



Technical Report

# **Simplify your data science journey with Amazon FSx for NetApp ONTAP and Iguazio**

Prabu Arjunan, NetApp  
Nick Schenone, Iguazio  
July 2022 | TR-4934

## **Abstract**

This document is intended to help customers set up their environment to simplify the data science journey with Amazon FSx for NetApp ONTAP and Iguazio.

## TABLE OF CONTENTS

<b>Solution overview .....</b>	<b>4</b>
<b>Architecture .....</b>	<b>4</b>
<b>Walkthrough .....</b>	<b>4</b>
Use cases.....	5
<b>Prerequisites .....</b>	<b>6</b>
<b>Deployment steps .....</b>	<b>6</b>
Create an Amazon VPC .....	6
Enable auto-assign IPv4 IP addresses on subnets .....	7
Create the FSx for ONTAP file system .....	8
Create the FSx for ONTAP file system using standard methos .....	8
Ensure VPC Connectivity to FSx security group.....	9
Create the Iguazio cluster.....	10
Deploy Trident Operator on Iguazio cluster .....	12
Define the backend and storage class for FSx .....	12
Ensure login access to SVM.....	13
Create the Trident backend and storage class for FSx.....	13
Dynamically provision FSx volumes by creating a PVC .....	14
Create Jupyter Service in Iguazio with attached FSx PVC .....	14
Run Kubernetes job with attached FSx PVC .....	19
<b>Use the NetApp DataOps Toolkit to perform data management operations .....</b>	<b>20</b>
Prerequisites.....	20
Install NetApp DataOps Toolkit for traditional environments .....	20
Example operation: Clone a volume using the NetApp DataOps Toolkit .....	21
Other operations.....	22
<b>Conclusion .....</b>	<b>22</b>
<b>Where to find additional information .....</b>	<b>22</b>
<b>Version history.....</b>	<b>23</b>

## LIST OF FIGURES

Figure 1) Iguazio integrating with NFS file shared running in FSx for ONTAP. ....	4
Figure 2) Seamless integration of NetApp storage, AWS cloud services, and Iguazio MLOps automation in the hybrid environment.....	5
Figure 3) VPC wizard settings (1).....	6

Figure 4) VPC wizard settings (2).....	7
Figure 5) Enable auto-assign IPv4 IP addresses on subnets 1.....	7
Figure 6) Enable auto-assign IPv4 IP addresses on subnets 2. ....	8
Figure 7) Create the Iguazio cluster. ....	10
Figure 8) Iguazio deployment confirmation.....	11

## Solution overview

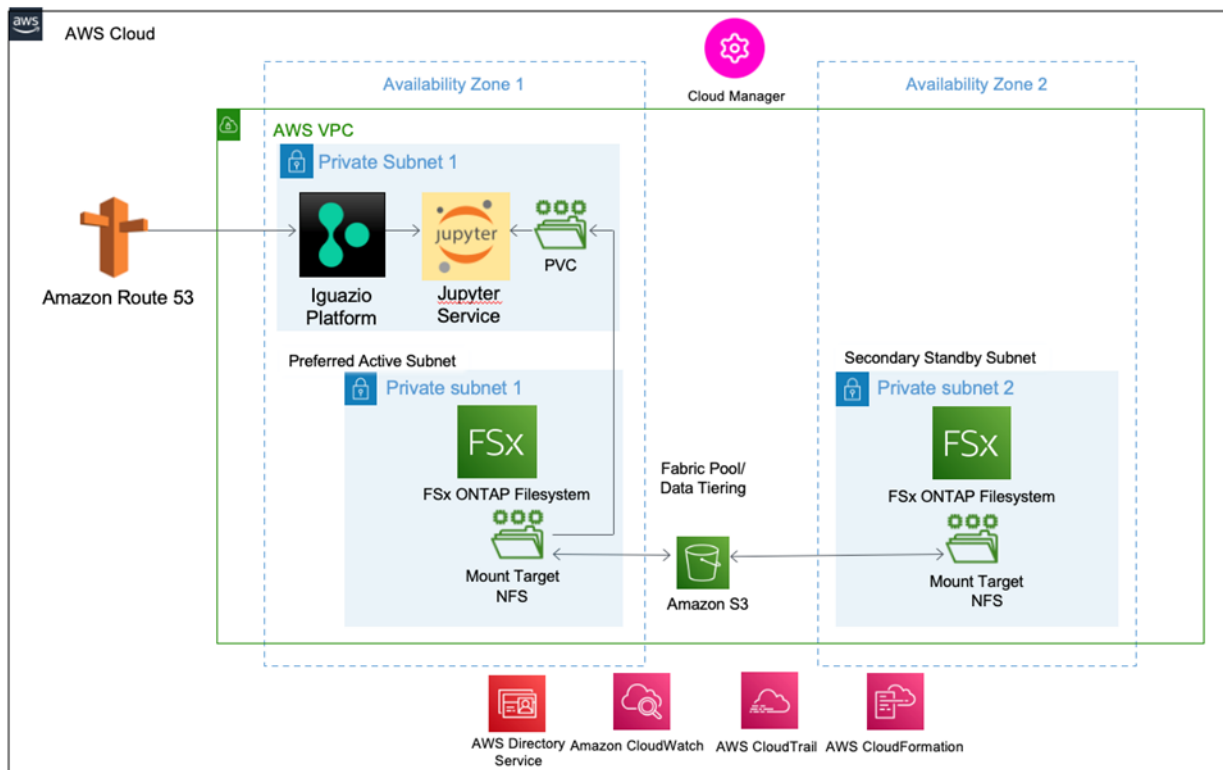
The Iguazio Data Science Platform is a fully integrated and secure data science platform (PaaS) that simplifies development, accelerates performance, facilitates collaboration, and addresses operational challenges. The FSx for ONTAP is a fully integrated managed storage built on the popular NetApp® ONTAP® data management software. The Iguazio Jupyter Notebook takes advantage of the data protection available FSx for ONTAP by mounting the FSx ONTAP volumes to the Jupyter Notebook or mounting them as a data directory for the training jobs.

The FSx for ONTAP file system is tightly integrated with NetApp Data Management Toolkit for traditional environments. The Iguazio Jupyter Notebook takes advantage of the data protection available in FSx for ONTAP.

The Iguazio Data Science Platform provides a complete data science workflow in a single ready-to-use platform that includes all the required building blocks for creating data science applications from research to production. One of the managed services is Jupyter Notebook. Each developer gets its deployment of a notebook container with the resources they need for development. Customers can assign the volume to their container to give them access to the FSX for ONTAP volume.

## Architecture

Figure 1) Iguazio integrating with NFS file shared running in FSx for ONTAP.



