Technical Report

Train model with high-performance storage using Amazon FSx for NetApp ONTAP and Domino Data Lab

Max Amende and Prabu Arjunan, NetApp
January 2023 | TR-4952

Abstract

This document describes how to create a Domino Dataset using the Amazon FSx for NetApp ONTAP file storage. Data scientists can take advantage of the high-performance file storage and data management features that come with Amazon FSx for ONTAP. The integration helps data scientists quickly create clones of datasets that they can reformat, normalize, and manipulate while preserving the original “gold-source” dataset.

TABLE OF CONTENTS

LIST OF FIGURES

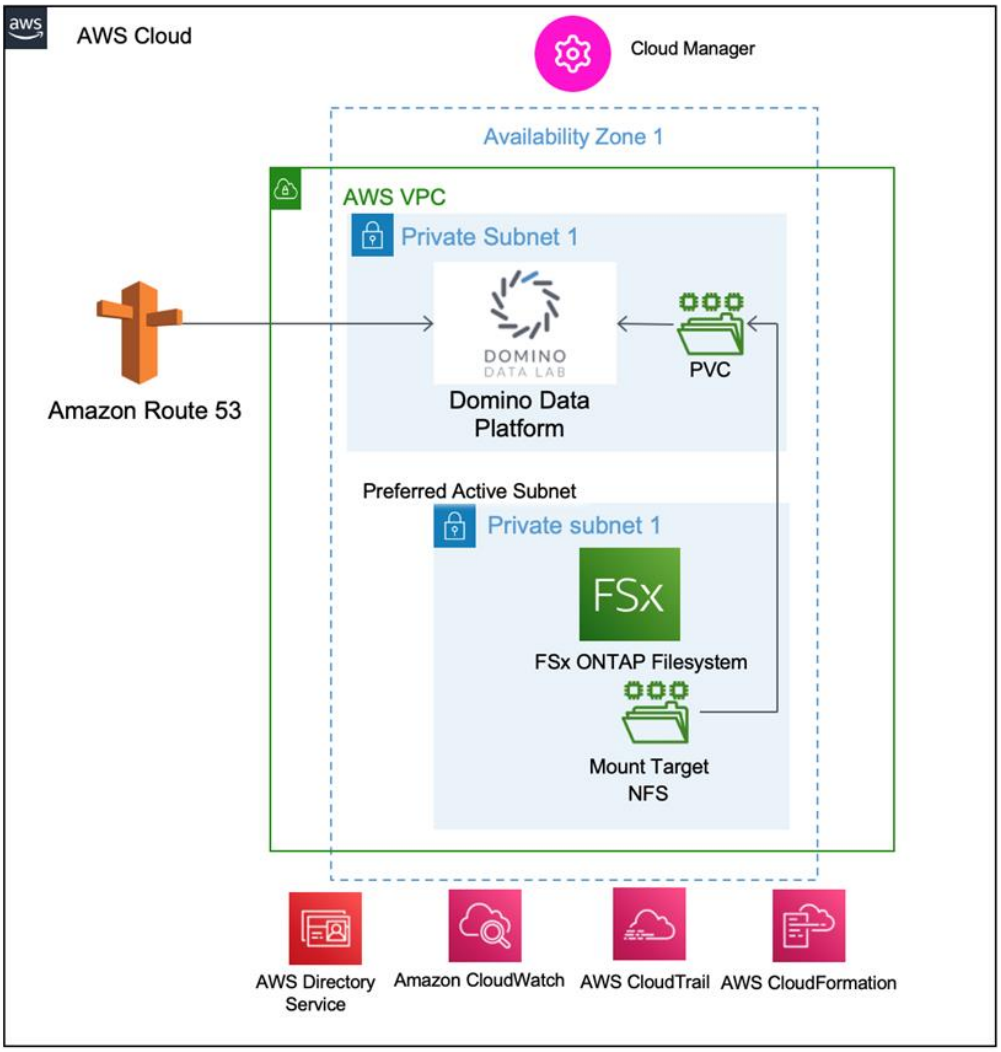
Solution overview

Domino's secure, scalable Enterprise MLOps platform gives data science teams a system of record to increase productivity through compounding knowledge and making work reproducible and reusable. It is an integrated model factory that lets you develop, deploy, and monitor models in one place using your preferred tools and languages. A self-service infrastructure portal provides one-click governed access to the data, tools, and computing you need. The Amazon FSx for NetApp ONTAP is a fully integrated managed storage built on NetApp's popular ONTAP® file system. The Domino Enterprise MLOps platform will allow businesses and organizations to run and manage artificial intelligence (AI) and machine learning (ML) and data science workloads easily on Amazon FSx for NetApp ONTAP without refactoring them.

Architecture

Figure 1 illustrates the solution architecture.

Figure 1) Domino integrating with NFS file shares running in Amazon FSx for ONTAP.



Walkthrough

This section walks you through the Domino integration with Amazon FSx for ONTAP. This walkthrough highlights two different ways of integrating the persistent volume with Domino Data Lab ML platform: the integration of launching a workspace backed by dataset created at FSx for NetApp ONTAP and how to use Amazon FSx for ONTAP volume as an external volume in the workspace.

Use cases

The solution integration from NetApp and Domino provides integration of NetApp storage, Amazon web Services (AWS) cloud services, and Domino MLOps platform. You can enjoy a seamless flow from storage to production and use the data in your ML and deep learning (DL) workflows, generating automated, reproducible pipelines that accelerate the deployment of AI in production and enable the continuous rollout of new AI services.

Accessing FSx for ONTAP managed storage from Domino workspaces

Domino offers the functionality to manage storage from Domino workspace. The workspace creation provisions and enables the FSx for ONTAP storage with the Jupyter Notebook. The option enables a data scientist to access the FSx for ONTAP managed storage from the Domino workspace.

Accessing the data hosted on FSx for ONTAP from a training job

Data scientists typically use working environments such as Jupyter Notebook for data wrangling and rapid prototyping. When it comes to training (large) models, this approach is not efficient, because you would have to block all the computing resources needed for the training during the phase of the data wrangling and rapid prototyping. Domino offers the functionality to schedule training jobs onto a compute cluster. This means that a data scientist can do all the data wrangling on a small compute instance. When it comes to training, the model or hyperparameter tuning switch to a larger more powerful instance, or even to multiple nodes.

Prerequisites

- AWS credentials that provide the necessary permissions to create the resources. In this example, we used administrator credentials.
- Amazon FSx for ONTAP file systems and volumes with valid credentials and a route to the Virtual Private Cloud (VPC) already created. To set up Amazon FSx for ONTAP, see the Getting started with Amazon FSx for NetApp ONTAP.
- AWS credentials that provide the necessary permissions to create the resources. In this example, we used administrator credentials.
- An API key and vault URL from Domino Data Lab for the cluster creation.

Note on connectivity

The most important part of this process is to ensure connectivity between the Domino Data Lab nodes and the FSx storage virtual machine (SVM) by using the appropriate VPC, subnet, route table, and security groups. There are many ways to accomplish this; however, in alignment with this guide, you will be installing in the same VPC and subnet.

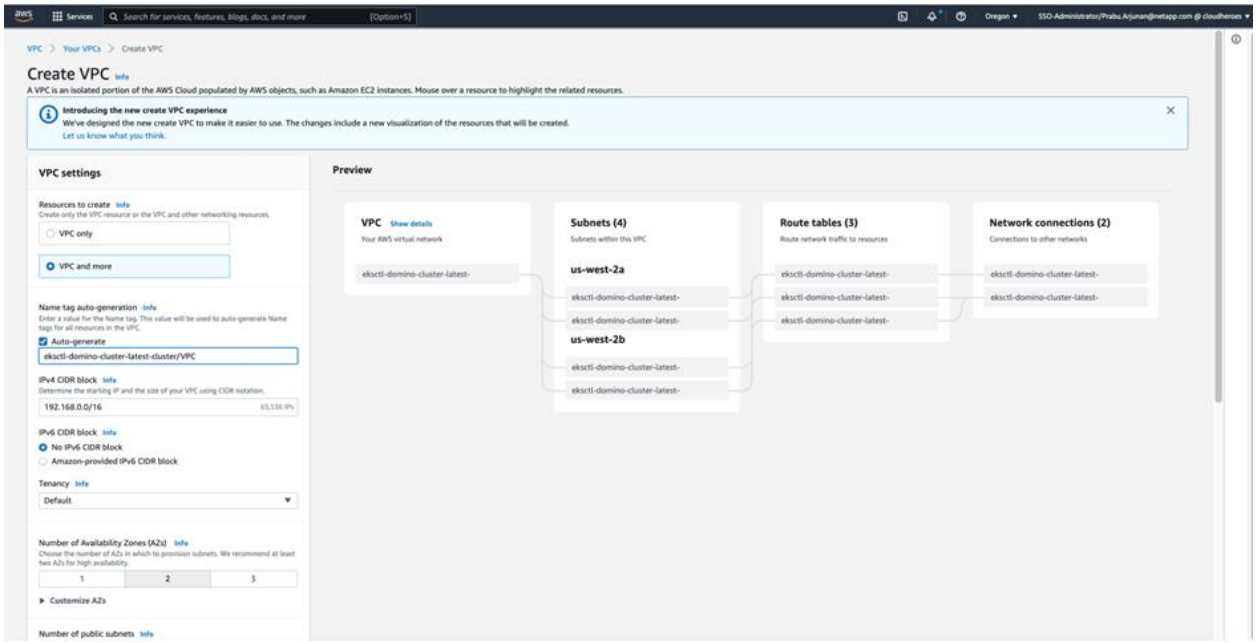
Deployment steps

Create an Amazon VPC

This step is optional because it is likely that a VPC already exists. Make sure there is connectivity between the Domino Data Lab nodes and the FSx SVM as explained in the “Note on connectivity” section.

