Data Pipelines, Data Lakes and Management

NetApp Solutions

NetApp
July 31, 2024

This PDF was generated from https://docs.netapp.com/us-en/netapp-solutions/ai/mlops_fsxn_s3_integration.html on July 31, 2024. Always check docs.netapp.com for the latest.

Table of Contents

Data Pipelines, Data Lakes and Management

AWS FSx for NetApp ONTAP (FSxN) for MLOps

This section delves into the practical application of AI infrastructure development, providing an end-to-end walkthrough of constructing an MLOps pipeline using FSxN. Comprising three comprehensive examples, it guides you to meet your MLOps needs via this powerful data management platform.

Author(s):
Jian Jian (Ken), Senior Data & Applied Scientist, NetApp

These articles focus on:

1. Part 1 - Integrating AWS FSx for NetApp ONTAP (FSxN) as a private S3 bucket into AWS SageMaker
2. Part 2 - Leveraging AWS FSx for NetApp ONTAP (FSxN) as a Data Source for Model Training in SageMaker
3. Part 3 - Building A Simplified MLOps Pipeline (CI/CT/CD)

By the end of this section, you will have gained a solid understanding of how to use FSxN to streamline MLOps processes.

Part 1 - Integrating AWS FSx for NetApp ONTAP (FSxN) as a private S3 bucket into AWS SageMaker

This section provides a guide on configuring FSxN as a private S3 bucket using AWS SageMaker.

Author(s):
Jian Jian (Ken), Senior Data & Applied Scientist, NetApp

Introduction

Using SageMaker as an example, this page provides guidance on configuring FSxN as a private S3 bucket.

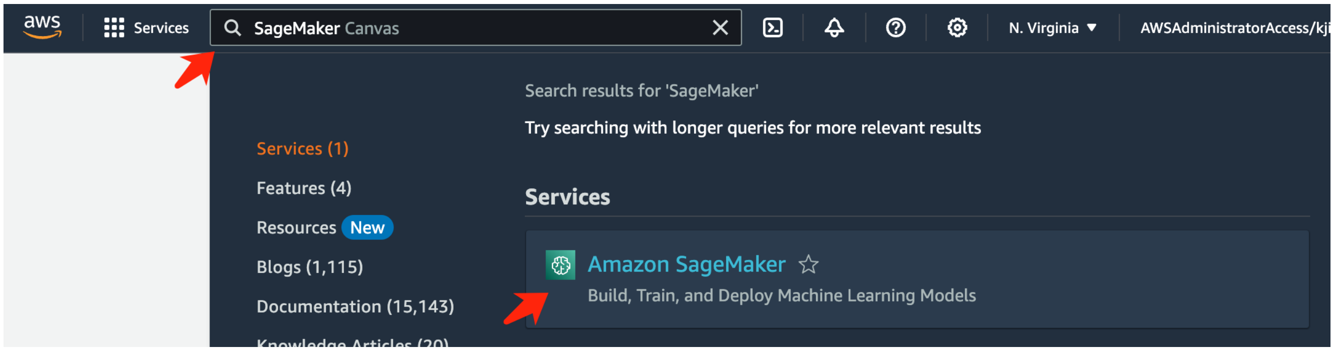
For more information about FSxN, please take a look at this presentation (Video Link)

User Guide

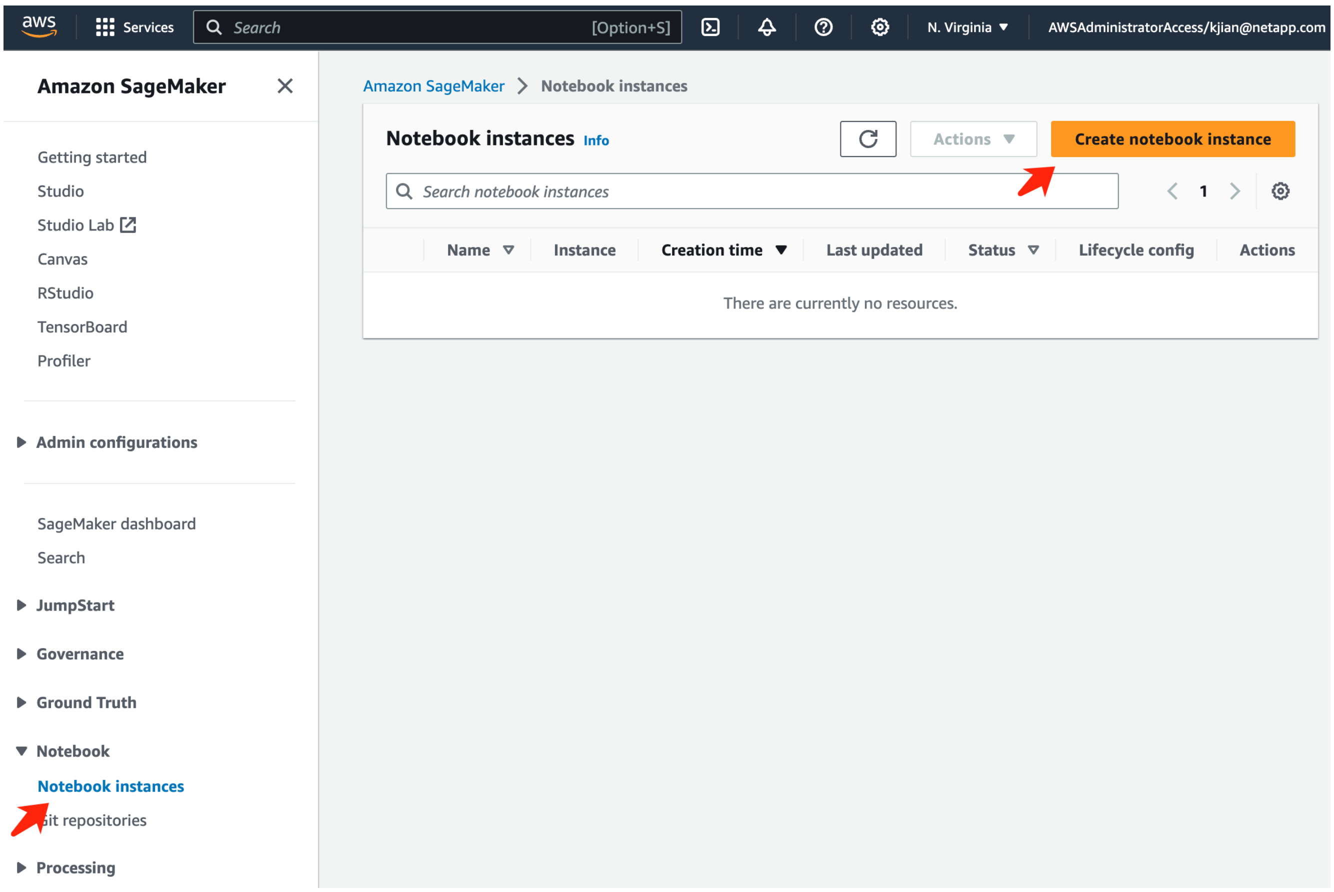
Server creation

Create a SageMaker Notebook Instance

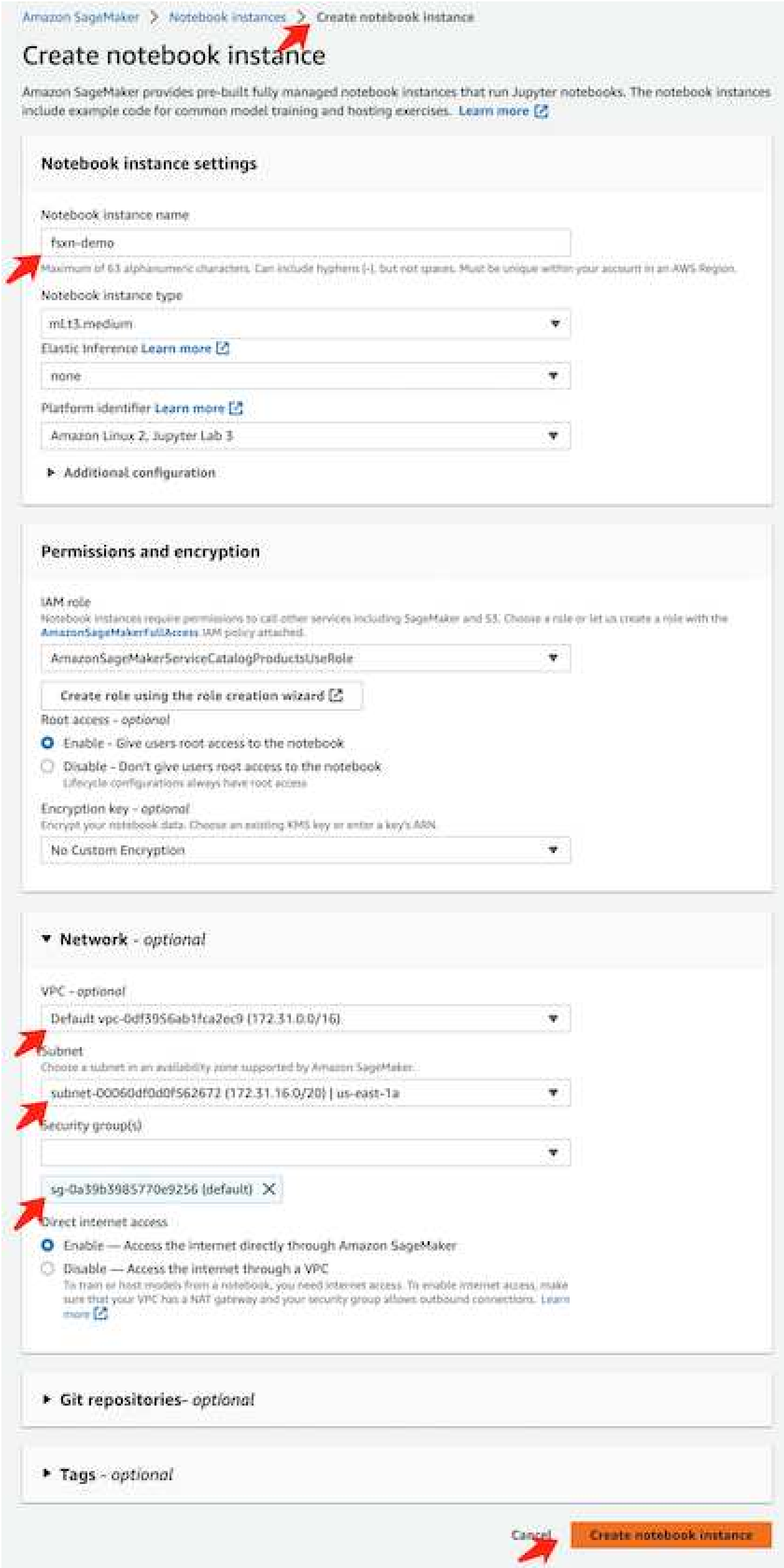
1. Open AWS console. In the search panel, search SageMaker and click the service **Amazon SageMaker**.



2. Open the **Notebook instances** under Notebook tab, click the orange button **Create notebook instance**.

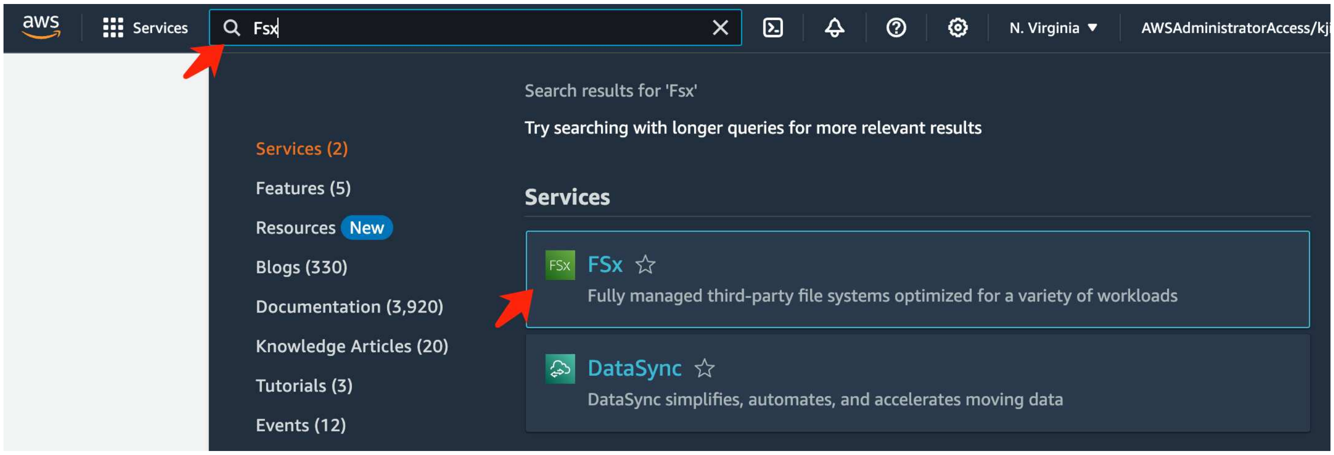


3. In the creation page,
Enter the **Notebook instance name**
Expand the **Network** panel
Leave other entries default and select a **VPC**, **Subnet**, and **Security group(s)**. (This **VPC** and **Subnet** will be used to create FSxN file system later)
Click the orange button **Create notebook instance** at the bottom right.

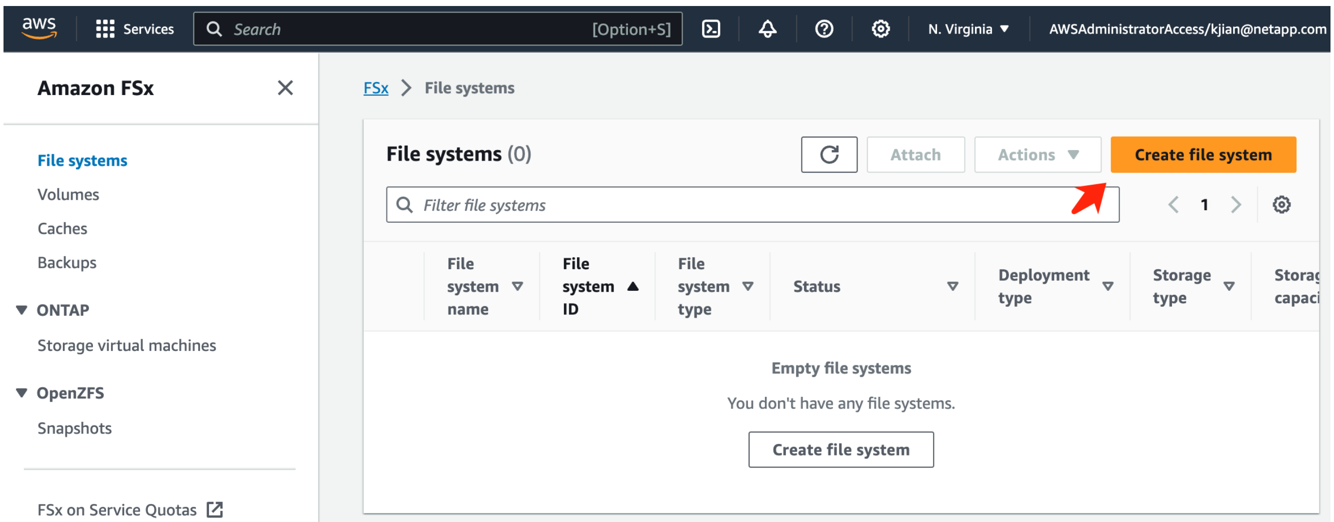


Create an FSxN File System

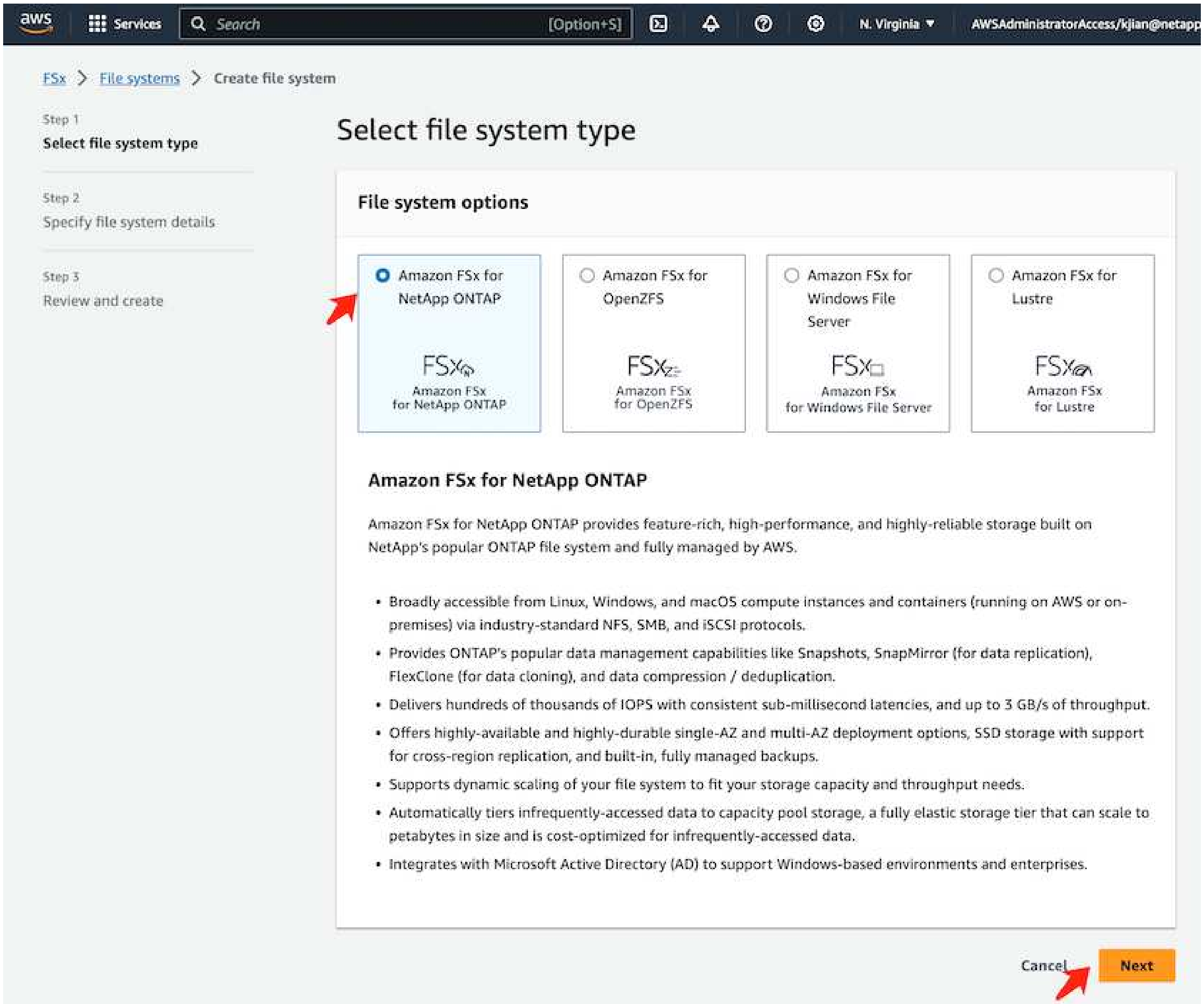
1. Open AWS console. In the search panel, search Fsx and click the service **FSx**.



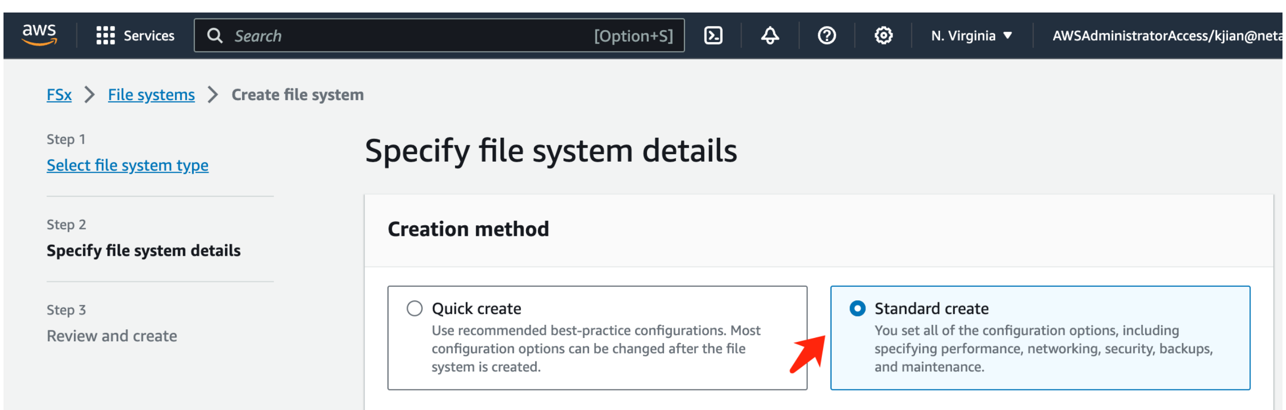
2. Click **Create file system**.



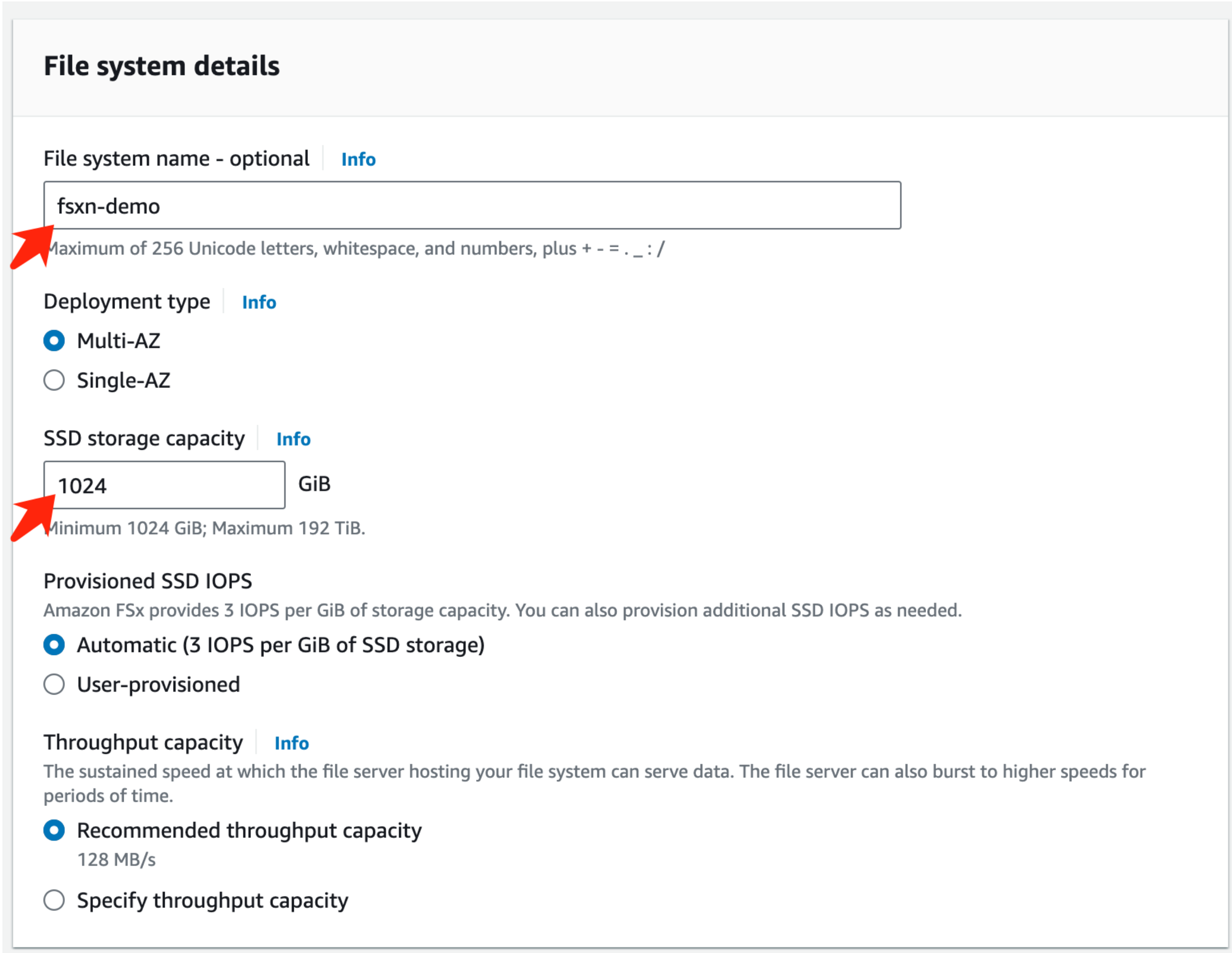
3. Select the first card **FSx for NetApp ONTAP** and click **Next**.



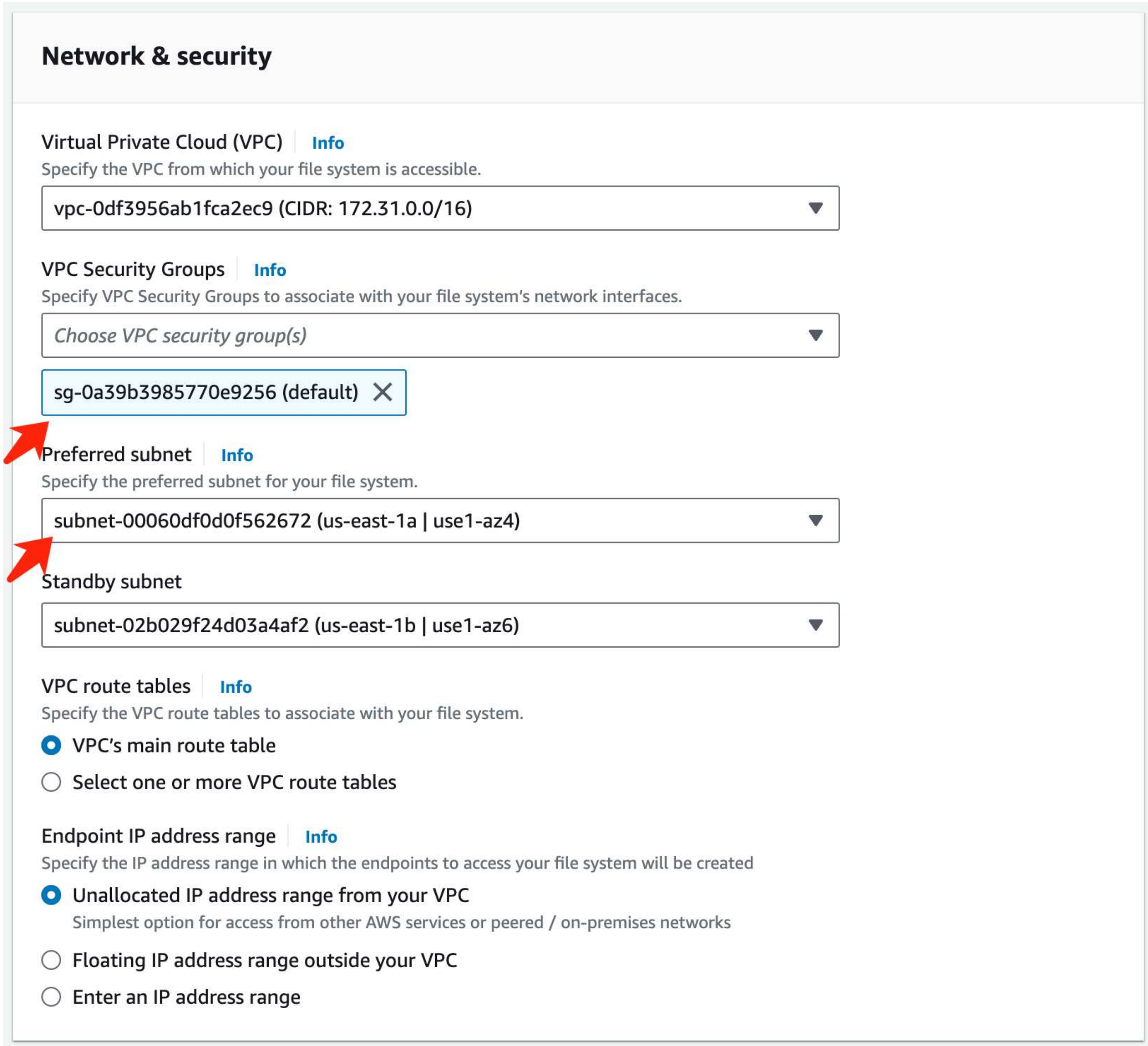
4. In the details configuration page.
    a. Select the **Standard create** option.



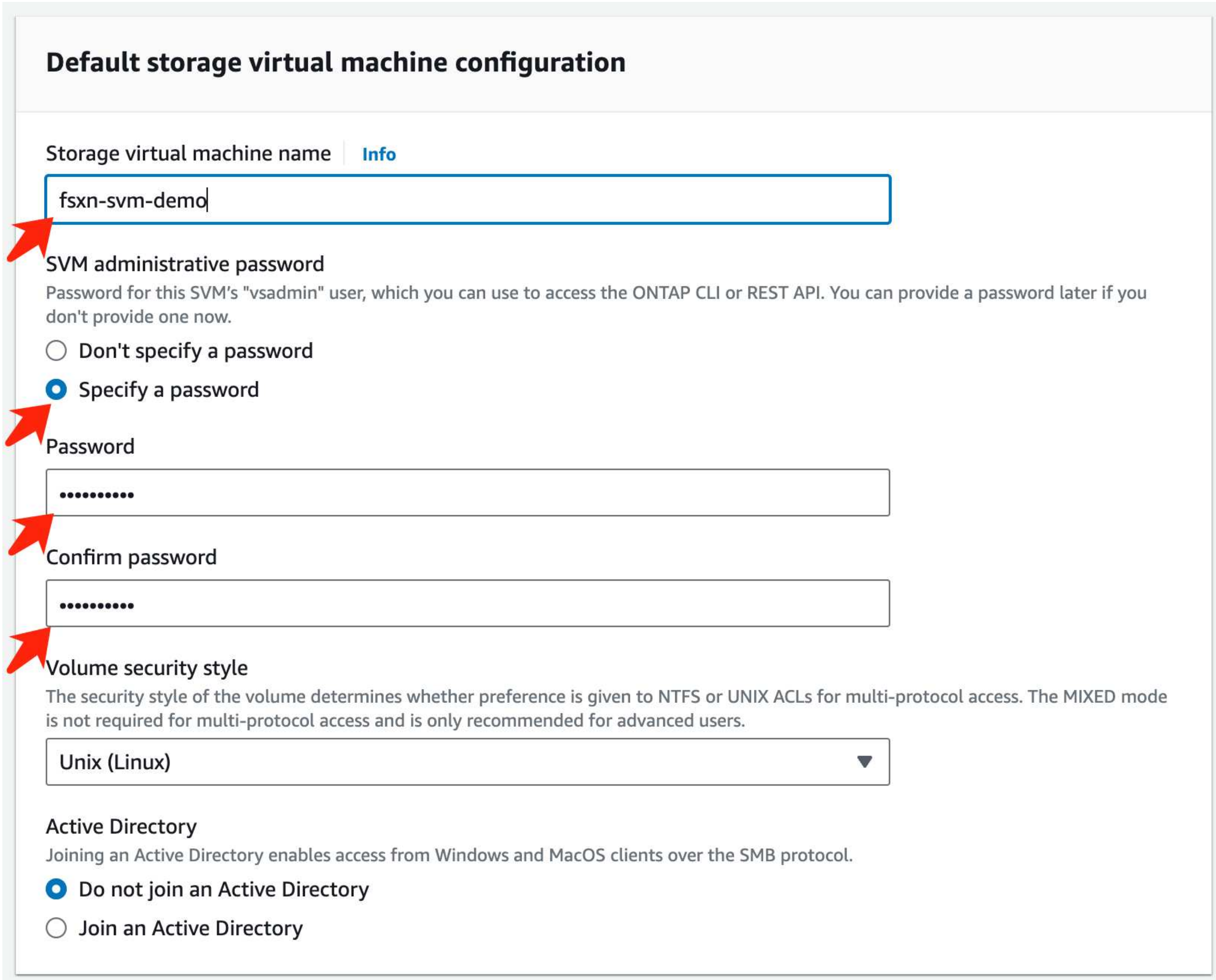
    b. Enter the **File system name** and the **SSD storage capacity**.



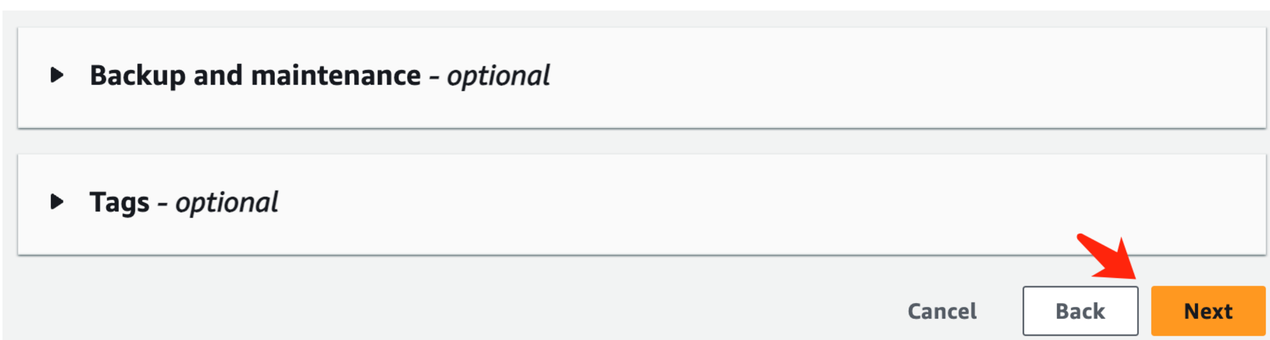
c. Make sure to use the **VPC** and **subnet** same to the **SageMaker Notebook** instance.