## Walkthrough

This section walks through Iguazio Notebook instances integration with NFS files shares running in FSx for ONTAP. This walkthrough highlights two separate ways to provision the volume: how to provision the

volume and use the volume with the Jupyter Notebook and how to use the data directory with the training data.

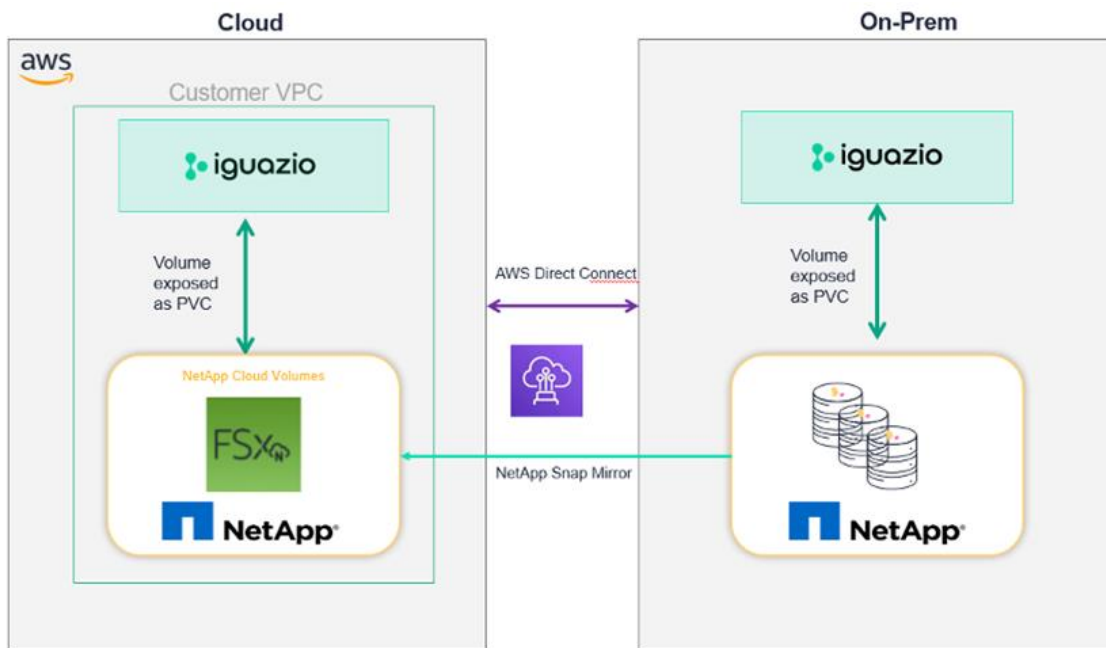
## Use cases

The NetApp and Iguazio solution provides seamless integration of NetApp storage, Amazon Web Service (AWS) cloud services, and Iguazio MLOps automation. You can enjoy a seamless flow from storage to production and use the data in your machine learning (ML) and deep learning (DL) workflows, generating automated, reproducible pipelines that accelerate the deployment of artificial intelligence (AI) in production and enable the continuous rollout of new AI services. You will receive the managed services directly from AWS in a few clicks from your AWS account. You do not need to obtain separate licenses, connectors, and so on. With this solution, you save time, effort, and resources and can focus on data science. The seamless hybrid deployments enable real-time use cases.

## Run Kubernetes Job with attached FSx PVC

Persistent storage Kubernetes offers applications in K8s a handy way to request and consume storage resources. The challenge for the data scientist is to use the data from the on-premises without migrating the data to S3. This challenge can be addressed by running a Kubernetes job with the attached FSx ONTAP PVC, where the data was snapmirrored (see Figure 2) from the on-premises NetApp storage to the FSx ONTAP volume in the cloud. The data scientist can provision the FSx volume to the training jobs and use the data that was replicated using the SnapMirror from the on-premises ONTAP.

**Figure 2) Seamless integration of NetApp storage, AWS cloud services, and Iguazio MLOps automation in the hybrid environment.**



## Create Jupyter Service in Iguazio with attached FSx PVC

Data scientists face several challenges today, one of them being the inability to create a space-efficient data volume that's an exact copy of an existing volume. By creating a Jupyter service in Iguazio with attached FSx ONTAP PVC, you can take advantage of the data protection features that come with 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.

## Prerequisites

- AWS credentials that provide the necessary permissions to create the resources. In this example, we use admin credentials.
- FSx for ONTAP file systems and volumes with valid credentials and a route to the Virtual Private Cloud (VPC) already created. To set up FSx for ONTAP, see [“Getting started with Amazon FSx for NetApp ONTAP.”](#)
- API key and Vault URL from Iguazio for cluster creation.

## Note on connectivity

The most important part of this process is ensuring connectivity between the Iguazio 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 connectivity, however, in this guide, we will be installing in the same VPC and subnet.

## Deployment steps

### Create an Amazon VPC

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

In this guide, we will be creating a VPC and subnets to install FSx and Iguazio through the AWS Console UI; however, feel free to use the tool of your choice. To create the VPC, use the VPC wizard with the settings shown in Figure 3 and Figure 4.

Figure 3) VPC wizard settings (1).

The screenshot shows the AWS VPC wizard settings page. On the left, the 'VPC settings' section includes options for 'Resources to create' (VPC only, VPC, subnets, etc.), 'Name tag auto-generation' (Auto-generate, iguazio-fsx), 'IPv4 CIDR block' (10.0.0.0/16, 65,536 IPs), 'IPv6 CIDR block' (No IPv6 CIDR block, Amazon-provided IPv6 CIDR block), 'Tenancy' (Default), and 'Availability Zones (AZs)' (1, 2, 3). On the right, the 'Preview' section shows a diagram of the VPC resources: 'VPC' (iguazio-fsx-vpc), 'Subnets (2)' (us-east-2a: iguazio-fsx-subnet-public1-us-east-2a, us-east-2b: iguazio-fsx-subnet-public2-us-east-2b), 'Route tables (1)' (iguazio-fsx-rtb-public), and 'Network connections (1)' (iguazio-fsx-igw).

Figure 4) VPC wizard settings (2).

Default

Availability Zones (AZs) [Info](#)  
Choose the number of AZs in which to provision subnets. We recommend at least two AZs for high availability.

1 2 3

Customize AZs

Number of public subnets [Info](#)  
The number of public subnets to add to your VPC. Use public subnets for web applications that need to be publicly accessible over the internet.

0 2

Number of private subnets [Info](#)  
The number of private subnets to add to your VPC. Use private subnets to secure backend resources that don't need public access.