Using this guide, you will be creating a VPC and subnets to install FSx and Domino data through the AWS console UI—but feel free to use the tool of your choice. To create a VPC, use the VPC wizard using the settings shown in Figure 2.

Figure 2) Create VPC wizard settings.



Create an Amazon FSx for NetApp ONTAP file system using standard methods

If you already have an existing FSx cluster, still read through this section to understand the necessary networking configuration regarding VPC, subnet, security group, and route table. You will need to make these configuration changes yourself.

If you do not have an existing FSx cluster, navigate to the FSx section in the AWS console and create a new ONTAP file system using the standard create, as described in Managing FSx for ONTAP file systems. Unless specified below, leave the values at their defaults:

- Creation method:
  - Creation method: Standard create
- File system details:
  - File system name: Customize for your environment(this value is optional, but recommended)
  - SSD storage capacity: at least 1024GB (FSx minimum)
- Network and security:
  - Virtual Private Cloud (VPC): The newly created VPC.
  - VPC Security Groups: Leave as the VPC default security group, but make a note of the security group for later.
  - Preferred subnet: One of the newly created subnets. Make a note of which subnet you choose as the preferred subnet; you will use it later during the Domino installation.
  - Standby subnet: One of the other newly created subnets.
  - VPC route tables: Select one or more VPC route tables. Select the corresponding route table for the newly created VPC.
- Security and encryption:
  - File system administrative password: Set a password of your choice; remember this password for later.
- Default SVM configuration:
  - SVM name: Name of your choice; you will use this value later.
  - Specify a password: Set a password of your choice; you will use this value later.

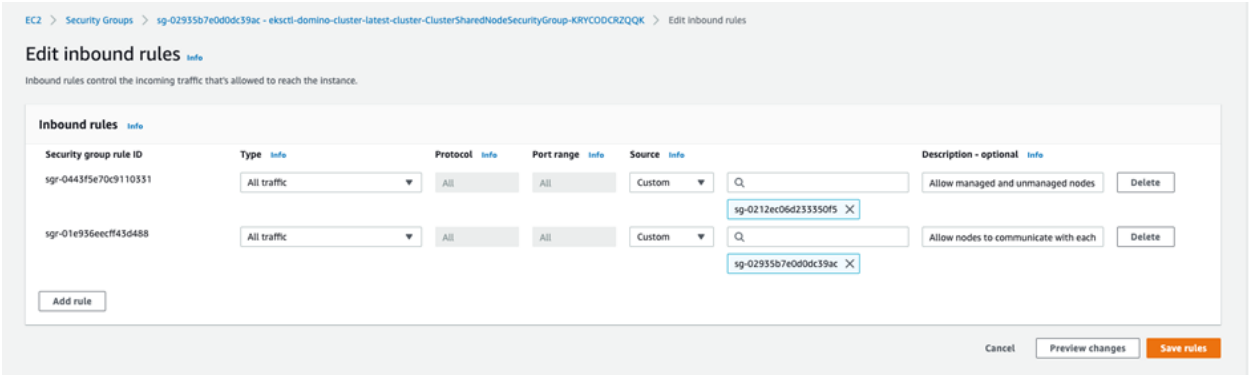
Ensure VPC connectivity to an FSx security group

As previously mentioned, there are many ways to ensure network connectivity in AWS—you must determine the best method for your needs. Using this guide, you will be installing FSx and Domino Data Lab into the same VPC and subnet.

Make sure that the VPC has access to the security group used during the FSx file system creation. One way to do this is to add an inbound rule to the security group, allowing all traffic within the VPC.

1. Get the IPv4 CIDR range for the newly created VPC.

2. Navigate to the security group used when creating the FSx file system. In this example, it is the default security group for the newly created VPC.
3. Add the following incoming rule to the security group (denoted VPC connectivity). This allows access from within the VPC.



Create a Domino cluster

Domino is a Kubernetes native platform. You can install it on many different kinds of infrastructure.

To install the Domino cluster, follow the instruction in the Domino Data Lab Installation Documentation.

Deploy a Trident operator on Domino cluster

Now that both the FSx file system and Domino cluster are deployed, connect them using the Trident operator. SSH into one of the Domino Data Lab nodes (see cluster creation output for data node IP addresses) and install the Trident operator through Helm, as described in this documentation.

You can also follow run the following commands:

```
wget https://github.com/NetApp/trident/releases/download/v22.01.1/trident-installer-
22.01.1.tar.gz

tar -xvf trident-installer-22.01.1.tar.gz

cd trident-installer/helm/

kubectl create ns trident
```

```
helm install trident trident-operator-22.01.1.tgz -n trident

cd ..
```

Define the back end and storage class for FSx

Define a Trident back end for FSx to use. Follow the full documentation for creating the back end, or use the following as a template:

```
# backend.json

{
    "version": 1,
    "storageDriverName": "ontap-nas",
    "backendName": " fsx-ontap",
    "managementLIF": "<SVM_MANAGEMENT_IP”,
    "svm": "<SVM_NAME>",
    "username": "vsadmin",
    "password": "<SVM_PASSWORD>"
}
```

The relevant fields are as follows:

- `storageDriverName`: `ontap-nas`
- `backendName`: Choose a name of your choice
- `managementLIF`: Management IP address of the SVM for your FSx file system
- `dataLIF`: Data LIF IP address of the SVM for your FSx file system (usually the same as the management LIF)
- `svm`: Name of the SVM for your FSx file system
- `username`: vsadmin
- `password`: Defined password for SVM when creating FSx file system

You also need a storage class for FSx to use. You can follow the full documentation for creating the storage class, or use the following as a template:

```
# storage-class-csi-nas.yaml

apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: trident-csi
provisioner: csi.trident.netapp.io
parameters:
  backendType: "ontap-nas"
  fsType: "ext4"
allowVolumeExpansion: True
reclaimPolicy: Retain

[ec2-user@ip-172-31-6-229 eks]$
```

The relevant fields are as follows:

- `metadata.name`: Customize for your environment, but you will be using this value later to create FSx volumes dynamically
- `parameters.backendType`: `ontap-nas`

Ensure log-in access to the SVM

To install the back end, you should be able to SSH into the SVM by using the IP address, `vsadmin` user, and the password defined during the installation process.

Run the following and enter the password defined during the installation process:

```
ssh vsadmin@SVM_MANAGEMENT_IP
```

If you are successful, continue to the “Create the Trident back end and storage class for FSx” section.

If you get a message that says: `Error: Account currently locked. Contact the storage administrator to unlock it.`, unlock the vsadmin account in order to proceed.

To unlock the account, log in to the FSX management console by running the following command and the file system administrative password defined during the file system creation:

```
ssh fsxadmin@FSX_MANAGEMENT_IP
```

1. Change the password for the vsadmin user, for example:

```
security login password -vserver <SVM_NAME> -username vsadmin
```

2. Unlock the vsadmin administrator account, for example:

```
security login unlock -vserver <SVM_NAME> -username vsadmin
```

You should now be able to SSH into the SVM by using the IP address, `vsadmin` user, and password defined during the installation process. Verify by running the command `ssh fsxadmin@FSX_MANAGEMENT_IP`.

Create the Trident back end and storage class for FSx

After your back end is defined, you can easily create it on the cluster by using `tridentctl`.

To create the back end, run the following command:

```
./tridentctl -n trident create backend -f backend.json
```

To create the storage class, run the following command:

```
kubectl create -f storage-class-basic.yaml
```

Dynamically provision the FSx volumes by creating PVC

At this point, the setup process is complete.

To follow the scope of this guide, you will now create a PVC to attach to a new Jupyter service later.

To create an FSx volume on-demand, define a PVC, for example:

```
# my-fsx-pvc.yaml

kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: basic
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 10Gi
  storageClassName: trident-csi
```

The relevant fields are as follows:

- `metadata.name`: Should match the previously created storage class (in this example, it is basic).

- `spec.accessModes`: Should be ReadWriteMany so that the PVC can be consumed by multiple sources (such as Jupyter service and training job).
- `spec.resources.requests.storage`: Size of the FSx volume to provision on-demand.
- `spec.storageClassName`: Customize for your environment; you will use this value when attaching the PVC to the Jupyter service.