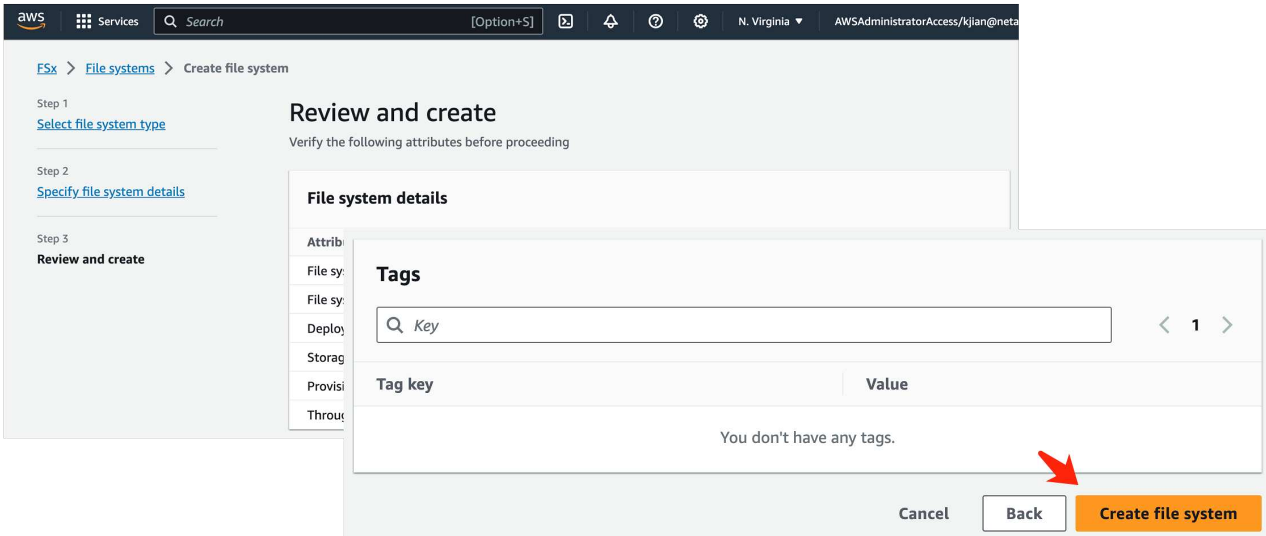
d. Enter the **Storage virtual machine** name and **Specify a password** for your SVM (storage virtual machine).



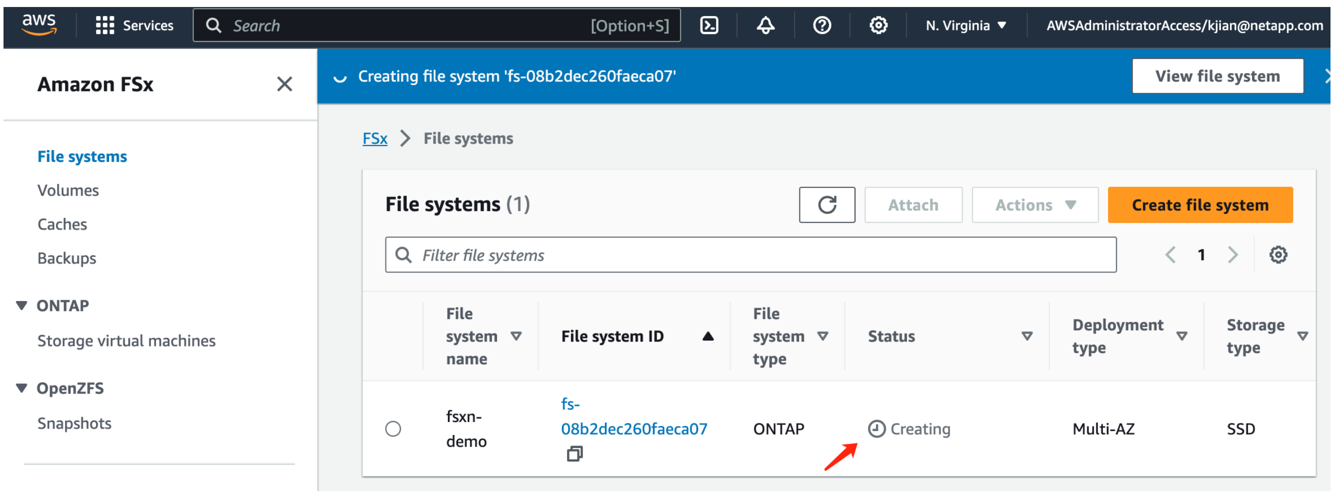
e. Leave other entries default and click the orange button **Next** at the bottom right.



f. Click the orange button **Create file system** at the bottom right of the review page.



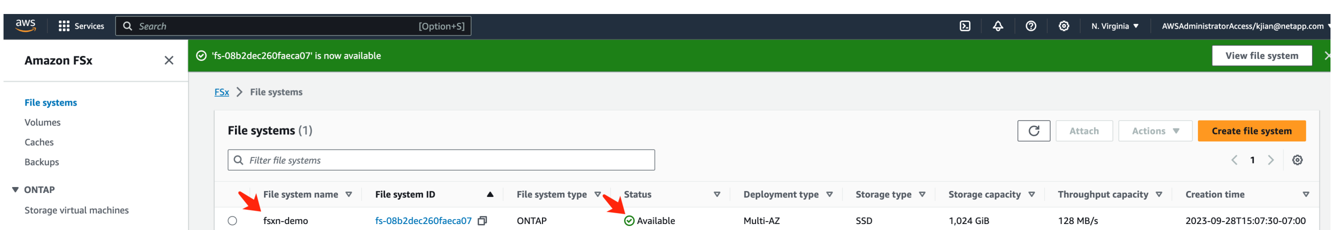
5. It may takes about **20-40 minutes** to spin up the FSx file system.



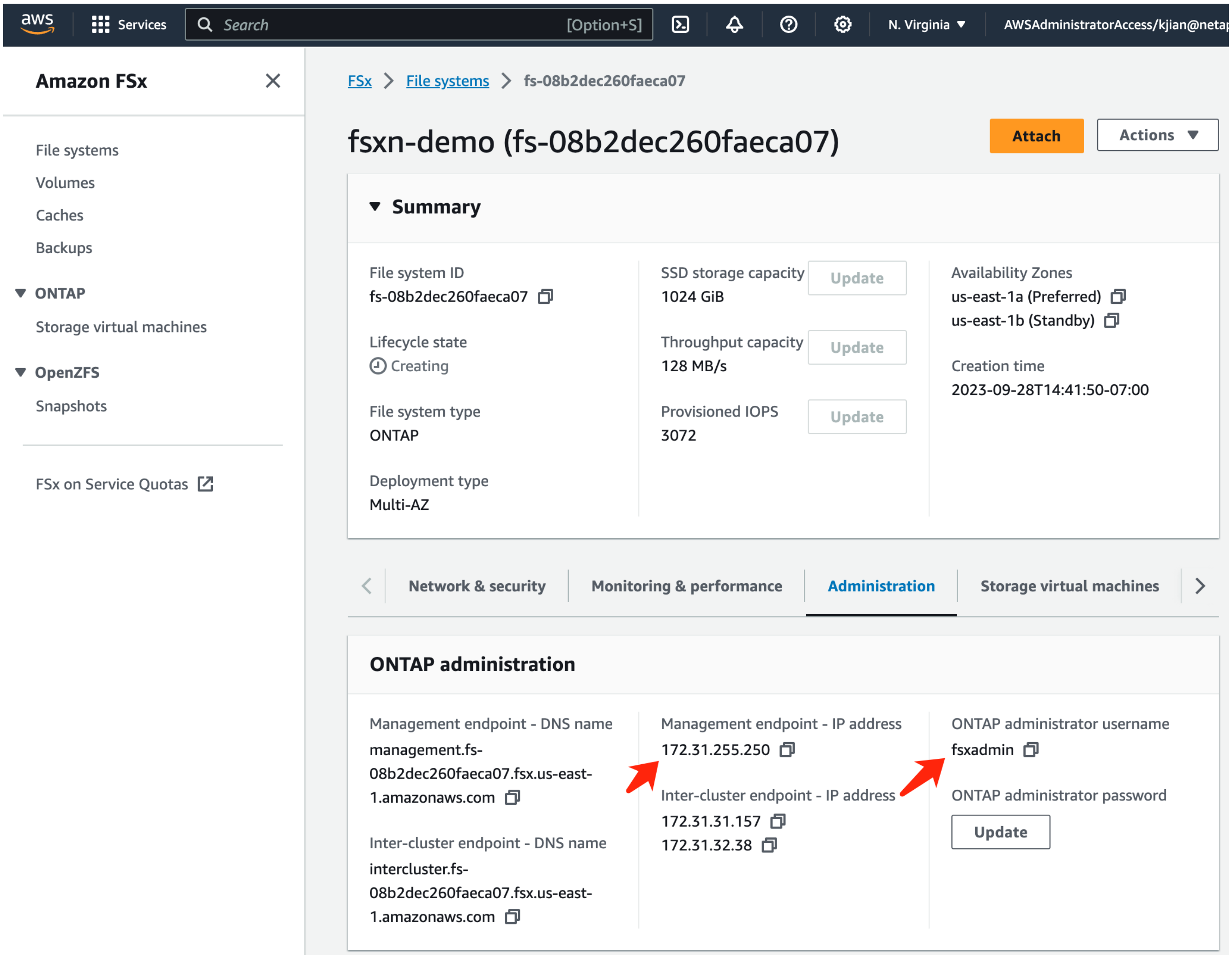
Server Configuration

ONTAP Configuration

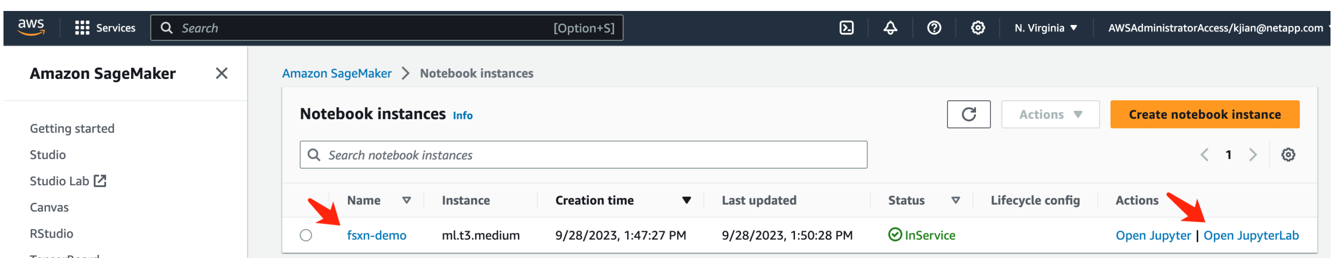
1. Open the created FSx file system. Please make sure the status is **Available**.



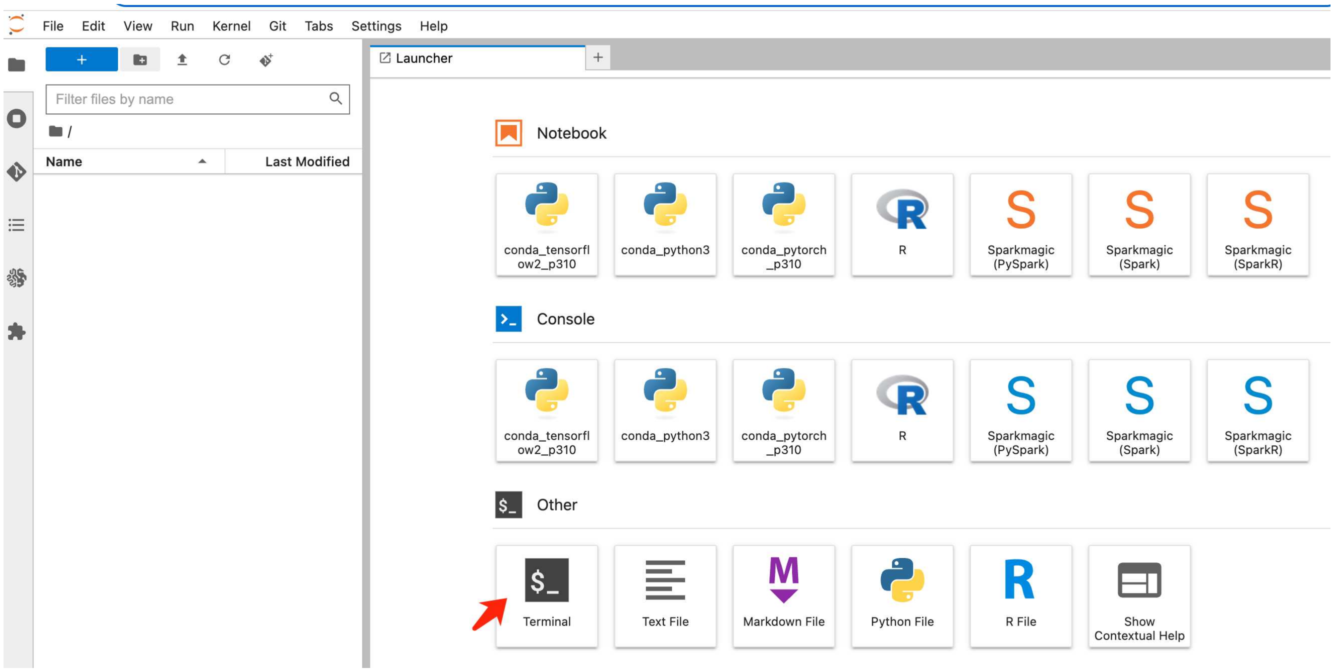
2. Select the **Administration** tab and keep the **Management endpoint - IP address** and **ONTAP administrator username**.



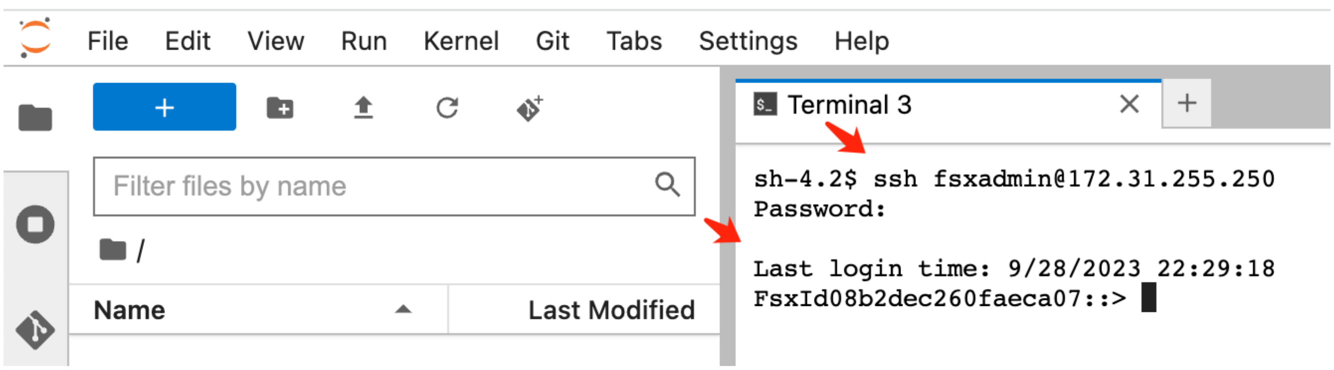
3. Open the created **SageMaker Notebook instance** and click **Open JupyterLab**.



4. In the Jupyter Lab page, open a new **Terminal**.

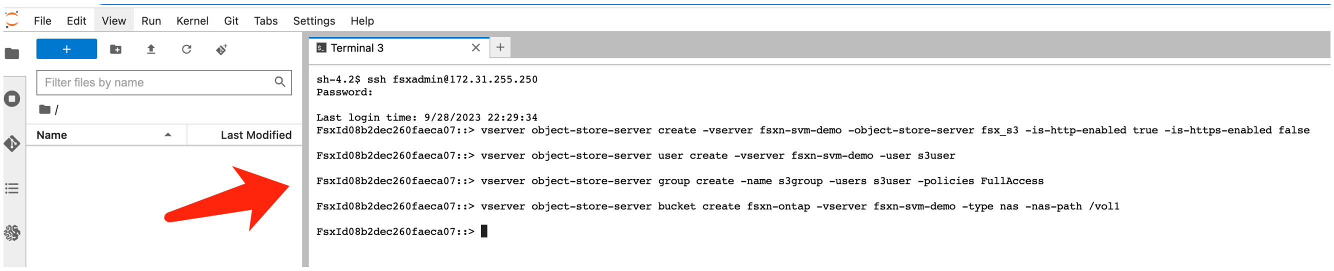


5. Enter the ssh command ssh <admin user name>@<ONTAP server IP> to login to the FSxN ONTAP file system. (The user name and IP address are retrieved from the step 2)
   Please use the password used when creating the **Storage virtual machine**.



6. Execute the commands in the following order.
   We use **fsxn-ontap** as the name for the **FSxN private S3 bucket name**.
   Please use the **storage virtual machine name** for the **-vserver** argument.

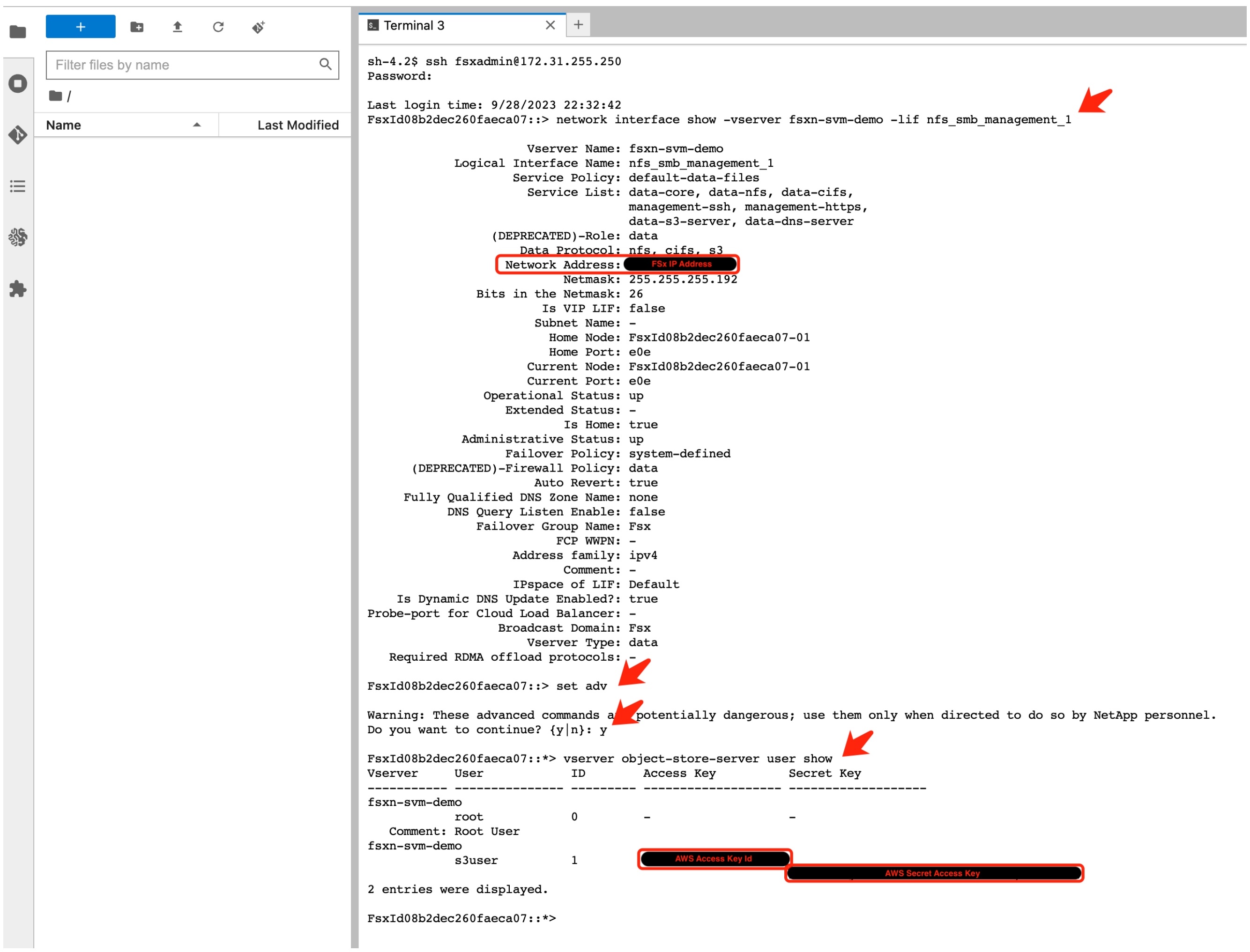
```
vserver object-store-server create -vserver fsxn-svm-demo -object-store
-server fsx_s3 -is-http-enabled true -is-https-enabled false

vserver object-store-server user create -vserver fsxn-svm-demo -user
s3user

vserver object-store-server group create -name s3group -users s3user
-policies FullAccess

vserver object-store-server bucket create fsxn-ontap -vserver fsxn-svm-
demo -type nas -nas-path /vol1
```



7. Execute the below commands to retrieve the endpoint IP and credentials for FSxN private S3.

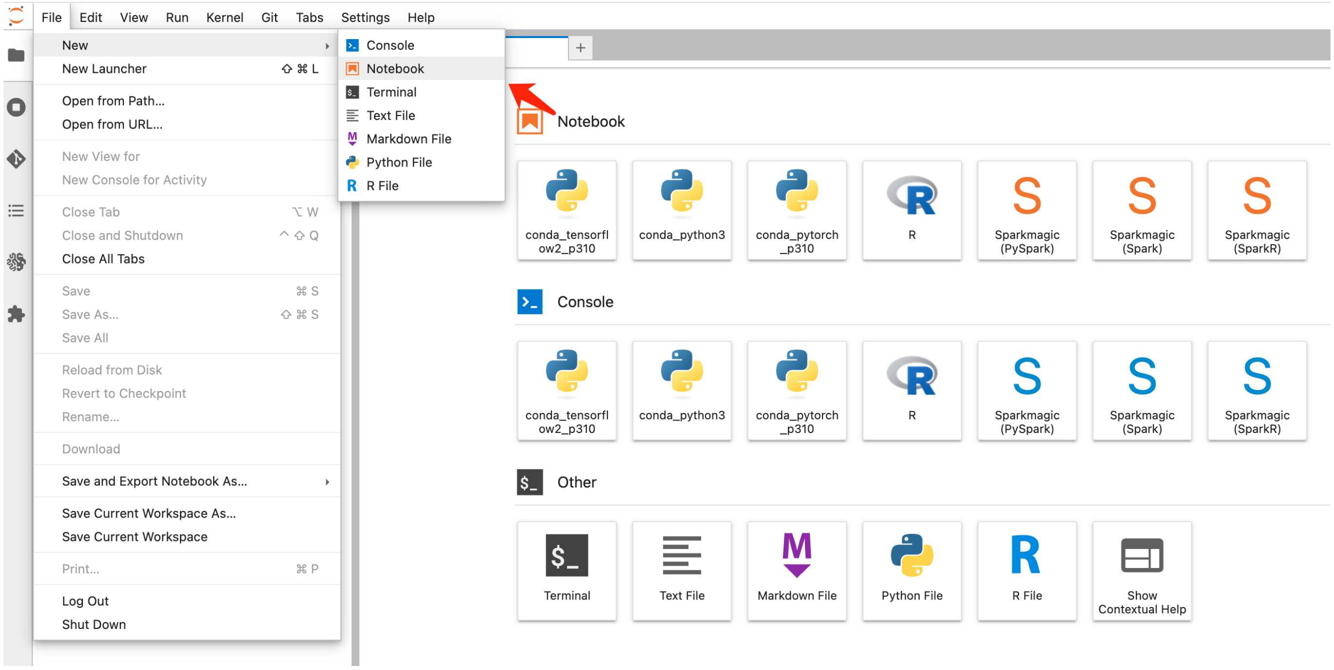
```
network interface show -vserver fsxn-svm-demo -lif nfs_smb_management_1

set adv

vserver object-store-server user show
```

8. Keep the endpoint IP and credential for future use.



Client Configuration

1. In SageMaker Notebook instance, create a new Jupyter notebook.



2. Use the below code as a work around solution to upload files to FSxN private S3 bucket.
   For a comprehensive code example please refer to this notebook.
   fsxn_demo.ipynb

```
# Setup configurations
# -------- Manual configurations --------
seed: int = 77                                          # Random
seed
bucket_name: str = 'fsxn-ontap'                         # The bucket
name in ONTAP
aws_access_key_id = '<Your ONTAP bucket key id>'        # Please get
this credential from ONTAP
aws_secret_access_key = '<Your ONTAP bucket access key>'    # Please get
this credential from ONTAP
fsx_endpoint_ip: str = '<Your FSxN IP address>'         # Please get
this IP address from FSXN
# -------- Manual configurations --------

# Workaround
## Permission patch
!mkdir -p vol1
!sudo mount -t nfs $fsx_endpoint_ip:/vol1 /home/ec2-user/SageMaker/vol1
!sudo chmod 777 /home/ec2-user/SageMaker/vol1

## Authentication for FSxN as a Private S3 Bucket
!aws configure set aws_access_key_id $aws_access_key_id
```

```
!aws configure set aws_secret_access_key $aws_secret_access_key

## Upload file to the FSxN Private S3 Bucket
%%capture
local_file_path: str = <Your local file path>

!aws s3 cp --endpoint-url http://$fsx_endpoint_ip /home/ec2-user
/SageMaker/$local_file_path  s3://$bucket_name/$local_file_path

# Read data from FSxN Private S3 bucket
## Initialize a s3 resource client
import boto3

# Get session info
region_name = boto3.session.Session().region_name

# Initialize Fsxn S3 bucket object
# --- Start integrating SageMaker with FSXN ---
# This is the only code change we need to incorporate SageMaker with
FSXN
s3_client: boto3.client = boto3.resource(
    's3',
    region_name=region_name,
    aws_access_key_id=aws_access_key_id,
    aws_secret_access_key=aws_secret_access_key,
    use_ssl=False,
    endpoint_url=f'http://{fsx_endpoint_ip}',
    config=boto3.session.Config(
        signature_version='s3v4',
        s3={'addressing_style': 'path'}
    )
)
# --- End integrating SageMaker with FSXN ---

## Read file byte content
bucket = s3_client.Bucket(bucket_name)

binary_data = bucket.Object(data.filename).get()['Body']
```

This concludes the integration between FSxN and the SageMaker instance.

Useful debugging checklist

- Ensure that the SageMaker Notebook instance and FSxN file system are in the same VPC.
- Remember to run the **set dev** command on ONTAP to set the privilege level to **dev**.

FAQ (As of Sep 27, 2023)

Q: Why am I getting the error "**An error occurred (NotImplemented) when calling the CreateMultipartUpload operation: The s3 command you requested is not implemented**" when uploading files to FSxN?

A: As a private S3 bucket, FSxN supports uploading files up to 100MB. When using the S3 protocol, files larger than 100MB are divided into 100MB chunks, and the 'CreateMultipartUpload' function is called. However, the current implementation of FSxN private S3 does not support this function.

Q: Why am I getting the error "**An error occurred (AccessDenied) when calling the PutObject operations: Access Denied**" when uploading files to FSxN?

A: To access the FSxN private S3 bucket from a SageMaker Notebook instance, switch the AWS credentials to the FSxN credentials. However, granting write permission to the instance requires a workaround solution that involves mounting the bucket and running the 'chmod' shell command to change the permissions.

Q: How can I integrate the FSxN private S3 bucket with other SageMaker ML services?

A: Unfortunately, the SageMaker services SDK does not provide a way to specify the endpoint for the private S3 bucket. As a result, FSxN S3 is not compatible with SageMaker services such as Sagemaker Data Wrangler, Sagemaker Clarify, Sagemaker Glue, Sagemaker Athena, Sagemaker AutoML, and others.

Part 2 - Leveraging AWS FSx for NetApp ONTAP (FSxN) as a Data Source for Model Training in SageMaker

This article is a tutorial on using AWS FSx for NetApp ONTAP (FSxN) for training PyTorch models in SageMaker, specifically for a tire quality classification project.

Author(s):
Jian Jian (Ken), Senior Data & Applied Scientist, NetApp

Introduction

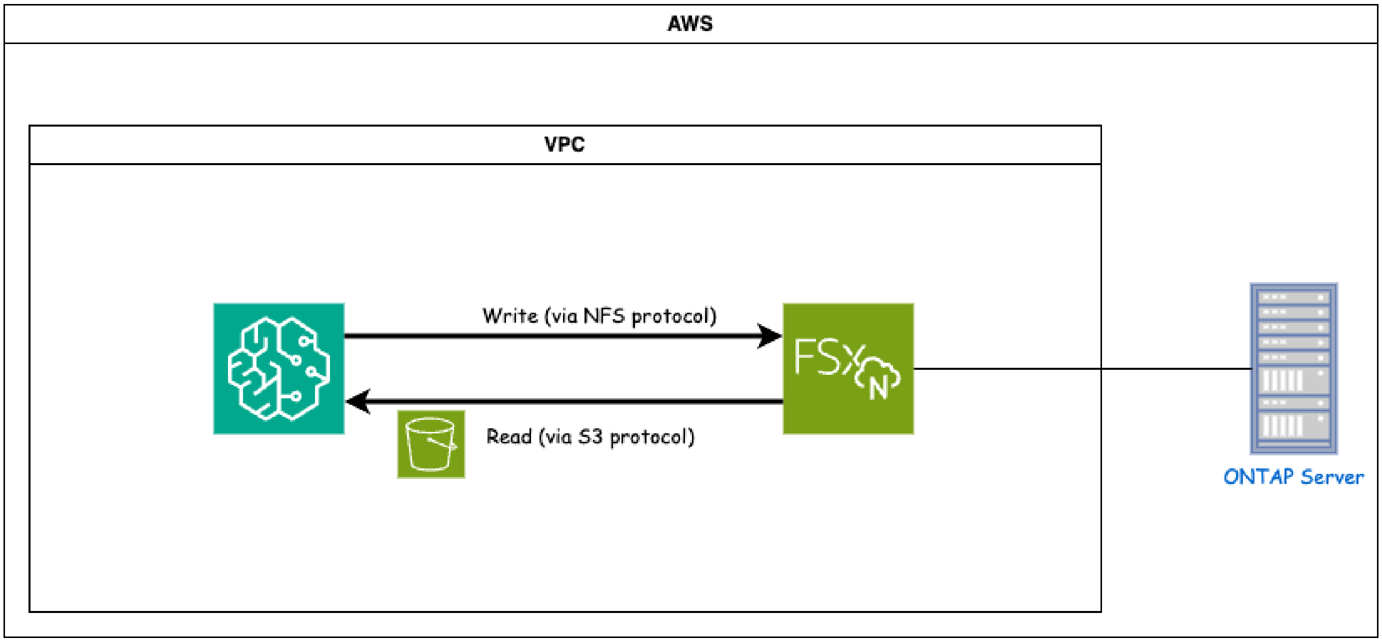
This tutorial offers a practical example of a computer vision classification project, providing hands-on experience in building ML models that utilize FSxN as the data source within the SageMaker environment. The project focuses on using PyTorch, a deep learning framework, to classify tire quality based on tire images. It emphasizes the development of machine learning models using FSxN as the data source in Amazon SageMaker.

What is FSxN

Amazon FSx for NetApp ONTAP is indeed a fully managed storage solution offered by AWS. It leverages NetApp's ONTAP file system to provide reliable and high-performance storage. With support for protocols like NFS, SMB, and iSCSI, it allows seamless access from different compute instances and containers. The service is designed to deliver exceptional performance, ensuring fast and efficient data operations. It also offers high availability and durability, ensuring that your data remains accessible and protected. Additionally, the storage capacity of Amazon FSx for NetApp ONTAP is scalable, allowing you to easily adjust it according to your needs.

Prerequisite

Network Environment



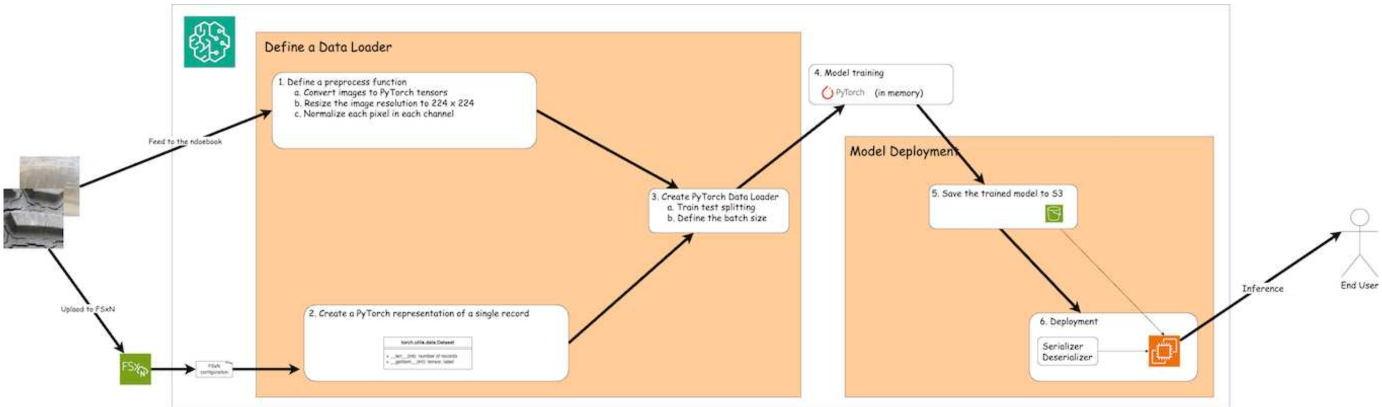
FSxN (Amazon FSx for NetApp ONTAP) is an AWS storage service. It includes a file system running on the NetApp ONTAP system and an AWS-managed system virtual machine (SVM) that connects to it. In the provided diagram, the NetApp ONTAP server managed by AWS is located outside the VPC. The SVM serves as the intermediary between SageMaker and the NetApp ONTAP system, receiving operation requests from SageMaker and forwarding them to the underlying storage. To access FSxN, SageMaker must be placed within the same VPC as the FSxN deployment. This configuration ensures communication and data access between SageMaker and FSxN.

Data Access

In real-world scenarios, data scientists typically utilize the existing data stored in FSxN to build their machine learning models. However, for demonstration purposes, since the FSxN file system is initially empty after creation, it is necessary to manually upload the training data. This can be achieved by mounting FSxN as a volume to SageMaker. Once the file system is successfully mounted, you can upload your dataset to the mounted location, making it accessible for training your models within the SageMaker environment. This approach allows you to leverage the storage capacity and capabilities of FSxN while working with SageMaker for model development and training.

