How Domino Addresses the Top 10 Questions of Data Science Platforms

Domino has built a fully open, flexible, functional data science platform for enterprises who need to scale their data science teams and make data science an organizational capability. Large leading enterprises in finance, life sciences, insurance, and technology use Domino today to power their model-driven businesses.

1. Where/how is Domino hosted?

Domino can be deployed in the Cloud or on-premise, all based on a single-code-base. The containerized architecture can be easily deployed onto any Docker-enabled infrastructure. Depending on enterprise deployment preferences, Domino can be deployed in three standard configurations:

- **Domino-hosted:** Domino is hosted in a single VPC reserved in Amazon Web Services (AWS).
- **Customer managed VPC**: Domino is hosted in customer's AWS VPC environment.
- **Bare Metal:** Domino is installed on customers' servers, wherever they are (e.g. in the enterprise data center, on another cloud provider, etc.).

2. How can Domino help me ensure that data scientists use tools and packages (open-source or proprietary) that have been approved?

Domino provide a completely open platform, on which enterprises can run any web-based data science tool (Jupyter, RStudio, SAS, H2O, Tensorflow, etc.). Administrators can control which software tools, packages, and versions are available in customer approved containerized template environments. To enable governance while enabling experimentation, administrators can assign which data scientists have access to which template environments and which environments can be used for production models. For deeper governance, one can lock down the environments (e.g. blocking certain network destinations or disallowing access to anything but the Domino APIs) from a single central node.

3. How does the platform handle the dynamic nature of data science work?

Domino provides access to governed, scalable compute with a single-click. Domino Compute Grid is a central, elastic hosting infrastructure upon which data scientists run their Docker container-based experiments and deploy their models. It supports both vertical and horizontal scalability. For vertical scalability, data scientists can easily increase the number of cores or RAM size in a drop-down menu. Similarly, specialized hardware such as GPUs are accessible with a single-click. For horizontal scalability, data scientists can run multiple experiments at the same time in the elastic infrastructure. Domino's Kubernetes based infrastructure provides a future-proof approach to accommodate future data science needs. With Domino's access to all the tools and compute resources they need, data scientists no longer resort to shadow IT (their own laptops) and the overall platform maintains high levels of user adoption.

4. How does the platform handle user security and increasingly complex governance requirements where data scientists have access to highly sensitive data?

Domino can support SAML 2.0 SSO (Single Sign-On) identity providers to meet customers' specific user authentication and security needs. Additionally, Domino provides capabilities to handle unique security aspects of models. Specifically, Domino's Reproducibility Engine captures every aspect of a model development (code, data, environments, tools, packages, parameters, and results) to enable full audibility of any project and user. Furthermore, administrators can assign users to roles and govern data science specific requirements: template environments, data, packages, and model projects. This ensures administrators can isolate the work of different teams according to their organizations and further assign the hardware tiers and environments to teams according to their budget and needs.

5. How does the platform help reduce regulatory and operational risks and help future-proof me from upcoming regulatory hurdles?

Domino provides capabilities for the detailed tracking of model provenance, creating a true system of record. All revisions of a project (code, data, environments, comments, software packages, parameters, and results) are preserved and tracked in the Knowledge Center. This unique tracking of each experiment is only available due to Domino's patented Reproducibility Engine, developed specifically for the needs of data science. This set of reproducible information eases compliance or regulatory needs, creating a full provenance of any model or project. Furthermore, this level of detail enables data scientists to go back in time on any model project, fork it to a new project, and create new models from previous building blocks - enabling the organization to adapt to new regulatory or market needs, without doing rework.

6. How does Domino compare/complement tools like Git and JIRA?

Domino was developed from the ground-up, specifically around the unique characteristics of research-based model development and deployment, while Git/JIRA were developed for linear software development processes. Domino provides a system-of-record (SOR) for models, including preserving every step of the process in the patent-pending Reproducibility Engine. Domino integrates with Git so that it can act as a storage location for final code in a model. But, models require more than code, so enterprises use Domino as the Model system of record. Furthermore, data scientists have shunned Git or Jira as a place to do data science work since it does not provide the tools necessary to make them successful, while Domino has become the de facto place for data science work in model-driven enterprises. This is because Domino includes capabilities specific to code-first data scientists like one-click access to elastic compute, access to any data science tool, automatic experiment tracking, reproducibility, collaboration, and model lifecycle management.