0 2 4

Customize subnets CIDR blocks

NAT gateways (1) [Info](#)  
Choose the number of Availability Zones (AZs) in which to create NAT gateways. Note that there is a charge for each NAT gateway.

None In 1 AZ 1 per AZ

VPC endpoints [Info](#)  
Endpoints can help reduce NAT gateway charges and improve security by accessing S3 directly from the VPC. By default, full access policy is used. You can customize this policy at any time.

None S3 Gateway

DNS options [Info](#)  
☐ Enable DNS hostnames  
☒ Enable DNS resolution

Cancel Create VPC

Preview

VPC [Show details](#)  
Your AWS virtual network  
iguzio-fsx-vpc

Subnets (2)  
Subnets within this VPC  
us-east-2a  
iguzio-fsx-subnet-public1-us-east-2a  
us-east-2b  
iguzio-fsx-subnet-public2-us-east-2b

Route tables (1)  
Route network traffic to resources  
iguzio-fsx-rtb-public

Network connections (1)  
Connections to other networks  
iguzio-fsx-igw

## Enable auto-assign IPv4 IP addresses on subnets

The Iguazio cluster requires that you enable auto-assign IPv4 IP addresses in the subnet if you are using public IP addresses. If you are not using public IP addresses, then this step is optional.

To enable this setting on the newly created subnets, navigate to the Subnets page and enable auto-assign public IPv4 address by following the steps for each of the new subnets.

Figure 5) Enable auto-assign IPv4 IP addresses on subnets 1.

New VPC Experience  
Tell us what you think

VPC Dashboard  
EC2 Global View **New**

Filter by VPC:  
Select a VPC

VIRTUAL PRIVATE CLOUD  
Your VPCs  
Subnets  
Route Tables  
Internet Gateways  
Egress Only Internet Gateways

Subnets (1/5) [Info](#)

Filter subnets

Name	Subnet ID	State	VPC
--	subnet-04bbacfd827386726	Available	vpc-0c51fa37673b2adb0
<input checked="" type="checkbox"/> iguzio-fsx-subnet-public2-us-east-2b	subnet-0fc2f08a588c4c5e9	Available	vpc-039181f78093b04cd
--	subnet-030cf298ead35f1f6	Available	vpc-0c51fa37673b2adb0
iguzio-fsx-subnet-public1-us-east-2a	subnet-094687eb0762aa77b	Available	vpc-039181f78093b04cd
--	subnet-0bd613616f715705e	Available	vpc-0c51fa37673b2adb0

Actions [Create subnet](#)

- View details
- Create flow log
- Edit subnet settings
- Edit IPv6 CIDRs
- Edit network ACL association
- Edit route table association
- Edit CIDR reservations
- Share subnet
- Manage tags
- Delete subnet

1.16.0/2  
6.0/20  
1.32.0/2  
1.0/20  
1.0.0/20

Figure 6) Enable auto-assign IPv4 IP addresses on subnets 2.

VPC > Subnets > subnet-0fc2f08a588c4c5e9 > Edit subnet settings

## Edit subnet settings [Info](#)

### Subnet

Subnet ID	Name
subnet-0fc2f08a588c4c5e9	iguazio-fsx-subnet-public2-us-east-2b

### Auto-assign IP settings [Info](#)

Enable the auto-assign IP settings to automatically request a public IPv4 or IPv6 address for a new network interface in this subnet.

☒ Enable auto-assign public IPv4 address [Info](#)

☐ Enable auto-assign customer-owned IPv4 address [Info](#)  
Option disabled because no customer owned pools found.

## Create the FSx for ONTAP file system

You can create the FSx for ONTAP file system using two options: the Quick Create or Standard Create option in the Amazon console. To create the FSx for ONTAP file system, see [Step 1: Create an Amazon FSx for NetApp ONTAP file system](#). For information about using the Standard Create option to create a file system with a customized configuration, see [Creating FSx for ONTAP file systems](#). The Quick Create option creates a file system with one SVM (fsx) and one volume (vol1). The volume has a junction path of /vol1.

## Create the FSx for 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 NetApp ONTAP file system using the Standard Create option (follow the instructions on the AWS page titled [Creating 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: Assign this optional value at your discretion, but it is recommended
  - SSD storage capacity: At least 1024GB (FSx minimum)
- Network and security:
  - VPC: The newly created VPC from above
  - VPC security groups: Leave as VPC default security group, but make note of the security group for later

- Preferred subnet: One of the subnets created above. Make a note of which subnet you choose as the preferred subnet; you will be using it later during Iguazio installation.
- Standby subnet: One of the other subnets created above.
- VPC route tables: Select one or more VPC route tables. Select the corresponding route table for the newly created VPC above.
- 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 name later).
  - Specify a password: Set a password of your choice (remember this password for later).

## Ensure VPC Connectivity to FSx security group

As previously mentioned, there are many ways to ensure network connectivity in AWS – you will need to determine the best way for your needs. In the examples in this guide, we install FSx and Iguazio into the same VPC and subnet.

We will need to ensure 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.

The screenshot shows the AWS Management Console interface for a VPC. On the left is a navigation menu with options like 'VPC Dashboard', 'EC2 Global View', and 'VIRTUAL PRIVATE CLOUD'. The main panel displays the details for the VPC 'vpc-039181f78093b04cd / iguazio-fsx-vpc'. The details are organized into a table with four columns: VPC ID, State, DNS hostnames, and DNS resolution. The VPC ID is 'vpc-039181f78093b04cd', the State is 'Available', DNS hostnames are 'Disabled', and DNS resolution is 'Enabled'. Other details include Tenancy (Default), DHCP options set (dopt-051d83c0644b99aa0), Main route table (rtb-09f05256b77a5a7e6), Main network ACL (acl-06b7d615fa5227a7d), Default VPC (No), IPv4 CIDR (10.0.0.0/16), IPv6 pool (–), Route 53 Resolver DNS Firewall rule groups (–), and Owner ID (669994010164).

VPC ID	State	DNS hostnames	DNS resolution
vpc-039181f78093b04cd	Available	Disabled	Enabled
Tenancy	DHCP options set	Main route table	Main network ACL
Default	dopt-051d83c0644b99aa0	rtb-09f05256b77a5a7e6	acl-06b7d615fa5227a7d
Default VPC	IPv4 CIDR	IPv6 pool	IPv6 CIDR
No	10.0.0.0/16	–	–
Route 53 Resolver DNS Firewall rule groups	Owner ID		
–	669994010164		

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 rule allows access from within the VPC.

VPC > Security Groups > sg-0cc5589ac85204ed2 - default > Edit inbound rules

## Edit inbound rules [Info](#)

Inbound rules control the incoming traffic that's allowed to reach the instance.