3. Create the PVC, for example:

```
kubectl create -n domino-compute -f my-fsx-pvc.yaml
```

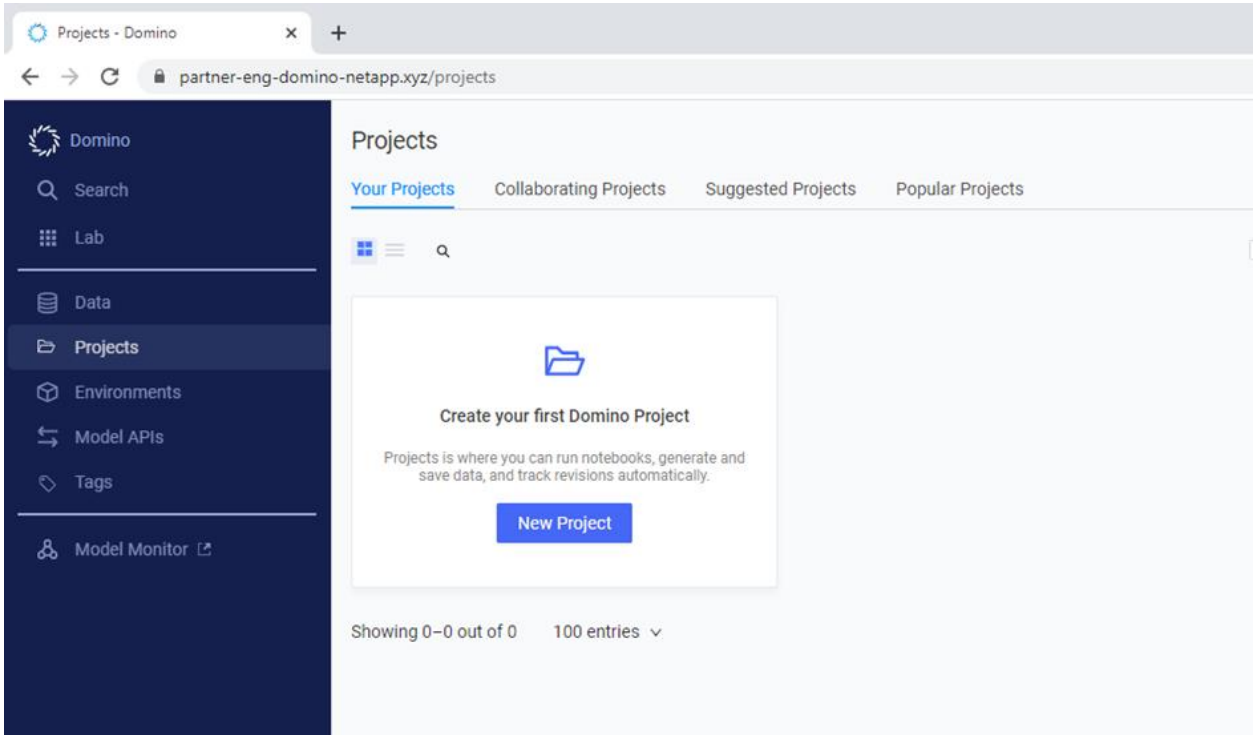
Note: The Kubernetes namespace must be `domino-compute`, because that is where the services are running in Domino Data Lab .

Access FSx for ONTAP managed storage from Domino workspaces

This section shows how data scientists can use and access the previously mounted FSx for ONTAP volumes.

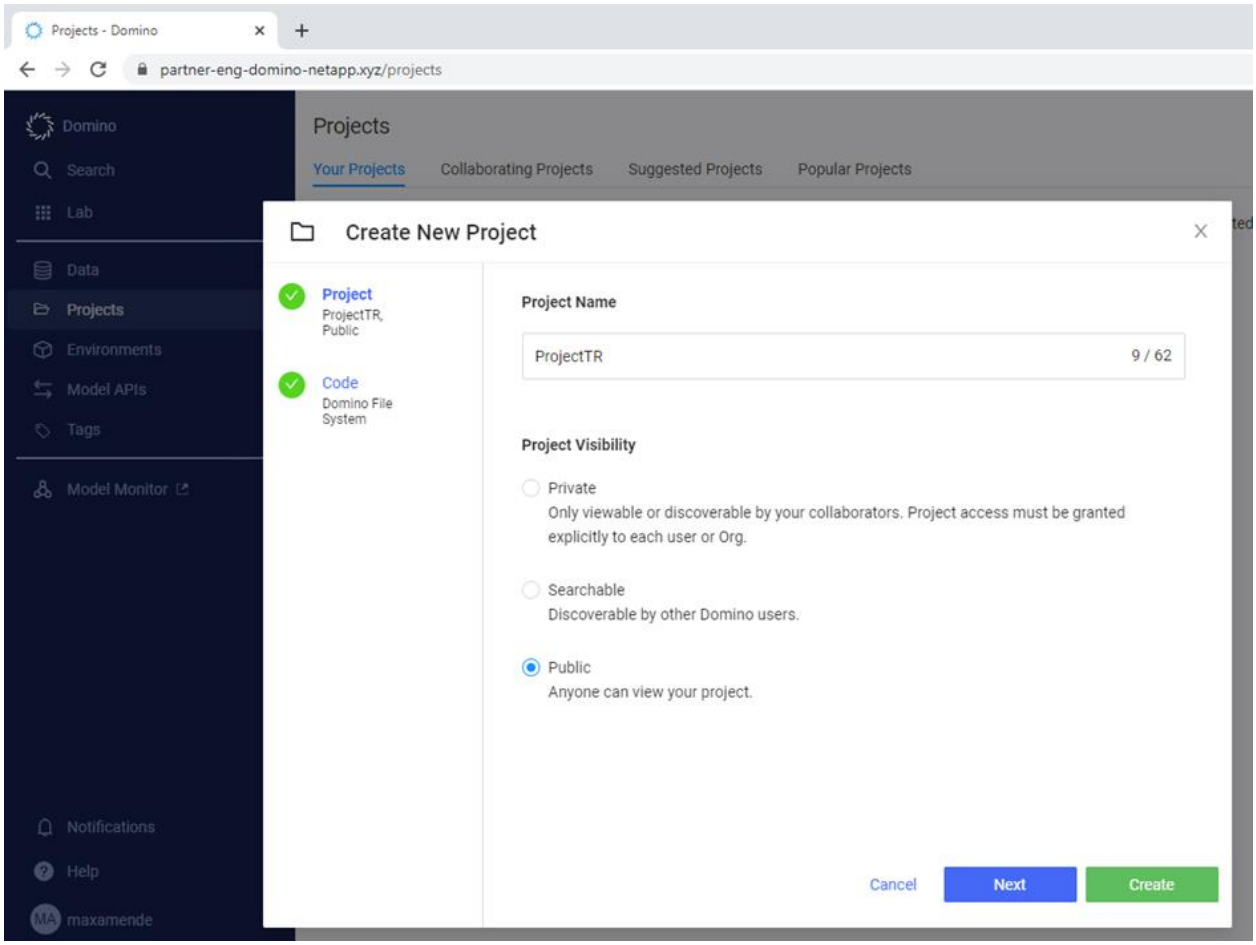
Create a project

1. Open and log in to the Domino workspace and select Projects. Click New Project.



In the opening window, provide a project name and select who should have access to the project. Click Create.

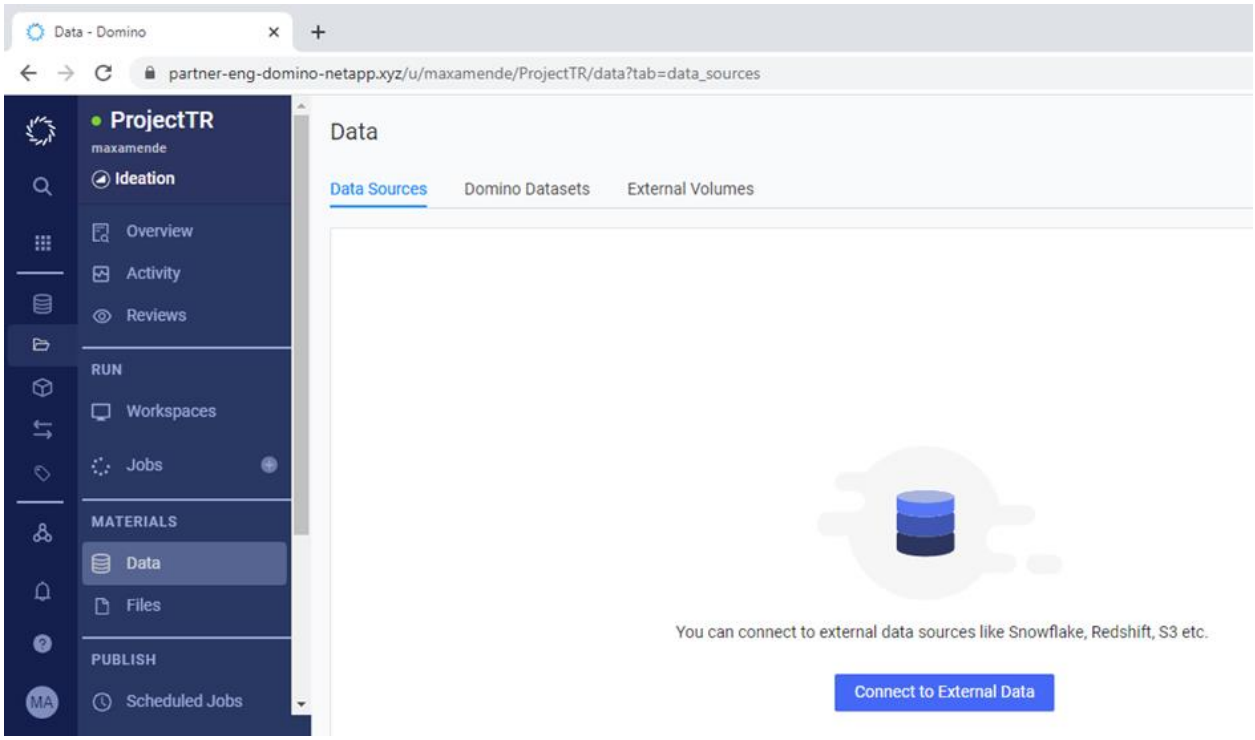
In this example, the Project is named ProjectTR and anyone who is in the same Domino environment is provided access.



At this point, Domino creates a new project space and attaches a new PVC using Trident.