The data reading process involves configuring FSxN as a private S3 bucket. To learn the detailed configuration instructions, please refer to Part 1 - Integrating AWS FSx for NetApp ONTAP (FSxN) as a private S3 bucket into AWS SageMaker

Integration Overview



The workflow of using training data in FSxN to build a deep learning model in SageMaker can be summarized into three main steps: data loader definition, model training, and deployment. At a high level, these steps form the foundation of an MLOps pipeline. However, each step involves several detailed sub-steps for a comprehensive implementation. These sub-steps encompass various tasks such as data preprocessing, dataset splitting, model configuration, hyperparameter tuning, model evaluation, and model deployment. These steps ensure a thorough and effective process for building and deploying deep learning models using training data from FSxN within the SageMaker environment.

Step-by-Step Integration

Data Loader

In order to train a PyTorch deep learning network with data, a data loader is created to facilitate the feeding of data. The data loader not only defines the batch size but also determines the procedure for reading and preprocessing each record within the batch. By configuring the data loader, we can handle the processing of data in batches, enabling training of the deep learning network.

The data loader consists of 3 parts.

Preprocessing Function

```
from torchvision import transforms

preprocess = transforms.Compose([
    transforms.ToTensor(),
    transforms.Resize((224,224)),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406],
        std=[0.229, 0.224, 0.225]
    )
])
```

The above code snippet demonstrates the definition of image preprocessing transformations using the **torchvision.transforms** module. In this turtorial, the preprocess object is created to apply a series of transformations. Firstly, the **ToTensor()** transformation converts the image into a tensor representation. Subsequently, the **Resize 224,224** transformation resizes the image to a fixed size of 224x224 pixels. Finally, the **Normalize()** transformation normalizes the tensor values by subtracting the mean and dividing by the standard deviation along each channel. The mean and standard deviation values used for normalization are commonly employed in pre-trained neural network models. Overall, this code prepares the image data for further processing or input into a pre-trained model by converting it to a tensor, resizing it, and normalizing the pixel values.

The PyTorch Dataset Class

```
import torch
from io import BytesIO
from PIL import Image


class FSxNImageDataset(torch.utils.data.Dataset):
    def __init__(self, bucket, prefix='', preprocess=None):
        self.image_keys = [
            s3_obj.key
            for s3_obj in list(bucket.objects.filter(Prefix=prefix).all())
        ]
        self.preprocess = preprocess

    def __len__(self):
        return len(self.image_keys)

    def __getitem__(self, index):
        key = self.image_keys[index]
        response = bucket.Object(key)

        label = 1 if key[13:].startswith('defective') else 0

        image_bytes = response.get()['Body'].read()
        image = Image.open(BytesIO(image_bytes))
        if image.mode == 'L':
            image = image.convert('RGB')

        if self.preprocess is not None:
            image = self.preprocess(image)
        return image, label
```

This class provides functionality to obtain the total number of records in the dataset and defines the method for reading data for each record. Within the ***getitem*** function, the code utilizes the boto3 S3 bucket object to retrieve the binary data from FSxN. The code style for accessing data from FSxN is similar to reading data from Amazon S3. The subsequent explanation delves into the creation process of the private S3 object **bucket**.

FSxN as a private S3 repository

```
seed = 77                                                   # Random seed
bucket_name = '<Your ONTAP bucket name>'                    # The bucket
name in ONTAP
aws_access_key_id = '<Your ONTAP bucket key id>'            # Please get
this credential from ONTAP
aws_secret_access_key = '<Your ONTAP bucket access key>'    # Please get
this credential from ONTAP
fsx_endpoint_ip = '<Your FSxN IP address>'                  # Please get
this IP address from FSXN
```

```
import boto3

# Get session info
region_name = boto3.session.Session().region_name

# Initialize Fsxn S3 bucket object
# --- Start integrating SageMaker with FSXN ---
# This is the only code change we need to incorporate SageMaker with FSXN
s3_client: boto3.client = boto3.resource(
    's3',
    region_name=region_name,
    aws_access_key_id=aws_access_key_id,
    aws_secret_access_key=aws_secret_access_key,
    use_ssl=False,
    endpoint_url=f'http://{fsx_endpoint_ip}',
    config=boto3.session.Config(
        signature_version='s3v4',
        s3={'addressing_style': 'path'}
    )
)
# s3_client = boto3.resource('s3')
bucket = s3_client.Bucket(bucket_name)
# --- End integrating SageMaker with FSXN ---
```

To read data from FSxN in SageMaker, a handler is created that points to the FSxN storage using the S3 protocol. This allows FSxN to be treated as a private S3 bucket. The handler configuration includes specifying the IP address of the FSxN SVM, the bucket name, and the necessary credentials. For a comprehensive explanation on obtaining these configuration items, please refer to the document at Part 1 - Integrating AWS FSx for NetApp ONTAP (FSxN) as a private S3 bucket into AWS SageMaker.

In the example mentioned above, the bucket object is used to instantiate the PyTorch dataset object. The dataset object will be further explained in the subsequent section.

The PyTorch Data Loader

```
from torch.utils.data import DataLoader
torch.manual_seed(seed)

# 1. Hyperparameters
batch_size = 64

# 2. Preparing for the dataset
dataset = FSxNImageDataset(bucket, 'dataset/tyre', preprocess=preprocess)

train, test = torch.utils.data.random_split(dataset, [1500, 356])

data_loader = DataLoader(dataset, batch_size=batch_size, shuffle=True)
```

In the example provided, a batch size of 64 is specified, indicating that each batch will contain 64 records. By combining the PyTorch **Dataset** class, the preprocessing function, and the training batch size, we obtain the data loader for training. This data loader facilitates the process of iterating through the dataset in batches during the training phase.

Model Training

```
from torch import nn


class TyreQualityClassifier(nn.Module):
    def __init__(self):
        super().__init__()
        self.model = nn.Sequential(
            nn.Conv2d(3,32,(3,3)),
            nn.ReLU(),
            nn.Conv2d(32,32,(3,3)),
            nn.ReLU(),
            nn.Conv2d(32,64,(3,3)),
            nn.ReLU(),
            nn.Flatten(),
            nn.Linear(64*(224-6)*(224-6),2)
        )
    def forward(self, x):
        return self.model(x)
```

```
import datetime

num_epochs = 2
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')

model = TyreQualityClassifier()
fn_loss = torch.nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=1e-3)


model.to(device)
for epoch in range(num_epochs):
    for idx, (X, y) in enumerate(data_loader):
        X = X.to(device)
        y = y.to(device)

        y_hat = model(X)

        loss = fn_loss(y_hat, y)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        current_time = datetime.datetime.now().strftime("%Y-%m-%d
%H:%M:%S")
        print(f"Current Time: {current_time} - Epoch [{epoch+1}/
{num_epochs}]- Batch [{idx + 1}] - Loss: {loss}", end='\r')
```

This code implements a standard PyTorch training process. It defines a neural network model called **TyreQualityClassifier** using convolutional layers and a linear layer to classify tire quality. The training loop iterates over data batches, computes the loss, and updates the model's parameters using backpropagation and optimization. Additionally, it prints the current time, epoch, batch, and loss for monitoring purposes.

Model Deployment

Deployment

```python
import io
import os
import tarfile
import sagemaker

# 1. Save the PyTorch model to memory
buffer_model = io.BytesIO()
traced_model = torch.jit.script(model)
torch.jit.save(traced_model, buffer_model)

# 2. Upload to AWS S3
sagemaker_session = sagemaker.Session()
bucket_name_default = sagemaker_session.default_bucket()
model_name = f'tyre_quality_classifier.pth'

# 2.1. Zip PyTorch model into tar.gz file
buffer_zip = io.BytesIO()
with tarfile.open(fileobj=buffer_zip, mode="w:gz") as tar:
    # Add PyTorch pt file
    file_name = os.path.basename(model_name)
    file_name_with_extension = os.path.split(file_name)[-1]
    tarinfo = tarfile.TarInfo(file_name_with_extension)
    tarinfo.size = len(buffer_model.getbuffer())
    buffer_model.seek(0)
    tar.addfile(tarinfo, buffer_model)

# 2.2. Upload the tar.gz file to S3 bucket
buffer_zip.seek(0)
boto3.resource('s3') \
    .Bucket(bucket_name_default) \
    .Object(f'pytorch/{model_name}.tar.gz') \
    .put(Body=buffer_zip.getvalue())
```

The code saves the PyTorch model to **Amazon S3** because SageMaker requires the model to be stored in S3 for deployment. By uploading the model to **Amazon S3**, it becomes accessible to SageMaker, allowing for the deployment and inference on the deployed model.

```python
import time
from sagemaker.pytorch import PyTorchModel
from sagemaker.predictor import Predictor
from sagemaker.serializers import IdentitySerializer
from sagemaker.deserializers import JSONDeserializer


class TyreQualitySerializer(IdentitySerializer):
```

```python
    CONTENT_TYPE = 'application/x-torch'

    def serialize(self, data):
        transformed_image = preprocess(data)
        tensor_image = torch.Tensor(transformed_image)

        serialized_data = io.BytesIO()
        torch.save(tensor_image, serialized_data)
        serialized_data.seek(0)
        serialized_data = serialized_data.read()

        return serialized_data


class TyreQualityPredictor(Predictor):
    def __init__(self, endpoint_name, sagemaker_session):
        super().__init__(
            endpoint_name,
            sagemaker_session=sagemaker_session,
            serializer=TyreQualitySerializer(),
            deserializer=JSONDeserializer(),
        )

sagemaker_model = PyTorchModel(
    model_data=f's3://{bucket_name_default}/pytorch/{model_name}.tar.gz',
    role=sagemaker.get_execution_role(),
    framework_version='2.0.1',
    py_version='py310',
    predictor_cls=TyreQualityPredictor,
    entry_point='inference.py',
    source_dir='code',
)

timestamp = int(time.time())
pytorch_endpoint_name = '{}-{}-{}'.format('tyre-quality-classifier', 'pt',
timestamp)
sagemaker_predictor = sagemaker_model.deploy(
    initial_instance_count=1,
    instance_type='ml.p3.2xlarge',
    endpoint_name=pytorch_endpoint_name
)
```

This code facilitates the deployment of a PyTorch model on SageMaker. It defines a custom serializer, **TyreQualitySerializer**, which preprocesses and serializes input data as a PyTorch tensor. The **TyreQualityPredictor** class is a custom predictor that utilizes the defined serializer and a **JSONDeserializer**. The code also creates a **PyTorchModel** object to specify the model's S3 location, IAM role, framework version, and entry point for inference. The code generates a timestamp and constructs an endpoint name based on the model and timestamp. Finally, the model is deployed using the deploy method, specifying the instance count, instance type, and generated endpoint name. This enables the PyTorch model to be deployed and accessible for inference on SageMaker.

Inference

```
image_object = list(bucket.objects.filter('dataset/tyre'))[0].get()
image_bytes = image_object['Body'].read()

with Image.open(with Image.open(BytesIO(image_bytes)) as image::
    predicted_classes = sagemaker_predictor.predict(image)

    print(predicted_classes)
```

This is the example of using the deployed endpoint to do the inference.

Part 3 - Building A Simplified MLOps Pipeline (CI/CT/CD)

This article provides a guide to building an MLOps pipeline with AWS services, focusing on automated model retraining, deployment, and cost optimization.

Author(s):
Jian Jian (Ken), Senior Data & Applied Scientist, NetApp

Introduction

In this tutorial, you will learn how to leverage various AWS services to construct a simple MLOps pipeline that encompasses Continuous Integration (CI), Continuous Training (CT), and Continuous Deployment (CD). Unlike traditional DevOps pipelines, MLOps requires additional considerations to complete the operational cycle. By following this tutorial, you will gain insights into incorporating CT into the MLOps loop, enabling continuous training of your models and seamless deployment for inference. The tutorial will guide you through the process of utilizing AWS services to establish this end-to-end MLOps pipeline.

Manifest

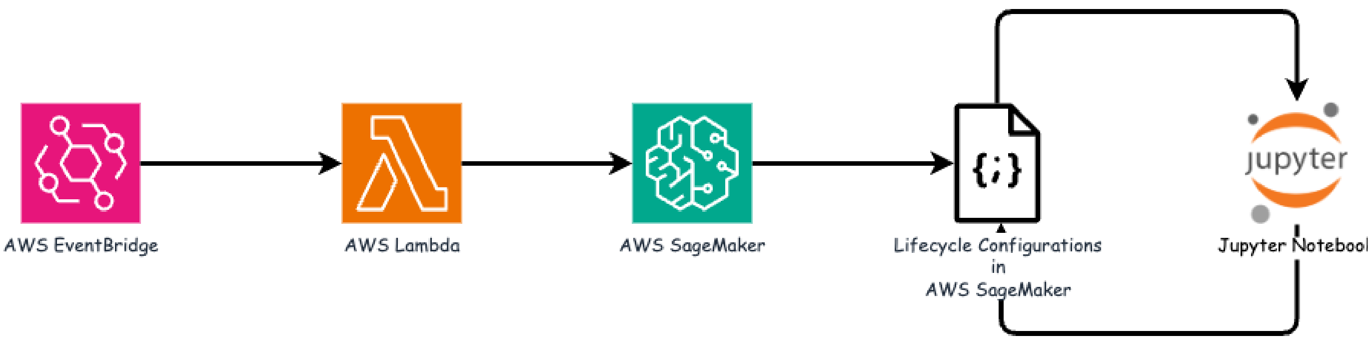
Functionality	Name	Comment
Data storage	AWS FSxN	Refer to Part 1 - Integrating AWS FSx for NetApp ONTAP (FSxN) as a private S3 bucket into AWS SageMaker.
Data science IDE	AWS SageMaker	This tutorial is based on the Jupyter notebook presented in Part 2 - Leveraging AWS FSx for NetApp ONTAP (FSxN) as a Data Source for Model Training in SageMaker.
Function to trigger the MLOps pipeline	AWS Lambda function	-

Functionality	Name	Comment
Cron job trigger	AWS EventBridge	-
Deep learning framework	PyTorch	-
AWS Python SDK	boto3	-
Programming Language	Python	v3.10

Prerequisite

- An pre-configured FSxN file system. This tutorial utilizes data stored in FSxN for the training process.
- A **SageMaker Notebook instance** that is configured to share the same VPC as the FSxN file system mentioned above.
- Before triggering the **AWS Lambda function**, ensure that the **SageMaker Notebook instance** is in **stopped** status.
- The **ml.g4dn.xlarge** instance type is required to leverage the GPU acceleration necessary for the computations of deep neural networks.

Architecture



This MLOps pipeline is a practical implementation that utilizes a cron job to trigger a serverless function, which in turn executes an AWS service registered with a lifecycle callback function. The **AWS EventBridge** acts as the cron job. It periodically invokes an **AWS Lambda function** responsible for retraining and redeploying the model. This process involves spinning up the **AWS SageMaker Notebook** instance to perform the necessary tasks.

Step-by-Step Configuration

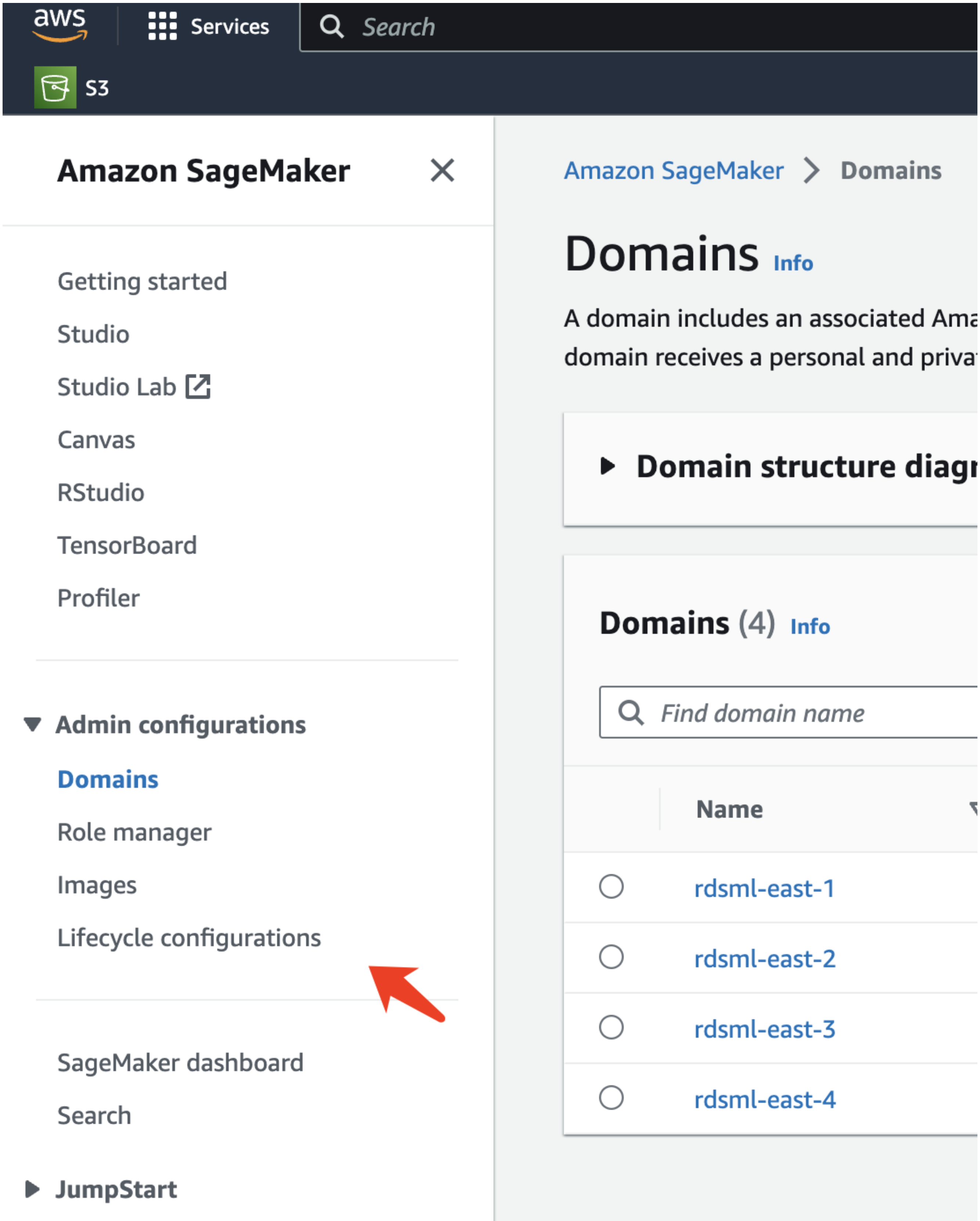
Lifecycle configurations

To configure the lifecycle callback function for the AWS SageMaker Notebook instance, you would utilize **Lifecycle configurations**. This service allow you to define the necessary actions to be performed during when spinning up the notebook instance. Specifically, a shell script can be implemented within the **Lifecycle configurations** to automatically shut down the notebook instance once the training and deployment processes are completed. This is a required configuration as the cost is one of the major consideration in MLOps.

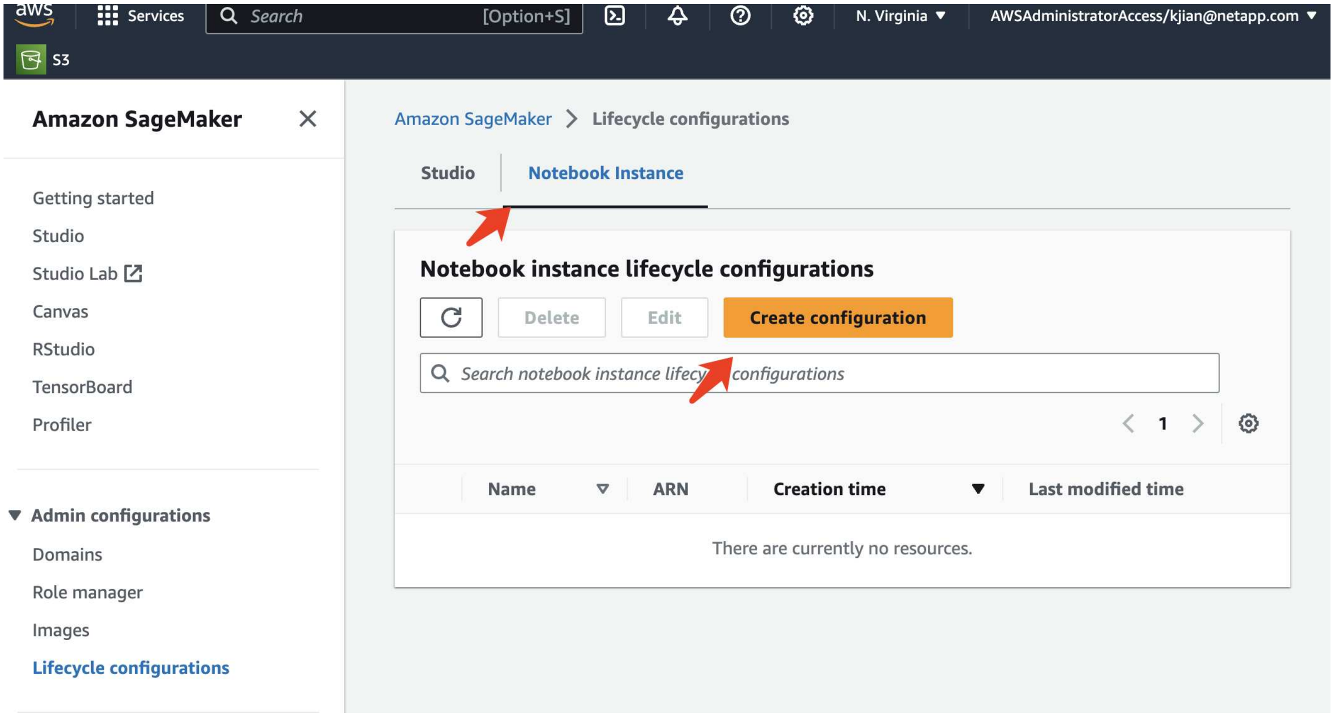
It's important to note that the configuration for **Lifecycle configurations** needs to be set up in advance. Therefore, it is recommended to prioritize configuring this aspect before proceeding with the other MLOps pipeline setup.

1. To set up a Lifecycle configurations, open the **Sagemaker** panel and navigate to **Lifecycle configurations**

under the section **Admin configurations**.



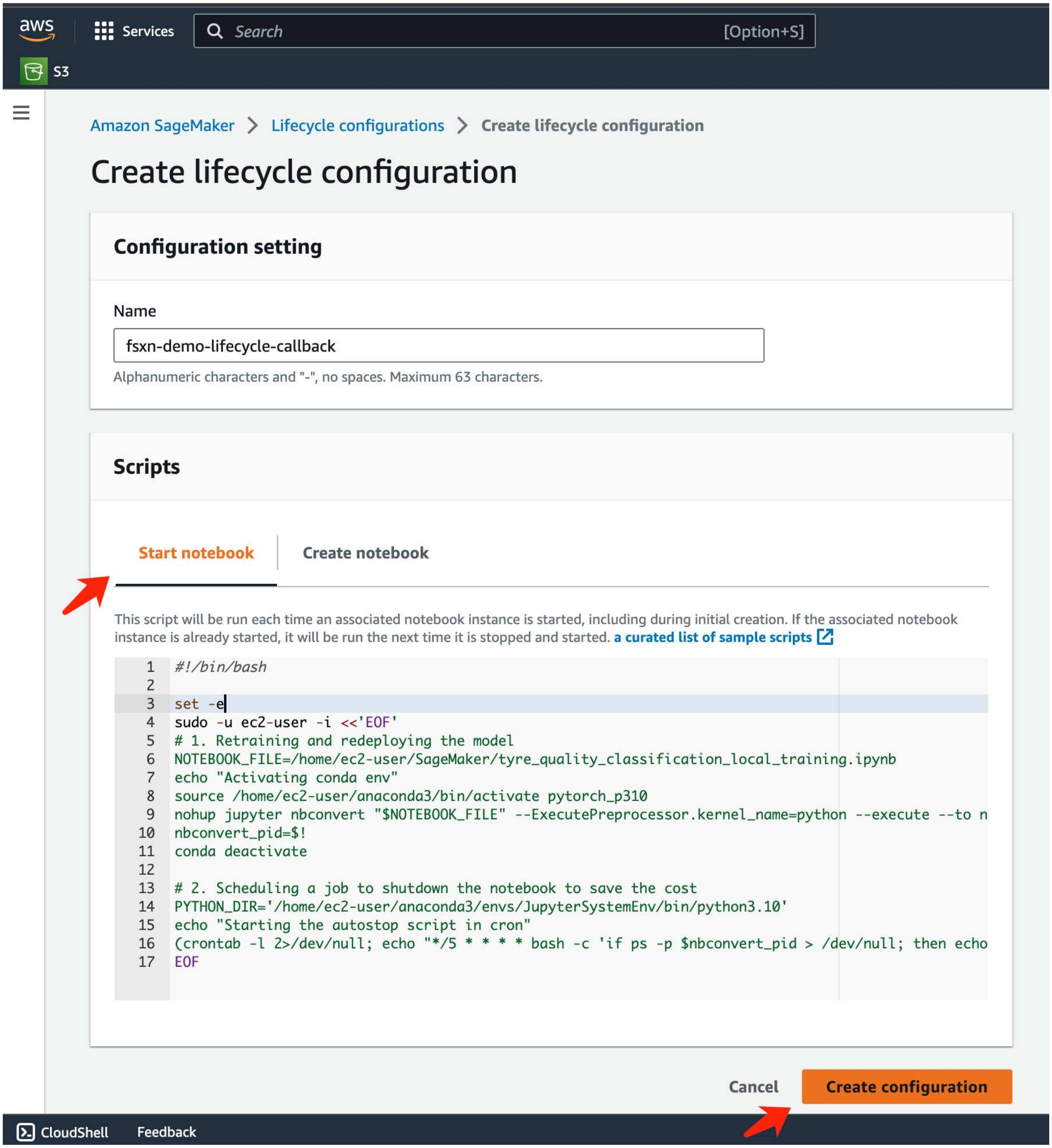
2. Select the **Notebook Instance** tab and click the **Create configuration** button



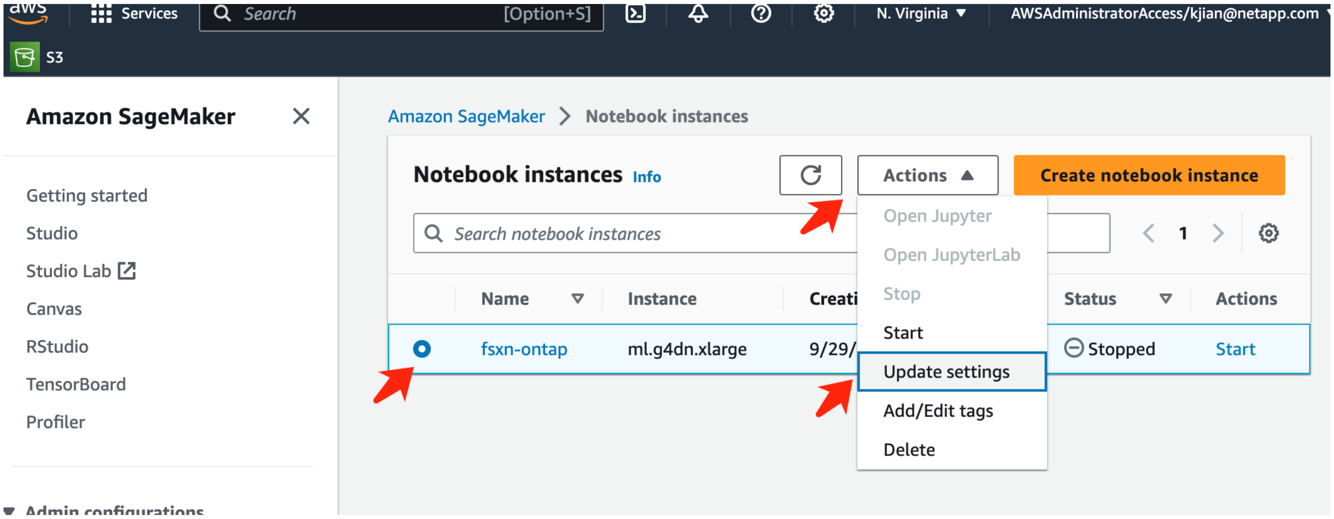
3. Paste the below code to the entry area.

```
#!/bin/bash

set -e
sudo -u ec2-user -i <<'EOF'
# 1. Retraining and redeploying the model
NOTEBOOK_FILE=/home/ec2-
user/SageMaker/tyre_quality_classification_local_training.ipynb
echo "Activating conda env"
source /home/ec2-user/anaconda3/bin/activate pytorch_p310
nohup jupyter nbconvert "$NOTEBOOK_FILE"
--ExecutePreprocessor.kernel_name=python --execute --to notebook &
nbconvert_pid=$!
conda deactivate

# 2. Scheduling a job to shutdown the notebook to save the cost
PYTHON_DIR='/home/ec2-
user/anaconda3/envs/JupyterSystemEnv/bin/python3.10'
echo "Starting the autostop script in cron"
(crontab -l 2>/dev/null; echo "*/5 * * * * bash -c 'if ps -p
$nbconvert_pid > /dev/null; then echo \"Notebook is still running.\" >>
/var/log/jupyter.log; else echo \"Notebook execution completed.\" >>
/var/log/jupyter.log; $PYTHON_DIR -c \"import boto3;boto3.client(
\'sagemaker\').stop_notebook_instance(NotebookInstanceName=get_notebook_
name())\" >> /var/log/jupyter.log; fi'") | crontab -
EOF
```

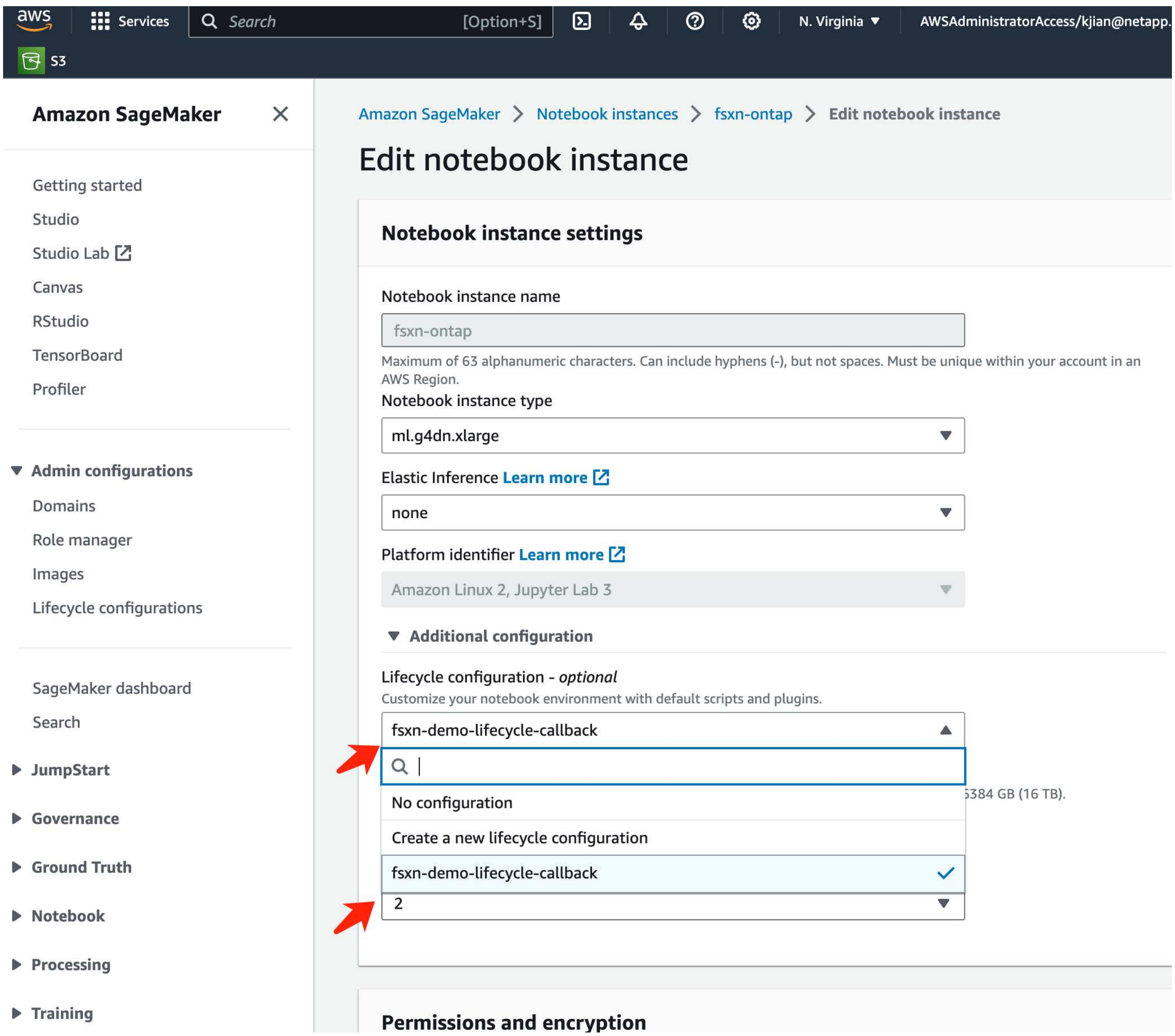
4. This script executes the Jupyter Notebook, which handles the retraining and redeployment of the model for inference. After the execution is complete, the notebook will automatically shut down within 5 minutes. To learn more about the problem statement and the code implementation, please refer to Part 2 - Leveraging AWS FSx for NetApp ONTAP (FSxN) as a Data Source for Model Training in SageMaker.



5. After the creation, navigate to Notebook instances, select the target instance, and click **Update settings** under Actions dropdown.