Security group rule ID	Type <a href="#">Info</a>	Protocol <a href="#">Info</a>	Port range <a href="#">Info</a>	Source <a href="#">Info</a>	Description - optional <a href="#">Info</a>	
sg-044f7fe7f2999e510	All traffic	All	All	Custom		Delete
-	All traffic	All	All	Custom	VPC connectivity	Delete

[Add rule](#)

Cancel [Preview changes](#) [Save rules](#)

## Create the Iguazio cluster

To install the Iguazio cluster, follow the [AWS Installation Documentation](#). This process entails using the web installer running on a provisioning machine that has access within the VPC. If your laptop or workstation does not have connectivity within the VPC, a small temporary EC2 instance is one way to achieve this.

The web installer will provision servers, install Kubernetes, and install the Iguazio MLOps platform for you.

**Note:** You need an API key and vault URL from Iguazio (as described in the prerequisites).

Follow the guide as described, filling in your own information where necessary. The main screen you should be concerned with regarding FSx connectivity is the Cloud screen (Figure 7).

**Figure 7) Create the Iguazio cluster.**

New System (local)

Installation Scenario General Data Cluster App Cluster **Cloud** Review

Placement Kind: Default

Region Name: us-east-2

VPC mode: Existing

VPC ID: vpc-039181f78093b04cd

CIDR: 10.0.0.0/16

Subnet IDs: subnet-094687eb0762aa77b, subnet-0fc2f08a588c4c5e9

Whitelisted CIDRs:

Installer CIDR: 18.218.163.79/32

Access Key ID:

Secret Access Key:

☐ Verbose Provisioning

BACK NEXT

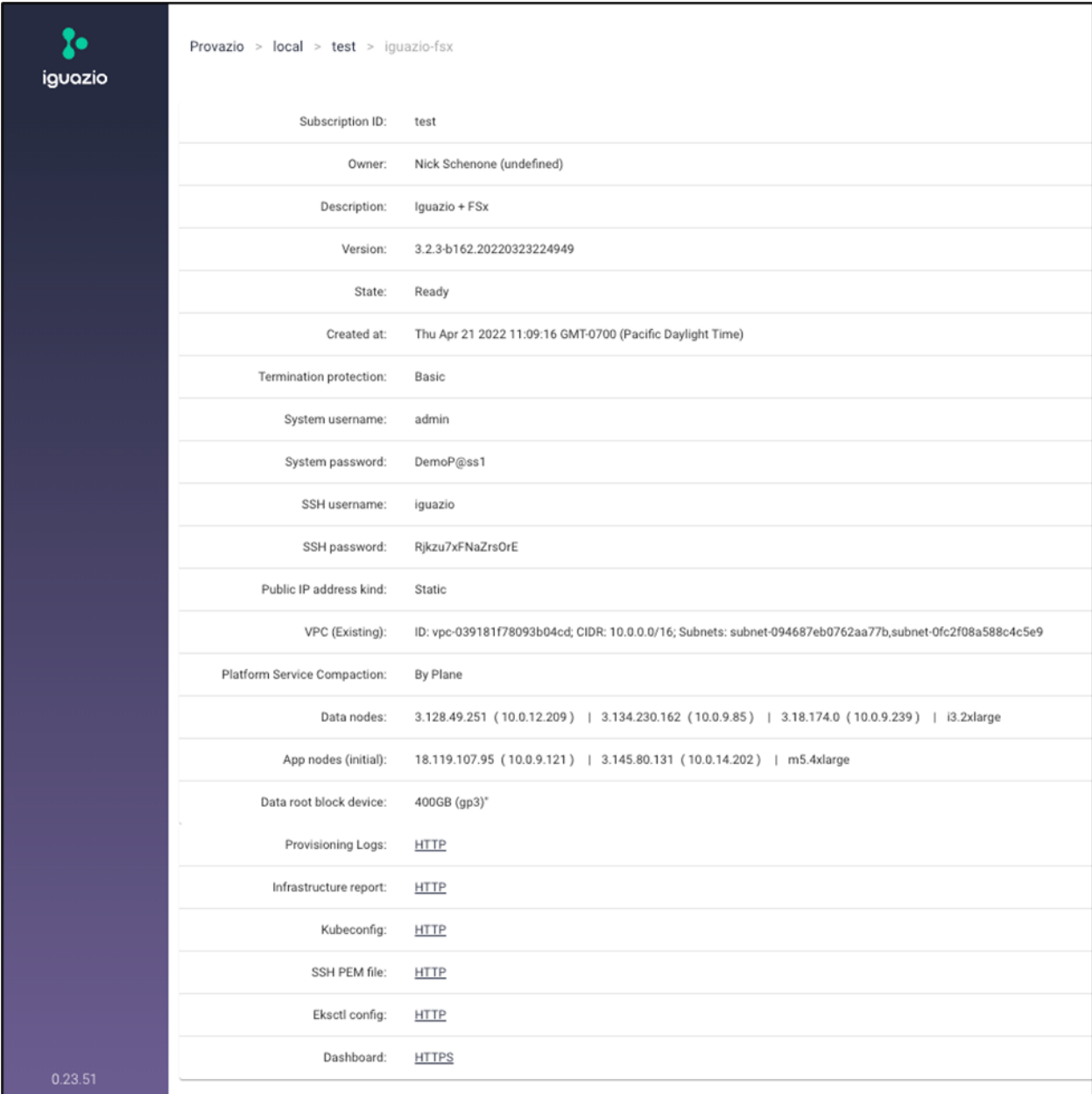
The relevant fields are as follows:

- **VPC Mode.** The existing value.

- **VPC-ID.** The ID of the VPC created earlier.
- **CIDR.** The IPv4 CIDR of the VPC created earlier.
- **Subnet IDs.** The two subnets created earlier.  
**Note:** The first listed subnet should be the primary subnet for FSx.
- **Installer CIDR.** The public IP address of your provisioning machine. In this example, a temporary EC2 instance in the same VPC.

After the Iguazio cluster is deployed, you will see a page that looks like the one in Figure 8.

