DOMINO

Running Complex Workloads Using On-Demand GPU-Accelerated Spark/RAPIDS Clusters

Nikolay Manchev
Head of Data Science for EMEA
Domino Data Lab

DOMINO

- Founded in 2013, created the Data Science Platform category
- Provides Data Science System-of-Record for the Enterprise
- Leader in Forrester's Wave on Notebook-Based ML Solutions
- 200+ employees across SF, NYC, Chicago, London, Bangalore

> “Domino successfully enables enterprises in their industrial-strength deployments.
> The platform truly helps data science teams orchestrate and streamline the ML workflow”

Gartner 2020 Magic Quadrant for Data Science and Machine Learning Platforms

Trusted by 20% of the Fortune 100

2 of largest	**2 of top-5**	**3 of top-5**	**4 of top-10**
global banks	health insurers	ratings agencies	pharma companies

BAYER Bristol Myers Squibb™ vmware®

TRANSAMERICA® Cigna Allstate You're in good hands.

MOODY'S ANALYTICS Red Hat BNP PARIBAS

S&P Global ZURICH LLOYDS BANKING GROUP

DELL easyJet Carnival®

Unleashing data science for leading enterprises

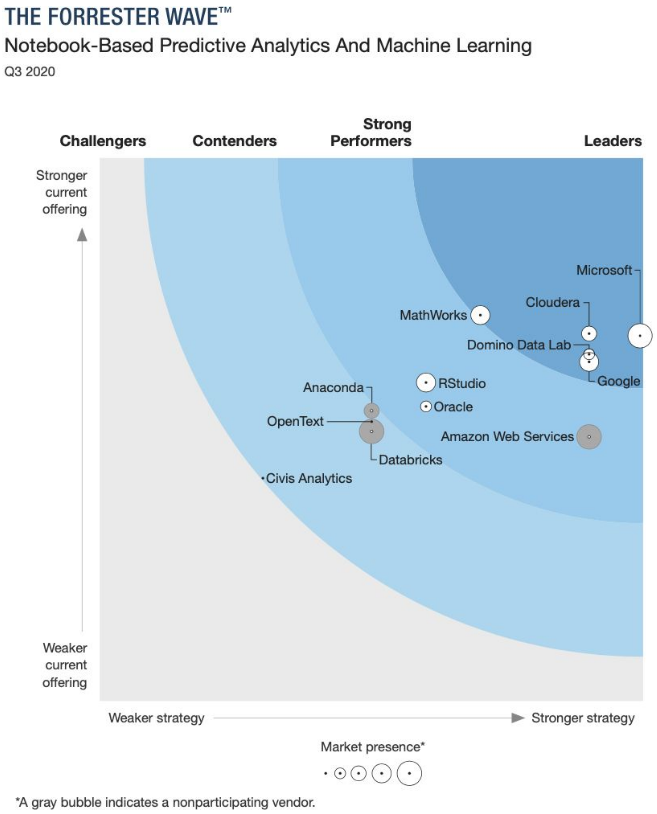
Domino is Recognized as a 'Leader' in the Industry

- Domino is the ONLY vendor to repeat as a 'Leader'
- Received the highest score in 'Model operations/ModelOps'
- Received perfect scores (5 out of 5) for:
  - 'Collaboration'
  - 'Platform Infrastructure'
  - 'Solution roadmap'
  - 'Ability to execute'
  - 'Enablement

A strong vision for what's needed today and in the future, backed by the resources to deliver on our promises and help make our customers successful