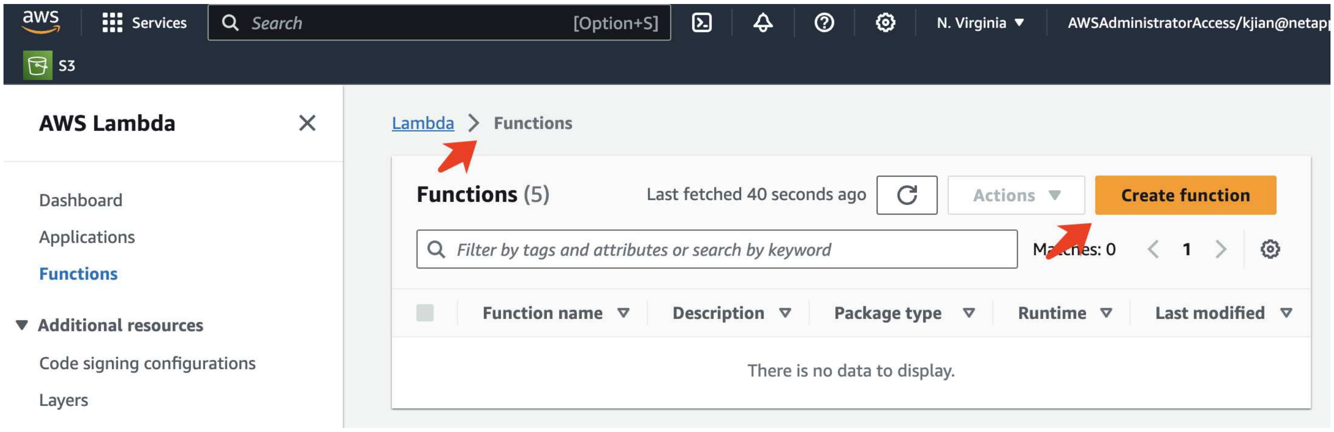
6. Select the created **Lifecycle configuration** and click **Update notebook instance**.



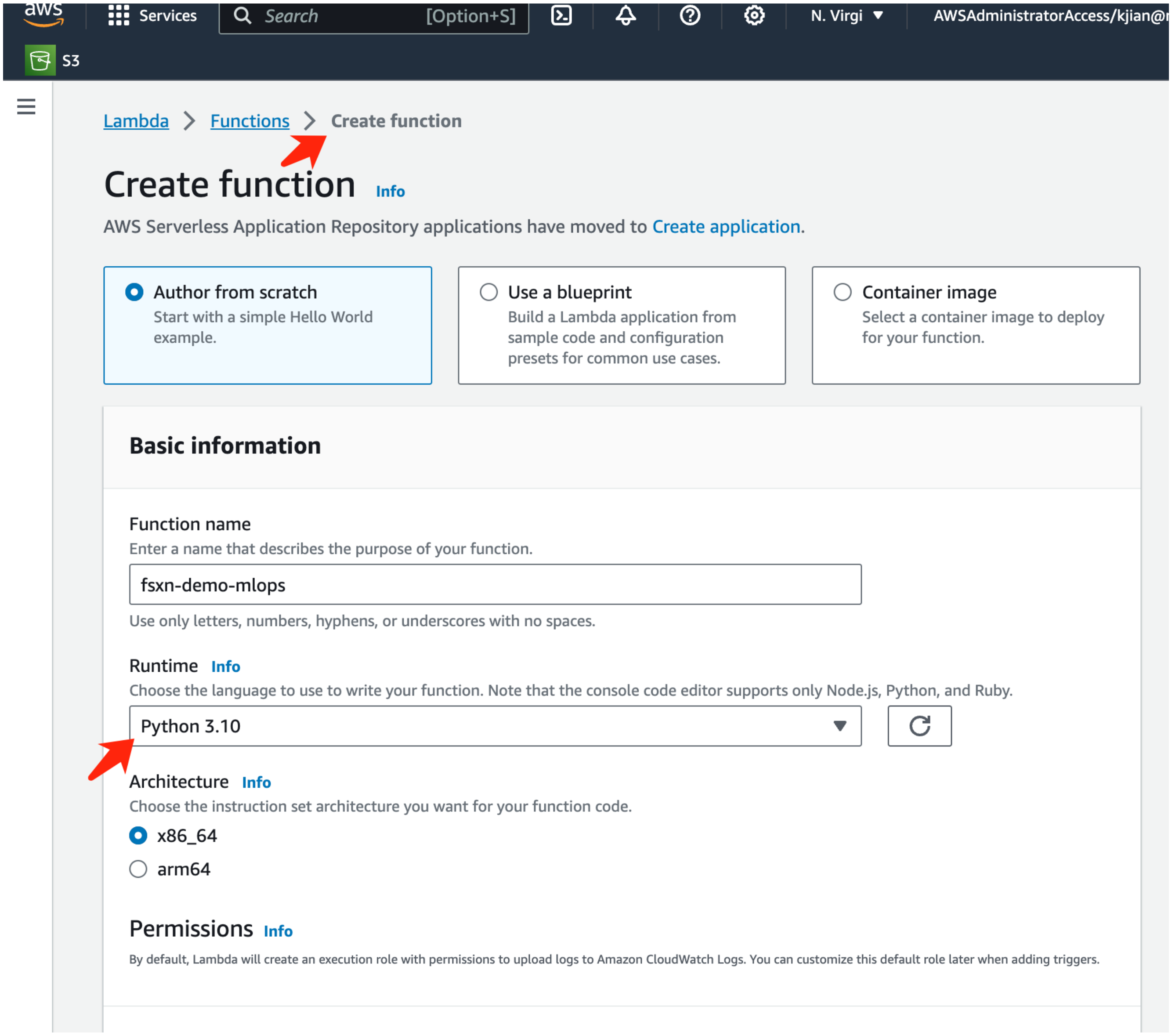
AWS Lambda serverless function

As mentioned earlier, the **AWS Lambda function** is responsible for spinning up the **AWS SageMaker Notebook instance**.

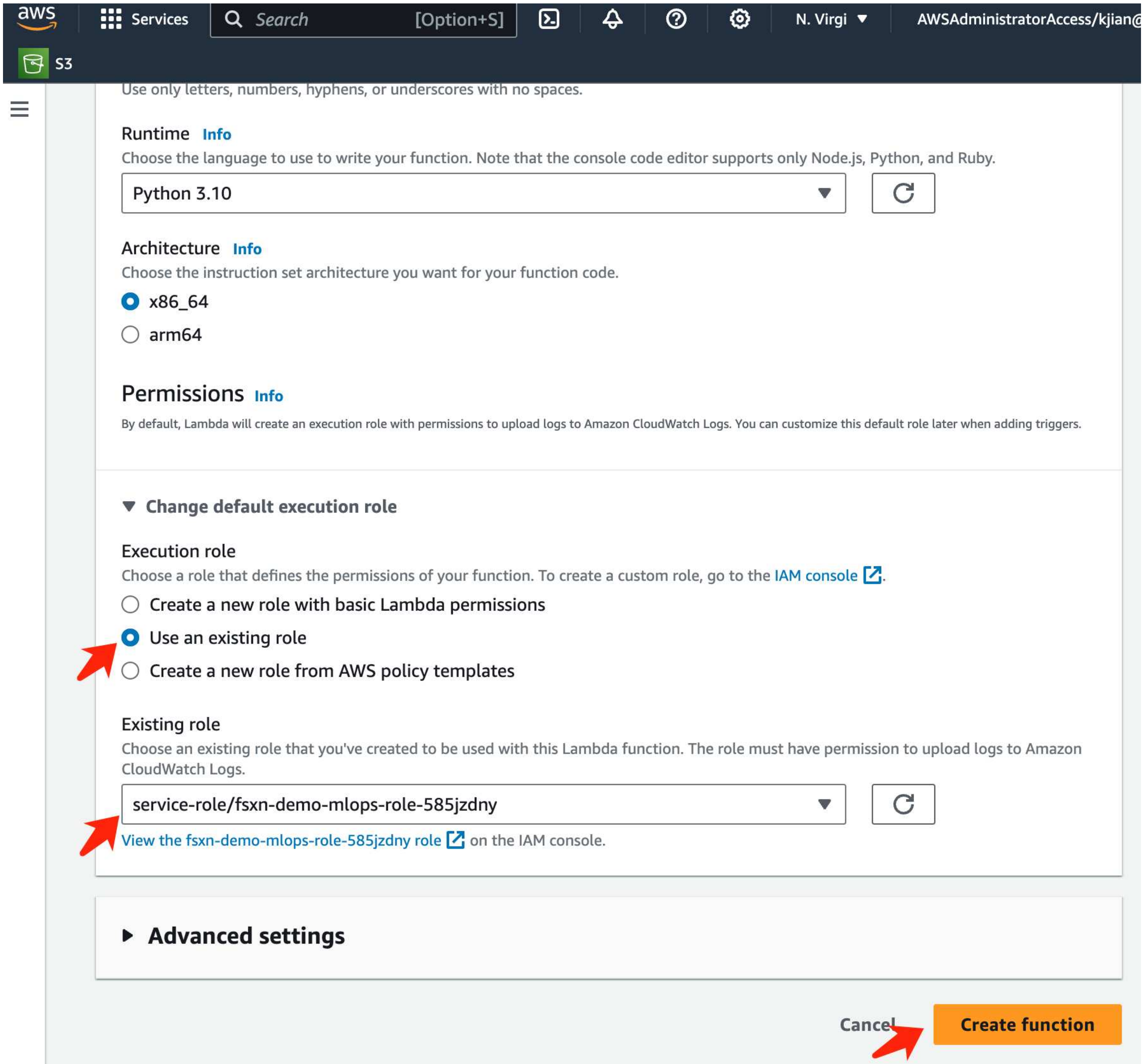
1. To create an **AWS Lambda function**, navigate to the respective panel, switch to the **Functions** tab, and click on **Create Function**.



2. Please file all required entries on the page and remember to switch the Runtime to **Python 3.10**.



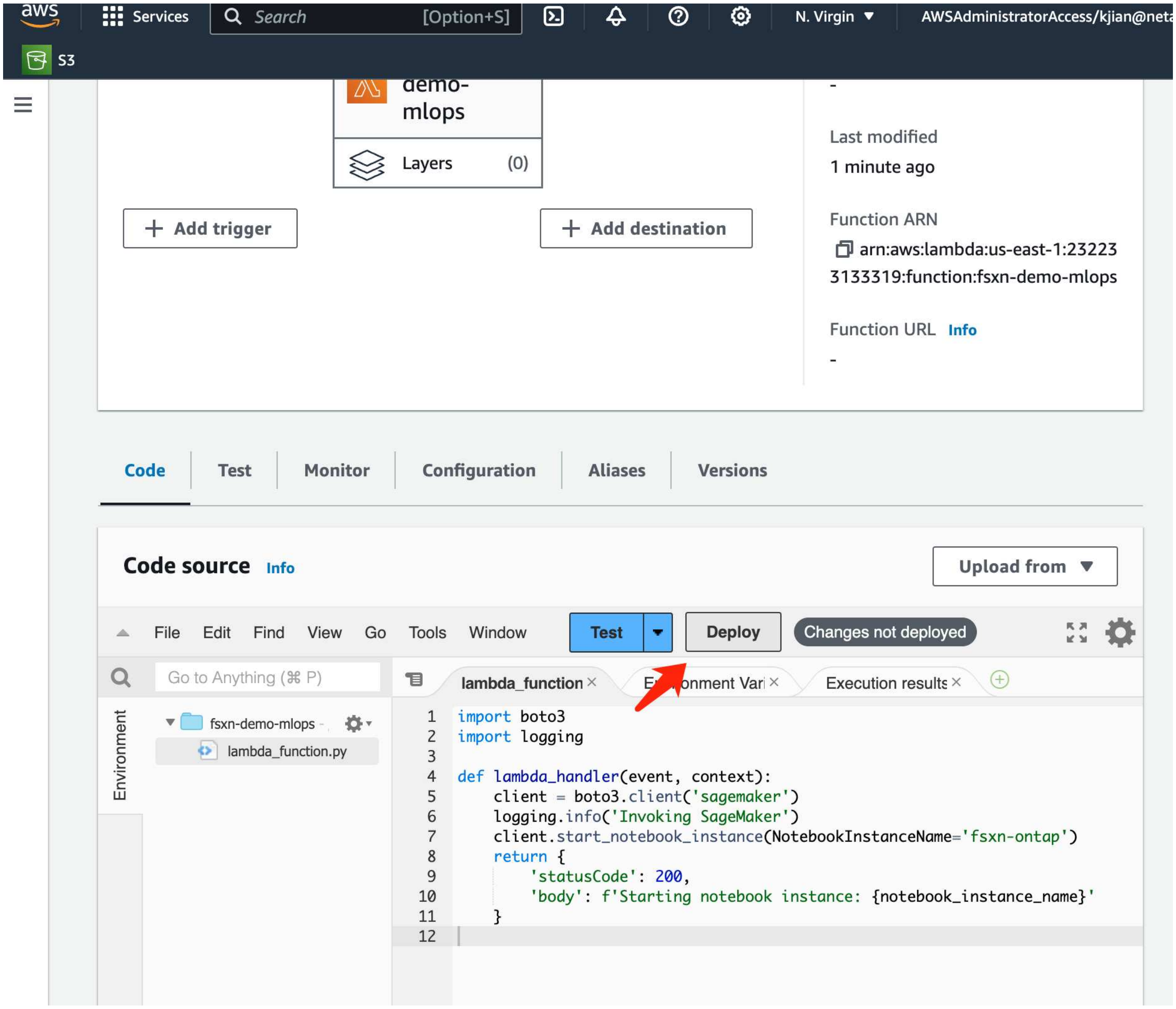
3. Please verify that the designated role has the required permission **AmazonSageMakerFullAccess** and click on the **Create function** button.



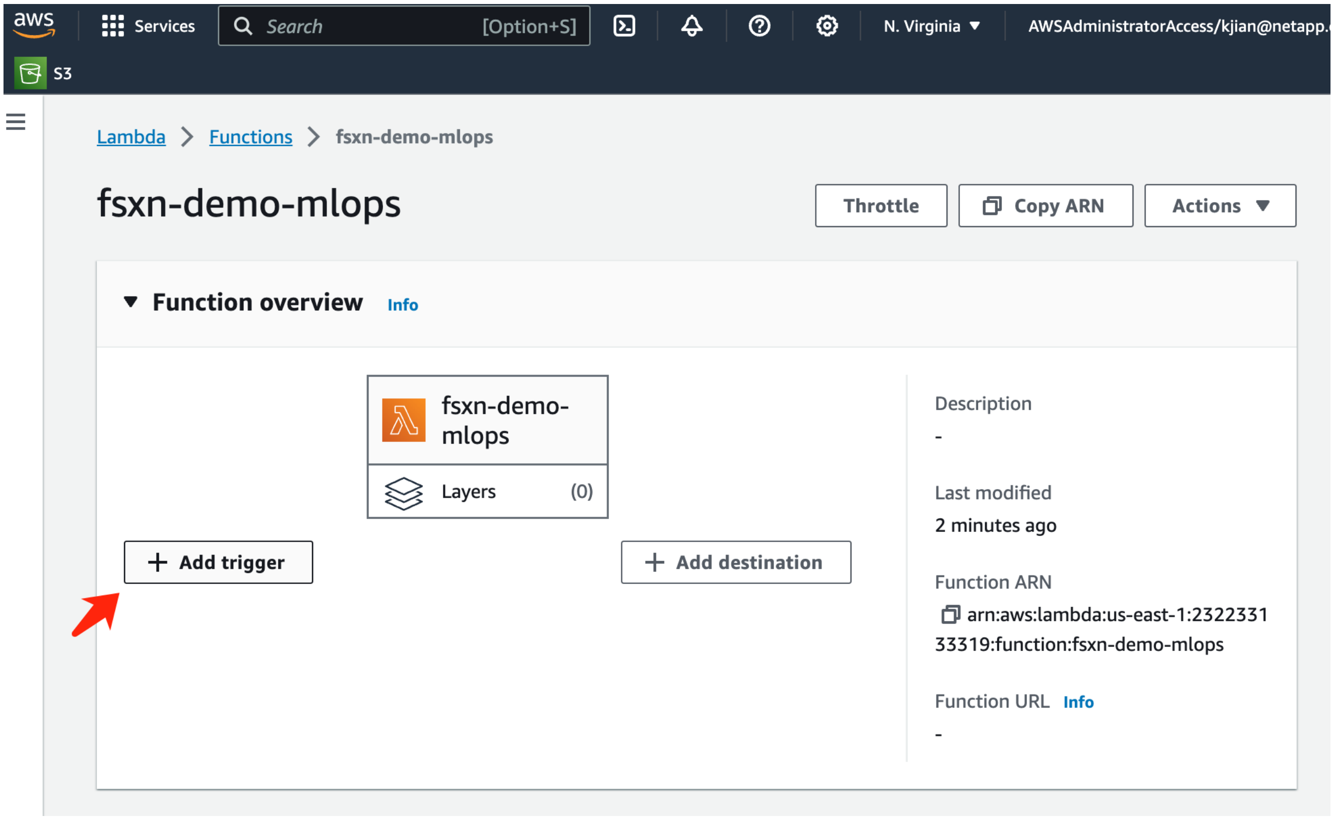
4. Select the created Lambda function. In the code tab, copy and paste the following code into the text area. This code starts the notebook instance named **fsxn-ontap**.

```
import boto3
import logging

def lambda_handler(event, context):
    client = boto3.client('sagemaker')
    logging.info('Invoking SageMaker')
    client.start_notebook_instance(NotebookInstanceName='fsxn-ontap')
    return {
        'statusCode': 200,
        'body': f'Starting notebook instance: {notebook_instance_name}'
    }
```

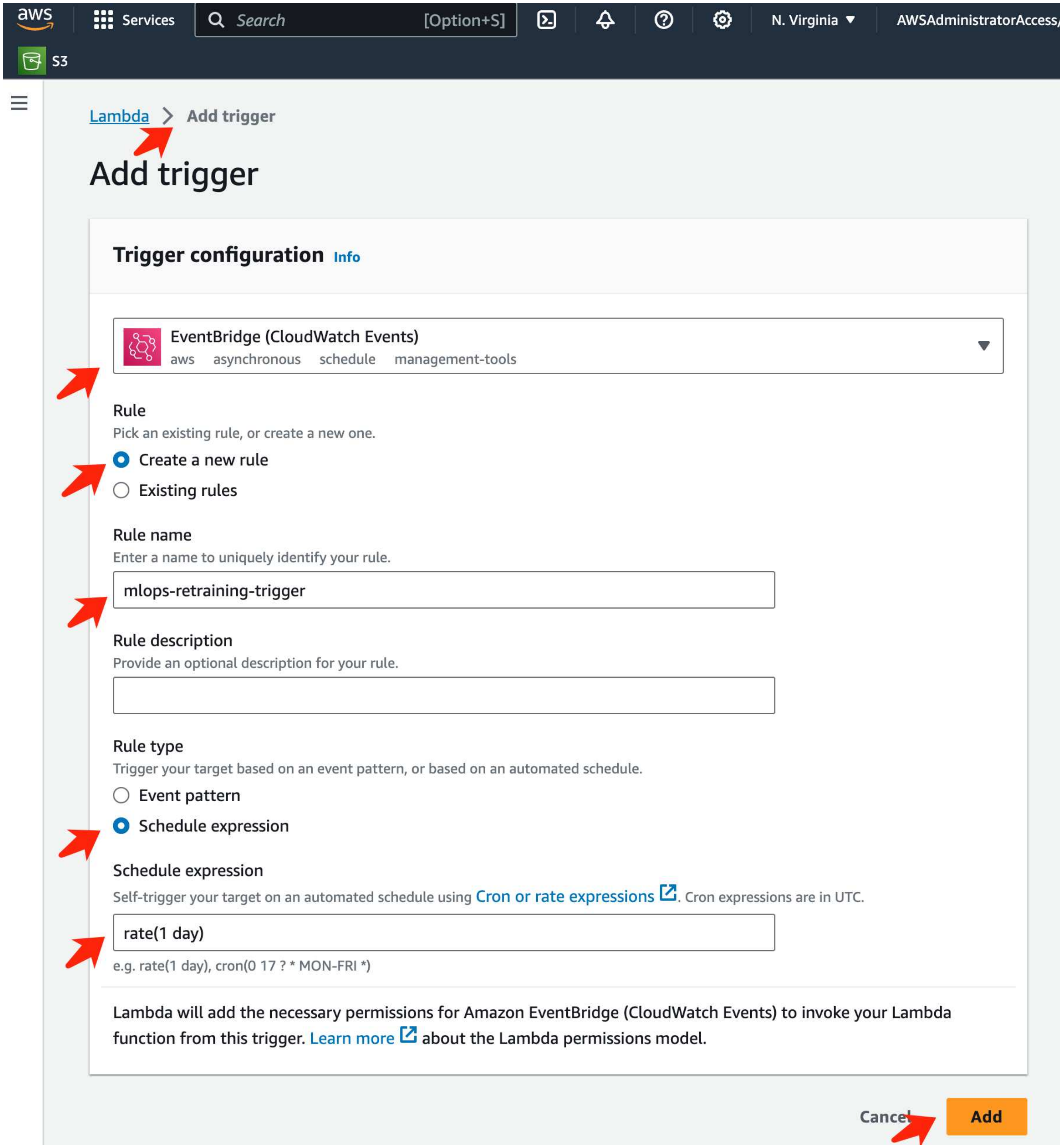
5. Click the **Deploy** button to apply this code change.



6. To specify how to trigger this AWS Lambda function, click on the Add Trigger button.



7. Select EventBridge from the dropdown menu, then click on the radio button labeled Create a new rule. In the schedule expression field, enter `rate(1 day)`, and click on the Add button to create and apply this new cron job rule to the AWS Lambda function.



After completing the two-step configuration, on a daily basis, the **AWS Lambda function** will initiate the **SageMaker Notebook**, perform model retraining using the data from the **FSxN** repository, redeploy the updated model to the production environment, and automatically shut down the **SageMaker Notebook instance** to optimize cost. This ensures that the model remains up to date.

This concludes the tutorial for developing an MLOps pipeline.

Hybrid Multicloud MLOps with Domino Data Lab and NetApp

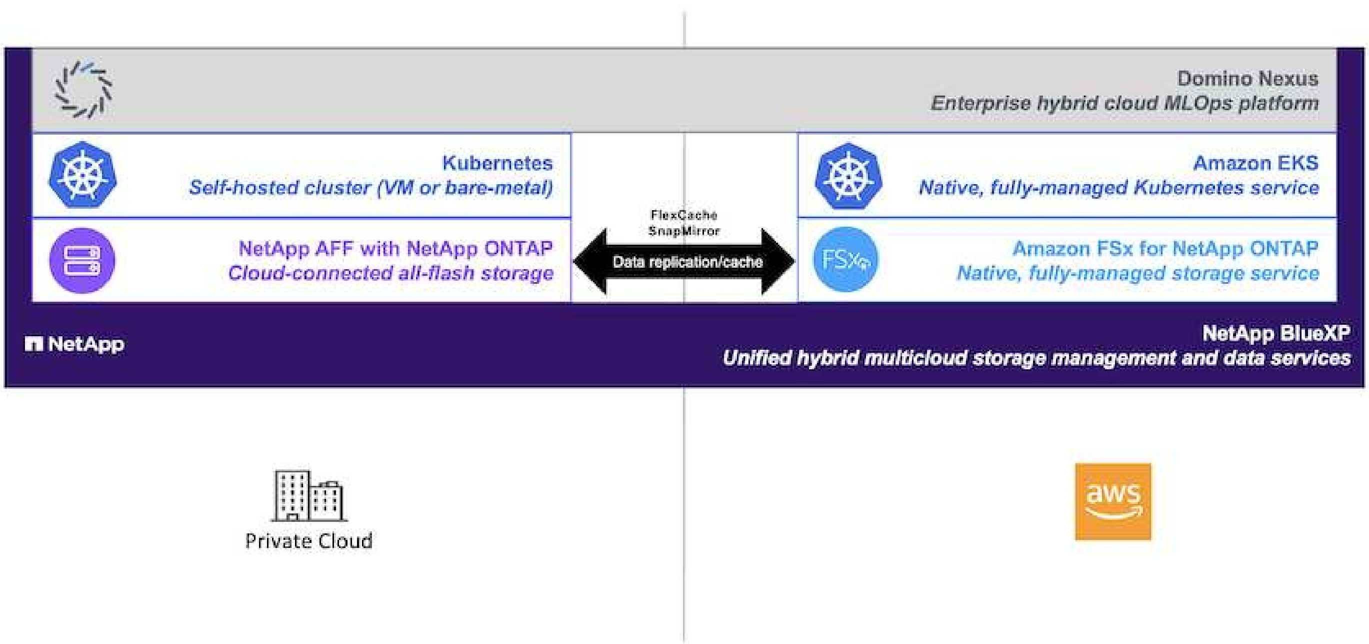
Hybrid Multicloud MLOps with Domino Data Lab and NetApp

Mike Oglesby, NetApp

Organizations all over the world are currently adopting AI to transform their businesses and processes. Because of this, AI-ready compute infrastructure is often in short supply. Enterprises are adopting hybrid multicloud MLOps architectures in order to take advantage of available compute environments across different regions, data centers, and clouds - balancing cost, availability, and performance.

Domino Nexus, from Domino Data Lab, is a unified MLOps control plane that lets you run data science and machine learning workloads across any compute cluster — in any cloud, region, or on-premises. It unifies data science silos across the enterprise, so you have one place to build, deploy, and monitor models. Likewise, NetApp's hybrid cloud data management capabilities enable you to bring your data to your jobs and workspaces, no matter where they are running. When you pair Domino Nexus with NetApp, you have the flexibility to schedule workloads across environments without having to worry about data availability. In other words, you have the ability to send your workloads and your data to the appropriate compute environment, enabling you to accelerate your AI deployments while navigating regulations around data privacy and sovereignty.

This solution demonstrates the deployment of a unified MLOps control plane incorporating an on-premises Kubernetes cluster and an Elastic Kubernetes Service (EKS) cluster running in Amazon Web Services (AWS).



Technology Overview

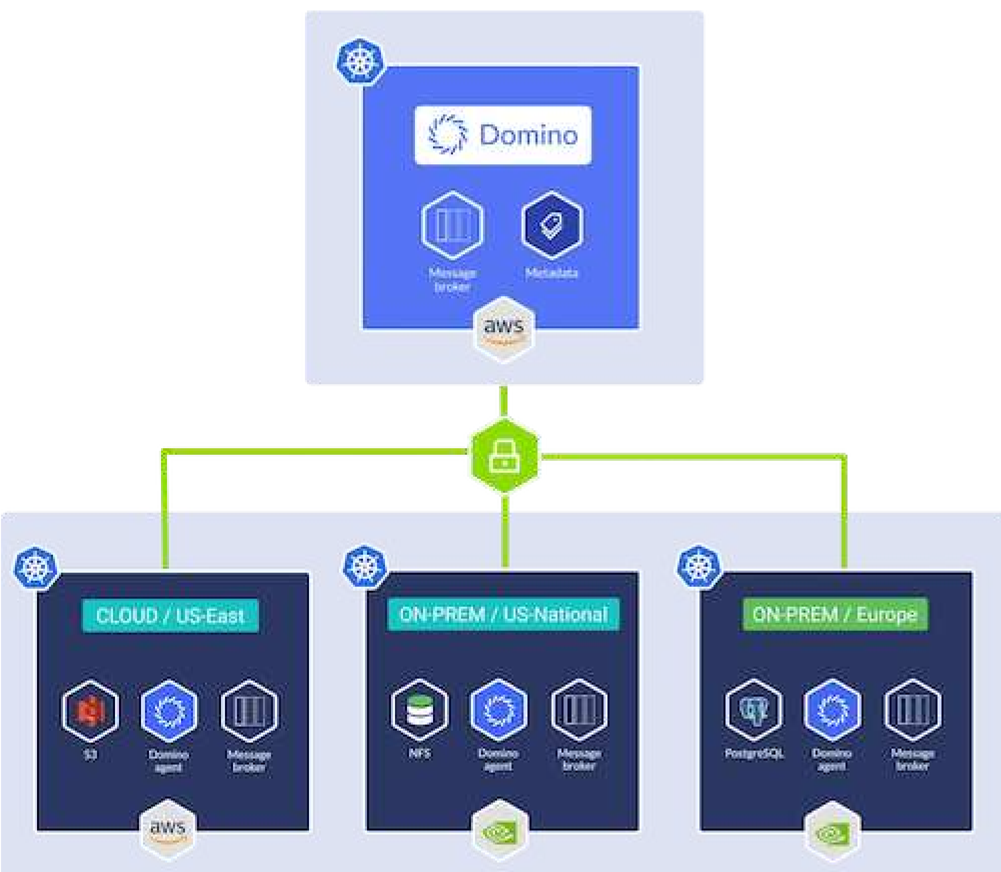
This section provides a technology overview for Hybrid Multicloud MLOps with Domino Data Lab and NetApp.

Domino Data Lab

Domino Data Lab powers model-driven businesses with its leading Enterprise AI platform trusted by over 20% of the Fortune 100. Domino accelerates the development and deployment of data science work while increasing collaboration and governance. With Domino, enterprises worldwide can develop better medicines, grow more productive crops, build better cars, and much more. Founded in 2013, Domino is backed by Coatue Management, Great Hill Partners, Highland Capital, Sequoia Capital and other leading investors.

Domino lets enterprises and their data scientists build, deploy and manage AI on a unified, end-to-end platform — fast, responsibly and cost-effectively. Teams can access all of the data, tools, compute, models, and projects they need across any environment, so they can collaborate, reuse past work, track models in production to improve accuracy, standardize with best practices, and make AI responsible and governed.

- **Open and Flexible:** Access the broadest ecosystem of open source and commercial tools, and infrastructure, for the best innovations and no vendor lock-in.
- **System of Record:** Central hub for AI operations and knowledge across the enterprise, enabling best practices, cross-functional collaboration, faster innovation, and efficiency.
- **Integrated:** Integrated workflows and automation — built for enterprise processes, controls, and governance — satisfy your compliance and regulatory needs.
- **Hybrid Multicloud:** Run AI workloads close to your data anywhere — on-premises, hybrid, any cloud or multi-cloud — for lower cost, optimal performance and compliance.



Domino Nexus

Domino Nexus is a single pane of glass that lets you run data science and machine learning workloads across any compute cluster — in any cloud, region, or on-premises. It unifies data science silos across the enterprise, so you have one place to build, deploy, and monitor models.

NetApp BlueXP

NetApp BlueXP unifies all of NetApp's storage and data services into a single tool that lets you build, protect, and govern your hybrid multicloud data estate. It delivers a unified experience for storage and data services across on-premises and cloud environments, and enables operational simplicity through the power of AIOps, with the flexible consumption parameters and integrated protection required for today's cloud-led world.

NetApp ONTAP

ONTAP 9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. You can also move data freely to wherever it is needed: the edge, the core, or the cloud. ONTAP 9 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

Simplify data management

Data management is crucial to enterprise IT operations and data scientists so that appropriate resources are used for AI applications and training AI/ML datasets. The following additional information about NetApp technologies is out of scope for this validation but might be relevant depending on your deployment.

ONTAP data management software includes the following features to streamline and simplify operations and reduce your total cost of operation:

- Inline data compaction and expanded deduplication. Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity. This applies to data stored locally and data tiered to the cloud.
- Minimum, maximum, and adaptive quality of service (AQoS). Granular quality of service (QoS) controls help maintain performance levels for critical applications in highly shared environments.
- NetApp FabricPool. Provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID storage solution. For more information about FabricPool, see TR-4598: FabricPool best practices.

Accelerate and protect data

ONTAP delivers superior levels of performance and data protection and extends these capabilities in the following ways:

- Performance and lower latency. ONTAP offers the highest possible throughput at the lowest possible latency.
- Data protection. ONTAP provides built-in data protection capabilities with common management across all platforms.
- NetApp Volume Encryption (NVE). ONTAP offers native volume-level encryption with both onboard and External Key Management support.
- Multitenancy and multifactor authentication. ONTAP enables sharing of infrastructure resources with the highest levels of security.

Future-proof infrastructure

ONTAP helps meet demanding and constantly changing business needs with the following features:

- Seamless scaling and nondisruptive operations. ONTAP supports the nondisruptive addition of capacity to existing controllers and to scale-out clusters. Customers can upgrade to the latest technologies, such as NVMe and 32Gb FC, without costly data migrations or outages.

- Cloud connection. ONTAP is the most cloud-connected storage management software, with options for software-defined storage and cloud-native instances in all public clouds.
- Integration with emerging applications. ONTAP offers enterprise-grade data services for next generation platforms and applications, such as autonomous vehicles, smart cities, and Industry 4.0, by using the same infrastructure that supports existing enterprise apps.

Amazon FSx for NetApp ONTAP

Amazon FSx for NetApp ONTAP is a first-party, fully managed AWS service that provides highly reliable, scalable, high-performing, and feature-rich file storage built on NetApp’s popular ONTAP file system. FSx for ONTAP combines the familiar features, performance, capabilities, and API operations of NetApp file systems with the agility, scalability, and simplicity of a fully managed AWS service.

NetApp Astra Trident

Astra Trident enables consumption and management of storage resources across all popular NetApp storage platforms, in the public cloud or on premises, including ONTAP (AFF, FAS, Select, Cloud, Amazon FSx for NetApp ONTAP), Element software (NetApp HCI, SolidFire), Azure NetApp Files service, and Cloud Volumes Service on Google Cloud. Astra Trident is a Container Storage Interface (CSI) compliant dynamic storage orchestrator that natively integrates with Kubernetes.

Kubernetes

Kubernetes is an open source, distributed, container orchestration platform that was originally designed by Google and is now maintained by the Cloud Native Computing Foundation (CNCF). Kubernetes enables the automation of deployment, management, and scaling functions for containerized applications, and is the dominant container orchestration platform in enterprise environments.