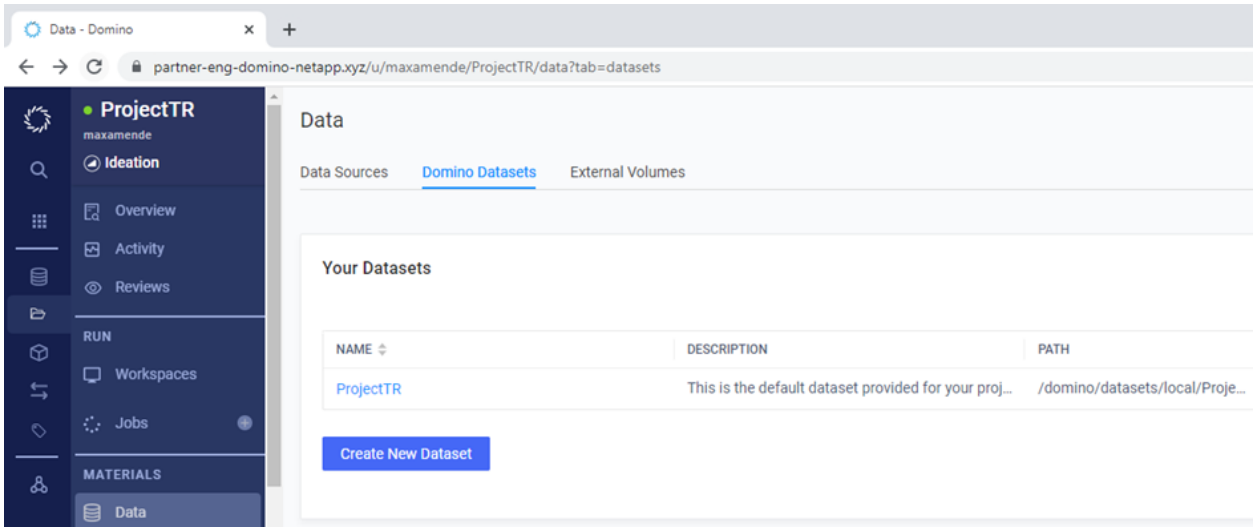
Place the demo data onto FSx for ONTAP

1. On the left side, click Data. The data source options for the project you just created are displayed.

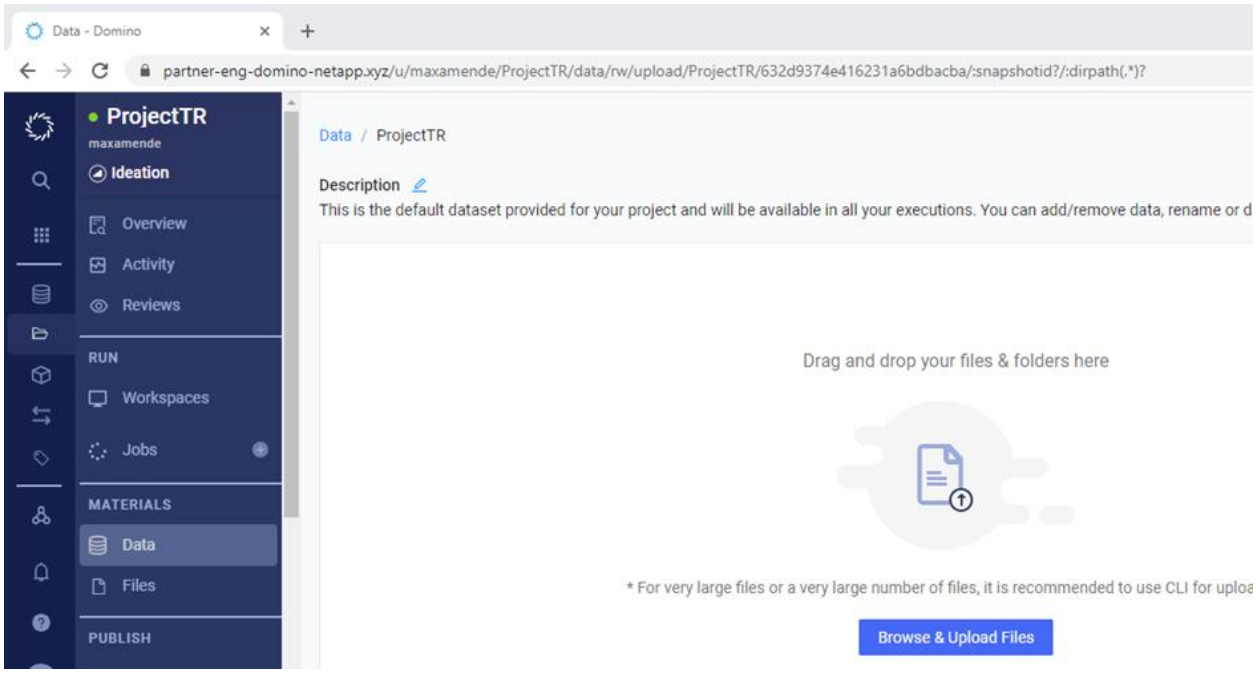


2. From the Data Sources Tab, connect your project to external data sources; for example, a NetApp StorageGRID-based S3 Bucket or a PostgreSQL Database managed by Intracluster.
3. From the External Volumes tab, mount other PVCs to your project.

   In this example, we selected the Domino Datasets tab.



4. The dataset that was created at the start of the project is displayed. This dataset is hosted on FSx for ONTAP. Click the name of the dataset. In this example, it's ProjectTR.

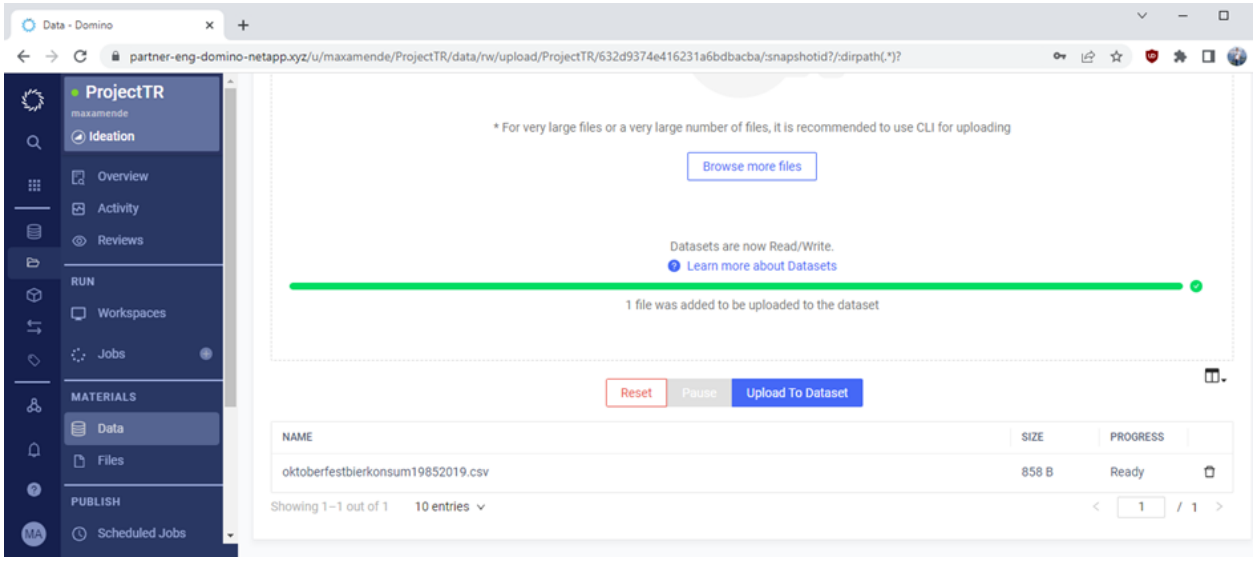


5. This gives us the option to upload data directly into the dataset. In this example, we used the following data to upload into the dataset.

   https://www.opengov-muenchen.de/dataset/8d6c8251-7956-4f92-8c96-f79106aab828/resource/56a0c3c8-c2ea-4b42-bbd2-21cb72d80803/download/oktoberfestbierkonsum19852019.csv

   Data source: `dl-de/by-2-0: Landeshauptstadt München – www.opengov-muenchen.de`

6. Download the file to your local computer, and then drag and drop the file into the field provided by Domino. Click Upload to Dataset.



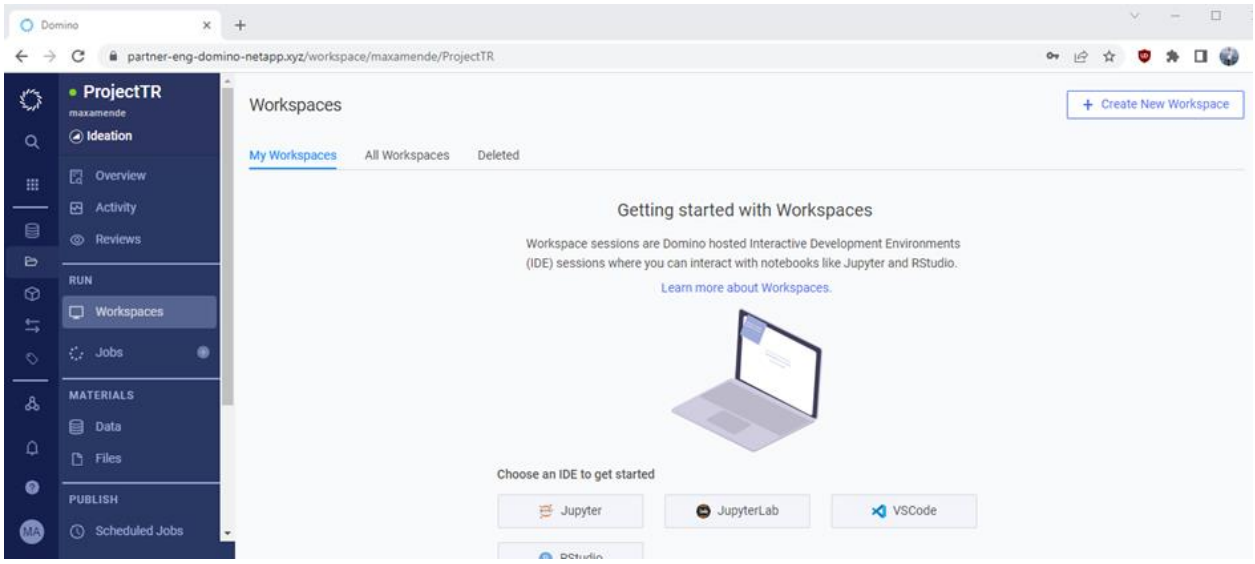
   After a few seconds, the file should be available.