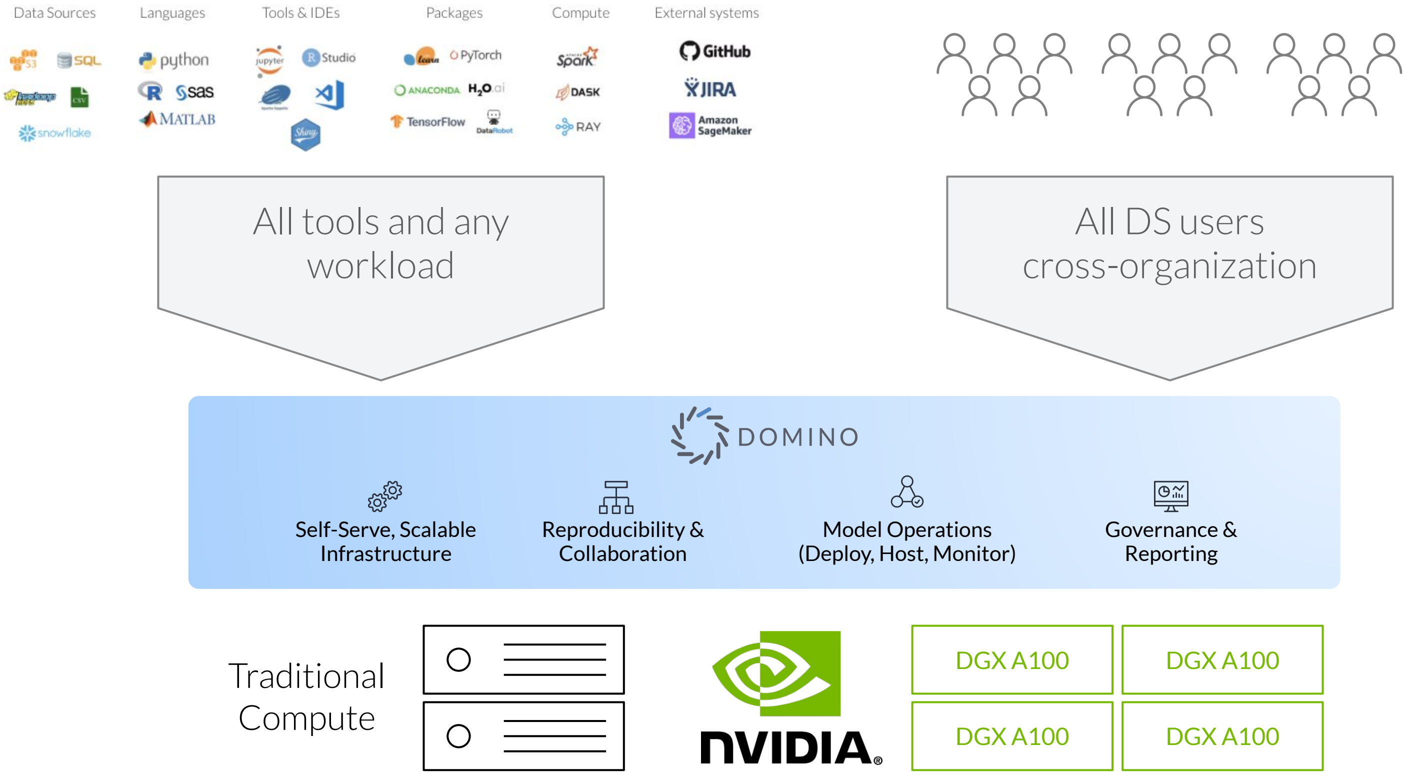
> "Domino provides an enterprise data science platform that supports the diversity of ML options that users need in today's rapidly expanding PAML ecosystem, with repeatability, discipline and governance"

FORRESTER® The Forrester Wave™: Notebook-Based Predictive Analytics And Machine Learning, Q3 2020



The only ***open*** data science platform

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



Data Sources

Languages

Tools & IDEs

Packages

Compute

External systems

All tools and any workload

All DS users cross-organization

DOMINO

Self-Serve, Scalable Infrastructure

Reproducibility & Collaboration

Model Operations (Deploy, Host, Monitor)

Governance & Reporting

Traditional Compute

DGX A100

DGX A100

DGX A100

DGX A100

Machine Learning on Spark

Distributed Machine Learning

Parallelize compute heavy workloads such as distributed training or hyper-parameter tuning

Take advantage of powerful machine learning algorithms from Spark MLlib

Interactive Exploratory Analysis

Efficiently load large data sets in distributed manner

Explore and understand the data using a familiar interface with Spark SQL

Featurization and Transformation

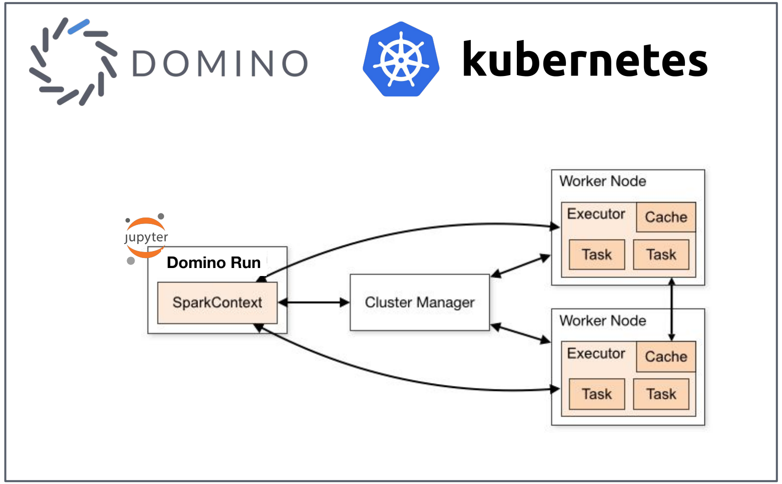
Sample, aggregate, and re-label large data sets