Amazon Elastic Kubernetes Service (EKS)

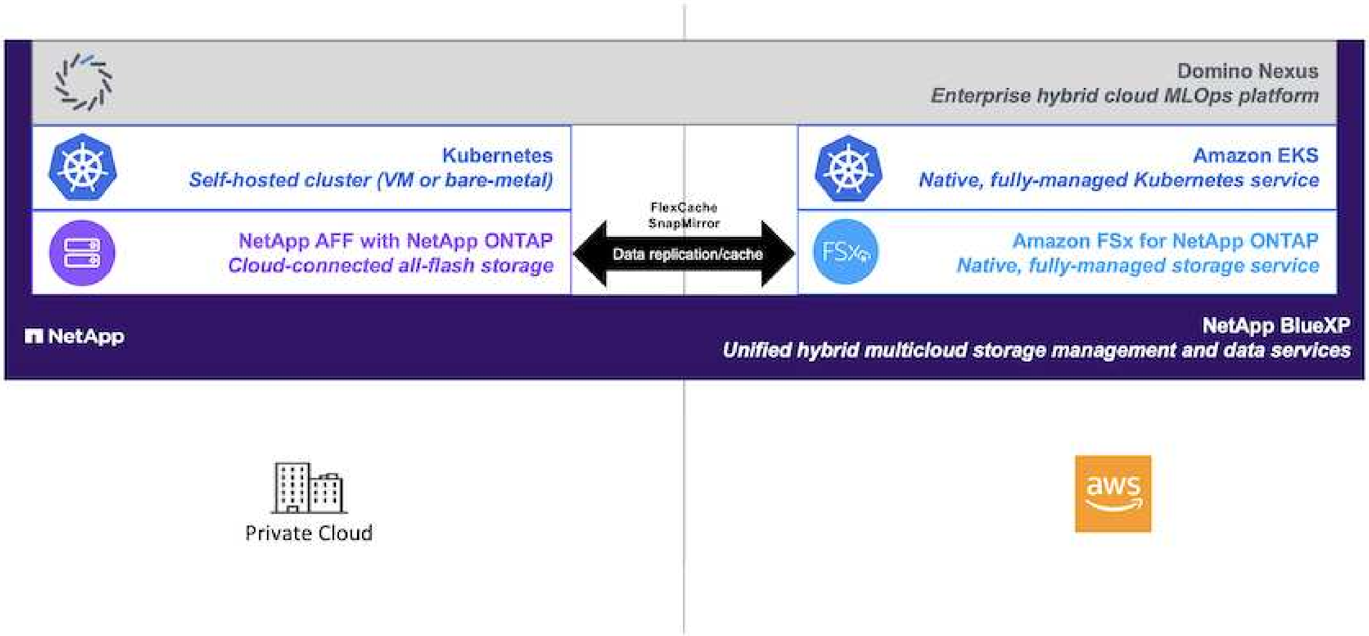
Amazon Elastic Kubernetes Service (Amazon EKS) is a managed Kubernetes service in the AWS cloud. Amazon EKS automatically manages the availability and scalability of the Kubernetes control plane nodes responsible for scheduling containers, managing application availability, storing cluster data, and other key tasks. With Amazon EKS, you can take advantage of all the performance, scale, reliability, and availability of AWS infrastructure, as well as integrations with AWS networking and security services.

Architecture

This solution combines Domino Nexus' hybrid multicloud workload scheduling capabilities with NetApp data services to create a unified hybrid cloud MLOps platform. See the following table for details.

Component	Name	Environment
MLOps Control Plane	Domino Enterprise AI Platform with Domino Nexus	AWS
MLOps Platform Compute Environments	Domino Nexus Data Planes	AWS, On-premises data center

Component	Name	Environment
On-premises Compute Platform	Kubernetes with NetApp Astra Trident	On-premises data center
Cloud Compute Platform	Amazon Elastic Kubernetes Service (EKS) with NetApp Astra Trident	AWS
On-premises Data Platform	NetApp storage appliance powered by NetApp ONTAP	On-premises data center
Cloud Data Platform	Amazon FSx for NetApp ONTAP	AWS



Initial Setup

This section describes the initial setup tasks that need to be performed in order to utilize Domino Nexus with NetApp data services in a hybrid environment incorporating an on-premises data center and AWS.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already performed the following tasks:

- You have already deployed and configured your on-premises NetApp ONTAP storage platform. For more information, refer to the NetApp product documentation.
- You have already provisioned an Amazon FSx for NetApp ONTAP instance in AWS. For more information, refer to the Amazon FSx for NetApp ONTAP product page.
- You have already provisioned a Kubernetes cluster in your on-premises data center. For more information, refer to the Domino admin guide.
- You have already provisioned an Amazon EKS cluster in AWS. For more information, refer to the Domino admin guide.
- You have installed NetApp Astra Trident in your on-premises Kubernetes cluster. Additionally, you have

configured this Trident instance to use your on-premises NetApp ONTAP storage platform when provisioning and managing storage resources. For more information, refer to the NetApp Astra Trident documentation.

- You have installed NetApp Astra Trident in your Amazon EKS cluster. Additionally, you have configured this Trident instance to use your Amazon FSx for NetApp ONTAP instance when provisioning and managing storage resources. For more information, refer to the NetApp Astra Trident documentation.
- You must have bi-directional network connectivity between your on-premises data center and your Virtual Private Cloud (VPC) in AWS. For more details on the various options for implementing this, refer to the Amazon Virtual Private Network (VPN) documentation.

Install the Domino Enterprise AI Platform in AWS

To install the Domino Enterprise MLOps Platform in AWS, follow the instructions outlined in Domino admin guide. You must deploy Domino in the same Amazon EKS cluster that you previously provisioned. Additionally, NetApp Astra Trident must already be installed and configured in this EKS cluster, and you must specify a Trident-managed storage class as the shared storage class in your domino.yml install configuration file.

> Refer to the Domino install configuration reference guide for details on how to specify a shared storage class in your domino.yml install configuration file.

> Technical Report TR-4952 walks through the deployment of Domino in AWS with Amazon FSx for NetApp ONTAP and may be a useful reference for troubleshooting any issues that arise.

Enable Domino Nexus

Next, you must enable Domino Nexus. Refer to the Domino admin guide for details.

Deploy a Domino Data Plane in your On-premises Data Center

Next, you must deploy a Domino Data Plane in your on-premises data center. You must deploy this data plane in the on-premises Kubernetes cluster that you previously provisioned. Additionally, NetApp Astra Trident must already be installed and configured in this Kubernetes cluster. Refer to the Domino admin guide for details.

Expose Existing NetApp Volumes to Domino

This section describes the tasks that need to be performed in order to expose existing NetApp ONTAP NFS volumes to the Domino MLOps platform. These same steps apply both on-premises and in AWS.

Why Expose NetApp ONTAP Volumes to Domino?

Using NetApp volumes in conjunction with Domino provides the following benefits:

- You can execute workloads against extremely large datasets by taking advantage of NetApp ONTAP's scale-out capabilities.
- You can execute workloads across multiple compute nodes without having to copy your data to the individual nodes.
- You can take advantage of NetApp's hybrid multicloud data movement and sync capabilities in order to access your data across multiple data centers and/or clouds.
- You want to be able to quickly and easily create a cache of your data in a different data center or cloud.

Expose Existing NFS Volumes that were not Provisioned by Astra Trident

If your existing NetApp ONTAP NFS volume was not provisioned by Astra Trident, follow the steps outlined in this sub-section.

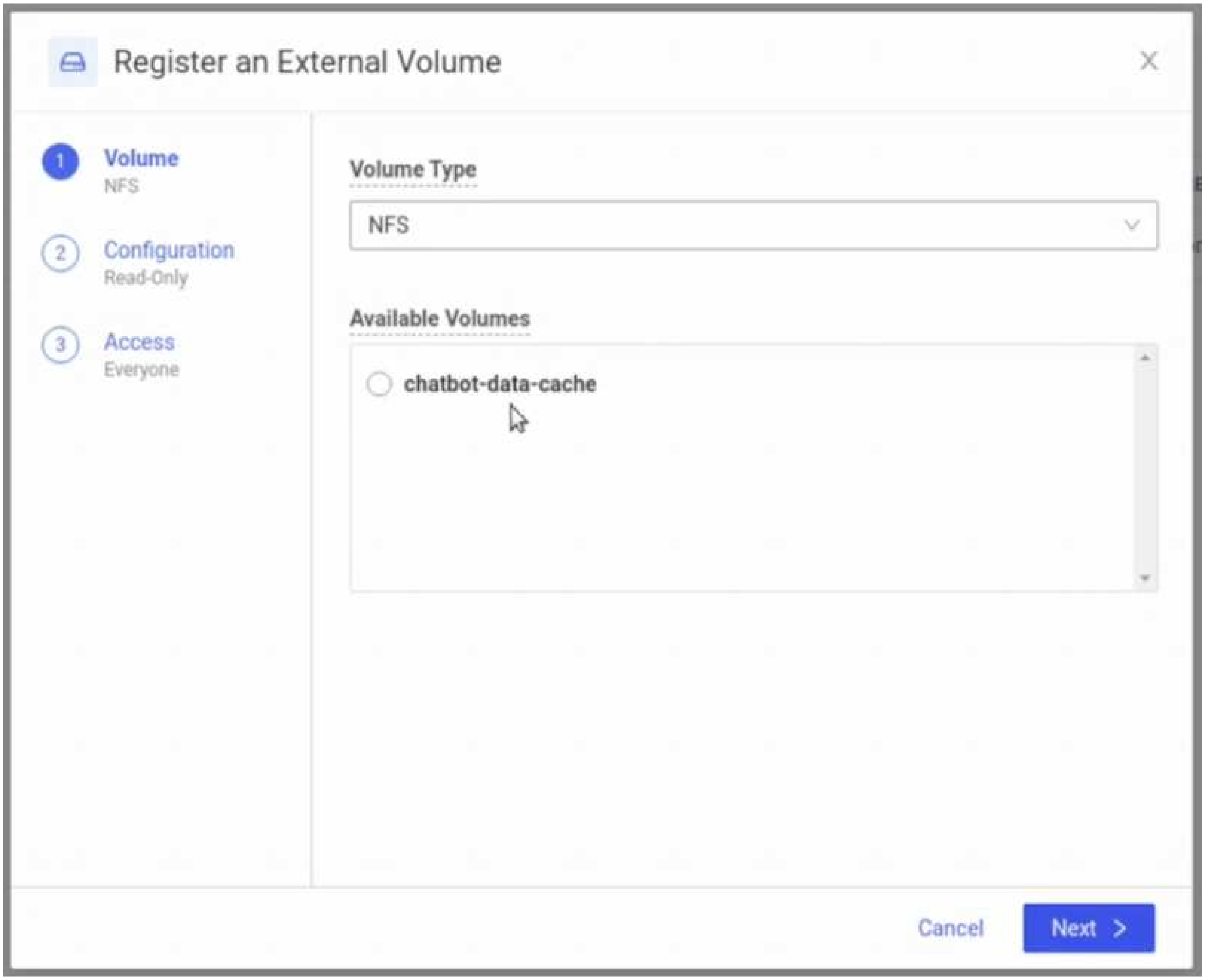
Create PV and PVC in Kubernetes

> ⓘ For on-premises volumes, create the PV and PVC in your on-premises Kubernetes cluster. For Amazon FSx for NetApp ONTAP volumes, create the PV and PVC in Amazon EKS.

First, you must create a persistent volume (PV) and persistent volume claim (PVC) in your Kubernetes cluster. To create the PV and PVC, use the NFS PV/PVC example from the Domino admin guide and update the values to reflect to your environment. Be sure to specify the correct values for the `namespace`, `nfs.path`, and `nfs.server` fields. Additionally, we recommend giving your PV and PVC unique names that represent that nature of the data that is stored on the corresponding ONTAP NFS volume. For example, if the volume contains images of manufacturing defects, you might name the PV, `pv-mfg-defect-images`, and the PVC, `pvc-mfg-defect-images`.

Register External Data Volume in Domino

Next, you must register an external data volume in Domino. To register an external data volume, refer to the instructions in the Domino admin guide. When registering the volume, be sure to select "NFS" from the 'Volume Type' drop-down menu. After selecting "NFS", you should see your PVC in the 'Available Volumes' list.



Expose Existing Volumes that were Provisioned by Astra Trident

If your existing volume was provisioned by Astra Trident, follow the steps outlined in this sub-section.

Edit Existing PVC

If your volume was provisioned by Astra Trident, then you already have a persistent volume claim (PVC) corresponding to your volume. In order to expose this volume to Domino, you must edit the PVC and add the following label to the list of labels in the `metadata.labels` field:

```
"dominodatalab.com/external-data-volume": "Generic"
```

Register External Data Volume in Domino

Next, you must register an external data volume in Domino. To register an external data volume, refer to the instructions in the Domino admin guide. When registering the volume, be sure to select "Generic" from the 'Volume Type' drop-down menu. After selecting "Generic", you should see your PVC in the 'Available Volumes' list.

Access the same Data Across Different Environments

This section describes the tasks that need to be performed in order to access the same data across different compute environments. In the Domino MLOps platform, compute environments are referred to "data planes." Follow the tasks outlined in this section if your data resides on a NetApp volume in one data plane, but you need to access it in another data plane. This type of scenario is often referred to as "bursting" or, when the destination environment is the cloud, "cloud bursting." This capability is often needed when dealing with constrained or over-subscribed compute resources. For example, if your on-premises compute cluster is over-subscribed, you may want to schedule workloads to the cloud where they can be started immediately.

There are two recommended options for accessing a NetApp volume that resides in a different data plane. These options are outlined in the sub-sections below. Choose one of these options depending on your specific requirements. The benefits and drawbacks of the two options are described in the following table.

Option	Benefits	Drawbacks
Option 1 - Cache	- Simpler workflow - Ability to cache a subset of data based on needs - Ability to write data back to source - No remote copy to manage	- Increased latency on initial data access as cache is hydrated.
Option 2 - Mirror	- Full copy of source volume - No increased latency due to cache hydration (after mirror operation is complete)	- Must wait for mirror operation to complete before accessing data - Must manage a remote copy - No ability to write back to source

Option 1 - Create a Cache of a Volume that Resides in a Different Data Plane

With NetApp FlexCache technology, you can create a cache of a NetApp volume that resides in a different data plane. For example, if you have a NetApp volume in your on-premises data plane, and you need to access that volume in your AWS data plane, you can create a cache of the volume in AWS. This section outlines the tasks that need to be performed in order to create a cache of a NetApp volume that resides in a different data plane.

Create FlexCache Volume in Destination Environment

> If the destination environment is your on-premises data center, you will create the FlexCache volume on your on-premises ONTAP system. If the destination environment is AWS, you will create the FlexCache volume on your Amazon FSx for NetApp ONTAP instance.

First, you must create a FlexCache volume in the destination environment.

We recommend using BlueXP to create the FlexCache volume. To create a FlexCache volume with BlueXP, follow the instructions outlined in the BlueXP volume caching documentation.

If you prefer not to use BlueXP, you can use ONTAP System Manager or the ONTAP CLI to create the FlexCache volume. To create a FlexCache volume with System Manager, refer to the instructions outlined in the ONTAP documentation. To create a FlexCache volume with the ONTAP CLI, refer to the instructions outlined in the ONTAP documentation.

If you wish to automate this process, you can use the BlueXP API, the ONTAP REST API, or the ONTAP Ansible collection.

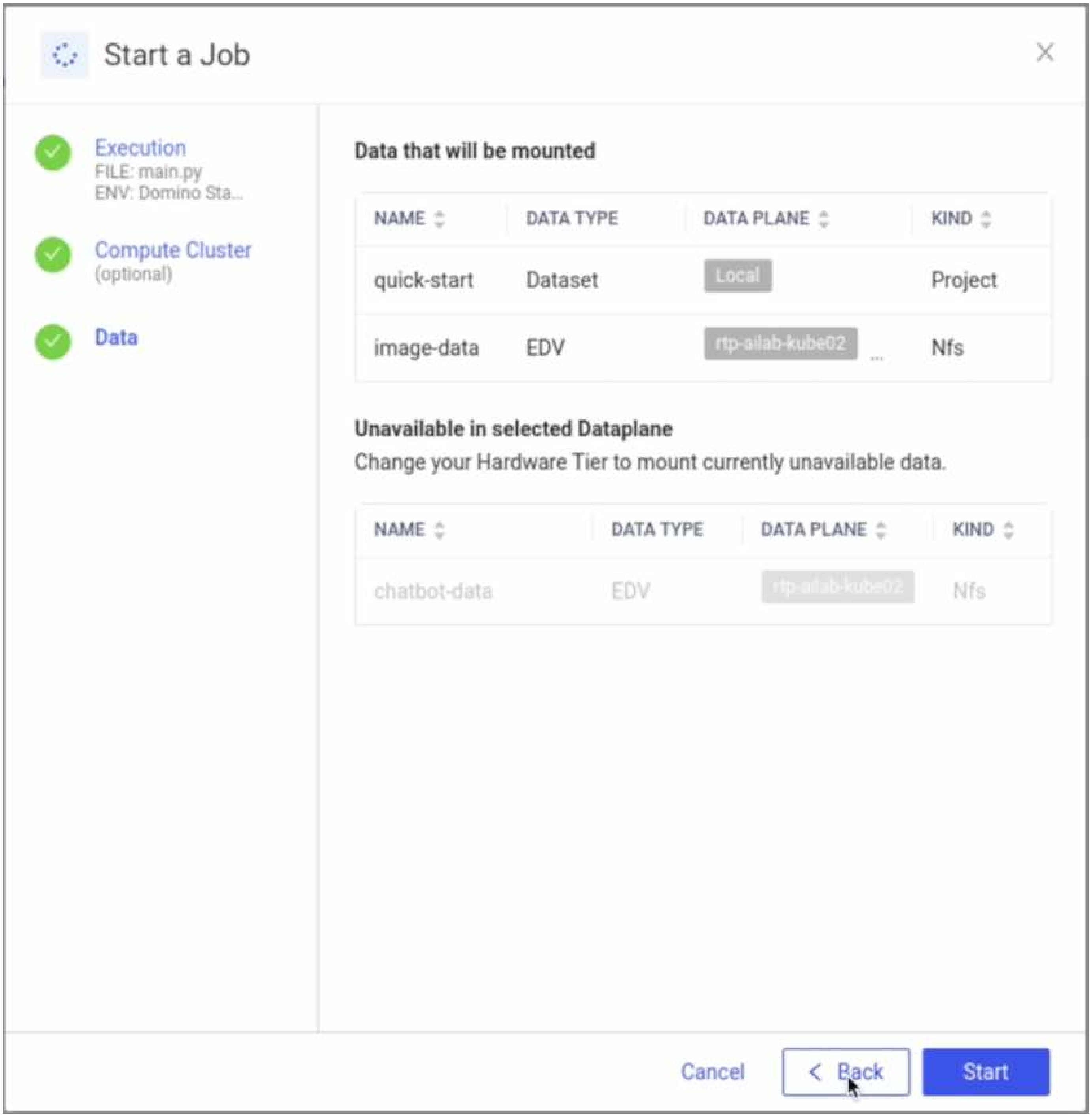
> System Manager is not available in Amazon FSx for NetApp ONTAP.

Expose FlexCache Volume to Domino

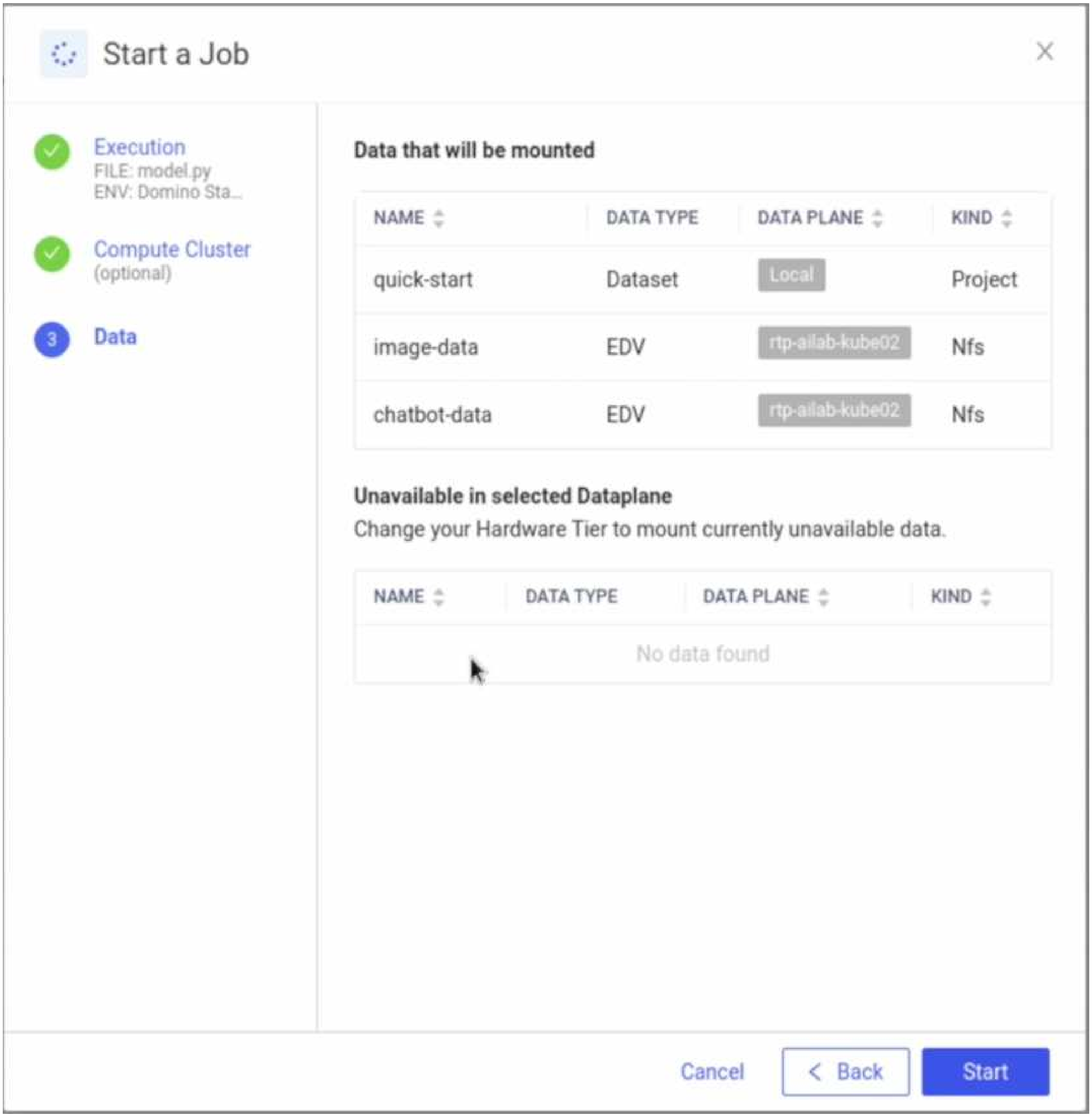
Next, you must expose the FlexCache volume to the Domino MLOps platform. To expose the FlexCache volume to Domino, follow the instructions outlined in the 'Expose Existing NFS Volumes that were not Provisioned by Astra Trident' sub-section of the 'Expose Existing NetApp Volumes to Domino' section of this solution.

Now, you will be able to mount the FlexCache volume when launching jobs and workspaces in the destination data plane as shown in the following screenshots.

Before Creating FlexCache Volume



After Exposing FlexCache Volume to Domino



Option 2 - Replicate a Volume that Resides in a Different Data Plane

With NetApp SnapMirror data replication technology, you can create a copy of a NetApp volume that resides in a different data plane. For example, if you have a NetApp volume in your on-premises data plane, and you need to access that volume in your AWS data plane, you can create a copy of the volume in AWS. This section outlines the tasks that need to be performed in order to create a copy of a NetApp volume that resides in a different data plane.

Create SnapMirror Relationship

First, you must create a SnapMirror relationship between your source volume and a new destination volume in the destination environment. Note that the destination volume will be created as part of the process of creating the SnapMirror relationship.

We recommend using BlueXP to create the SnapMirror relationship. To create a SnapMirror relationship with BlueXP, follow the instructions outlined in the BlueXP replication documentation.

If you prefer not to use BlueXP, you can use ONTAP System Manager or the ONTAP CLI to create the SnapMirror relationship. To create a SnapMirror relationship with System Manager, refer to the instructions outlined in the ONTAP documentation. To create a SnapMirror relationship with the ONTAP CLI, refer to the instructions outlined in the ONTAP documentation.

If you wish to automate this process, you can use the BlueXP API, the ONTAP REST API, or the ONTAP Ansible collection.

> System Manager is not available in Amazon FSx for NetApp ONTAP.

Break SnapMirror Relationship

Next, you must break the SnapMirror relationship in order to activate the destination volume for data access. Wait until the initial replication is complete before performing this step.

> You can determine whether or not the replication is complete by checking the mirror state in BlueXP, ONTAP System Manager, or the ONTAP CLI. When the replication is complete, the mirror state will be "snapmirrored".

We recommend using BlueXP to break the SnapMirror relationship. To break a SnapMirror relationship with BlueXP, follow the instructions outlined in the BlueXP replication documentation.

If you prefer not to use BlueXP, you can use ONTAP System Manager or the ONTAP CLI to break the SnapMirror relationship. To break a SnapMirror relationship with System Manager, refer to the instructions outlined in the ONTAP documentation. To break a SnapMirror relationship with the ONTAP CLI, refer to the instructions outlined in the ONTAP documentation.

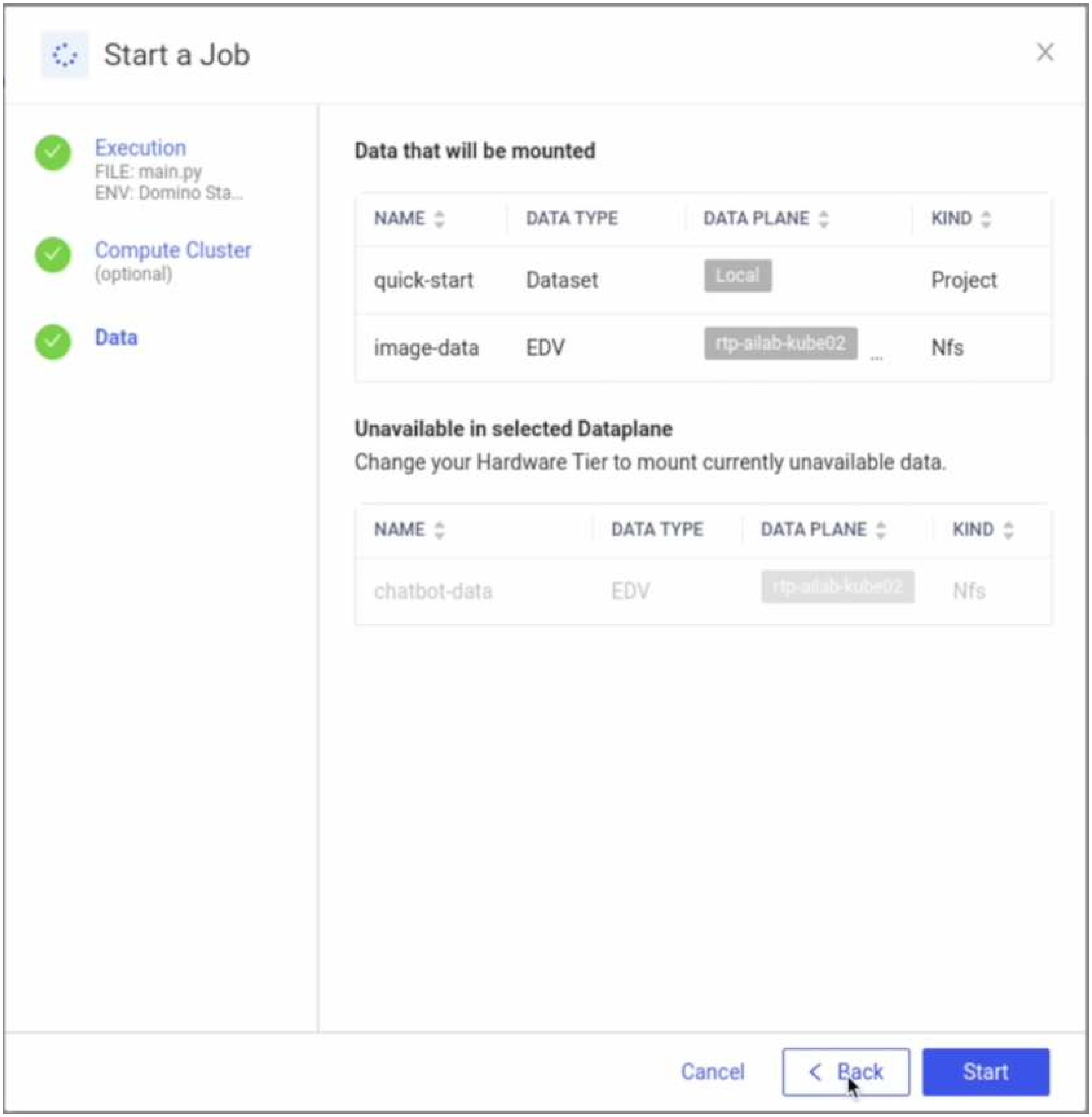
If you wish to automate this process, you can use the BlueXP API, the ONTAP REST API, or the ONTAP Ansible collection.

Expose Destination Volume to Domino

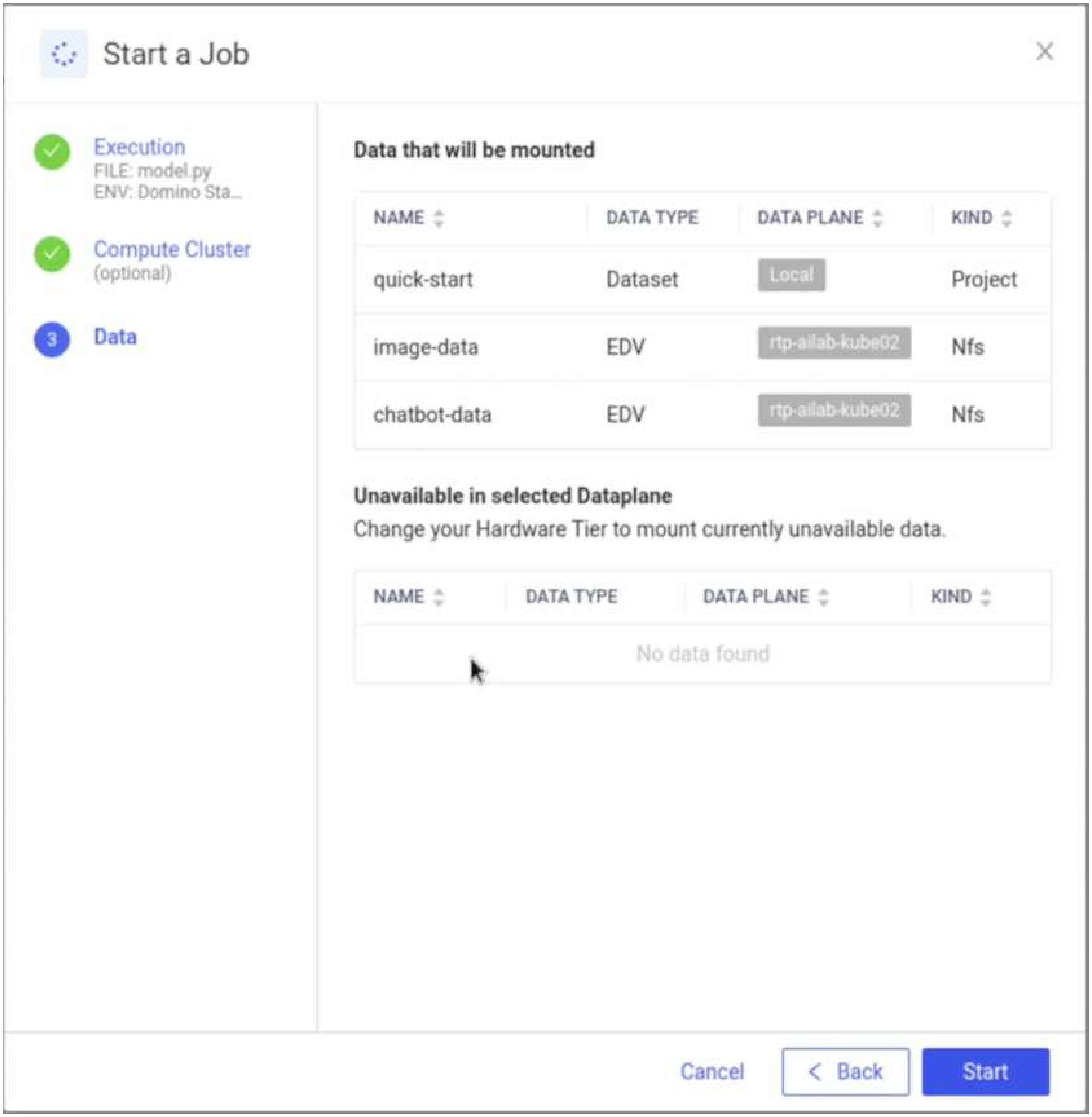
Next, you must expose the destination volume to the Domino MLOps platform. To expose the destination volume to Domino, follow the instructions outlined in the 'Expose Existing NFS Volumes that were not Provisioned by Astra Trident' sub-section of the 'Expose Existing NetApp Volumes to Domino' section of this solution.

Now, you will be able to mount the destination volume when launching jobs and workspaces in the destination data plane as shown in the following screenshots.

