and more. Domino reduces the need for IT to support data scientists spinning up instances by making it self-service, while giving IT the ability to monitor and create constraints when needed.
Domino scales EC2 resources to allow data scientists to run multiple experiments in parallel. All experimental results are tracked and revisioned automatically in S3.

Domino’s core functionality:

1. Compute Environment Management lets users define shared, reusable, versioned environments using Docker, with the necessary packages and configuration to run data science tasks.

2. Through a simple UI, data scientists can choose any type of EC2 instance and run code in three ways:
   - One-click access to interactive workspaces, where they can develop models (e.g., Jupyter, RStudio, SAS Studio, Matlab)
   - One-click submission of batch experiments
   - Simple job scheduling for recurring tasks like reporting or data ingestion

3. Users can run as many tasks as they wish, and Domino will elastically scale compute resources behind the scenes. This capability dramatically accelerates the pace of experimentation, allowing data scientists to discover and deliver results to the business faster without constant task-level support from IT.

4. As this happens, Domino automatically tracks the work — code, data, results, parameters — from these experiments and work sessions, and keeps it all stored centrally to facilitate sharing, discussion, and reuse.

5. Finally, Domino lets data scientists publish or deploy their work, sharing it with the business either as dashboards, scheduled reports, or APIs. By reducing the time it takes to go from a working model to model-driven business decisions, data science can truly make an impact.

These core capabilities elegantly solve many challenges enterprises face when migrating data science workloads to cloud infrastructure.
Domino's Architecture on AWS

Domino combines multiple core AWS services, industry standard open-source components, and a proprietary user experience layer to create an enterprise workbench tailored to data science workflows.

**EC2 and Docker** dynamically manage scalable compute resources to meet the demand for interactive workspaces, batch experiments, and hosts for serving production models.

**EFS and EBS** provide fast, scalable access to large file-based systems.

**S3 and git** to automatically revision and secure data science artifacts, including code, data sets, and results.

**ELB and Kubernetes**, to provide robust horizontal scalability and high availability across the architecture.

**Docker**, to provide flexible, secure, revisioned, and shareable environments for running data science experiments.
Data scientists need access to machines where they can run computationally intensive analyses without DevOps or IT bottlenecks.

Self-Service Data Science Workbench

Data Scientists often work in multiple languages like R, Python, SAS, or MATLAB, using integrated development environments (IDEs) such as Jupyter Notebooks or RStudio. They need access to machines where they can run computationally intensive analyses without DevOps or IT bottlenecks.

Note that these use cases are not all big data workflows. They often involve small or medium data sets with complex computational techniques.

The Problem

While AWS has simplified the work to spin up a cloud server for engineers, many data scientists are not well equipped to manage their infrastructure needs. For example, to spin up a server in AWS, a data scientist needs to:

1. Create an EC2 instance.
2. Choose an AMI.
3. Configure security groups.
4. SSH into the new machine to move files there.
5. Make sure the right packages are installed for data science work (e.g., Python and R packages).
6. Start up and test the server process if using a web-based tool like Jupyter. Networking and security rules may not be appropriately configured to allow access to the web server — requiring the process to be restarted from step #3.
7. Find a way to extract the work from that server when it is finished running— or forget it, putting the work at risk of being lost.

Countless tutorials exist online walking the user through a similar process. This cumbersome and time-consuming process can be a significant drain on productivity.

The above process only spins up one machine. To try this on a bigger machine, or if the machine gets terminated, a data scientist would need to repeat the entire process.
How Domino Helps

Domino automates the process of creating an EC2 instance. It enables data scientists to deploy a self-contained, reproducible Domino Compute Environment to those machines (i.e., the set of packages necessary to run advanced workloads), move their files to that machine, create all the network configuration to expose the workspace, and automatically preserve work in a central place—the Domino File Store which is backed by Amazon S3. When a data scientist wants to work in Domino, with one click 1, they select the infrastructure they want to use (EC2 instance type), and with another click 2, select any web-based IDE they want to spin up (e.g., RStudio, JupyterLab, Tensorboard, and SAS) when the data scientist is done working, Domino automatically persists their work to a central place and stops the EC2 instance 3.
Horizontal Scaling for Parallel Experimentation

Spinning up one instance is nice, but many data science workloads benefit from parallel experimentation. For example, you may have a model training task or simulation that takes an hour to run on high-end hardware and you want to try a few variations of it. Perhaps you’re testing a trading strategy and want to test it on multiple stocks, or you’re running a Monte Carlo simulation and you want to try multiple values for a random seed.

On local hardware, you would do these tasks serially, or you would compete for fixed resources with your colleagues. With the cloud, you can create multiple instances, assuming data science teams can easily do this without intensive IT intervention.

Domino makes it easy for data scientists to run parallel experimentation on elastic compute resources. From their desktop via our CLI, or from our web UI, data scientists can run as many experiments as they see fit — Domino will spin up EC2 instances for each one, spin them all down when they complete, and automatically detect and store results from each one.
Cost Controls and Usage Monitoring / Reporting

Imagine you’ve given your data scientists access to create cloud compute instances. Maybe you’ve even automated the ability for them to create new instances to make it easier for them to try bigger machines or experiment in parallel.

Now you have a new challenge: cost management. Data scientists leave machines running. When they are powerful machines, this can add up. Consider a large, high-memory instance type, with dozens of cores, a workhorse for data science workflows. At ~$4/hour, it’s $100/day; leave it running for three months and that’s a bill for almost $200,000.

If you have automated the process of creating new machines, this problem gets worse: now one person can create multiple machines, which they may intentionally or accidentally keep running for an extended period.