Optimal performance may require a practitioner who is skilled in tuning Spark

Spark on Domino

On-Demand Spark

Natively orchestrated directly on Domino controlled infrastructure



DOMINO

kubernetes

jupyter

Domino Run

SparkContext

Cluster Manager

Worker Node

Executor

Cache

Task

Task

Worker Node

Executor

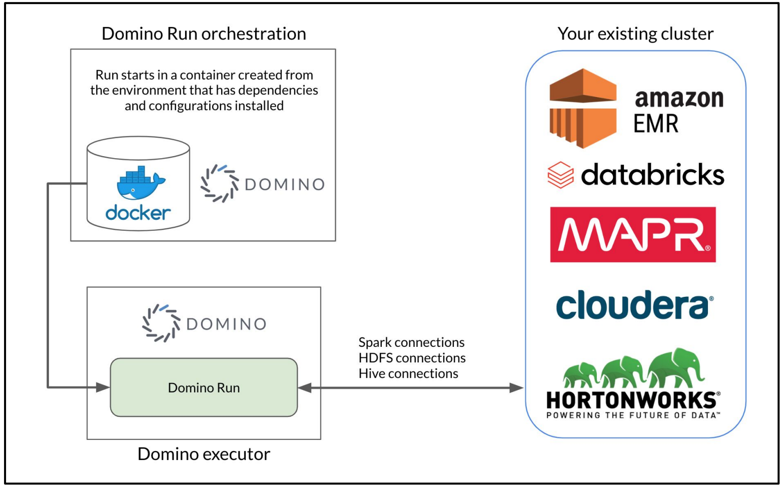
Cache

Task

Task

External Spark

Leverage existing long-lived Spark deployments in the enterprise



Domino Run orchestration

Run starts in a container created from the environment that has dependencies and configurations installed