Before Creating SnapMirror Relationship



After Exposing Destination Volume to Domino



Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and/or websites:

- Domino Data Lab

  https://domino.ai

- Domino Nexus

  https://domino.ai/platform/nexus

- NetApp BlueXP

  https://bluexp.netapp.com

- NetApp ONTAP data management software

  https://www.netapp.com/data-management/ontap-data-management-software/

- NetApp AI Solutions

  https://www.netapp.com/artificial-intelligence/

Acknowledgments

- Josh Mineroff, Director of SA for Tech Alliances, Domino Data Lab
- Nicholas Jablonski, Field CTO, Domino Data Lab
- Prabu Arjunan, Solution Architect, NetApp
- Brian Young, Global Alliance Director, Technology Alliance Partners, NetApp

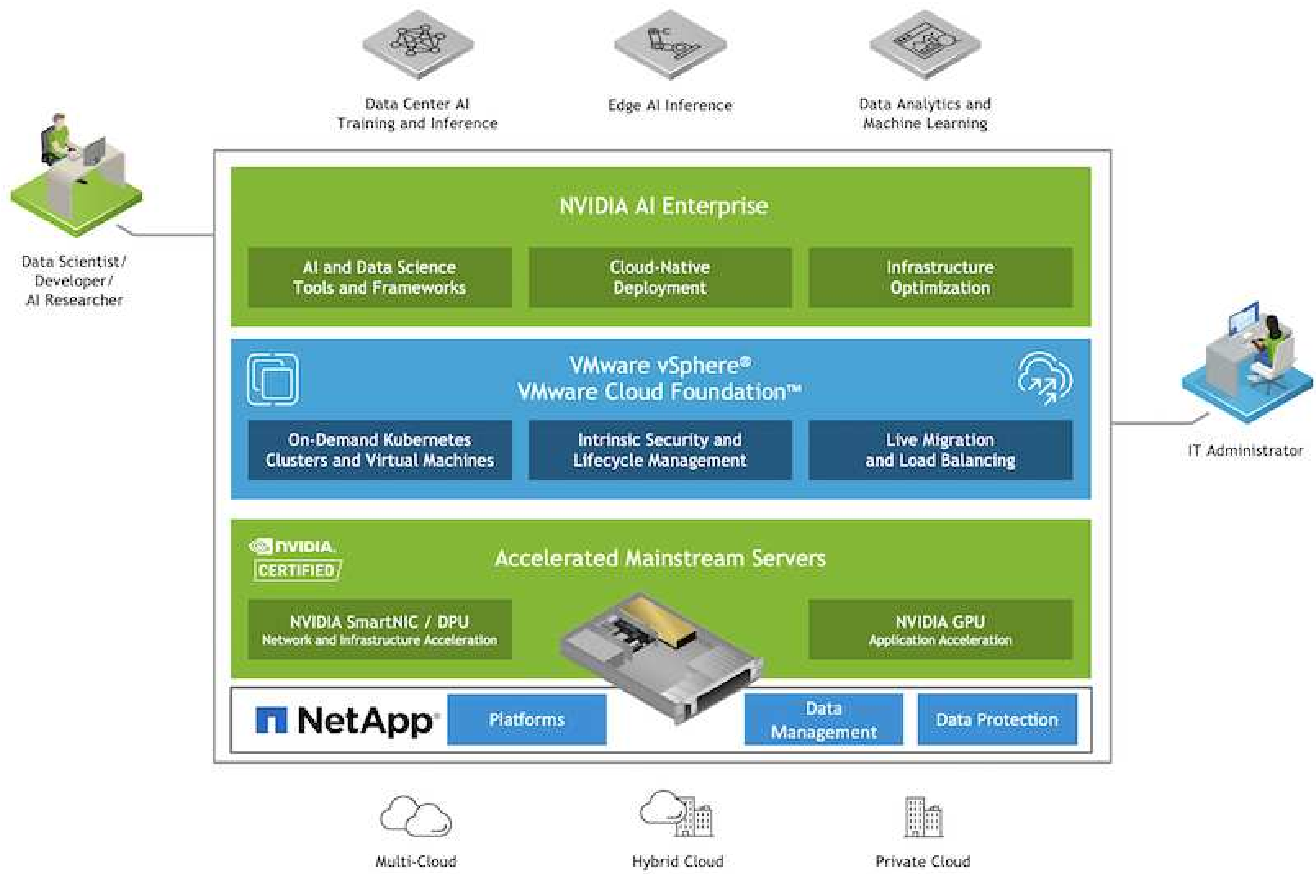
NVIDIA AI Enterprise with NetApp and VMware

NVIDIA AI Enterprise with NetApp and VMware

Mike Oglesby, NetApp

For IT architects and admins, AI tooling can be complicated and unfamiliar. Additionally, many AI platforms are not enterprise-ready. NVIDIA AI Enterprise, powered by NetApp and VMware, was created to deliver a streamlined, enterprise-class AI architecture.

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. This software facilitates the simple and rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud environment. NVIDIA AI Enterprise, powered by NetApp and VMware, delivers enterprise-class AI workload and data management in a simplified, familiar package.



Technology Overview

This section provides a technology overview for NVIDIA AI Enterprise with NetApp and VMware.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. This software facilitates the simple and rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud environment.

NVIDIA GPU Cloud (NGC)

NVIDIA NGC hosts a catalog of GPU-optimized software for AI practitioners to develop their AI solutions. It also provides access to various AI services including NVIDIA Base Command for model training, NVIDIA Fleet Command to deploy and monitor models, and the NGC Private Registry for securely accessing and managing proprietary AI software. Also, NVIDIA AI Enterprise customers can request support through the NGC portal.

VMware vSphere

VMware vSphere is VMware's virtualization platform, which transforms data centers into aggregated computing infrastructures that include CPU, storage, and networking resources. vSphere manages these infrastructures as a unified operating environment, and provides administrators with the tools to manage the data centers that participate in that environment.

The two core components of vSphere are ESXi and vCenter Server. ESXi is the virtualization platform where administrators create and run virtual machines and virtual appliances. vCenter Server is the service through which administrators manage multiple hosts connected in a network and pool host resources.

NetApp ONTAP

ONTAP 9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. You can also move data freely to wherever it is needed: the edge, the core, or the cloud. ONTAP 9 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

Simplify data management

Data management is crucial to enterprise IT operations and data scientists so that appropriate resources are used for AI applications and training AI/ML datasets. The following additional information about NetApp technologies is out of scope for this validation but might be relevant depending on your deployment.

ONTAP data management software includes the following features to streamline and simplify operations and reduce your total cost of operation:

- Inline data compaction and expanded deduplication. Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity. This applies to data stored locally and data tiered to the cloud.
- Minimum, maximum, and adaptive quality of service (AQoS). Granular quality of service (QoS) controls help maintain performance levels for critical applications in highly shared environments.
- NetApp FabricPool. Provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID storage solution. For more information about FabricPool, see TR-4598: FabricPool best practices.

Accelerate and protect data

ONTAP delivers superior levels of performance and data protection and extends these capabilities in the following ways:

- Performance and lower latency. ONTAP offers the highest possible throughput at the lowest possible latency.
- Data protection. ONTAP provides built-in data protection capabilities with common management across all platforms.
- NetApp Volume Encryption (NVE). ONTAP offers native volume-level encryption with both onboard and External Key Management support.
- Multitenancy and multifactor authentication. ONTAP enables sharing of infrastructure resources with the highest levels of security.

Future-proof infrastructure

ONTAP helps meet demanding and constantly changing business needs with the following features:

- Seamless scaling and nondisruptive operations. ONTAP supports the nondisruptive addition of capacity to existing controllers and to scale-out clusters. Customers can upgrade to the latest technologies, such as NVMe and 32Gb FC, without costly data migrations or outages.

- Cloud connection. ONTAP is the most cloud-connected storage management software, with options for software-defined storage (ONTAP Select) and cloud-native instances (NetApp Cloud Volumes Service) in all public clouds.
- Integration with emerging applications. ONTAP offers enterprise-grade data services for next generation platforms and applications, such as autonomous vehicles, smart cities, and Industry 4.0, by using the same infrastructure that supports existing enterprise apps.

NetApp DataOps Toolkit

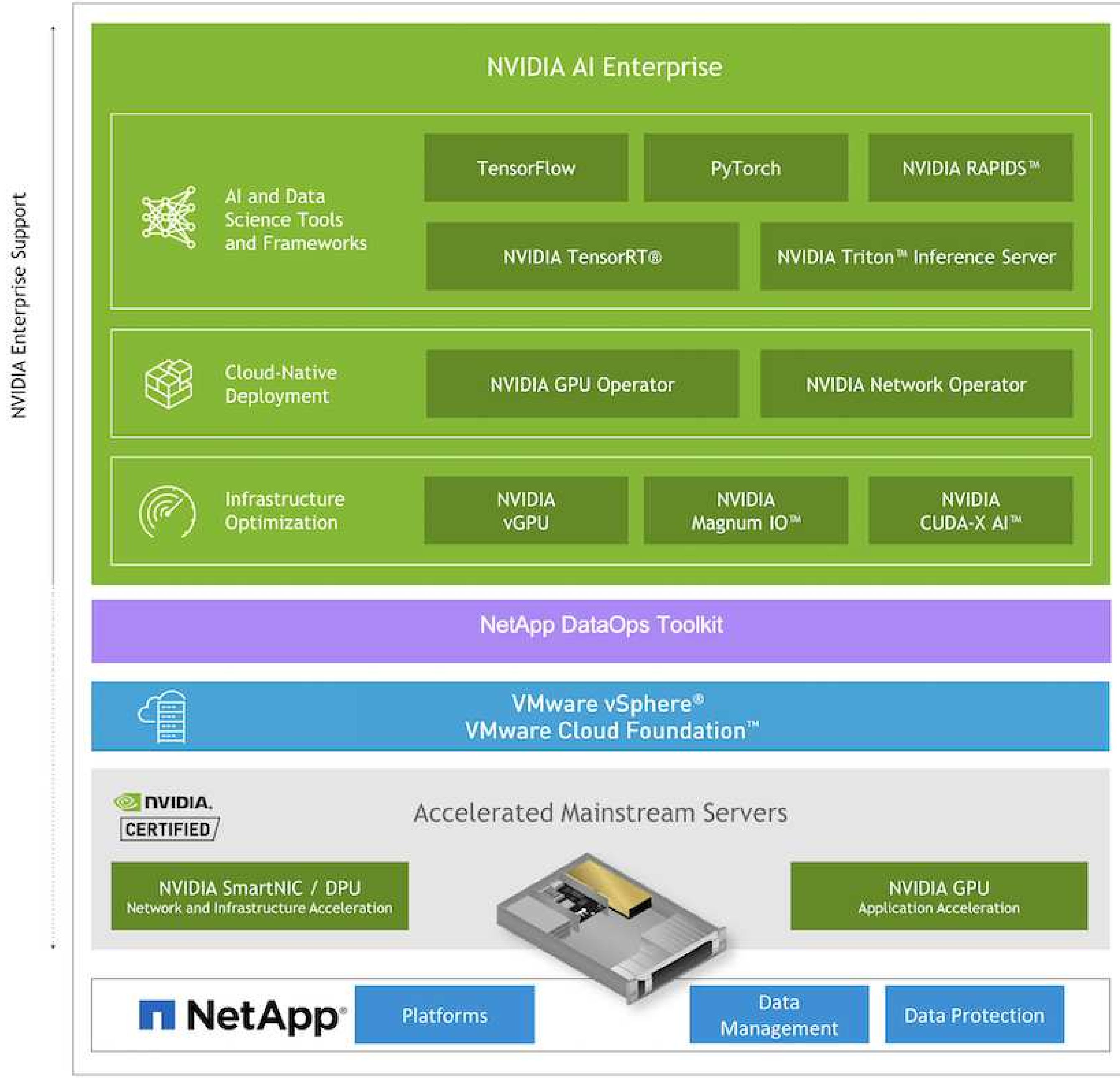
The NetApp DataOps Toolkit is a Python-based tool that simplifies the management of development/training workspaces and inference servers that are backed by high-performance, scale-out NetApp storage. Key capabilities include:

- Rapidly provision new high-capacity JupyterLab workspaces that are backed by high-performance, scale-out NetApp storage.
- Rapidly provision new NVIDIA Triton Inference Server instances that are backed by enterprise-class NetApp storage.
- Near-instaneously clone high-capacity JupyterLab workspaces in order to enable experimentation or rapid iteration.
- Near-instaneously save snapshots of high-capacity JupyterLab workspaces for backup and/or traceability/baselining.
- Near-instaneously provision, clone, and snapshot high-capacity, high-performance data volumes.

Architecture

This solution builds upon a proven and familiar architecture featuring NetApp, VMware, and NVIDIA-Certified Systems. See the following table for details.

Component	Details
AI and Data Analytics Software	NVIDIA AI Enterprise for VMware
Virtualization Platform	VMware vSphere
Compute Platform	NVIDIA-Certified Systems
Data Management Platform	NetApp ONTAP



Initial Setup

This section describes the initial setup tasks that need to be performed in order to utilize NVIDIA AI Enterprise with NetApp and VMware.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already deployed VMware vSphere and NetApp ONTAP. Refer to the NVIDIA AI Enterprise Product Support Matrix for details on supported vSphere versions. Refer to the NetApp and VMware solution documentation for details on deploying VMware vSphere with NetApp ONTAP.

Install NVIDIA AI Enterprise Host Software

To install the NVIDIA AI Entrprise host software, follow the instructions outlined in sections 1-4 in the NVIDIA AI Enterprise Quick Start Guide.

Utilize NVIDIA NGC Software

This section describes the tasks that need to be performed in order to utilize NVIDIA NGC enterprise software within an NVIDIA AI Enterprise environment.

Setup

This section describes the initial setup tasks that need to be performed in order to utilize NVIDIA NGC enterprise software within an NVIDIA AI Enterprise environment.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already deployed the NVIDIA AI Entrprise host software by following the instructions outlined on the Initial Setup page.

Create an Ubuntu Guest VM with vGPU

First, you must create an Ubuntu 20.04 guest VM with vGPU. To create an Ubuntu 20.04 guest VM with vGPU, follow the instructions outline in the NVIDIA AI Enterprise Deployment Guide.

Download and Install NVIDIA Guest Software

Next, you must install the required NVIDIA guest software within the guest VM that you created in the previous step. To download and install the required NVIDIA guest software within the guest VM, follow the instructions outlined in sections 5.1-5.4 in the NVIDIA AI Enterprise Quick Start Guide.

> ⓘ When performing the verification tasks outlined in section 5.4, you may need to use a different CUDA container image version tag as the CUDA container image has been updated since the writing of the guide. In our validation, we used 'nvidia/cuda:11.0.3-base-ubuntu20.04'.

Download AI/Analytics Framework Container(s)

Next, you must download needed AI or analytics framework container images from NVIDIA NGC so that they will be available within your guest VM. To download framework containers within the guest VM, follow the instructions outlined in the NVIDIA AI Enterprise Deployment Guide.

Install and Configure the NetApp DataOps Toolkit

Next, you must install the NetApp DataOps Toolkit for Traditional Environemnts within the guest VM. The NetApp DataOps Toolkit can be used to manage scale-out data volumes on your ONTAP system directly from the terminal within the guest VM. To install the NetApp DataOps Toolkit within the guest VM, perform the following tasks.

1. Install pip.

```
$ sudo apt update
$ sudo apt install python3-pip
$ python3 -m pip install netapp-dataops-traditional
```

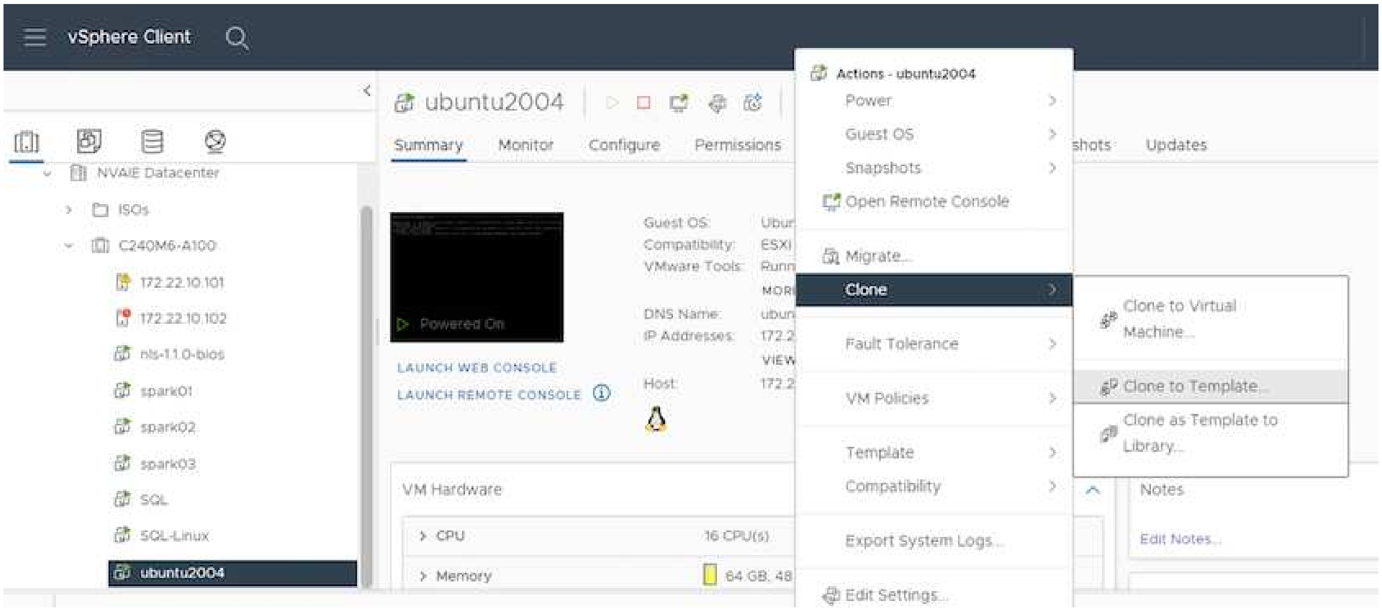
2. Log out of the guest VM terminal and then log back in.
3. Configure the NetApp DataOps Toolkit. In order to complete this step, you will need API access details for your ONTAP system. You may need to obtain these from your storage admin.

```
$ netapp_dataops_cli.py config

Enter ONTAP management LIF hostname or IP address (Recommendation: Use
SVM management interface): 172.22.10.10
Enter SVM (Storage VM) name: NVAIE-client
Enter SVM NFS data LIF hostname or IP address: 172.22.13.151
Enter default volume type to use when creating new volumes
(flexgroup/flexvol) [flexgroup]:
Enter export policy to use by default when creating new volumes
[default]:
Enter snapshot policy to use by default when creating new volumes
[none]:
Enter unix filesystem user id (uid) to apply by default when creating
new volumes (ex. '0' for root user) [0]:
Enter unix filesystem group id (gid) to apply by default when creating
new volumes (ex. '0' for root group) [0]:
Enter unix filesystem permissions to apply by default when creating new
volumes (ex. '0777' for full read/write permissions for all users and
groups) [0777]:
Enter aggregate to use by default when creating new FlexVol volumes:
aff_a400_01_NVME_SSD_1
Enter ONTAP API username (Recommendation: Use SVM account): admin
Enter ONTAP API password (Recommendation: Use SVM account):
Verify SSL certificate when calling ONTAP API (true/false): false
Do you intend to use this toolkit to trigger BlueXP Copy and Sync
operations? (yes/no): no
Do you intend to use this toolkit to push/pull from S3? (yes/no): no
Created config file: '/home/user/.netapp_dataops/config.json'.
```

Create a Guest VM template

Lastly, you must create a VM template based on your guest VM. You will be able to use this template to quickly create guest VMs for utilizing NVIDIA NGC software.

To create a VM template based on your guest VM, log into VMware vSphere, righ-click on the guest VM name, choose 'Clone', choose 'Clone to Template…', and then follow the wizard.



Example Use Case - TensorFlow Training Job

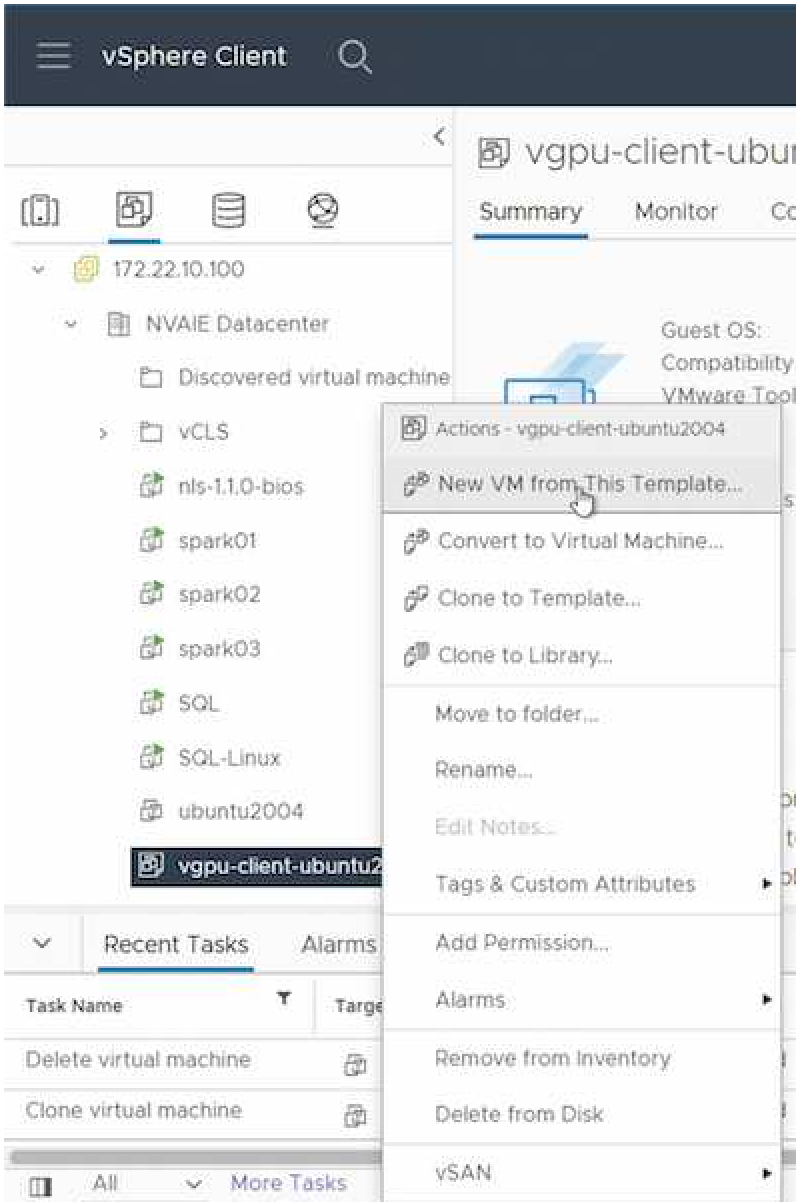
This section describes the tasks that need to be performed in order to execute a TensorFlow training job within an NVIDIA AI Enterprise environment.

Prerequisites

Before you perform the steps that are outlined in this section, we assume that you have already created a guest VM template by following the instructions outlined on the Setup page.

Create Guest VM from Template

First, you must create a new guest VM from the template that you created in the previous section. To create a new guest VM from your template, log into VMware vSphere, righ-click on the template name, choose 'New VM from This Template…', and then follow the wizard.



Create and Mount Data Volume

Next, you must create a new data volume on which to store your training dataset. You can quickly create a new data volume using the NetApp DataOps Toolkit. The example command that follows shows the creation of a volume named 'imagenet' with a capacity of 2 TB.

```
$ netapp_dataops_cli.py create vol -n imagenet -s 2TB
```

Before you can populate your data volume with data, you must mount it within the guest VM. You can quickly mount a data volume using the NetApp DataOps Toolkit. The example command that follows shows the mouting of the volume that was created in the previous step.

```
$ sudo -E netapp_dataops_cli.py mount vol -n imagenet -m ~/imagenet
```

Populate Data Volume

After the new volume has been provisioned and mounted, the training dataset can be retrieved from the source location and placed on the new volume. This typically will involve pulling the data from an S3 or Hadoop data lake and sometimes will involve help from a data engineer.

Execute TensorFlow Training Job

Now, you are ready to execute your TensorFlow training job. To execute your TensorFlow training job, perform the following tasks.

1. Pull the NVIDIA NGC enterprise TensorFlow container image.

```
$ sudo docker pull nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3
```

2. Launch an instance of the NVIDIA NGC enterprise TensorFlow container. Use the '-v' option to attach your data volume to the container.

```
$ sudo docker run --gpus all -v ~/imagenet:/imagenet -it --rm
nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3
```

3. Execute your TensorFlow training program within the container. The example command that follows shows the execution of an example ResNet-50 training program that is included in the container image.

```
$ python ./nvidia-examples/cnn/resnet.py --layers 50 -b 64 -i 200 -u
batch --precision fp16 --data_dir /imagenet/data
```

Where to Find Additional Information

To learn more about the information described in this document, refer to the following documents and/or websites:

- NetApp ONTAP data management software — ONTAP information library

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- NetApp DataOps Toolkit

  https://github.com/NetApp/netapp-dataops-toolkit

- NVIDIA AI Enterprise with VMware

  https://www.nvidia.com/en-us/data-center/products/ai-enterprise/vmware/^]

Acknowledgments

- Bobby Oommen, Sr. Manager, NetApp

* Ramesh Isaac, Systems Administrator, NetApp
* Roney Daniel, Technical Marketing Engineer, NetApp

TR-4851: NetApp StorageGRID data lake for autonomous driving workloads - Solution design

David Arnette, NetApp

TR-4851 demonstrates the use of NetApp StorageGRID object storage as a data repository and management system for machine learning (ML) and deep learning (DL) software development. This paper describes the data flow and requirements in autonomous vehicle software development and the StorageGRID features that streamline the data lifecycle. This solution applies to any multistage data pipeline workflow that is typical in ML and DL development processes.

TR-4851: NetApp StorageGRID data lake for autonomous driving workloads - Solution design

Open Source MLOps with NetApp

Open Source MLOps with NetApp

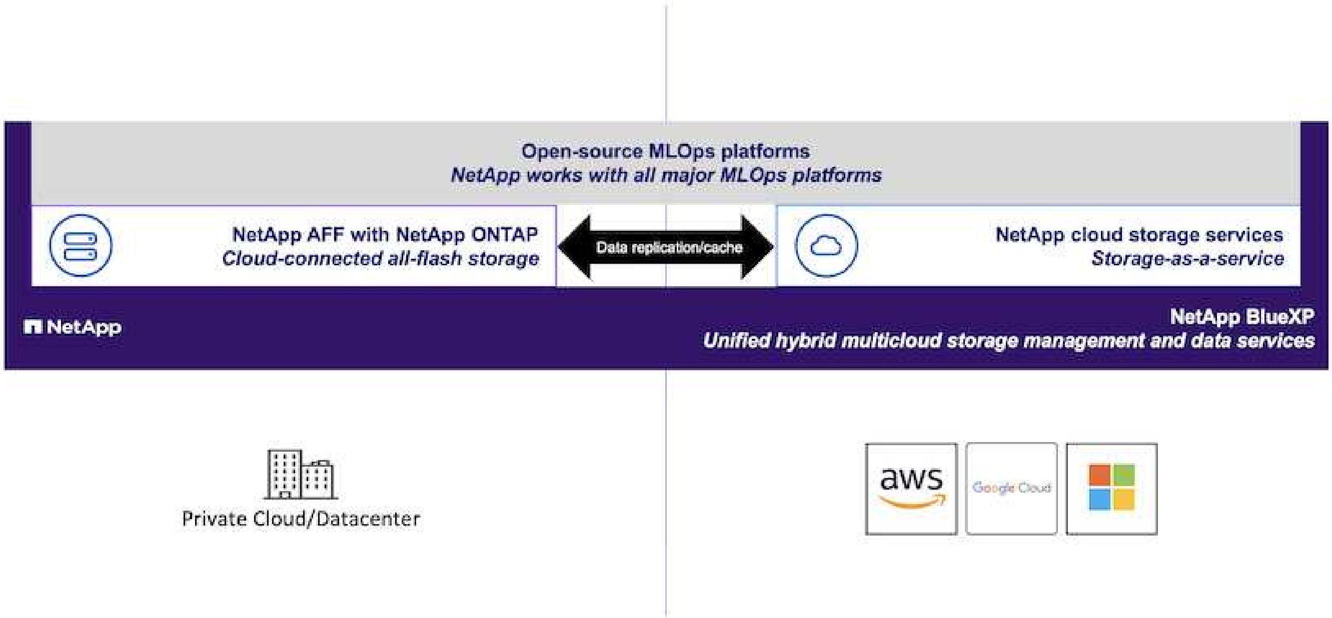
Mike Oglesby, NetApp
Mohan Acharya, NetApp

Companies and organizations of all sizes and across many industries are turning to artificial intelligence (AI), machine learning (ML), and deep learning (DL) to solve real-world problems, deliver innovative products and services, and to get an edge in an increasingly competitive marketplace. As organizations increase their use of AI, ML, and DL, they face many challenges, including workload scalability and data availability. This solution demonstrates how you can address these challenges by pairing NetApp data management capabilities with popular open-source tools and frameworks.

This solution is intended to demonstrate several different open-source tools and frameworks that can be incorporated into an MLOps workflow. These different tools and frameworks can be used together or by themselves depending on the requirements and use case.

The following tools/frameworks are covered in this solution:

* Apache Airflow
* Kubeflow



Technology Overview

This section focuses on the technology overview for OpenSource MLOps with NetApp.

Artificial Intelligence

AI is a computer science discipline in which computers are trained to mimic the cognitive functions of the human mind. AI developers train computers to learn and to solve problems in a manner that is similar to, or even superior to, humans. Deep learning and machine learning are subfields of AI. Organizations are increasingly adopting AI, ML, and DL to support their critical business needs. Some examples are as follows:

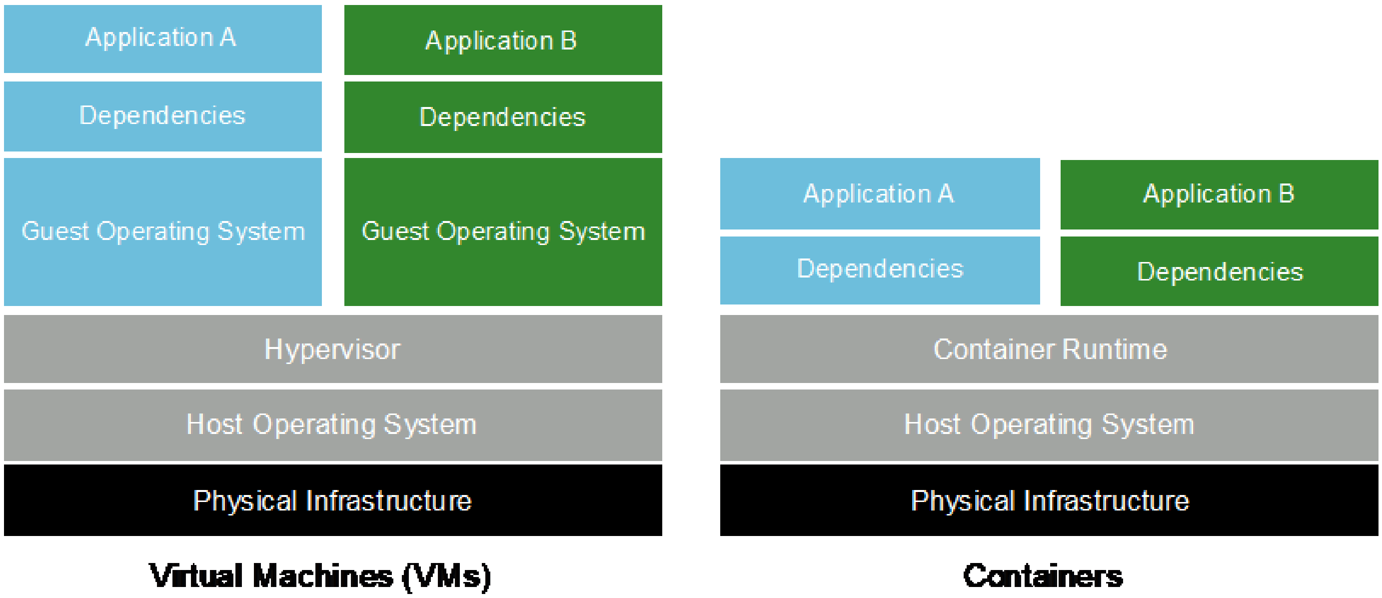
- Analyzing large amounts of data to unearth previously unknown business insights
- Interacting directly with customers by using natural language processing
- Automating various business processes and functions

Modern AI training and inference workloads require massively parallel computing capabilities. Therefore, GPUs are increasingly being used to execute AI operations because the parallel processing capabilities of GPUs are vastly superior to those of general-purpose CPUs.

Containers

Containers are isolated user-space instances that run on top of a shared host operating system kernel. The adoption of containers is increasing rapidly. Containers offer many of the same application sandboxing benefits that virtual machines (VMs) offer. However, because the hypervisor and guest operating system layers that VMs rely on have been eliminated, containers are far more lightweight. The following figure depicts a visualization of virtual machines versus containers.

Containers also allow the efficient packaging of application dependencies, run times, and so on, directly with an application. The most commonly used container packaging format is the Docker container. An application that has been containerized in the Docker container format can be executed on any machine that can run Docker containers. This is true even if the application's dependencies are not present on the machine because all dependencies are packaged in the container itself. For more information, visit the [Docker website](https://www.docker.com/).



Kubernetes

Kubernetes is an open source, distributed, container orchestration platform that was originally designed by Google and is now maintained by the Cloud Native Computing Foundation (CNCF). Kubernetes enables the automation of deployment, management, and scaling functions for containerized applications. In recent years, Kubernetes has emerged as the dominant container orchestration platform. For more information, visit the Kubernetes website.

NetApp Astra Trident

Astra Trident enables consumption and management of storage resources across all popular NetApp storage platforms, in the public cloud or on premises, including ONTAP (AFF, FAS, Select, Cloud, Amazon FSx for NetApp ONTAP), Element software (NetApp HCI, SolidFire), Azure NetApp Files service, and Cloud Volumes Service on Google Cloud. Astra Trident is a Container Storage Interface (CSI) compliant dynamic storage orchestrator that natively integrates with Kubernetes.

NetApp DataOps Toolkit

The NetApp DataOps Toolkit is a Python-based tool that simplifies the management of development/training workspaces and inference servers that are backed by high-performance, scale-out NetApp storage. Key capabilities include:

- Rapidly provision new high-capacity workspaces that are backed by high-performance, scale-out NetApp storage.
- Near-instaneously clone high-capacity workspaces in order to enable experimentation or rapid iteration.
- Near-instaneously save snapshots of high-capacity workspaces for backup and/or traceability/baselining.
- Near-instaneously provision, clone, and snapshot high-capacity, high-performance data volumes.

Kubeflow

Kubeflow is an open source AI and ML toolkit for Kubernetes that was originally developed by Google. The Kubeflow project makes deployments of AI and ML workflows on Kubernetes simple, portable, and scalable. Kubeflow abstracts away the intricacies of Kubernetes, allowing data scientists to focus on what they know best — data science. See the following figure for a visualization. Kubeflow is a good open-source option for organizations that prefer an all-in-one MLOps platform. For more information, visit the Kubeflow website.