Create a new workspace

In this example, the data loaded into the dataset. It's now time to create a new workspace.

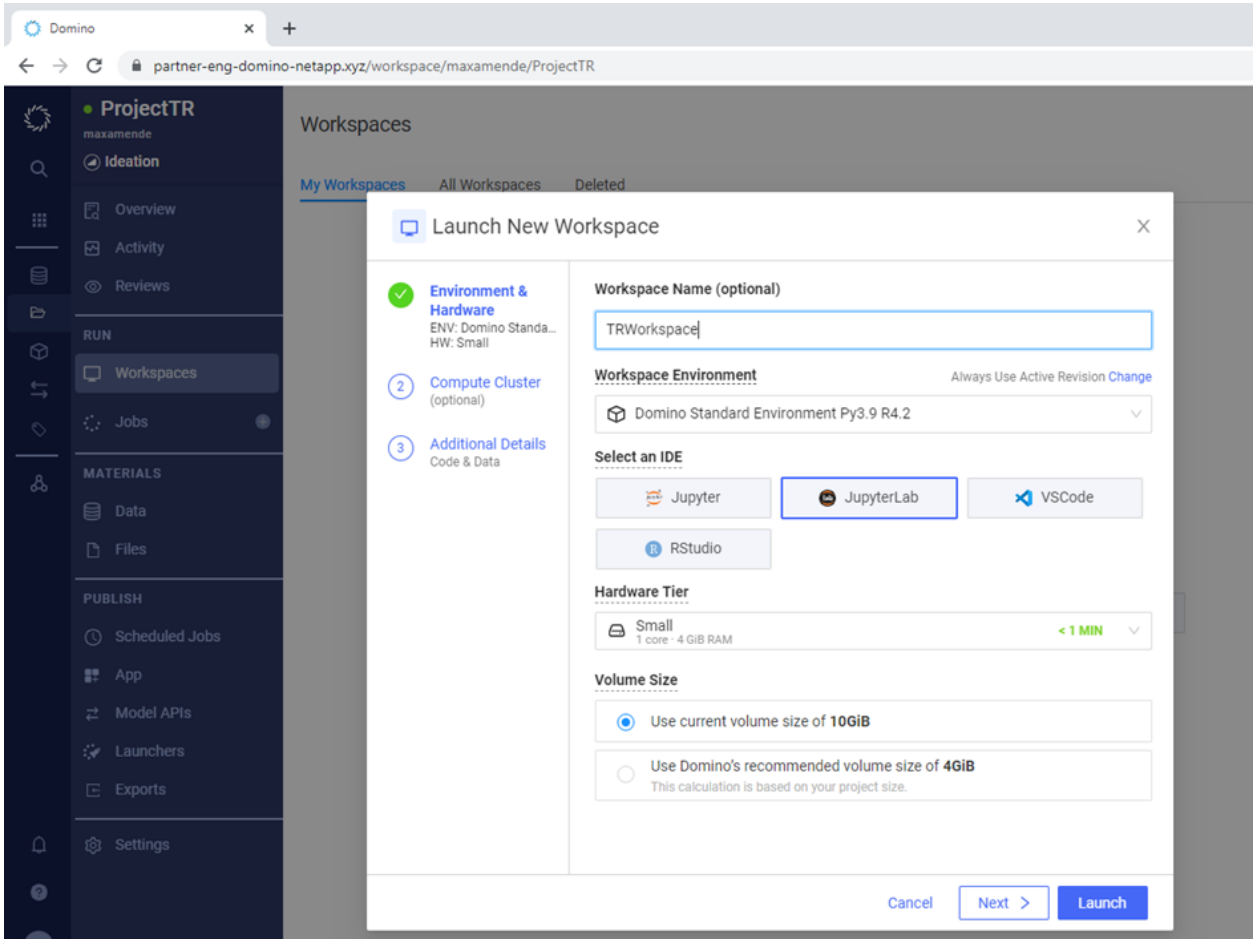
Workspaces are in Domino, the place where data scientists can work and wrangle with the data by getting access to a working environment.

1. From the left page, click Workspaces and click + Create New Workspace.



2. In the then opening window, define a name for your workspace, select a workspace environment, an integrated development environment (IDE), a hardware tier, and a volume size.

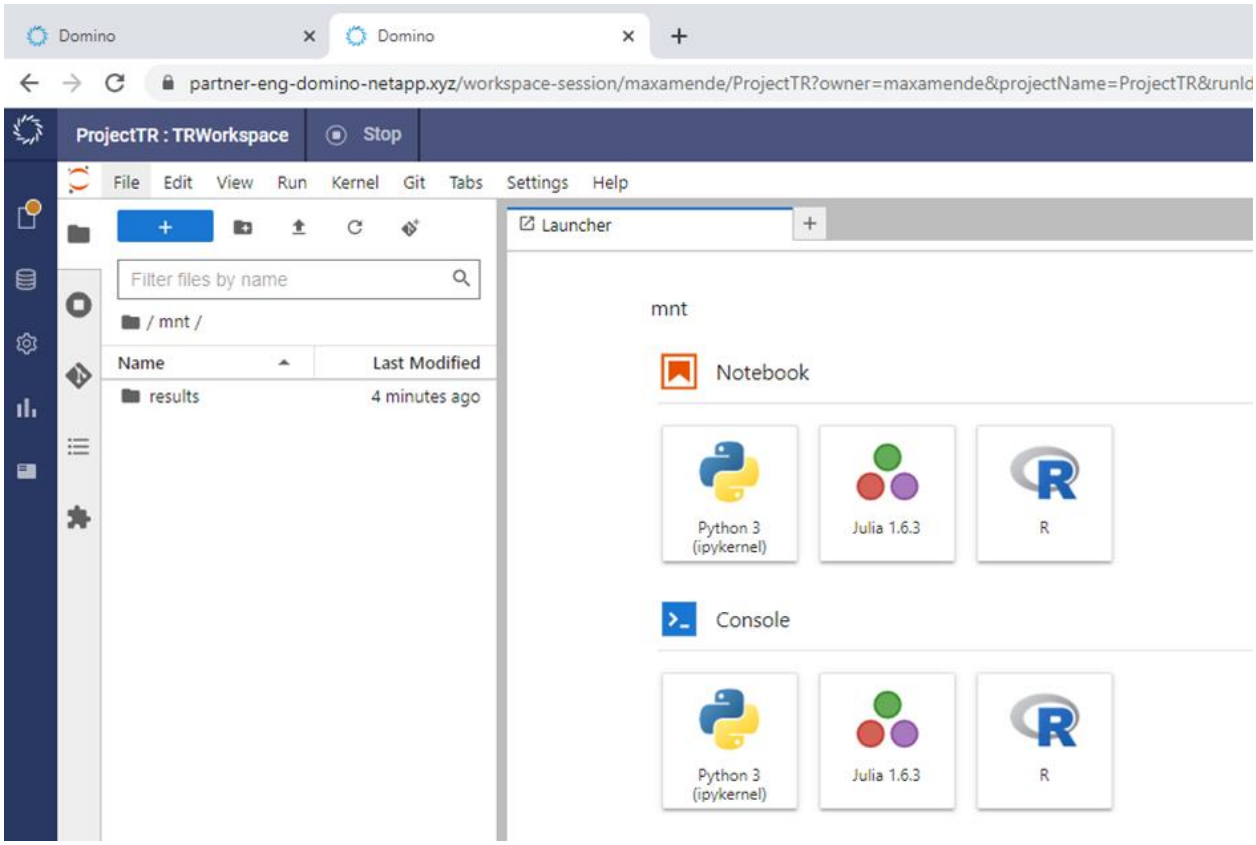
   In this example, the workspace is named TRWorkspace. Select JupyterLab as the IDE. The other options can be left as is. Click Launch.