docker

DOMINO

DOMINO

Domino Run

Domino executor

Spark connections
HDFS connections
Hive connections

Your existing cluster

amazon EMR

databricks

MAPR

cloudera

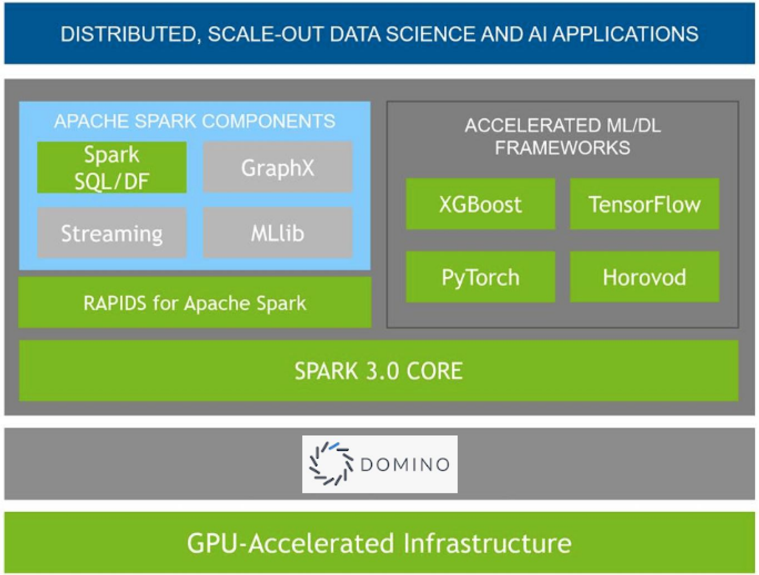
HORTONWORKS
POWERING THE FUTURE OF DATA

Faster Execution Time

Streamline Analytics to AI

Reduced Infrastructure Costs

SPARK 3.x



DISTRIBUTED, SCALE-OUT DATA SCIENCE AND AI APPLICATIONS

APACHE SPARK COMPONENTS

Spark SQL/DF

GraphX

Streaming

MLlib

RAPIDS for Apache Spark

ACCELERATED ML/DL FRAMEWORKS

XGBoost

TensorFlow

PyTorch

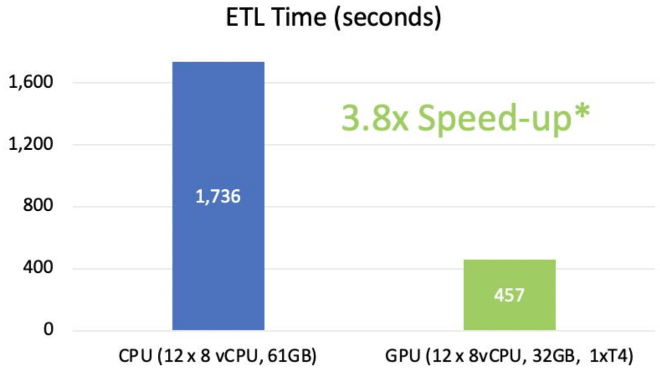
Horovod

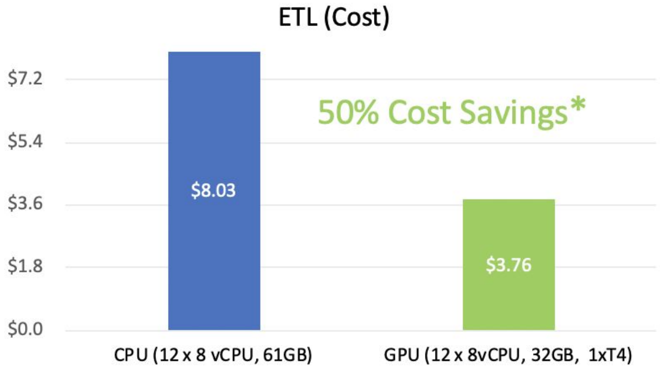
SPARK 3.0 CORE

DOMINO

GPU-Accelerated Infrastructure

RAPIDS Accelerator for Apache Spark





RAPIDS - https://github.com/rapidsai

RAPIDS Accelerator for Spark - https://github.com/NVIDIA/spark-rapids

How does it work

Apache Spark 3.0+ lets users replace the backend for SQL and DataFrame operations

RAPIDS Accelerator for Apache Spark

- replaces SQL operations with GPU accelerated versions
- fall back to Spark CPU
- GPU to GPU transfers (via UCX)

Key configuration parameters

```
spark.executor.resource.gpu.amount=1*
spark.task.resource.gpu.amount=1

spark.executor.resource.gpu.discoveryScript=./getGpusResources.sh
```

* set according to `spark.executor.cores`

Prereqs

- Spark 3.0+
- *cudf* compliant GPUs
- one GPU per executor
- *cudf* jar and RAPIDS accelerator jar available on the cluster

How does it work in Domino

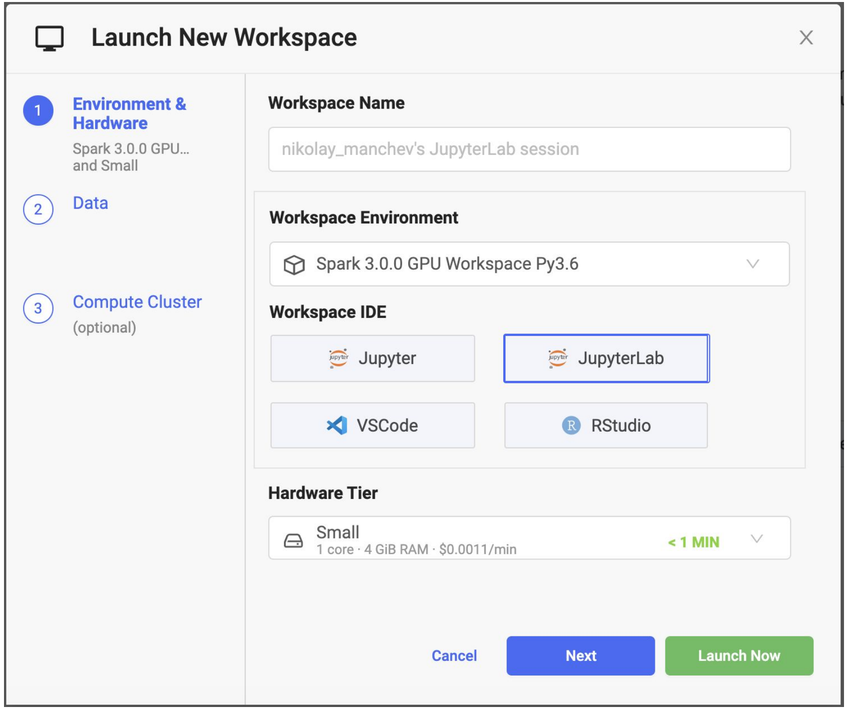
We need to configure 2 compute environments