**Figure 8) Iguazio deployment confirmation.**



Subscription ID:	test
Owner:	Nick Schenone (undefined)
Description:	Iguazio + FSx
Version:	3.2.3-b162.20220323224949
State:	Ready
Created at:	Thu Apr 21 2022 11:09:16 GMT-0700 (Pacific Daylight Time)
Termination protection:	Basic
System username:	admin
System password:	DemoP@ss1
SSH username:	iguazio
SSH password:	Rjkzu7xFNaZrsOrE
Public IP address kind:	Static
VPC (Existing):	ID: vpc-039181f78093b04cd; CIDR: 10.0.0.0/16; Subnets: subnet-094687eb0762aa77b,subnet-0fc2f08a588c4c5e9
Platform Service Compaction:	By Plane
Data nodes:	3.128.49.251 ( 10.0.12.209 )   3.134.230.162 ( 10.0.9.85 )   3.18.174.0 ( 10.0.9.239 )   i3.2xlarge
App nodes (initial):	18.119.107.95 ( 10.0.9.121 )   3.145.80.131 ( 10.0.14.202 )   m5.4xlarge
Data root block device:	400GB (gp3)*
Provisioning Logs:	<a href="#">HTTP</a>
Infrastructure report:	<a href="#">HTTP</a>
Kubeconfig:	<a href="#">HTTP</a>
SSH PEM file:	<a href="#">HTTP</a>
Eksctl config:	<a href="#">HTTP</a>
Dashboard:	<a href="#">HTTPS</a>

## Deploy Trident Operator on Iguazio cluster

Now that both the FSx file system and Iguazio cluster are deployed, connect them using the Trident Operator. SSH into one of the Iguazio data nodes (see the 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 backend and storage class for FSx

We will need to define a Trident backend for FSx to use. You can follow the full [documentation](#) for creating the backend, or use the following as a template:

```
# backend.json

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

The relevant fields are as follows:

- **storageDriverName:** ONTAP-nas
- **backendName:** Assign this value at your discretion.
- **managementLIF:** The management IP address of the SVM for your FSx file system.
- **svm:** The name of the SVM for your FSx file system.
- **username:** vsadmin
- **password:** Defined a password for the SVM when creating the FSx file system.

We will 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-basic.yaml

apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: basic
provisioner: csi.trident.netapp.io
parameters:
  backendType: "ONTAP-nas"
```

```
fsType: "__FILESYSTEM_TYPE__"
```

The relevant fields are as follows:

- `metadata.name`: Assign this optional value as your discretion, but you will be using this later to create FSx volumes dynamically.
- `parameters.backendType`: ONTAP-nas

## Ensure login access to SVM

To install the backend, you should be able to SSH into the SVM using the IP address, vsadmin user, and password defined during installation.

1. Try the following and enter the password defined during installation:

```
ssh vsadmin@SVM_MANAGEMENT_IP
```

2. If you are successful, proceed to the "Create the Trident backend and storage class for FSx" section.
3. If you get a message that says: `Error: Account currently locked. Contact the storage administrator to unlock it.`, unlock the vsadmin account to proceed.
4. To unlock the account, log in to the FSx management console with the following command and the file system administrative password defined during file system creation:

```
ssh fsxadmin@FSX_MANAGEMENT_IP
```

5. [Change the password](#) for the vsadmin user; for example:

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

6. [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 installation. Verify with the same command from before.

## Create the Trident backend and storage class for FSx

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

**Note:** You should still be SSH'd into one of the Iguazio data nodes and in the same directory as the previously downloaded the `tridentctl` tool.

1. To create the backend, run the following command:

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

2. To create the storage class, run the following command:

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

## Dynamically provision FSx volumes by creating a PVC

At this point, the setup process is complete. Creating a PVC can be done now or later on-demand within the Iguazio MLOps platform.

For the sake of this guide, create a PVC now to attach to a new Jupyter service later.

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

```
# my-fsx-pvc.yaml
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: my-fsx-pvc
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 1Gi
  storageClassName: basic
```

The relevant fields are as follows:

- `metadata.name`: Assign a name; you will use this value when attaching the PVC to the Jupyter service.
- `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`: The size of FSx volume to provision on demand.
- `spec.storageClassName`: This value should match the previously created storage class (in this example, it is `basic`).

2. Create the PVC; for example:

```
kubectl create -n default-tenant -f my-fsx-pvc.yaml
```

**Note:** The Kubernetes namespace needs to be `default-tenant`, because that is where the services are running in Iguazio.

## Create Jupyter Service in Iguazio with attached FSx PVC

Now you can easily create a Jupyter service in Iguazio and attach the newly created FSx PVC.

1. In the Iguazio platform, navigate to the Services tab and select New Service in the top-right corner.

Services									
<input type="checkbox"/>	Name ↑	Running User ↑	Version ↑	CPU (cores) ↑	Memory ↑	API	Status ↑	Logs	
<input type="checkbox"/>	<b>authenticator</b> Type: OAuth2 (OIDC) Au...		2.30.0	266μ	11.65 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>docker-registry</b> Type: Docker Registry		2.7.1	93μ	8.14 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>framesd</b> Type: V3IO Frames		0.9.6	205μ	20.01 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>log-forwarder</b> Type: Log forwarder		7.9.2	0	0 bytes		Standby	N/A	⋮
<input type="checkbox"/>	<b>mlrun</b> Type: MLRun (Beta)		0.10.1	9m	899.88 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>monitoring</b> Type: Monitoring		2.22.0	1m	47.17 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>mpi-operator</b> Type: Horovod		0.2.3	675μ	14.61 MB		Running	N/A	⋮
<input type="checkbox"/>	<b>nuclio</b> Type: Nuclio		1.6.32	7m	94.38 MB		Running	N/A	⋮

2. Select Jupyter as the service to create with your desired name.

### Create a new service

1

2

3

BASIC SETTINGS

COMMON PARAMETERS

CUSTOM PARAMETERS

Configure your service

**Service type \***

Jupyter Notebook

**Service name \***

jupyter-fsx

42

**Description**

Write your description...

?

☒ Enabled

NEXT STEP >

3. Add Kubernetes resources and share with additional users as desired. For this guide, leave everything at the default values.

## Create a new service

1 
2 
3

BASIC SETTINGS
COMMON PARAMETERS
CUSTOM PARAMETERS

**Scale to zero**

☐ Enabled

**Resources**

For more information about the resource parameters, see [Kubernetes documentation](#).

The memory and CPU configurations are applied to each replica.

	Request		Limit	
Memory	<input type="text"/> GB		<input type="text"/> GB	
CPU	Request <input type="text" value="Example: 1500"/> millicpu		Limit <input type="text" value="Example: 1500"/> millicpu	
GPU			Limit <input type="text" value="Unlimited"/>	

**Running User \***

Username

☐ Shared

< PREVIOUS STEP
NEXT STEP >

#### 4. Add the newly created FSx PVC.

**Note:** The name of the PVC matches the name in the YAML definition. For convenience, the mount path should start with `/User` because that is the home directory in the Jupyter service.

### Create a new service

✓

✓

3

BASIC SETTINGS

COMMON PARAMETERS

CUSTOM PARAMETERS

Flavor

Full stack without GPU

Spark

None

Create new...