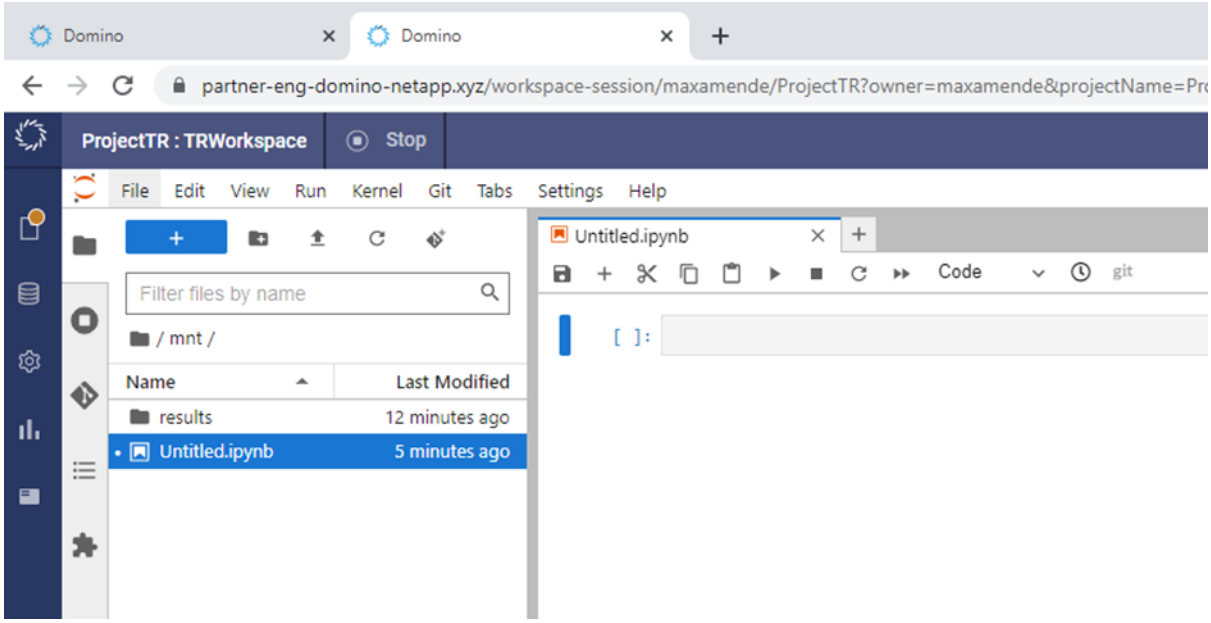
   Initializing the new workspace might take a few seconds.

Access the data hosted on FSx for ONTAP from a workspace

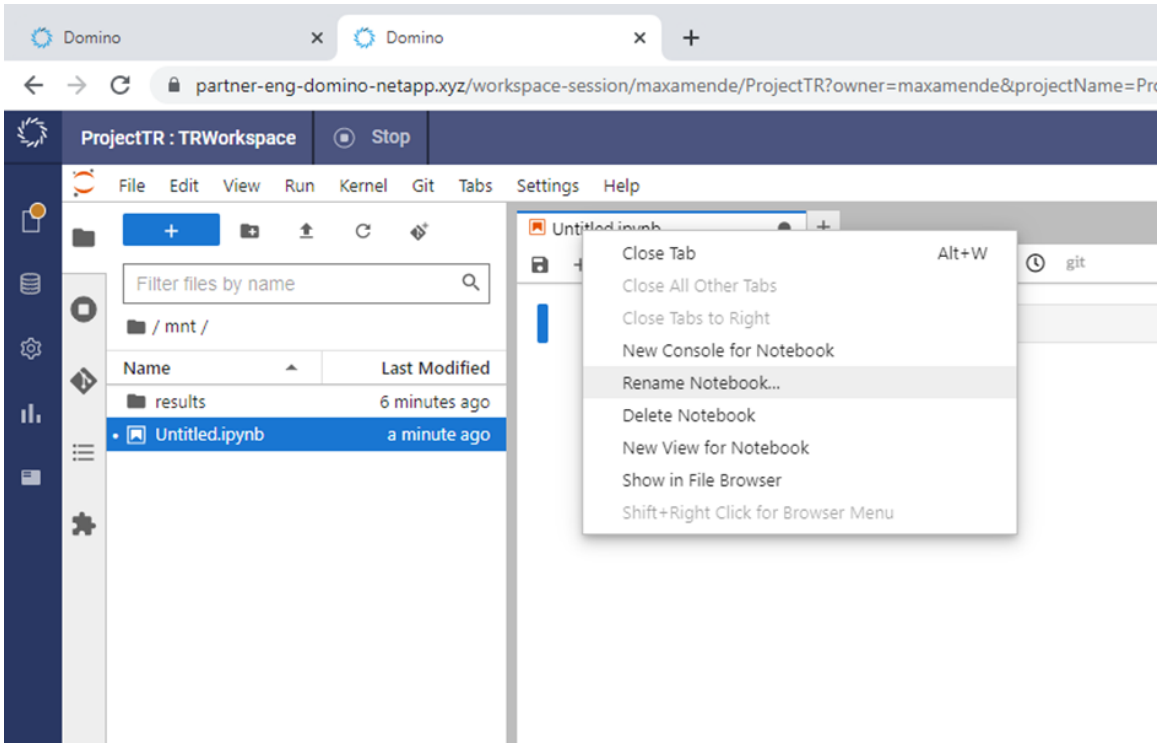
1. In the newly created workspace, you should see a JupyterLab environment.



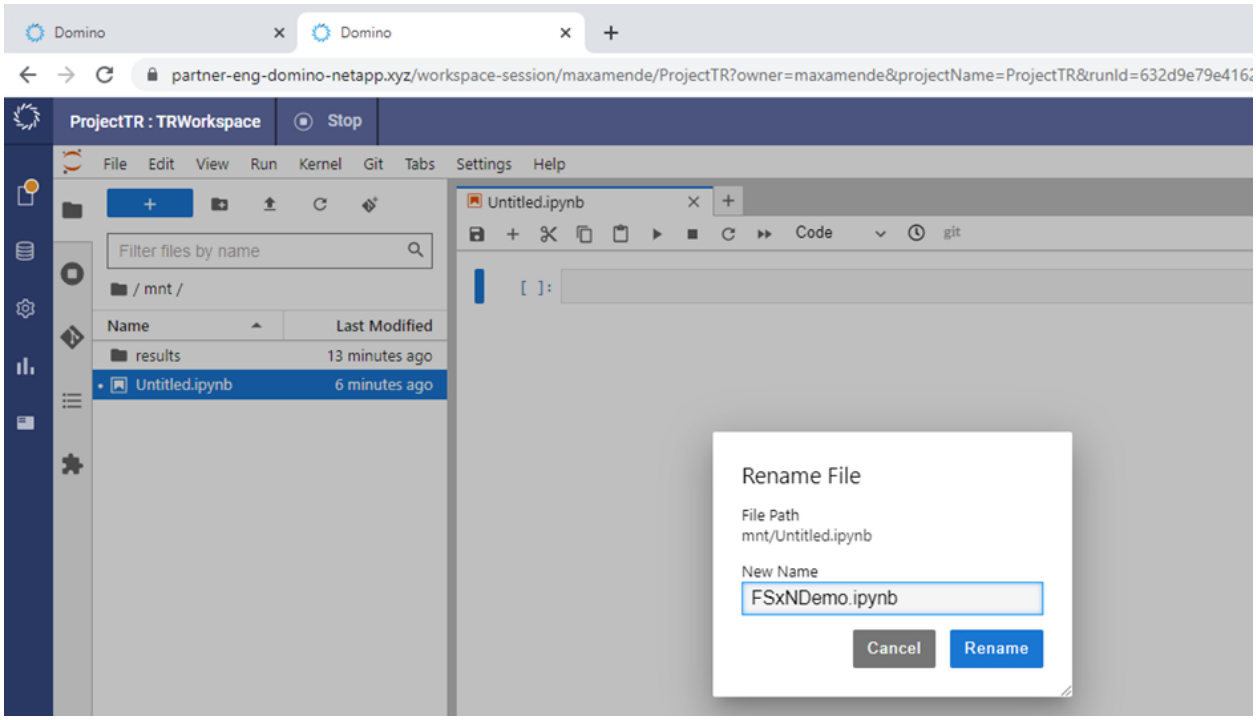
2. Under Notebook, select Python 3 (ipykernel). This creates a new Jupyter Notebook managed by Domino.



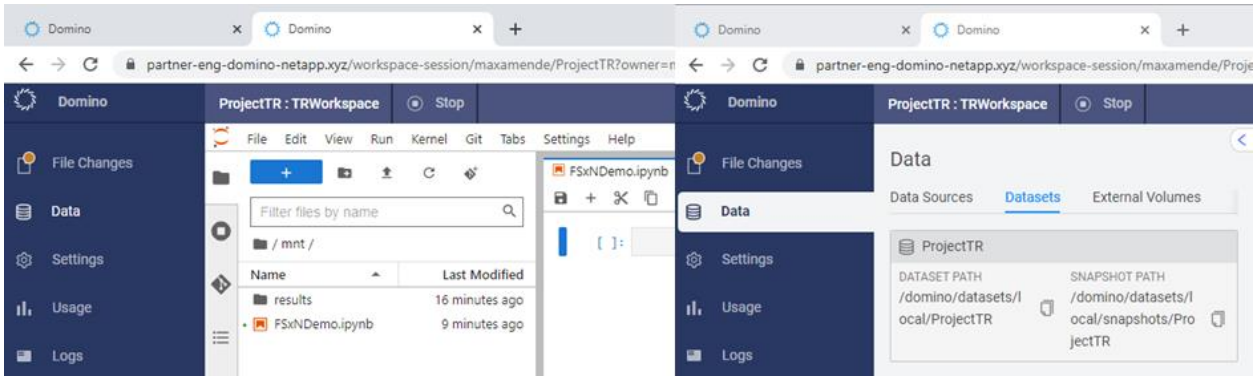
3. To work better with the Notebook, right-click the name of the Notebook and select Rename Notebook.



In this example, we called the Notebook `FSxNDemo.ipynb`.