- PySpark Workspace Compute Environment
  - Spark 3.0.0 and Hadoop 3.2.1
  - Dockerfile is available here
- Spark Executor image
  - Spark 3.0.0
  - Nvidia Cuda drivers / libraries
  - Spark RAPIDS plugin
  - GPU discovery script

GPU specific configuration (spark-defaults.conf)

```
spark.task.cpus=1
spark.task.resource.gpu.amount=0.25 # Set to 1/HW_TIER_CPUS
spark.executor.resource.gpu.amount=1 # Number of GPUs in HW Tier
...
```

How does it work in Domino



Launch New Workspace

1 Environment & Hardware
Spark 3.0.0 GPU... and Small

2 Data

3 Compute Cluster (optional)

Workspace Name

nikolay_manchev's JupyterLab session

Workspace Environment

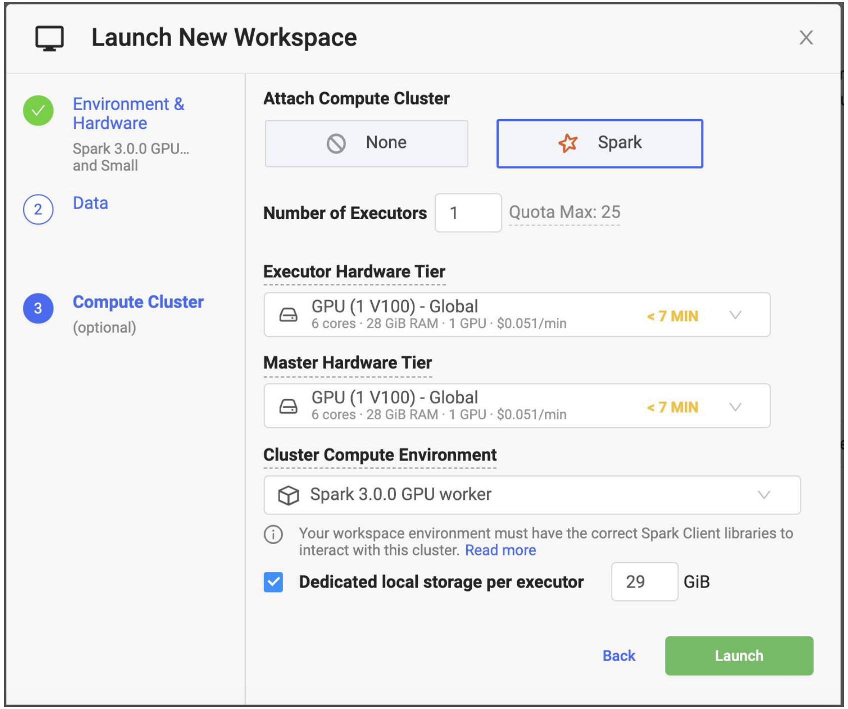
Spark 3.0.0 GPU Workspace Py3.6

Workspace IDE

Jupyter | JupyterLab | VSCode | RStudio

Hardware Tier

Small
1 core · 4 GiB RAM · $0.0011/min — < 1 MIN



Launch New Workspace

✓ Environment & Hardware
Spark 3.0.0 GPU... and Small

2 Data

3 Compute Cluster (optional)

Attach Compute Cluster

None | Spark

Number of Executors 1 Quota Max: 25

Executor Hardware Tier

GPU (1 V100) - Global
6 cores · 28 GiB RAM · 1 GPU · $0.051/min — < 7 MIN

Master Hardware Tier

GPU (1 V100) - Global
6 cores · 28 GiB RAM · 1 GPU · $0.051/min — < 7 MIN

Cluster Compute Environment

Spark 3.0.0 GPU worker

ⓘ Your workspace environment must have the correct Spark Client libraries to interact with this cluster. Read more

☑ **Dedicated local storage per executor** 29 GiB

Demo

MLOPS WITH DOMINO AND NVIDIA DGX SYSTEMS

- Effortless access to compute resources
- Simplified workflow and deployment
- Records of experiments
- Collaboration and sharing
- Model management and version control
- Utilizes GPU-accelerated containers from NVIDIA NGC
- Fully tested and certified on NVIDIA DGX systems as part of the **DGX-Ready Software program**

NVIDIA

DOMINO

https://www.dominodatalab.com/partners/nvidia/

DOMINO

Thank You

Nikolay Manchev

@nikolaymanchev
nikolay.manchev@dominodatalab.com