7. What data does the platform provide access to? And how does it handle the data versioning requirements of data science?

Domino provides access to all the various different types of data across the enterprise. With the power of elastic compute, Domino allows data scientists to train models on datasets of all sizes. Domino comes with:

- High-speed data connectors to various sources such as Hadoop, Spark, S3, relational databases, flat files, NoSQL databases, and more.
- Automatic versioning of data in Domino Data Sets during experimentation process to preserve full reproducibility. This also provides fast access to snapshots of cleansed data to expedite the data engineering process. These data sets can be placed on-premises NFS, AWS EFS, or other cost-effective file storage.
- Big Data compute framework integration with Spark to run PySpark/SparkR jobs from Domino Executors for large data sets. Domino Executors act as edge node clients to big data environments. Domino also captures all the Spark code written and versions it in The Reproducibility Engine so users can share Spark code, work with colleagues, and preserve code for reuse.

8. How does the platform enable user-friendly, enterprise-ready model operations (ModelOps)?

Domino provides the mechanisms for data scientists to easily deploy production-grade models in the right mode for the appropriate business use case without requiring IT to rewrite model code. In Domino, models can be deployed as ad hoc reports, scheduled reports, lightweight, interactive apps (i.e. Flask or Shiny apps), or web forms. Domino can also deploy models as APIs (batch or real time) with high availability and low latency for machine consumers. Domino API deployment scales horizontally and vertically with Domino Compute Grid and has the ability to leverage whatever compute is available at that time. Once deployed, Domino provides a closed-loop workflow allowing data scientists to version models, deploy to test, and migrate to production. Lastly, Domino Model Launchpad provides a single place where all models are hosted, enabling end-users and data scientists to collaborate. This reduces friction in model iteration, model deployment, and improves the efficiency of ModelOps.

9. How does Domino help govern cloud infrastructure costs and plan for future technology needs?

Domino provides the visibility and controls necessary to ensure compute resources are properly allocated and consumed by the data science teams. With Domino Control Center, administrators can balance the flexibility demanded by data scientists with the realities of IT budgets. In particular, administrators can:

- Configure the type and number of hardware nodes available at a given time.
- Automate save and shutdown, along with notifications, of long-running jobs.
- Track and monitor usage metrics for cost attribution across departments, users, and projects to ensure spend is being utilized on high value projects.

With such visibility, IT and business line leaders can perform more accurate budgeting and planning and ensure data science projects are cost effectively using the right hardware. Additionally, data science leaders also have complete visibility into hardware costs and software tools used, so they can collaborate with IT leaders to create proper governed template environments, properly allocate compute based on project ROI, and also govern how specific software packages are used.

10. How does Domino work with traditional software development processes?

Domino was custom built for the unique characteristics of the model development lifecycle, but was also built to fit into enterprise IT processes. For example, Domino includes seamless, comprehensive integration with popular Git solutions including GitHub, GitHub Enterprise, and Bitbucket. Domino users can source and store code in a best-fit version control system without compromise. Domino will version all unique model assets like data, packages, environments, discussion threads, parameters, and results in the Reproducibility Engine. This integration allows IT departments to maintain a single place for code (Git) while leveraging Domino for model provenance. Additionally, Domino provides a workflow process to manage versioning of models, enabling one to setup and govern a process to move models from development to validation (test) to production, with full revert capabilities.

Domino is a secure, scalable, and centralized data science platform for developing, validating, delivering, and monitoring models with full auditability, governance, and transparency. Purposefully designed and developed for enterprise data science, Domino integrates into existing infrastructures, creates a system of record for models, all while accelerating data science team work.