Kubeflow Pipelines

Kubeflow Pipelines are a key component of Kubeflow. Kubeflow Pipelines are a platform and standard for defining and deploying portable and scalable AI and ML workflows. For more information, see the official Kubeflow documentation.

Jupyter Notebook Server

A Jupyter Notebook Server is an open source web application that allows data scientists to create wiki-like documents called Jupyter Notebooks that contain live code as well as descriptive test. Jupyter Notebooks are widely used in the AI and ML community as a means of documenting, storing, and sharing AI and ML projects. Kubeflow simplifies the provisioning and deployment of Jupyter Notebook Servers on Kubernetes. For more information on Jupyter Notebooks, visit the Jupyter website. For more information about Jupyter Notebooks within the context of Kubeflow, see the official Kubeflow documentation.

Katib

Katib is a Kubernetes-native project for automated machine learning (AutoML). Katib supports hyperparameter tuning, early stopping and neural architecture search (NAS). Katib is the project which is agnostic to machine learning (ML) frameworks. It can tune hyperparameters of applications written in any language of the users' choice and natively supports many ML frameworks, such as TensorFlow, MXNet, PyTorch, XGBoost, and others. Katib supports a lot of various AutoML algorithms, such as Bayesian optimization, Tree of Parzen Estimators, Random Search, Covariance Matrix Adaptation Evolution Strategy, Hyperband, Efficient Neural Architecture Search, Differentiable Architecture Search and many more. For more information about Jupyter Notebooks within the context of Kubeflow, see the official Kubeflow documentation.

Apache Airflow

Apache Airflow is an open-source workflow management platform that enables programmatic authoring, scheduling, and monitoring for complex enterprise workflows. It is often used to automate ETL and data pipeline workflows, but it is not limited to these types of workflows. The Airflow project was started by Airbnb but has since become very popular in the industry and now falls under the auspices of The Apache Software Foundation. Airflow is written in Python, Airflow workflows are created via Python scripts, and Airflow is designed under the principle of "configuration as code." Many enterprise Airflow users now run Airflow on top of Kubernetes.

Directed Acyclic Graphs (DAGs)

In Airflow, workflows are called Directed Acyclic Graphs (DAGs). DAGs are made up of tasks that are executed in sequence, in parallel, or a combination of the two, depending on the DAG definition. The Airflow scheduler executes individual tasks on an array of workers, adhering to the task-level dependencies that are specified in the DAG definition. DAGs are defined and created via Python scripts.

NetApp ONTAP

ONTAP 9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. You can also move data freely to wherever it is needed: the edge, the core, or the cloud. ONTAP 9 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

Simplify data management

Data management is crucial to enterprise IT operations and data scientists so that appropriate resources are used for AI applications and training AI/ML datasets. The following additional information about NetApp technologies is out of scope for this validation but might be relevant depending on your deployment.

ONTAP data management software includes the following features to streamline and simplify operations and reduce your total cost of operation:

- Inline data compaction and expanded deduplication. Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity. This applies to data stored locally and data tiered to the cloud.
- Minimum, maximum, and adaptive quality of service (AQoS). Granular quality of service (QoS) controls help maintain performance levels for critical applications in highly shared environments.
- NetApp FabricPool. Provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID storage solution. For more information about FabricPool, see TR-4598: FabricPool best practices.

Accelerate and protect data

ONTAP delivers superior levels of performance and data protection and extends these capabilities in the following ways:

- Performance and lower latency. ONTAP offers the highest possible throughput at the lowest possible latency.
- Data protection. ONTAP provides built-in data protection capabilities with common management across all platforms.
- NetApp Volume Encryption (NVE). ONTAP offers native volume-level encryption with both onboard and External Key Management support.
- Multitenancy and multifactor authentication. ONTAP enables sharing of infrastructure resources with the highest levels of security.

Future-proof infrastructure

ONTAP helps meet demanding and constantly changing business needs with the following features:

- Seamless scaling and nondisruptive operations. ONTAP supports the nondisruptive addition of capacity to existing controllers and to scale-out clusters. Customers can upgrade to the latest technologies, such as NVMe and 32Gb FC, without costly data migrations or outages.
- Cloud connection. ONTAP is the most cloud-connected storage management software, with options for software-defined storage and cloud-native instances in all public clouds.
- Integration with emerging applications. ONTAP offers enterprise-grade data services for next generation platforms and applications, such as autonomous vehicles, smart cities, and Industry 4.0, by using the same infrastructure that supports existing enterprise apps.

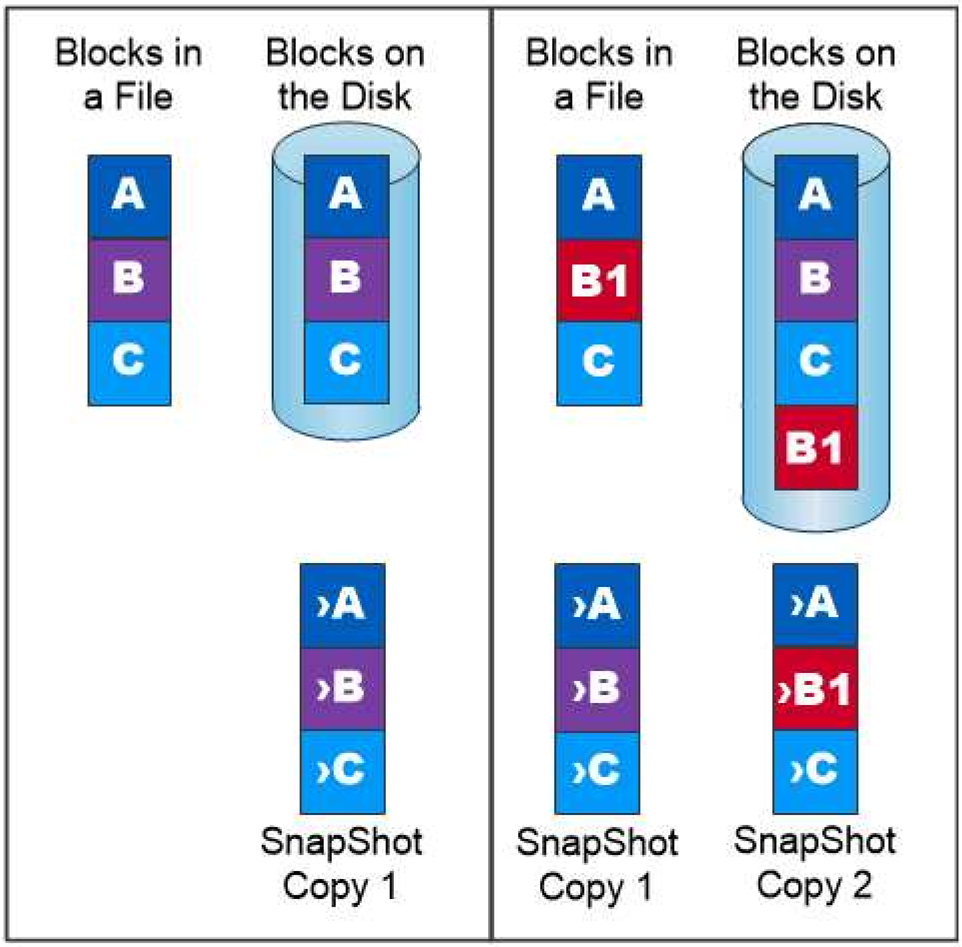
NetApp Snapshot Copies

A NetApp Snapshot copy is a read-only, point-in-time image of a volume. The image consumes minimal storage space and incurs negligible performance overhead because it only records changes to files create since the last Snapshot copy was made, as depicted in the following figure.

Snapshot copies owe their efficiency to the core ONTAP storage virtualization technology, the Write Anywhere

File Layout (WAFL). Like a database, WAFL uses metadata to point to actual data blocks on disk. But, unlike a database, WAFL does not overwrite existing blocks. It writes updated data to a new block and changes the metadata. It’s because ONTAP references metadata when it creates a Snapshot copy, rather than copying data blocks, that Snapshot copies are so efficient. Doing so eliminates the seek time that other systems incur in locating the blocks to copy, as well as the cost of making the copy itself.

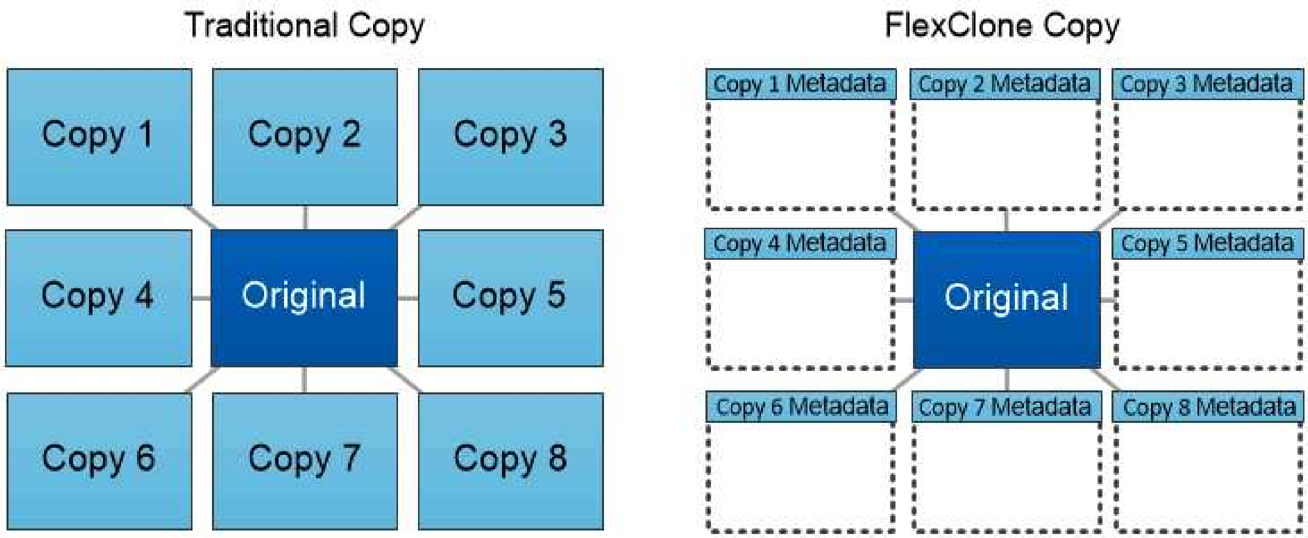
You can use a Snapshot copy to recover individual files or LUNs or to restore the entire contents of a volume. ONTAP compares pointer information in the Snapshot copy with data on disk to reconstruct the missing or damaged object, without downtime or a significant performance cost.



A Snapshot copy records only changes to the active file system since the last Snapshot copy.

NetApp FlexClone Technology

NetApp FlexClone technology references Snapshot metadata to create writable, point-in-time copies of a volume. Copies share data blocks with their parents, consuming no storage except what is required for metadata until changes are written to the copy, as depicted in the following figure. Where traditional copies can take minutes or even hours to create, FlexClone software lets you copy even the largest datasets almost instantaneously. That makes it ideal for situations in which you need multiple copies of identical datasets (a development workspace, for example) or temporary copies of a dataset (testing an application against a production dataset).



FlexClone copies share data blocks with their parents, consuming no storage except what is required for metadata.

NetApp SnapMirror Data Replication Technology

NetApp SnapMirror software is a cost-effective, easy-to-use unified replication solution across the data fabric. It replicates data at high speeds over LAN or WAN. It gives you high data availability and fast data replication for applications of all types, including business critical applications in both virtual and traditional environments. When you replicate data to one or more NetApp storage systems and continually update the secondary data, your data is kept current and is available whenever you need it. No external replication servers are required. See the following figure for an example of an architecture that leverages SnapMirror technology.

SnapMirror software leverages NetApp ONTAP storage efficiencies by sending only changed blocks over the network. SnapMirror software also uses built-in network compression to accelerate data transfers and reduce network bandwidth utilization by up to 70%. With SnapMirror technology, you can leverage one thin replication data stream to create a single repository that maintains both the active mirror and prior point-in-time copies, reducing network traffic by up to 50%.

NetApp BlueXP Copy and Sync

BlueXP Copy and Sync is a NetApp service for rapid and secure data synchronization. Whether you need to transfer files between on-premises NFS or SMB file shares, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, AWS S3, AWS EFS, Azure Blob, Google Cloud Storage, or IBM Cloud Object Storage, BlueXP Copy and Sync moves the files where you need them quickly and securely.

After your data is transferred, it is fully available for use on both source and target. BlueXP Copy and Sync can sync data on-demand when an update is triggered or continuously sync data based on a predefined schedule. Regardless, BlueXP Copy and Sync only moves the deltas, so time and money spent on data replication is minimized.

BlueXP Copy and Sync is a software as a service (SaaS) tool that is extremely simple to set up and use. Data transfers that are triggered by BlueXP Copy and Sync are carried out by data brokers. BlueXP Copy and Sync data brokers can be deployed in AWS, Azure, Google Cloud Platform, or on-premises.

NetApp XCP

NetApp XCP is client-based software for any-to-NetApp and NetApp-to-NetApp data migrations and file system insights. XCP is designed to scale and achieve maximum performance by utilizing all available system resources to handle high-volume datasets and high-performance migrations. XCP helps you to gain complete visibility into the file system with the option to generate reports.

NetApp XCP is available in a single package that supports NFS and SMB protocols. XCP includes a Linux binary for NFS data sets and a windows executable for SMB data sets.

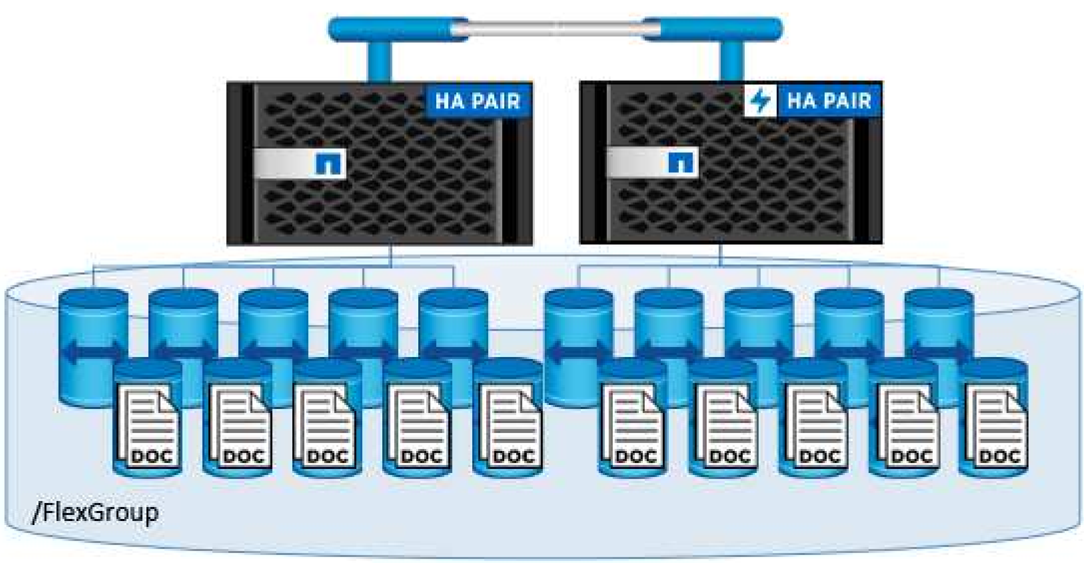
NetApp XCP File Analytics is host-based software that detects file shares, runs scans on the file system, and provides a dashboard for file analytics. XCP File Analytics is compatible with both NetApp and non-NetApp systems and runs on Linux or Windows hosts to provide analytics for NFS and SMB-exported file systems.

NetApp ONTAP FlexGroup Volumes

A training dataset can be a collection of potentially billions of files. Files can include text, audio, video, and other forms of unstructured data that must be stored and processed to be read in parallel. The storage system must store large numbers of small files and must read those files in parallel for sequential and random I/O.

A FlexGroup volume is a single namespace that comprises multiple constituent member volumes, as shown in the following figure. From a storage administrator viewpoint, a FlexGroup volume is managed and acts like a NetApp FlexVol volume. Files in a FlexGroup volume are allocated to individual member volumes and are not striped across volumes or nodes. They enable the following capabilities:

- FlexGroup volumes provide multiple petabytes of capacity and predictable low latency for high-metadata workloads.
- They support up to 400 billion files in the same namespace.
- They support parallelized operations in NAS workloads across CPUs, nodes, aggregates, and constituent FlexVol volumes.



Architecture

This solution is not dependent on specific hardware. The solution is compatible with any NetApp physical storage appliance, software-defined instance, or cloud service, that is supported by Trident. Examples include a NetApp AFF storage system, Amazon FSx for NetApp ONTAP, Azure NetApp Files, or a NetApp Cloud Volumes ONTAP instance. Additionally, the solution can be implemented on any Kubernetes cluster as long as the Kubernetes version used is supported by Kubeflow and NetApp Astra Trident. For a list of Kubernetes versions that are supported by Kubeflow, see the see the official Kubeflow documentation. For a list of Kubernetes versions that are supported by Trident, see the Trident documentation. See the following tables for details on the environment that was used to validate the solution.

Software Component	Version
Apache Airflow	2.0.1
Apache Airflow Helm Chart	8.0.8
Kubeflow	1.7, deployed via deployKF 0.1.1
Kubernetes	1.26
NetApp Astra Trident	23.07

Support

NetApp does not offer enterprise support for Apache Airflow, Kubeflow, or Kubernetes. If you are interested in a fully supported MLOps platform, contact NetApp about fully supported MLOps solutions that NetApp offers jointly with partners.

NetApp Astra Trident Configuration

Example Astra Trident Backends for NetApp AIPod Deployments

Before you can use Astra Trident to dynamically provision storage resources within your Kubernetes cluster, you must create one or more Trident Backends. The examples that follow represent different types of Backends that you might want to create if you are deploying components of this solution on a NetApp AIPod. For more information about Backends, see the Astra Trident documentation.

1. NetApp recommends creating a FlexGroup-enabled Trident Backend for your AIPod.

   The example commands that follow show the creation of a FlexGroup-enabled Trident Backend for an AIPod storage virtual machine (SVM). This Backend uses the `ontap-nas-flexgroup` storage driver. ONTAP supports two main data volume types: FlexVol and FlexGroup. FlexVol volumes are size-limited (as of this writing, the maximum size depends on the specific deployment). FlexGroup volumes, on the other hand, can scale linearly to up to 20PB and 400 billion files, providing a single namespace that greatly simplifies data management. Therefore, FlexGroup volumes are optimal for AI and ML workloads that rely on large amounts of data.

   If you are working with a small amount of data and want to use FlexVol volumes instead of FlexGroup volumes, you can create Trident Backends that use the `ontap-nas` storage driver instead of the `ontap-nas-flexgroup` storage driver.

```
$ cat << EOF > ./trident-backend-aipod-flexgroups-iface1.json
{
    "version": 1,
    "storageDriverName": "ontap-nas-flexgroup",
    "backendName": "aipod-flexgroups-iface1",
    "managementLIF": "10.61.218.100",
    "dataLIF": "192.168.11.11",
    "svm": "ontapai_nfs",
    "username": "admin",
    "password": "ontapai"
}
EOF
$ tridentctl create backend -f ./trident-backend-aipod-flexgroups-
iface1.json -n trident
+-------------------------+---------------------
+--------------------------------------+--------+---------+
|          NAME           |   STORAGE DRIVER    |                 UUID
| STATE  | VOLUMES |
+-------------------------+---------------------
+--------------------------------------+--------+---------+
| aipod-flexgroups-iface1 | ontap-nas-flexgroup | b74cbddb-e0b8-40b7-
b263-b6da6dec0bdd | online |       0 |
+-------------------------+---------------------
+--------------------------------------+--------+---------+
$ tridentctl get backend -n trident
+-------------------------+---------------------
+--------------------------------------+--------+---------+
|          NAME           |   STORAGE DRIVER    |                 UUID
| STATE  | VOLUMES |
+-------------------------+---------------------
+--------------------------------------+--------+---------+
| aipod-flexgroups-iface1 | ontap-nas-flexgroup | b74cbddb-e0b8-40b7-
b263-b6da6dec0bdd | online |       0 |
+-------------------------+---------------------
+--------------------------------------+--------+---------+
```

2. NetApp also recommends creating a FlexVol- enabled Trident Backend. You may want to use FlexVol volumes for hosting persistent applications, storing results, output, debug information, and so on. If you want to use FlexVol volumes, you must create one or more FlexVol- enabled Trident Backends. The example commands that follow show the creation of a single FlexVol- enabled Trident Backend.

```
$ cat << EOF > ./trident-backend-aipod-flexvols.json
{
    "version": 1,
    "storageDriverName": "ontap-nas",
    "backendName": "aipod-flexvols",
    "managementLIF": "10.61.218.100",
    "dataLIF": "192.168.11.11",
    "svm": "ontapai_nfs",
    "username": "admin",
    "password": "ontapai"
}
EOF
$ tridentctl create backend -f ./trident-backend-aipod-flexvols.json -n
trident
+--------------------------+---------------------
+--------------------------------------+--------+---------+
|            NAME          |   STORAGE DRIVER    |                 UUID
| STATE  | VOLUMES |
+--------------------------+---------------------
+--------------------------------------+--------+---------+
| aipod-flexvols           | ontap-nas           | 52bdb3b1-13a5-4513-a9c1-
52a69657fabe | online |       0 |
+--------------------------+---------------------
+--------------------------------------+--------+---------+
$ tridentctl get backend -n trident
+--------------------------+---------------------
+--------------------------------------+--------+---------+
|            NAME          |   STORAGE DRIVER    |                 UUID
| STATE  | VOLUMES |
+--------------------------+---------------------
+--------------------------------------+--------+---------+
| aipod-flexvols           | ontap-nas           | 52bdb3b1-13a5-4513-a9c1-
52a69657fabe | online |       0 |
| aipod-flexgroups-iface1 | ontap-nas-flexgroup | b74cbddb-e0b8-40b7-b263-
b6da6dec0bdd | online |       0 |
+--------------------------+---------------------
+--------------------------------------+--------+---------+
```

Example Kubernetes StorageClasses for NetApp AIPod Deployments

Before you can use Astra Trident to dynamically provision storage resources within your Kubernetes cluster, you must create one or more Kubernetes StorageClasses. The examples that follow represent different types of StorageClasses that you might want to create if you are deploying components of this solution on a NetApp AIPod. For more information about StorageClasses, see the Astra Trident documentation.

1. NetApp recommends creating a StorageClass for the FlexGroup-enabled Trident Backend that you created in the section Example Astra Trident Backends for NetApp AIPod Deployments, step 1. The example commands that follow show the creation of multiple StorageClasses that corresponds to the two example Backend that was created in the section Example Astra Trident Backends for NetApp AIPod Deployments, step 1 - one that utilizes NFS over RDMA and one that does not.

   So that a persistent volume isn't deleted when the corresponding PersistentVolumeClaim (PVC) is deleted, the following example uses a `reclaimPolicy` value of `Retain`. For more information about the `reclaimPolicy` field, see the official Kubernetes documentation.

   Note: The following example StorageClasses use a maximum transfer size of 262144. To use this maximum transfer size, you must configure the maximum transfer size on your ONTAP system accordingly. Refer to the ONTAP documentation for details.

   Note: To use NFS over RDMA, you must configure NFS over RDMA on your ONTAP system. Refer to the linkhttps://docs.netapp.com/us-en/ontap/nfs-rdma/[ONTAP documentation] for details.

   Note: In the following example, a specific Backend is not specified in the storagePool field in StorageClass definition file.

```
$ cat << EOF > ./storage-class-aipod-flexgroups-retain.yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: aipod-flexgroups-retain
provisioner: csi.trident.netapp.io
mountOptions: ["vers=4.1", "nconnect=16", "rsize=262144",
"wsize=262144"]
parameters:
  backendType: "ontap-nas-flexgroup"
  storagePools: "aipod-flexgroups-iface1:.*"
reclaimPolicy: Retain
EOF
$ kubectl create -f ./storage-class-aipod-flexgroups-retain.yaml
storageclass.storage.k8s.io/aipod-flexgroups-retain created
$ cat << EOF > ./storage-class-aipod-flexgroups-retain-rdma.yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: aipod-flexgroups-retain-rdma
provisioner: csi.trident.netapp.io
mountOptions: ["vers=4.1", "proto=rdma", "max_connect=16",
"rsize=262144", "wsize=262144"]
parameters:
  backendType: "ontap-nas-flexgroup"
  storagePools: "aipod-flexgroups-iface1:.*"
reclaimPolicy: Retain
EOF
$ kubectl create -f ./storage-class-aipod-flexgroups-retain-rdma.yaml
storageclass.storage.k8s.io/aipod-flexgroups-retain-rdma created
$ kubectl get storageclass
NAME                             PROVISIONER             AGE
aipod-flexgroups-retain          csi.trident.netapp.io   0m
aipod-flexgroups-retain-rdma     csi.trident.netapp.io   0m
```

2. NetApp also recommends creating a StorageClass that corresponds to the FlexVol-enabled Trident Backend that you created in the section Example Astra Trident Backends for AIPod Deployments, step 2. The example commands that follow show the creation of a single StorageClass for FlexVol volumes.

   Note: In the following example, a particular Backend is not specified in the storagePool field in StorageClass definition file. When you use Kubernetes to administer volumes using this StorageClass, Trident attempts to use any available backend that uses the `ontap-nas` driver.

```
$ cat << EOF > ./storage-class-aipod-flexvols-retain.yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: aipod-flexvols-retain
provisioner: netapp.io/trident
parameters:
  backendType: "ontap-nas"
reclaimPolicy: Retain
EOF
$ kubectl create -f ./storage-class-aipod-flexvols-retain.yaml
storageclass.storage.k8s.io/aipod-flexvols-retain created
$ kubectl get storageclass
NAME                             PROVISIONER              AGE
aipod-flexgroups-retain          csi.trident.netapp.io    0m
aipod-flexgroups-retain-rdma     csi.trident.netapp.io    0m
aipod-flexvols-retain            csi.trident.netapp.io    0m
```

Kubeflow

Kubeflow Deployment

This section describes the tasks that you must complete to deploy Kubeflow in your Kubernetes cluster.

Prerequisites

Before you perform the deployment exercise that is outlined in this section, we assume that you have already performed the following tasks:

1. You already have a working Kubernetes cluster, and you are running a version of Kubernetes that is supported by the Kubeflow version that you intend to deploy. For a list of supported Kubernetes versions, refer to the dependencies for your Kubeflow version in the official Kubeflow documentation.
2. You have already installed and configured NetApp Astra Trident in your Kubernetes cluster. For more details on Astra Trident, refer to the Astra Trident documentation.

Set Default Kubernetes StorageClass

Before you deploy Kubeflow, we recommend designating a default StorageClass within your Kubernetes cluster. The Kubeflow deployment process may attempt to provision new persistent volumes using the default StorageClass. If no StorageClass is designated as the default StorageClass, then the deployment may fail. To designate a default StorageClass within your cluster, perform the following task from the deployment jump host. If you have already designated a default StorageClass within your cluster, then you can skip this step.

1. Designate one of your existing StorageClasses as the default StorageClass. The example commands that follow show the designation of a StorageClass named `ontap-ai-flexvols-retain` as the default StorageClass.

The `ontap-nas-flexgroup` Trident Backend type has a minimum PVC size that is fairly large. By default, Kubeflow attempts to provision PVCs that are only a few GBs in size. Therefore, you should not designate a StorageClass that utilizes the `ontap-nas-flexgroup` Backend type as the default StorageClass for the purposes of Kubeflow deployment.

```
$ kubectl get sc
NAME                                PROVISIONER             AGE
ontap-ai-flexgroups-retain          csi.trident.netapp.io   25h
ontap-ai-flexgroups-retain-iface1   csi.trident.netapp.io   25h
ontap-ai-flexgroups-retain-iface2   csi.trident.netapp.io   25h
ontap-ai-flexvols-retain            csi.trident.netapp.io   3s
$ kubectl patch storageclass ontap-ai-flexvols-retain -p '{"metadata":
{"annotations":{"storageclass.kubernetes.io/is-default-class":"true"}}}'
storageclass.storage.k8s.io/ontap-ai-flexvols-retain patched
$ kubectl get sc
NAME                                 PROVISIONER             AGE
ontap-ai-flexgroups-retain           csi.trident.netapp.io   25h
ontap-ai-flexgroups-retain-iface1    csi.trident.netapp.io   25h
ontap-ai-flexgroups-retain-iface2    csi.trident.netapp.io   25h
ontap-ai-flexvols-retain (default)   csi.trident.netapp.io   54s
```

Kubeflow Deployment Options

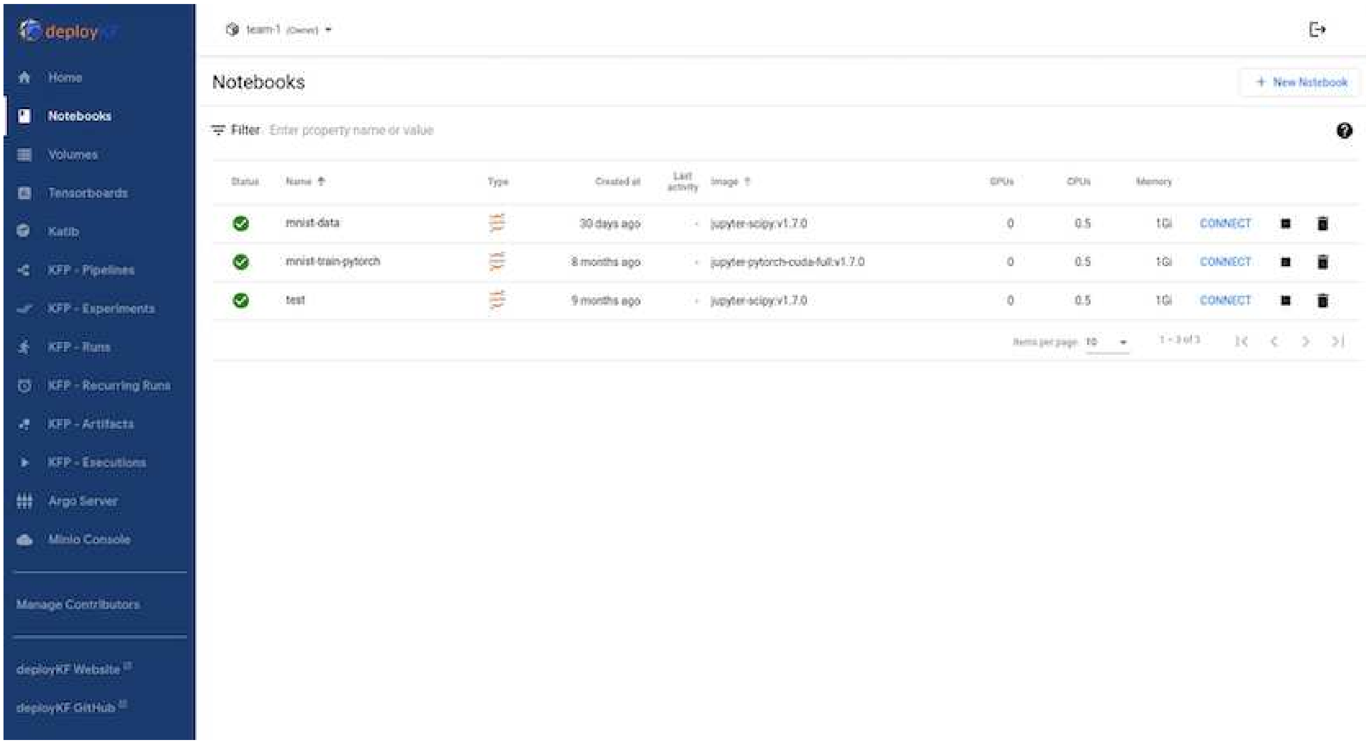
There are many different options for deploying Kubeflow. Refer to the official Kubeflow documentation for a list of deployment options, and choose the option that is the best fit for your needs.

For validation purposes, we deployed Kubeflow 1.7 using deployKF 0.1.1.

Example Kubeflow Operations and Tasks

Provision a Jupyter Notebook Workspace for Data Scientist or Developer Use

Kubeflow is capable of rapidly provisioning new Jupyter Notebook servers to act as data scientist workspaces. For more information about Jupyter Notebooks within the Kubeflow context, see the official Kubeflow documentation.



Use the NetApp DataOps Toolkit with Kubeflow

The NetApp Data Science Toolkit for Kubernetes can be used in conjunction with Kubeflow. Using the NetApp Data Science Toolkit with Kubeflow provides the following benefits:

- Data scientists can perform advanced NetApp data management operations, such as creating snapshots and clones, directly from within a Jupyter Notebook.
- Advanced NetApp data management operations, such as creating snapshots and clones, can be incorporated into automated workflows using the Kubeflow Pipelines framework.

Refer to the Kubeflow Examples section within the NetApp Data Science Toolkit GitHub repository for details on using the toolkit with Kubeflow.

Example Workflow - Train an Image Recognition Model Using Kubeflow and the NetApp DataOps Toolkit

This section describes the steps involved in training and deploying a Neural Network for Image Recognition using Kubeflow and the NetApp DataOps Toolkit. This is intended to serve as an example to show a training job that incorporates NetApp storage.