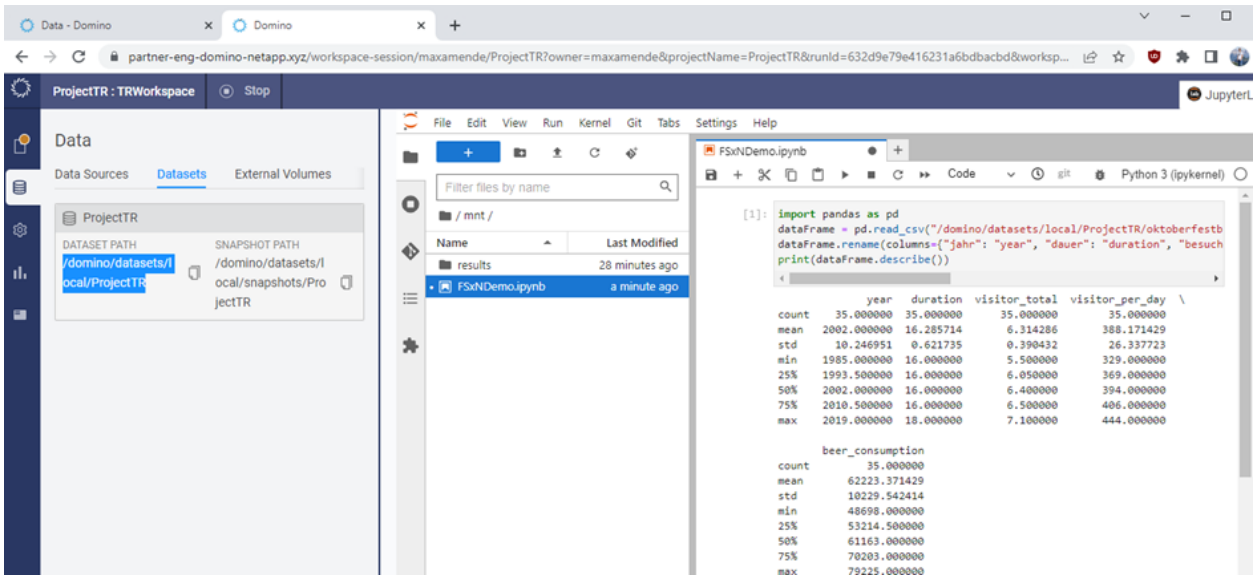
4. Move your mouse over the Database symbol on the left and click Data.



5. Click Datasets. The file that was previously uploaded is located under Dataset Path.

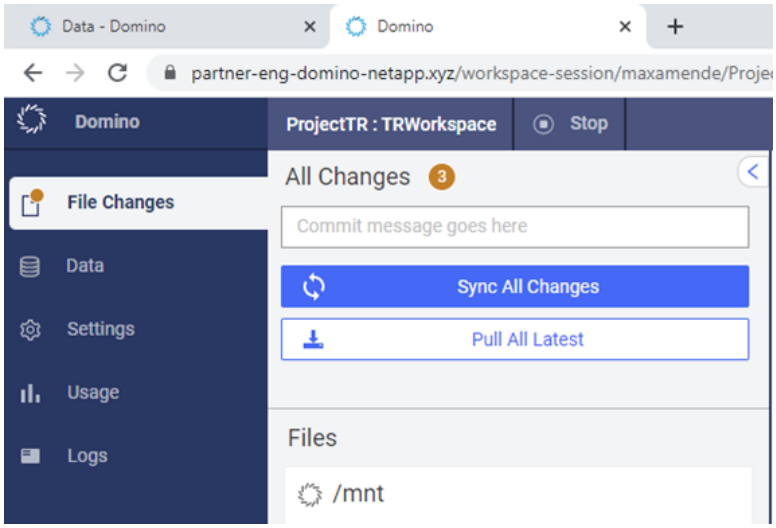
   You can now access the files in the dataset as if they were local on your system. If you followed the steps in this document, you can paste the following code into a cell of the Jupyter Notebook and execute it.

```
import pandas as pd
dataFrame = pd.read_csv("/domino/datasets/local/ProjectTR/oktoberfestbierkonsum19852019.csv")
dataFrame.rename(columns={"jahr": "year", "dauer": "duration", "besucher_gesamt":
"visitor_total", "besucher_tag": "visitor_per_day", "bier_konsum": "beer_consumption"},
inplace=True)
print(dataFrame.describe())
```



6. Hover over the File symbol on the left, click File Changes, and click Sync All Changes.

   Synching the data from the workspace allows you to access the newly created file from other parts of the project.



Access the data hosted on FSx for ONTAP from a training job

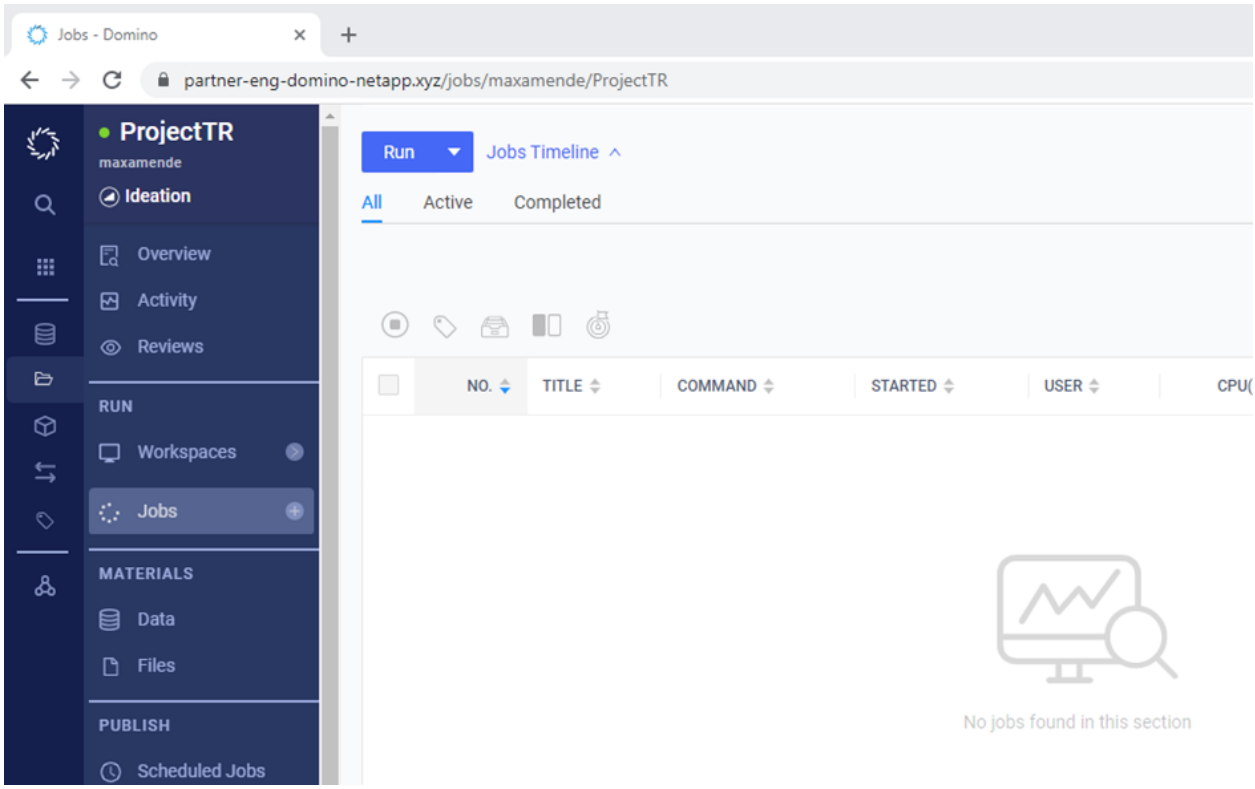
Data scientists typically use working environments such as Jupyter Notebooks for data wrangling and rapid prototyping. When it comes to training (large) models, this approach is not efficient, because you would have to block all the computing resources needed for the training during the phase of the data wrangling and rapid prototyping.

Instead, Domino offers the functionality to schedule training jobs onto a compute cluster. This means that a data scientist can do all the data wrangling on a small compute instance, and when it comes to training the model or hyperparameter tuning switch to a larger more powerful instance, or even to multiple nodes.

This document shows you how to access data from datasets hosted on FSx for ONTAP from training jobs. Therefore, use the Notebook from the section titled, “Access the data hosted on FSx for ONTAP from a workspace.”

Note: The Notebook does not contain any model training, but the process is the same as if you would train a model.