Environment Variables

+ Create a new environment variable

Persistent Volume Claims (PVCs)

Name ?

Mount Path

my-fsx-pvc

/User/fsx

+ Add PVC

< PREVIOUS STEP

CREATE SERVICE

5. From there, select Apply Changes.

Services

NEW SERVICE

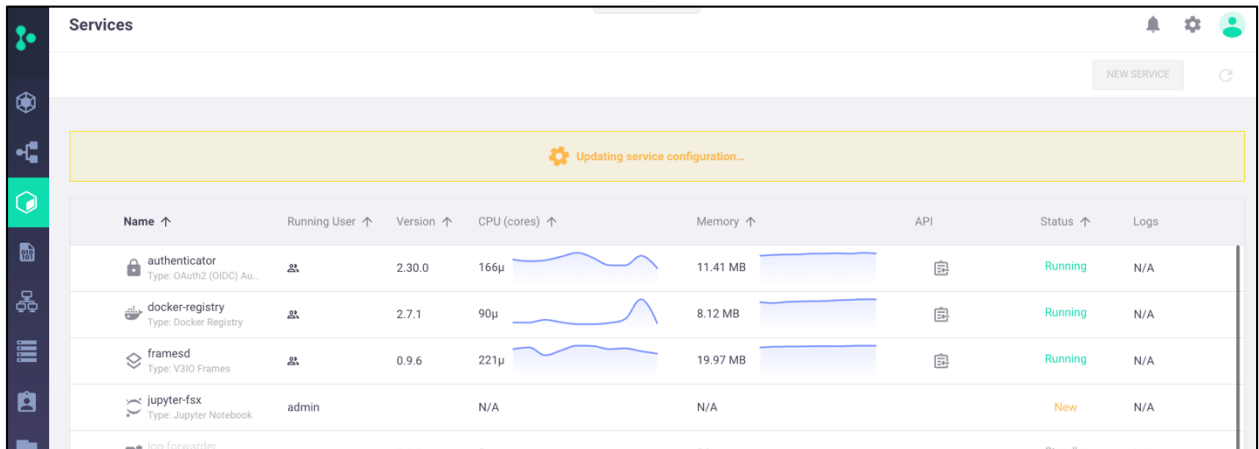
Pending changes detected. The changes will take effect only after they're applied.

APPLY CHANGES

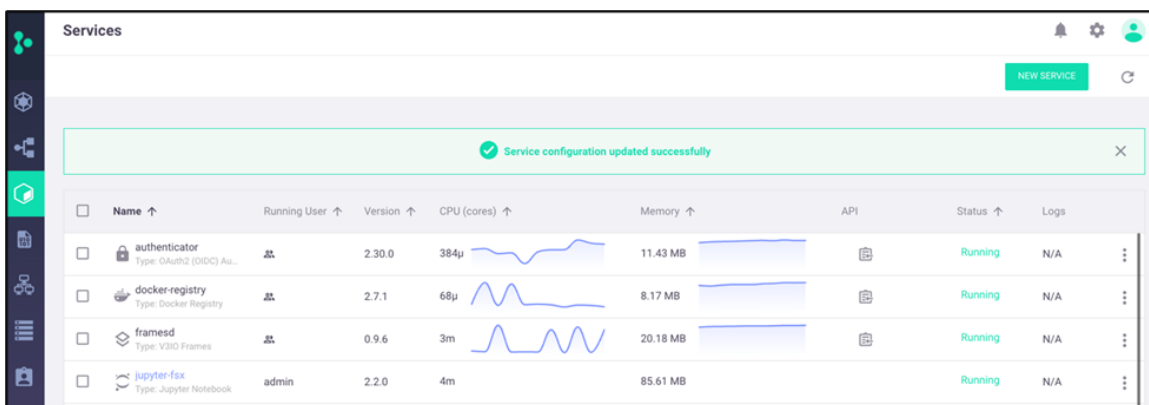
DISCARD

<input type="checkbox"/>	Name ↑	Running User ↑	Version ↑	CPU (cores) ↑	Memory ↑	API	Status ↑	Logs
<input type="checkbox"/>	<div>authenticator</div> <div>Type: OAuth2 (OIDC) Au...</div>		2.30.0	166μ	11.41 MB		Running	N/A
<input type="checkbox"/>	<div>docker-registry</div> <div>Type: Docker Registry</div>		2.7.1	90μ	8.12 MB		Running	N/A
<input type="checkbox"/>	<div>framesd</div> <div>Type: V3IO Frames</div>		0.9.6	221μ	19.97 MB		Running	N/A
<input type="checkbox"/>	<div>jupyter-fsx</div> <div>Type: Jupyter Notebook</div>	admin	N/A	N/A	N/A		New	N/A