Domino provides IT with the visibility and management to curtail skyrocketing costs and contain waste, helping keep your organization’s public cloud investments under control. Domino’s features allow IT to reduce, monitor, and manage usage costs, without adding friction to valuable Data Science work. Combined, these advantages can dramatically reduce your total cost of ownership of cloud infrastructure.
Automatic Shutdown of Unused Resources

Domino provides multiple features that reduce overall costs of cloud infrastructure.

1. Domino automatically spins down instances once a researcher’s script or interactive work (e.g., in Jupyter) is done executing.

2. If there are long-running jobs (e.g., 12+ hours), Domino sends emails automatically alerting the researcher in case this was unintentional. Domino can also be configured to stop sessions automatically.

Monitoring and Transparency of Resource Use

The Admin Dashboard shows the current state of the system. It is like the EC2 Console but provides much more context about the work happening on the machine.

With Domino, you can see which user started up an environment, which project and what code they’re running on it, and even whether the machine’s CPU is idle or still working.

From here, administrators can easily detect and, if appropriate, kill runaway jobs or irresponsible resource use.
Resource Limits

Domino makes it easy to see who's using what resources and apply limits, create special-purpose queues, and prioritize work.

As a system administrator, it's essential to know how your cloud infrastructure is used and to prevent runaway costs. Because Domino acts as the orchestration layer that automates elastic compute resources, it can provide an elegant, integrated solution to these challenges.

Domino's Administration UI provides a single interface to limit and monitor machine utilization. For example, it's easy to:

- Configure a maximum number of instances to be created at a time for any given class of EC2 Instance Type (e.g., r4.4xlarge). If users request more than this number, their workloads will queue until other work winds down.
- Configure how much time a machine should stay idle before Domino automatically stops and terminates it, striking a balance between the time users will wait for their tasks to begin running and overall cost management.

Cost Attribution and Chargebacks

Since cloud infrastructure is often a shared resource for multiple departments, IT leaders need to be able to attribute fees from their cloud provider back to a specific group. Domino makes this easy in two ways:

1. Download reports showing compute time, hardware tiers, and cost breakdowns by user, project, and time period.

2. Domino offers first-class integration with EC2 instance tagging. IT teams who use tags to generate cost reports through other systems can assign custom tags to instances that Domino creates on behalf of their researchers.
Environment Management and Container Orchestration

To stay on the cutting edge of data science techniques, data scientists want easy access to the latest open source packages. This desire poses a challenge for IT administrators.

On the one hand, if data scientists are continually installing or upgrading tools on their machines, they will waste a lot of their time on properly setting up their environments. Worse, there is no guarantee that their work can be used in the future or by their colleagues, because environment dependencies may not be captured and preserved alongside the work.

On the other hand, IT admins and operators need to ensure environments are consistent, stable and reliable. These valid but seemingly opposing desires can create a bottleneck for experimentation and drain team resources.

On top of that, data scientists face challenges reproducing and reusing past work built for older machine images that may no longer exist.

How Does Domino Help

Domino’s Compute Environment Management capabilities let you create reusable, shareable, and revisioned environments using Docker images under the hood. When Domino spins up new machines to run experiments and interactive workspaces as described above, it also handles Docker container orchestration, moving the appropriate Compute Environment to the underlying machine before running the user’s code.

As Docker containers, these Compute Environments can be modified without changing the underlying machine images, ensuring that other users are not affected and systems will not be destabilized.

Better still, support teams and researchers can always reproduce their past work and get up and running quickly, avoiding environment setup challenges that typically plague data science teams.
**Turnkey GPU Support**

Despite demonstrated performance benefits of GPU-enhanced libraries and modeling techniques, when a cloud provider offers GPU instances through EC2, there are hurdles to using them productively for data science workloads.

Installing and configuring the right hardware drivers and properly configuring data science packages can distract researchers from higher-value work.

Building on top of the Compute Environment, Domino provides Docker images pre-configured with drivers and packages to work with modern GPU hardware. With two clicks, data scientists can spin up state-of-the-art GPU machines that are ready to use for their real workloads. Once infeasible use cases like machine vision are now within grasp of even small data science teams without demanding IT resources.
Deploy to Production at Scale

Data science projects often stall when moving to production. Models are frequently translated from R and Python into “enterprise languages” like Java or .NET, delaying business impact by months.

Domino on the cloud can drastically reduce the friction around model deployment supporting web-scale traffic.

Domino provides a scalable model deployment platform, backed by Kubernetes. With one click, a data scientist can deploy a model as a production-grade API so business applications can easily consume it. Domino handles horizontal scalability, high-availability, security, and other concerns that should not trouble IT leaders.
Security: Best Practices Out of the Box

AWS already offers secure environments and certifications across multiple dimensions, including physical security, network security, staffing, and more. However, even when using a cloud provider, IT providers need to make a number of decisions and configuration choices to avoid security pitfalls.

Domino comes configured with security best practices out of the box. Our customers — financial services institutions, insurance companies, life sciences companies — all use Domino to store their most valuable intellectual property, as Domino consistently meets their stringent security requirements.

Security Features and Configurations Baked into Domino

- Encryption of data in transit throughout all parts of the architecture, including transmitting code and data from users’ machines to cloud compute resources
- Encryption of data at rest throughout all parts of the architecture
- Support for LDAP integration
- Two-factor authentication
- Containerization of users’ code so it can’t access underlying host resources
- Secure configuration management, including mechanisms for securely storing connection strings or other values necessary to connect to data sources

For more information visit dominodatalab.com