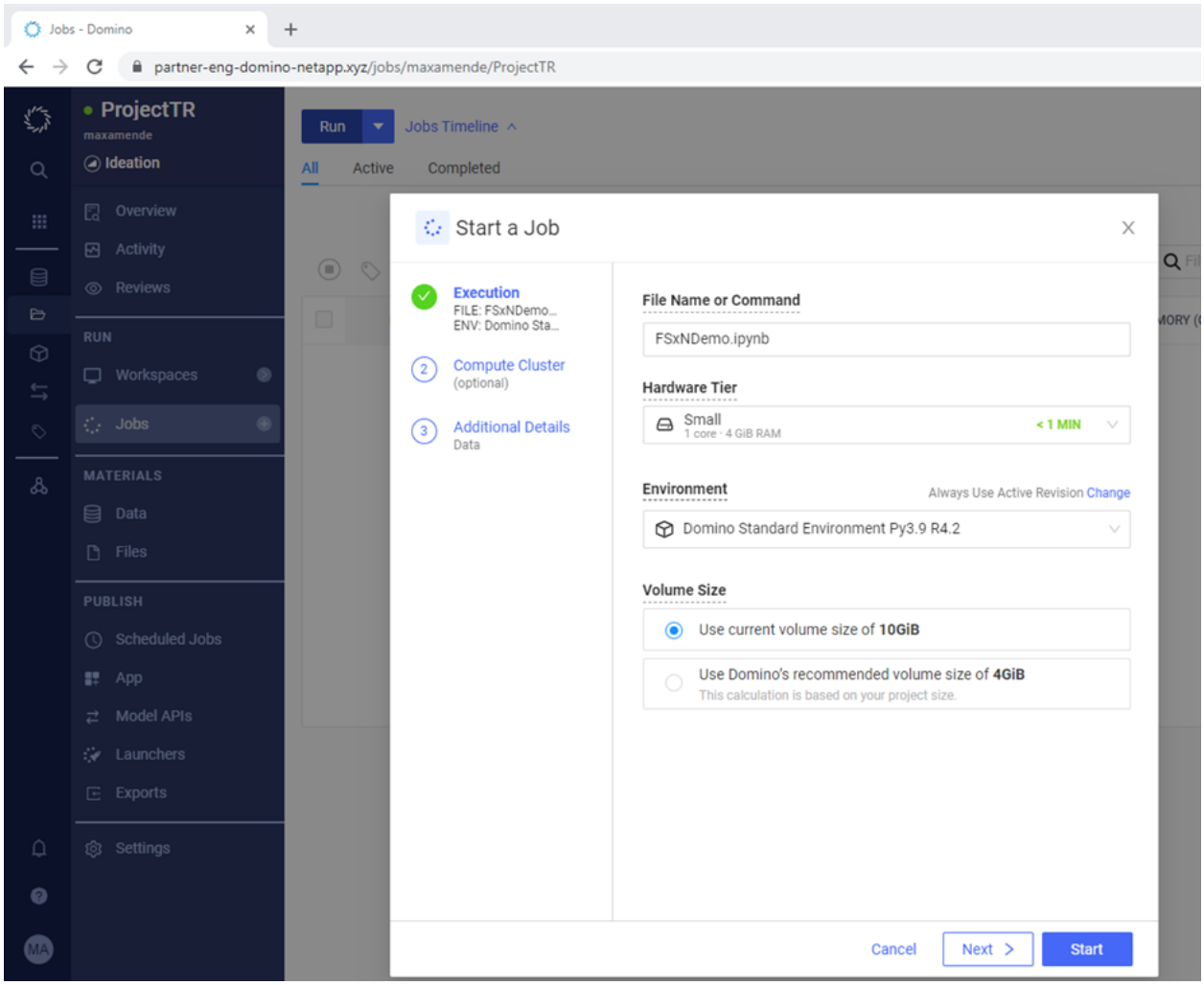
1. Switch back from the Workspace overview to the Project overview and select Jobs.



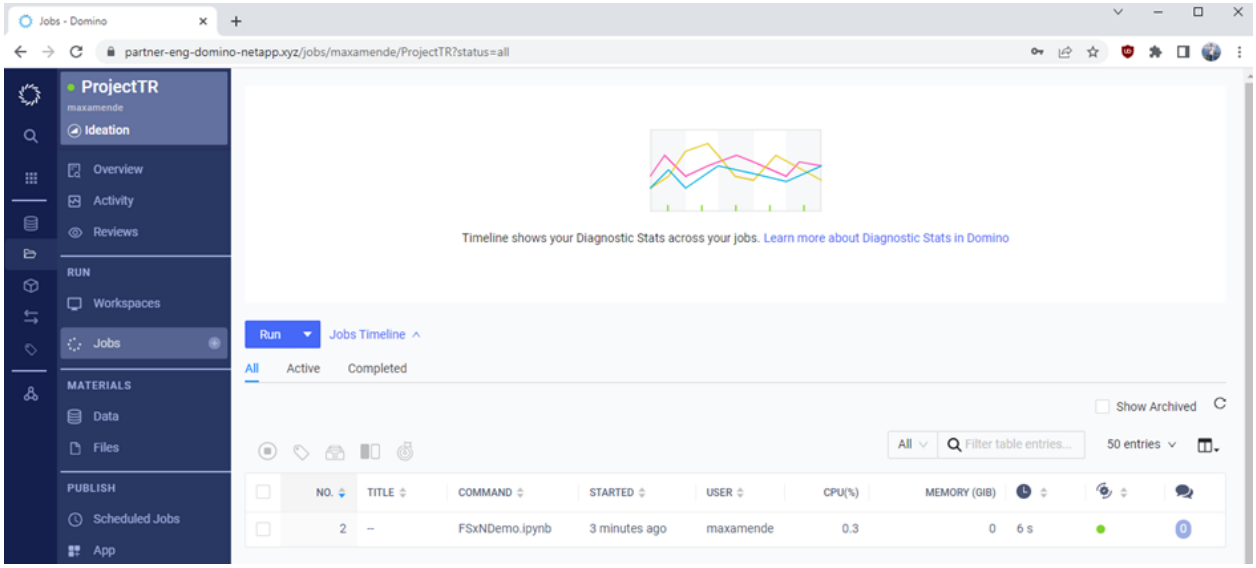
2. Click Run. In the opening window, select the file name of your Notebook. When writing the first letters of your Notebook name, Domino should automatically try to autocomplete the name of the Notebook. If you followed the steps in this document, the name of the Notebook should be `FSxNDemo.ipynb`.

   If you cannot find the Notebook from the field, make sure that you followed all the steps in the section titled, “Accessing FSx for ONTAP managed storage from Domino workspaces,” including the last step of the section titled, “Access the data hosted on FSx for ONTAP from a workspace” where you synched all changes.

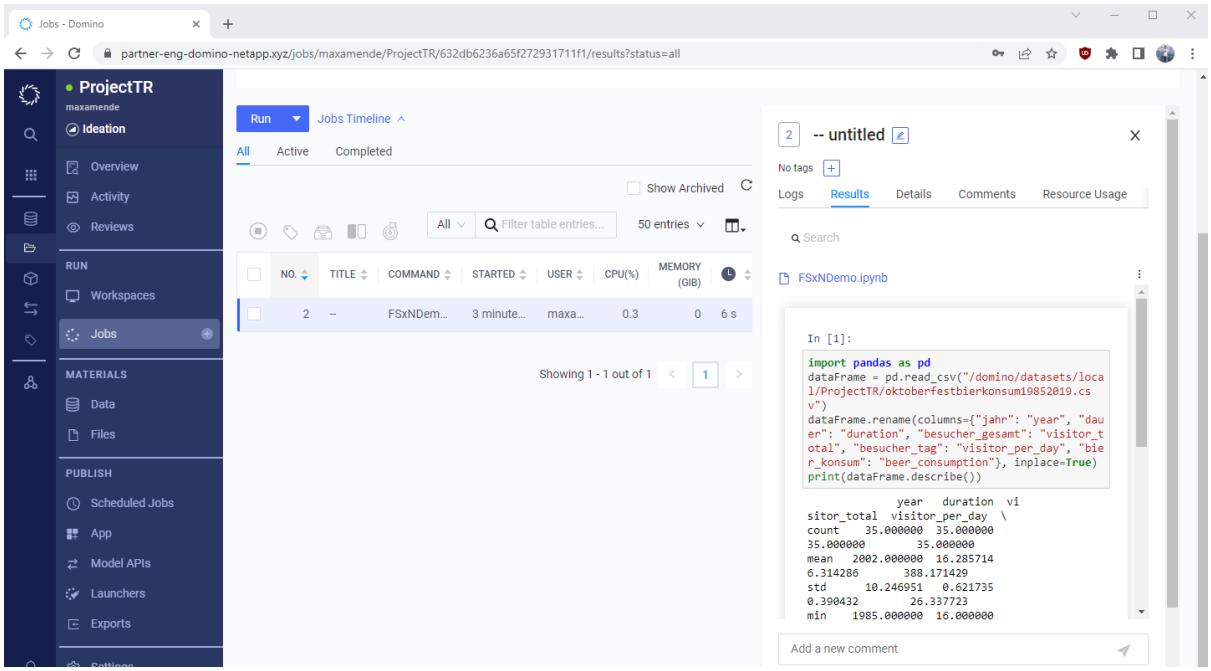
3. Leave the other values as default and click Start.



Domino is now spinning up a node in Amazon Elastic Kubernetes Service (EKS) and deploys the Notebook on it. After a few minutes, the job should be finished.



4. Click on the job. You should see that Domino executed the Notebook, accessing the dataset lying on FSx for ONTAP coming to the same result as in the section titled, “Access the data hosted on FSx for ONTAP from a workspace.”



After the job is successfully executed, Domino automatically scales down the EKS node again.

Conclusion

Customers can benefit from taking advantage of the enterprise-level performance, scalability for high-performance workloads, and data protection features provided by the FSx for ONTAP. This is a simple end-to-end solution for deploying and managing large-scale AI applications in hybrid and real-time environments. The solution brings automation of the data science process and 12x acceleration in the deployment of AI products. This solution helps data scientists run the Kubernetes job with the attached FSx for ONTAP PVC. They can also spin up a Jupyter Notebook with the persistent volume and the traditional data operations (DataOps) Toolkit. Data scientists are not required to redo the environment every time. They can destroy the environment whenever they do not need it, because the data resides in the persistent volume.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- Amazon FSx for NetApp ONTAP documentation
  Amazon FSx for NetApp ONTAP documentation
- Domino documentation
  Dominodata product documentation
- NetApp Product Documentation
  https://www.netapp.com/support-and-training/documentation/

Version history

Version	Date	Document version history
Version 1.0	January 2023	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright information

TR-4952-0123

NetApp