6. While the service is creating, this dialog at the top is displayed.

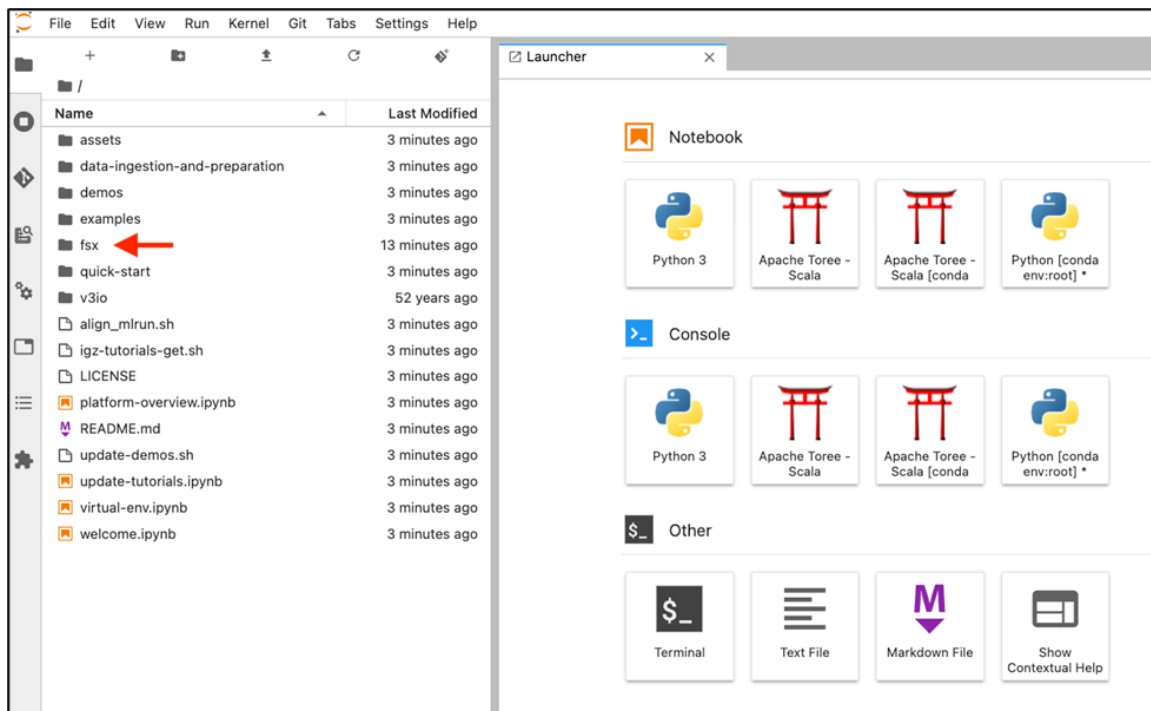


7. After the service is successfully created, this dialog at the top is displayed.



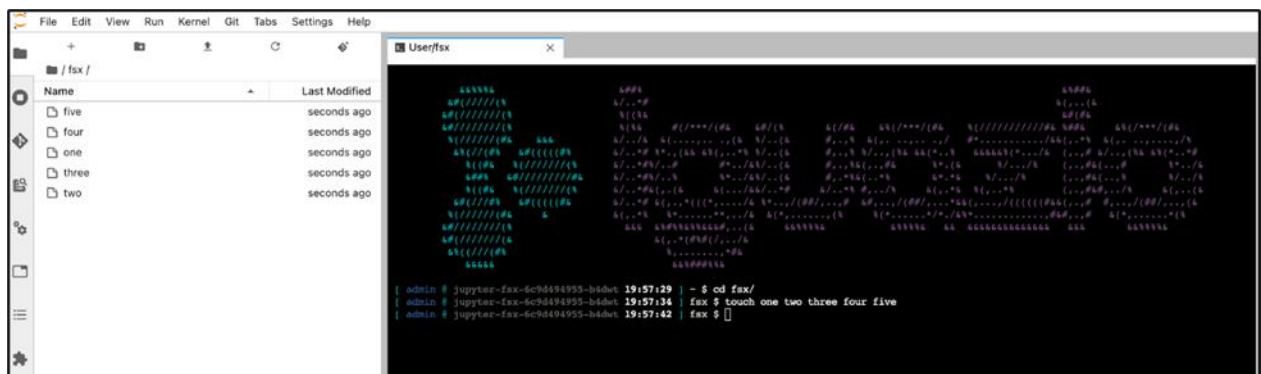
You will also be able to select the new Jupyter service, which open in a new tab.

**Note:** The `fsx` directory is in the file browser. This is our dynamically provisioned FSx volume:



## Run Kubernetes job with attached FSx

1. To showcase the ability to mount the FSx file system, create some files in the Jupyter terminal; for example:



2. Create a simple script to list the files in the directory that you mount to the job. When you create the job, mount the FSx volume to /mnt/my\_fsx\_mount.

```
# my_job.py

import os

def print_fsx(context):
    print("FSX VOLUME CONTENTS", os.listdir("/mnt/my_fsx_mount"))
```

3. Use Iguazio's ability to easily containerize and run code on Kubernetes to showcase the volume mount within the job.

```
[1]: import mlrun

[2]: job = mlrun.code_to_function(
    name="my-fsx-print-job",
    filename="my_job.py",
    kind="job",
    image="mlrun/mlrun",
    handler="print_fsx"
)

[3]: job = job.apply(mlrun.auto_mount(pvc_name="my-fsx-pvc", volume_mount_path="/mnt/my_fsx_mount")) # Mount FSx

[4]: job.run() # Run job on K8s cluster

> 2022-04-21 20:18:50,897 [info] starting run my-fsx-print-job-print_fsx uid=b35cca9c41b54d2d8a058bee5d125f2e DB=http://mlrun-api:8080
> 2022-04-21 20:18:51,055 [info] Job is running in the background, pod: my-fsx-print-job-print-fsx-ltk7s
FSX VOLUME CONTENTS ['one', 'two', 'three', 'four', 'five']
> 2022-04-21 20:19:19,128 [info] run executed, status=completed
final state: completed

project uid iter start state name labels inputs parameters results artifacts
default ...5d125f2e 0 Apr 21 20:19:19 completed my-fsx-print-job-print-fsx v3io_useradmin kind:job owneradmin mlfunction_version:0.10.1 host:my-fsx-print-job-print-fsx-16.7s

> to track results use the .show() or .logs() methods or click here to open in UI
> 2022-04-21 20:19:20,429 [info] run executed, status=completed
[4]: <mlrun.model.Run0bject at 0x7f59c241c550>
```

Although this is a trivial example, this example could easily be expanded to loading a large dataset for a training job or a utilizing the shared filesystem to store models and artifacts.

## Use the NetApp DataOps Toolkit to perform data management operations

### Prerequisites

Access a Jupyter Notebook terminal.

### Install NetApp DataOps Toolkit for traditional environments

To install the NetApp DataOps toolkit, complete the following steps:

1. Follow [https://github.com/NetApp/netapp-dataops-toolkit/tree/main/netapp\\_dataops\\_traditional](https://github.com/NetApp/netapp-dataops-toolkit/tree/main/netapp_dataops_traditional) and install netapp-dataops-traditional.

```
sh-4.2$ python3 -m pip install netapp-dataops-traditional
```

2. Import the functions. See [Advanced: Importable Library of Functions](#).

```

sh-4.2$
sh-4.2$ python3 -m pip install netapp-dataops-traditional
Collecting netapp-dataops-traditional
  Downloading netapp_dataops_traditional-2.1.0-py3-none-any.whl (22 kB)
Requirement already satisfied: requests in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from netapp-dataops-traditional) (2.26.0)
Requirement already satisfied: pandas in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from netapp-dataops-traditional) (0.22.0)
Requirement already satisfied: boto3 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from netapp-dataops-traditional) (1.18.45)
Collecting netapp-ontap
  Downloading netapp_ontap-9.9.1.0-py3-none-any.whl (14.7 MB)
    14.7 MB 21.8 MB/s
Collecting tabulate
  Downloading tabulate-0.8.9-py3-none-any.whl (25 kB)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from boto3->netapp-dataops-traditional) (0.10.0)
Requirement already satisfied: botocore<1.22.0,>=1.21.45 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from boto3->netapp-dataops-traditional) (1.21.45)
Requirement already satisfied: s3transfer<0.6.0,>=0.5.0 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from boto3->netapp-dataops-traditional) (0.5.0)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from botocore<1.22.0,>=1.21.45->boto3->netapp-dataops-traditional) (1.26.6)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from botocore<1.22.0,>=1.21.45->boto3->netapp-dataops-traditional) (2.8.2)
Requirement already satisfied: six>=1.5 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from python-dateutil<3.0.0,>=2.1->botocore<1.22.0,>=1.21.45->boto3->netapp-dataops-traditional) (1.16.0)
Collecting requests-toolbelt>=0.9.1
  Downloading requests_toolbelt-0.9.1-py2.py3-none-any.whl (54 kB)
    54 kB 5.2 MB/s
Collecting marshmallow>=3.2.1
  Downloading marshmallow-3.14.0-py3-none-any.whl (47 kB)
    47 kB 7.8 MB/s
Requirement already satisfied: certifi>=2017.4.17 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from requests->netapp-dataops-traditional) (2021.5.30)
Requirement already satisfied: charset-normalizer<2.0.0 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from requests->netapp-dataops-traditional) (2.0.6)
Requirement already satisfied: idna<4,>=2.5 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from requests->netapp-dataops-traditional) (3.1)
Requirement already satisfied: numpy>=1.9.0 in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from pandas->netapp-dataops-traditional) (1.19.5)
Requirement already satisfied: pytz>=2011k in ./anaconda3/envs/JupyterSystemEnv/lib/python3.6/site-packages (from pandas->netapp-dataops-traditional) (2021.1)
Installing collected packages: requests-toolbelt, marshmallow, tabulate, netapp-ontap, netapp-dataops-traditional
Successfully installed marshmallow-3.14.0 netapp-dataops-traditional-2.1.0 netapp-ontap-9.9.1.0 requests-toolbelt-0.9.1 tabulate-0.8.9
sh-4.2$

```

- After the NetApp DataOps Toolkit for traditional environments is installed, you can execute the following command and verify the installation.

```

sh-4.2$ netapp_dataops_cli.py
Error: Invalid command.

The NetApp DataOps Toolkit is a Python library that makes it simple for data scientists and data engineers to perform various data management tasks, such as provisioning a new data volume, near-instantaneously cloning a data volume, and near-instantaneously snapshotting a data volume for traceability/baselining.

Basic Commands:

    config                Create a new config file (a config file is required to perform other commands)
    help                  Print help text.
    version               Print version details.

Data Volume Management Commands:
Note: To view details regarding options/arguments for a specific command, run the command with the '-h' or '--help' option.

    clone volume          Create a new data volume that is an exact copy of an existing volume.
    create volume         Create a new data volume.
    delete volume         Delete an existing data volume.
    list volumes          List all data volumes.
    mount volume          Mount an existing data volume locally. Note: on Linux hosts - must be run as root.
    unmount volume        Unmount an existing data volume. Note: on Linux hosts - must be run as root.

Snapshot Management Commands:
Note: To view details regarding options/arguments for a specific command, run the command with the '-h' or '--help' option.

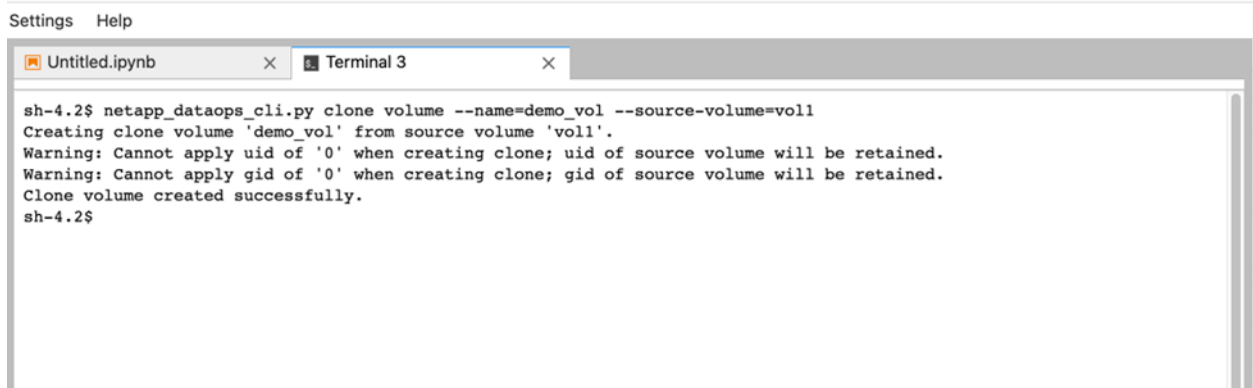
    create snapshot       Create a new snapshot for a data volume.
    delete snapshot       Delete an existing snapshot for a data volume.
    list snapshots        List all snapshots for a data volume.
    restore snapshot       Restore a snapshot for a data volume (restore the volume to its exact state at the time that the snapshot was created).

```

## Example operation: Clone a volume using the NetApp DataOps Toolkit

You can use the NetApp DataOps Toolkit to clone the mounted volume. The following command will help to clone the volume in use.

```
sh-4.2$ netapp_dataops_cli.py clone volume --name=demo_vol --source-volume=vol1
Creating clone volume 'demo_vol' from source volume 'vol1'.
Clone volume created successfully.
sh-4.2$
```



## Other operations

You can also use the NetApp DataOps Toolkit to perform other operations. For example, you can use the NetApp DataOps Toolkit to create a read-only snapshot of the mounted volume for traceability purposes. For more information, see the toolkit's "create snapshot" operation.

## Conclusion

Customers can benefit by taking advantage of the enterprise-level performance, scalability for high-performance workloads, and data protection features provided by 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. This solution brings automation of the data science process and twelve times acceleration in the deployment of AI products. This solution helps data scientists run a Kubernetes job with the attached FSx ONTAP PVC. They have the ability to snapmirror the data from on-premises ONTAP, which eliminates the migration of data. Data scientists can also create a Jupyter Notebook with the persistent volume and the traditional NetApp DataOps Toolkit; they do not have to redo the environment every time. Data scientists can destroy the environment whenever it is not needed because the data resides in the persistent volume. A traditional NetApp DataOps Toolkit provides the data protection features required.

## Where to find additional information

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

- NetApp Product Documentation  
<https://www.netapp.com/support-and-training/documentation/>
- Amazon FSx for NetApp ONTAP documentation  
<https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/what-is-fsx-ontap.html>
- Iguazio Product Page  
<https://www.iguazio.com/>
- Iguazio Product Documentation  
<https://www.iguazio.com/docs/latest-release/>

## Version history

Version	Date	Document version history
Version 1.0	June 2022	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**

Copyright © 2022 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

### **Trademark information**

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

TR-4934-0722