Prerequisites

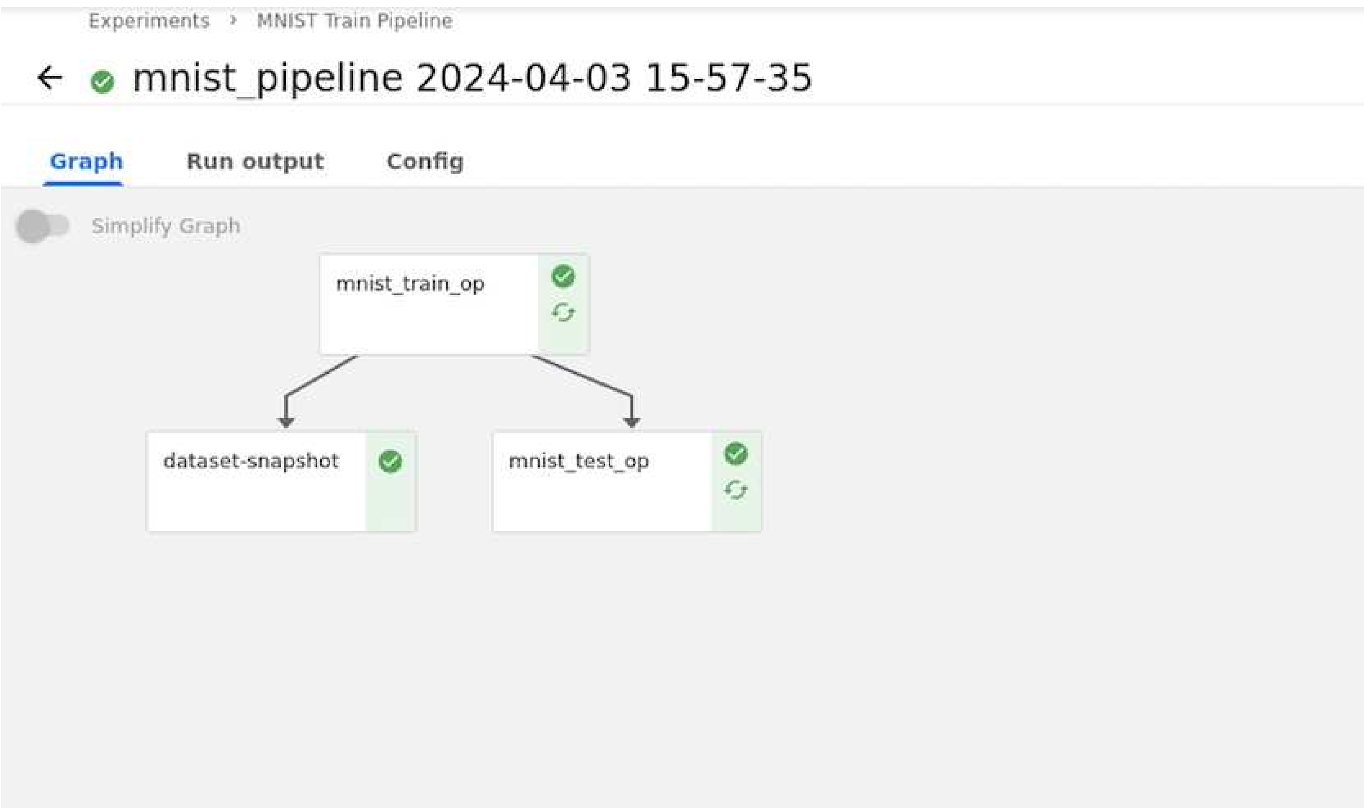
Create a Dockerfile with the required configurations to use for the train and test steps within the Kubeflow pipeline.
Here is an example of a Dockerfile -

```
FROM pytorch/pytorch:latest
RUN pip install torchvision numpy scikit-learn matplotlib tensorboard
WORKDIR /app
COPY . /app
COPY train_mnist.py /app/train_mnist.py
CMD ["python", "train_mnist.py"]
```

Depending on your requirements, install all required libraries and packages needed to run the program. Before you train the Machine Learning model, it is assumed that you already have a working Kubeflow deployment.

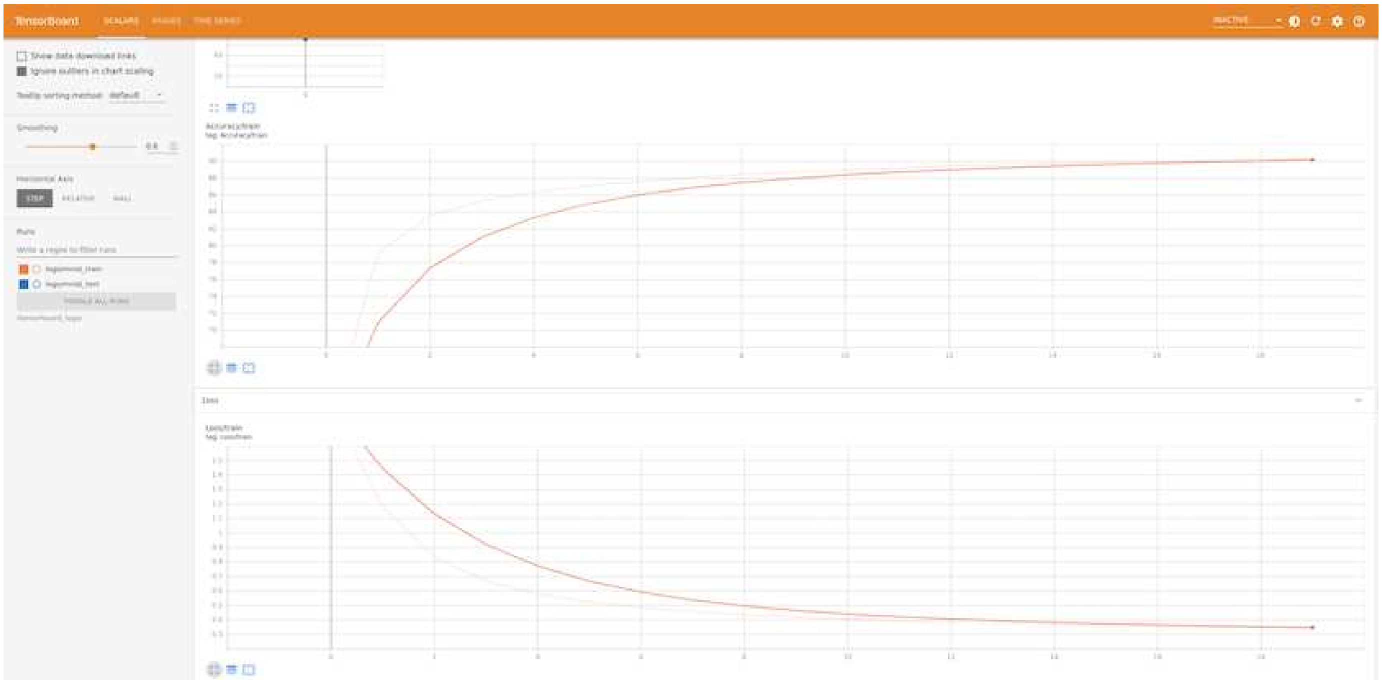
Train a Small NN on MNIST Data Using PyTorch and Kubeflow Pipelines

We use the example of a small Neural Network trained on MNIST data. The MNIST dataset consists of handwritten images of digits from 0-9. The images are 28x28 pixels in size. The dataset is divided into 60,000 train images and 10,000 validation images. The Neural Network used for this experiment is a 2-layer feedforward network. Training is executed using Kubeflow Pipelines. Refer to the documentation here for more information. Our Kubeflow pipeline incorporates the docker image from the Prerequisites section.



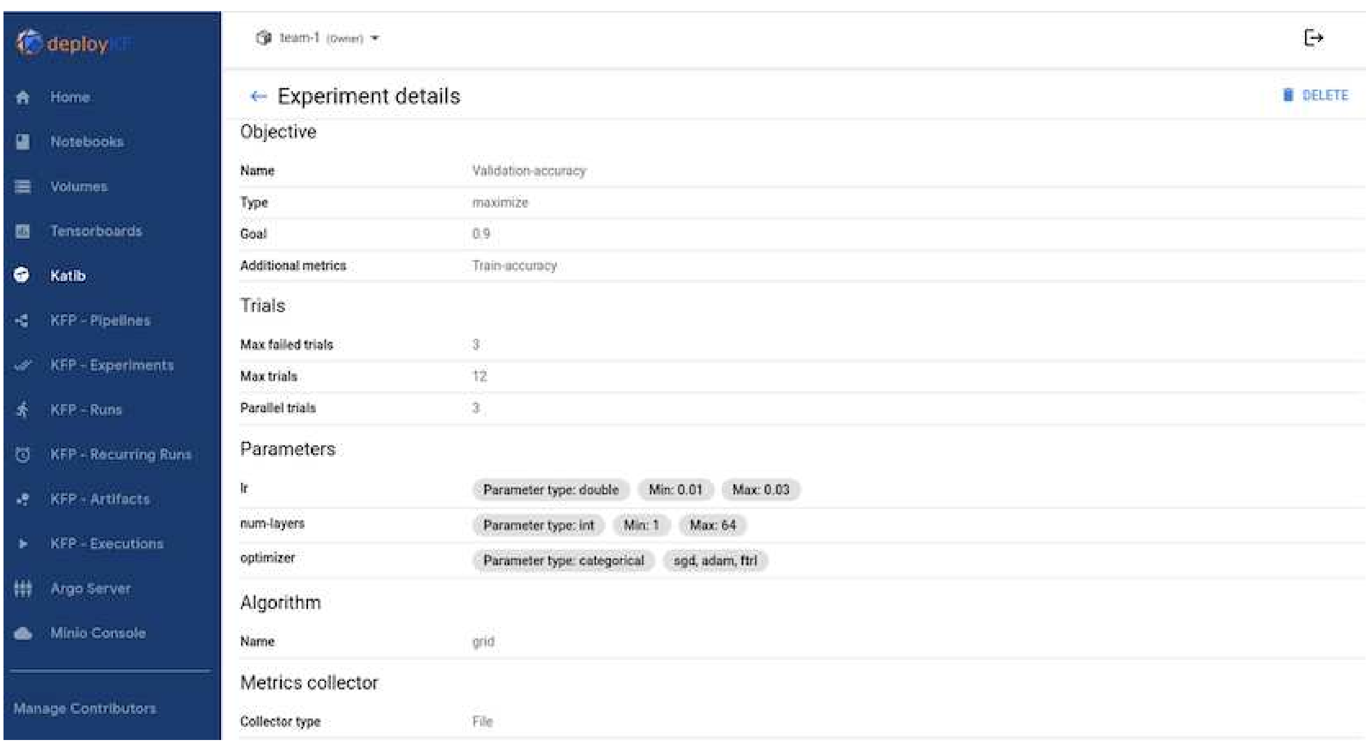
Visualize Results Using Tensorboard

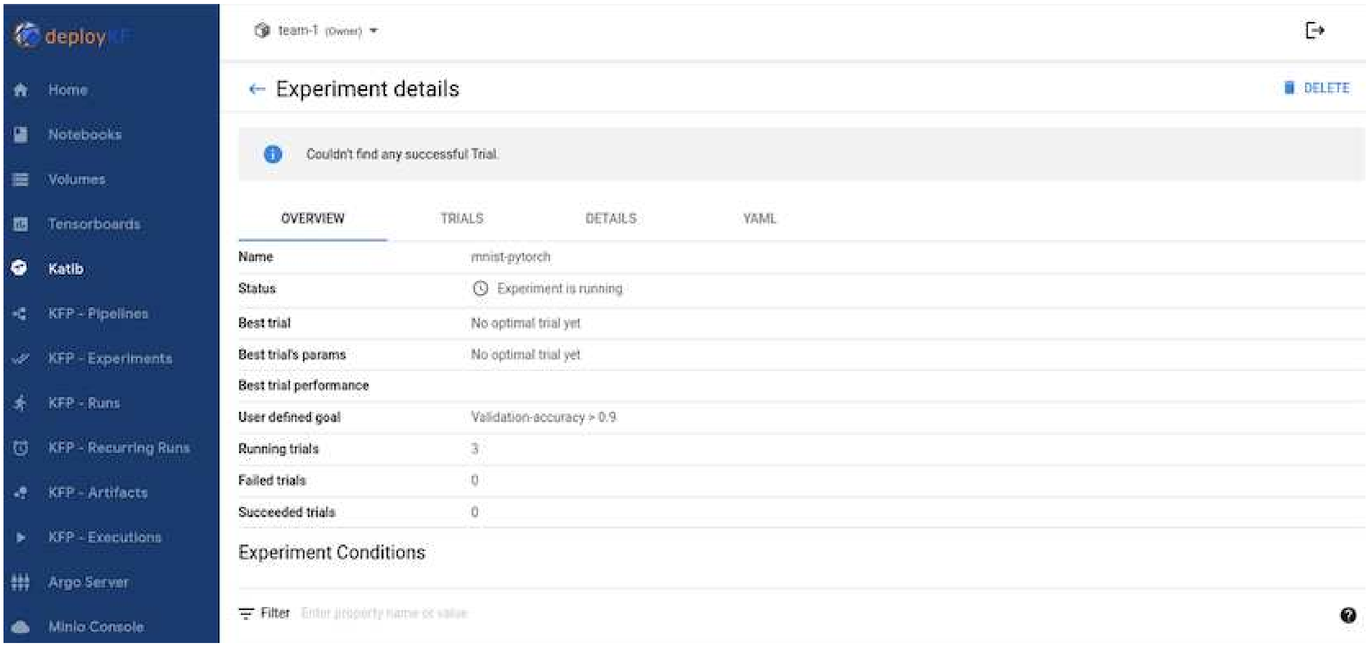
Once the model is trained, we can visualize the results using Tensorboard. Tensorboard is available as a feature on the Kubeflow Dashboard. You can create a custom tensorboard for your job. An example below shows the plot of training accuracy vs. number of epochs and training loss vs. number of epochs.



Experiment with Hyperparameters Using Katib

[Katib](https://www.kubeflow.org/docs/components/katib/) is a tool within Kubeflow that can be used to experiment with the model hyperparameters. To create an experiment, define a desired metric/goal first. This is usually the test accuracy. Once the metric is defined, choose hyperparameters that you would like to play around with (optimizer/learning_rate/number of layers). Katib does a hyperparameter sweep with the user-defined values to find the best combination of parameters that satisfy the desired metric. You can define these parameters in each section in the UI. Alternatively, you could define a **YAML** file with the necessary specifications. Below is an illustration of a Katib experiment -





Use NetApp Snapshots to Save Data for Traceability

During the model training, we may want to save a snapshot of the training dataset for traceability. To do this, we can add a snapshot step to the pipeline as shown below. To create the snapshot, we can use the NetApp DataOps Toolkit for Kubernetes.

```
@dsl.pipeline(
    name = 'MNIST Classification Pipeline',
    description = 'Train a simple NN for classification'
)
def mnist_pipeline():
    mnist_train_task = mnist_train_op()
    mnist_train_task.apply(
        kfp.onprem.mount_pvc('mnist-data', 'mnist-data-vol', '/mnt/data/')
    )

    mnist_test_task = mnist_test_op()
    mnist_test_task.apply(
        kfp.onprem.mount_pvc('mnist-data', 'mnist-data-vol', '/mnt/data/')
    )

    volume_snapshot_name = "mnist-pytorch-snapshot"
    dataset_snapshot = dsl.ContainerOp(
        name="dataset-snapshot",
        image="python:3.9",
        command=["/bin/bash", "-c"],
        arguments=["\
            python3 -m pip install netapp-dataops-k8s && \
            echo '" + volume_snapshot_name + "' > /volume_snapshot_name.txt && \
            netapp_dataops_k8s_cli.py create volume-snapshot --pvc-name=" + "mnist-data" + " --snapshot-name=" + str(volume_snapshot_name) + " --namespace={{workflow.namespace}}"],
        file_outputs={"volume_snapshot_name": "/volume_snapshot_name.txt"}
    )
    mnist_test_task.after(mnist_train_task)
    dataset_snapshot.after(mnist_train_task)
```

Refer to the NetApp DataOps Toolkit example for Kubeflow for more information.

Apache Airflow

Apache Airflow Deployment

This section describes the tasks that you must complete to deploy Airflow in your Kubernetes cluster.

> It is possible to deploy Airflow on platforms other than Kubernetes. Deploying Airflow on platforms other than Kubernetes is outside of the scope of this solution.

Prerequisites

Before you perform the deployment exercise that is outlined in this section, we assume that you have already performed the following tasks:

1. You already have a working Kubernetes cluster.
2. You have already installed and configured NetApp Astra Trident in your Kubernetes cluster. For more details on Astra Trident, refer to the Astra Trident documentation.

Install Helm

Airflow is deployed using Helm, a popular package manager for Kubernetes. Before you deploy Airflow, you must install Helm on the deployment jump host. To install Helm on the deployment jump host, follow the installation instructions in the official Helm documentation.

Set Default Kubernetes StorageClass

Before you deploy Airflow, you must designate a default StorageClass within your Kubernetes cluster. The Airflow deployment process attempts to provision new persistent volumes using the default StorageClass. If no StorageClass is designated as the default StorageClass, then the deployment fails. To designate a default StorageClass within your cluster, follow the instructions outlined in the Kubeflow Deployment section. If you have already designated a default StorageClass within your cluster, then you can skip this step.

Use Helm to Deploy Airflow

To deploy Airflow in your Kubernetes cluster using Helm, perform the following tasks from the deployment jump host:

1. Deploy Airflow using Helm by following the deployment instructions for the official Airflow chart on the Artifact Hub. The example commands that follow show the deployment of Airflow using Helm. Modify, add, and/or remove values in the `custom- values.yaml` file as needed depending on your environment and desired configuration.

```
$ cat << EOF > custom-values.yaml
###################################
# Airflow - Common Configs
###################################
airflow:
  ## the airflow executor type to use
  ##
  executor: "CeleryExecutor"
  ## environment variables for the web/scheduler/worker Pods (for
airflow configs)
  ##
  #
###################################
# Airflow - WebUI Configs
###################################
web:
  ## configs for the Service of the web Pods
  ##
```

```
  service:
    type: NodePort
###################################
# Airflow - Logs Configs
###################################
logs:
  persistence:
    enabled: true
###################################
# Airflow - DAGs Configs
###################################
dags:
  ## configs for the DAG git repository & sync container
  ##
  gitSync:
    enabled: true
    ## url of the git repository
    ##
    repo: "git@github.com:mboglesby/airflow-dev.git"
    ## the branch/tag/sha1 which we clone
    ##
    branch: master
    revision: HEAD
    ## the name of a pre-created secret containing files for ~/.ssh/
    ##
    ## NOTE:
    ## - this is ONLY RELEVANT for SSH git repos
    ## - the secret commonly includes files: id_rsa, id_rsa.pub,
known_hosts
    ## - known_hosts is NOT NEEDED if `git.sshKeyscan` is true
    ##
    sshSecret: "airflow-ssh-git-secret"
    ## the name of the private key file in your `git.secret`
    ##
    ## NOTE:
    ## - this is ONLY RELEVANT for PRIVATE SSH git repos
    ##
    sshSecretKey: id_rsa
    ## the git sync interval in seconds
    ##
    syncWait: 60
EOF
$ helm install airflow airflow-stable/airflow -n airflow --version 8.0.8
--values ./custom-values.yaml
...
Congratulations. You have just deployed Apache Airflow!
```

```
1. Get the Airflow Service URL by running these commands:
   export NODE_PORT=$(kubectl get --namespace airflow -o
jsonpath="{.spec.ports[0].nodePort}" services airflow-web)
   export NODE_IP=$(kubectl get nodes --namespace airflow -o
jsonpath="{.items[0].status.addresses[0].address}")
   echo http://$NODE_IP:$NODE_PORT/
2. Open Airflow in your web browser
```

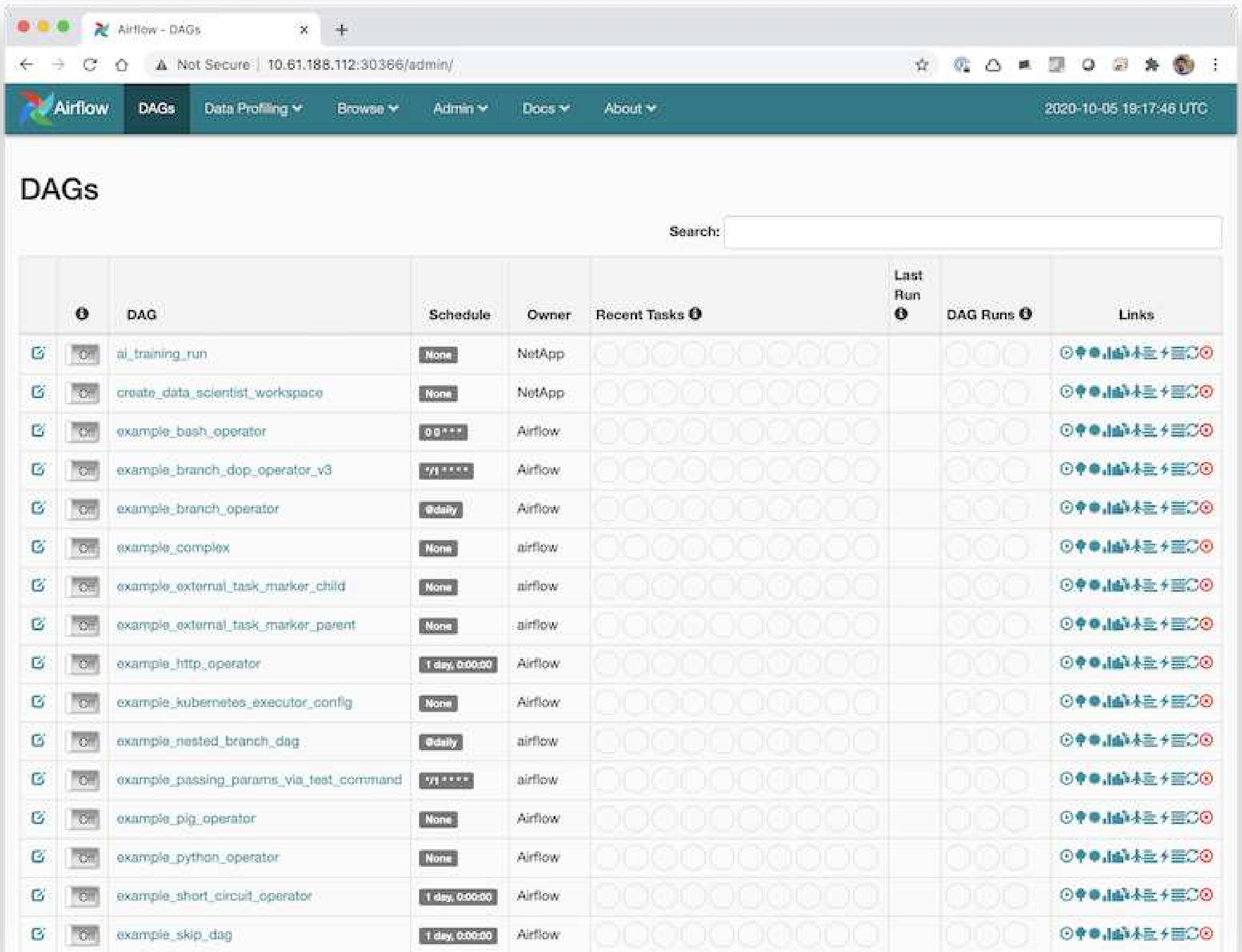
2. Confirm that all Airflow pods are up and running. It may take a few minutes for all pods to start.

```
$ kubectl -n airflow get pod
NAME                                READY   STATUS    RESTARTS   AGE
airflow-flower-b5656d44f-h8qjk      1/1     Running   0          2h
airflow-postgresql-0                1/1     Running   0          2h
airflow-redis-master-0              1/1     Running   0          2h
airflow-scheduler-9d95fcdf9-clf4b   2/2     Running   2          2h
airflow-web-59c94db9c5-z7rg4        1/1     Running   0          2h
airflow-worker-0                    2/2     Running   2          2h
```

3. Obtain the Airflow web service URL by following the instructions that were printed to the console when you deployed Airflow using Helm in step 1.

```
$ export NODE_PORT=$(kubectl get --namespace airflow -o
jsonpath="{.spec.ports[0].nodePort}" services airflow-web)
$ export NODE_IP=$(kubectl get nodes --namespace airflow -o
jsonpath="{.items[0].status.addresses[0].address}")
$ echo http://$NODE_IP:$NODE_PORT/
```

4. Confirm that you can access the Airflow web service.



Use the NetApp DataOps Toolkit with Airflow

The NetApp DataOps Toolkit for Kubernetes can be used in conjunction with Airflow. Using the NetApp DataOps Toolkit with Airflow enables you to incorporate NetApp data management operations, such as creating snapshots and clones, into automated workflows that are orchestrated by Airflow.

Refer to the Airflow Examples section within the NetApp DataOps Toolkit GitHub repository for details on using the toolkit with Airflow.

Example Astra Trident Operations

This section includes examples of various operations that you may want to perform with Astra Trident.

Import an Existing Volume

If there are existing volumes on your NetApp storage system/platform that you want to mount on containers within your Kubernetes cluster, but that are not tied to PVCs in the cluster, then you must import these volumes. You can use the Trident volume import functionality to import these volumes.

The example commands that follow show the importing of a volume named `pb_fg_all`. For more information about PVCs, see the official Kubernetes documentation. For more information about the volume import functionality, see the Trident documentation.

An `accessModes` value of `ReadOnlyMany` is specified in the example PVC spec files. For more information about the `accessMode` field, see the official Kubernetes documentation.

```
$ cat << EOF > ./pvc-import-pb_fg_all-iface1.yaml
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pb-fg-all-iface1
  namespace: default
spec:
  accessModes:
    - ReadOnlyMany
  storageClassName: ontap-ai-flexgroups-retain-iface1
EOF
$ tridentctl import volume ontap-ai-flexgroups-iface1 pb_fg_all -f ./pvc-
import-pb_fg_all-iface1.yaml -n trident
+--------------------------------+--------
+------------------------------------+----------
+------------------------------------------+--------+---------+
|          NAME                  |  SIZE  |       STORAGE CLASS
| PROTOCOL |             BACKEND UUID                         | STATE  |
MANAGED |
+--------------------------------+--------
+------------------------------------+----------
+------------------------------------------+--------+---------+
| default-pb-fg-all-iface1-7d9f1 | 10 TiB | ontap-ai-flexgroups-retain-
iface1 | file     | b74cbddb-e0b8-40b7-b263-b6da6dec0bdd | online | true
|
+--------------------------------+--------
+------------------------------------+----------
+------------------------------------------+--------+---------+
$ tridentctl get volume -n trident
+----------------------------------+---------
+------------------------------------+----------
+--------------------------------------+--------+---------+
|                NAME              |  SIZE   |           STORAGE CLASS
| PROTOCOL |             BACKEND UUID             | STATE  | MANAGED |
+----------------------------------+---------
+------------------------------------+----------
+--------------------------------------+--------+---------+
| default-pb-fg-all-iface1-7d9f1   | 10 TiB  | ontap-ai-flexgroups-retain-
iface1 | file     | b74cbddb-e0b8-40b7-b263-b6da6dec0bdd | online | true
```

```
|
+---------------------------------+---------
+----------------------------------+----------
+-------------------------------------+--------+---------+
$ kubectl get pvc
NAME                    STATUS   VOLUME                                   CAPACITY
ACCESS MODES   STORAGECLASS                        AGE
pb-fg-all-iface1      Bound    default-pb-fg-all-iface1-7d9f1
10995116277760   ROX            ontap-ai-flexgroups-retain-iface1   25h
```

Provision a New Volume

You can use Trident to provision a new volume on your NetApp storage system or platform.

Provision a New Volume Using kubectl

The following example commands show the provisioning of a new FlexVol volume using kubectl.

An `accessModes` value of `ReadWriteMany` is specified in the following example PVC definition file. For more information about the `accessMode` field, see the official Kubernetes documentation.

```
$ cat << EOF > ./pvc-tensorflow-results.yaml
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: tensorflow-results
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 1Gi
  storageClassName: ontap-ai-flexvols-retain
EOF
$ kubectl create -f ./pvc-tensorflow-results.yaml
persistentvolumeclaim/tensorflow-results created
$ kubectl get pvc
NAME                              STATUS    VOLUME
CAPACITY         ACCESS MODES   STORAGECLASS                        AGE
pb-fg-all-iface1                  Bound     default-pb-fg-all-iface1-7d9f1
10995116277760   ROX            ontap-ai-flexgroups-retain-iface1   26h
tensorflow-results                Bound     default-tensorflow-results-
2fd60   1073741824      RWX            ontap-ai-flexvols-retain
25h
```

Provision a New Volume Using the NetApp DataOps Toolkit

You can also use the NetApp DataOps Toolkit for Kubernetes to provision a new volume on your NetApp storage system or platform. The NetApp DataOps Toolkit for Kubernetes utilizes Trident to provision volumes but simplifies the process for the user. Refer to the documentation for details.

Example High-performance Jobs for AIPod Deployments

Execute a Single-Node AI Workload

To execute a single-node AI and ML job in your Kubernetes cluster, perform the following tasks from the deployment jump host. With Trident, you can quickly and easily make a data volume, potentially containing petabytes of data, accessible to a Kubernetes workload. To make such a data volume accessible from within a Kubernetes pod, simply specify a PVC in the pod definition.

> This section assumes that you have already containerized (in the Docker container format) the specific AI and ML workload that you are attempting to execute in your Kubernetes cluster.

1. The following example commands show the creation of a Kubernetes job for a TensorFlow benchmark workload that uses the ImageNet dataset. For more information about the ImageNet dataset, see the ImageNet website.

   This example job requests eight GPUs and therefore can run on a single GPU worker node that features eight or more GPUs. This example job could be submitted in a cluster for which a worker node featuring eight or more GPUs is not present or is currently occupied with another workload. If so, then the job remains in a pending state until such a worker node becomes available.

   Additionally, in order to maximize storage bandwidth, the volume that contains the needed training data is mounted twice within the pod that this job creates. Another volume is also mounted in the pod. This second volume will be used to store results and metrics. These volumes are referenced in the job definition by using the names of the PVCs. For more information about Kubernetes jobs, see the official Kubernetes documentation.

   An `emptyDir` volume with a `medium` value of `Memory` is mounted to `/dev/shm` in the pod that this example job creates. The default size of the `/dev/shm` virtual volume that is automatically created by the Docker container runtime can sometimes be insufficient for TensorFlow's needs. Mounting an `emptyDir` volume as in the following example provides a sufficiently large `/dev/shm` virtual volume. For more information about `emptyDir` volumes, see the official Kubernetes documentation.

   The single container that is specified in this example job definition is given a `securityContext > privileged` value of `true`. This value means that the container effectively has root access on the host. This annotation is used in this case because the specific workload that is being executed requires root access. Specifically, a clear cache operation that the workload performs requires root access. Whether or not this `privileged: true` annotation is necessary depends on the requirements of the specific workload that you are executing.

   ```
   $ cat << EOF > ./netapp-tensorflow-single-imagenet.yaml
   apiVersion: batch/v1
   kind: Job
   metadata:
   ```

```
  name: netapp-tensorflow-single-imagenet
spec:
  backoffLimit: 5
  template:
    spec:
      volumes:
      - name: dshm
        emptyDir:
          medium: Memory
      - name: testdata-iface1
        persistentVolumeClaim:
          claimName: pb-fg-all-iface1
      - name: testdata-iface2
        persistentVolumeClaim:
          claimName: pb-fg-all-iface2
      - name: results
        persistentVolumeClaim:
          claimName: tensorflow-results
      containers:
      - name: netapp-tensorflow-py2
        image: netapp/tensorflow-py2:19.03.0
        command: ["python", "/netapp/scripts/run.py", "--
dataset_dir=/mnt/mount_0/dataset/imagenet", "--dgx_version=dgx1", "--
num_devices=8"]
        resources:
          limits:
            nvidia.com/gpu: 8
        volumeMounts:
        - mountPath: /dev/shm
          name: dshm
        - mountPath: /mnt/mount_0
          name: testdata-iface1
        - mountPath: /mnt/mount_1
          name: testdata-iface2
        - mountPath: /tmp
          name: results
        securityContext:
          privileged: true
      restartPolicy: Never
EOF
$ kubectl create -f ./netapp-tensorflow-single-imagenet.yaml
job.batch/netapp-tensorflow-single-imagenet created
$ kubectl get jobs
NAME                                          COMPLETIONS   DURATION   AGE
netapp-tensorflow-single-imagenet             0/1           24s        24s
```

2. Confirm that the job that you created in step 1 is running correctly. The following example command confirms that a single pod was created for the job, as specified in the job definition, and that this pod is currently running on one of the GPU worker nodes.

```
$ kubectl get pods -o wide
NAME                                                   READY   STATUS
RESTARTS   AGE
IP               NODE            NOMINATED NODE
netapp-tensorflow-single-imagenet-m7x92          1/1     Running     0
3m    10.233.68.61    10.61.218.154   <none>
```

3. Confirm that the job that you created in step 1 completes successfully. The following example commands confirm that the job completed successfully.

```
$ kubectl get jobs
NAME                                                 COMPLETIONS   DURATION
AGE
netapp-tensorflow-single-imagenet                    1/1           5m42s
10m
$ kubectl get pods
NAME                                                       READY   STATUS
RESTARTS   AGE
netapp-tensorflow-single-imagenet-m7x92                    0/1     Completed
0          11m
$ kubectl logs netapp-tensorflow-single-imagenet-m7x92
[netapp-tensorflow-single-imagenet-m7x92:00008] PMIX ERROR: NO-
PERMISSIONS in file gds_dstore.c at line 702
[netapp-tensorflow-single-imagenet-m7x92:00008] PMIX ERROR: NO-
PERMISSIONS in file gds_dstore.c at line 711
Total images/sec = 6530.59125
================ Clean Cache !!! ==================
mpirun -allow-run-as-root -np 1 -H localhost:1 bash -c 'sync; echo 1 >
/proc/sys/vm/drop_caches'
=========================================
mpirun -allow-run-as-root -np 8 -H localhost:8 -bind-to none -map-by
slot -x NCCL_DEBUG=INFO -x LD_LIBRARY_PATH -x PATH python
/netapp/tensorflow/benchmarks_190205/scripts/tf_cnn_benchmarks/tf_cnn_be
nchmarks.py --model=resnet50 --batch_size=256 --device=gpu
--force_gpu_compatible=True --num_intra_threads=1 --num_inter_threads=48
--variable_update=horovod --batch_group_size=20 --num_batches=500
--nodistortions --num_gpus=1 --data_format=NCHW --use_fp16=True
--use_tf_layers=False --data_name=imagenet --use_datasets=True
--data_dir=/mnt/mount_0/dataset/imagenet
--datasets_parallel_interleave_cycle_length=10
--datasets_sloppy_parallel_interleave=False --num_mounts=2
--mount_prefix=/mnt/mount_%d --datasets_prefetch_buffer_size=2000
--datasets_use_prefetch=True --datasets_num_private_threads=4
--horovod_device=gpu >
/tmp/20190814_105450_tensorflow_horovod_rdma_resnet50_gpu_8_256_b500_ima
genet_nodistort_fp16_r10_m2_nockpt.txt 2>&1
```

4. **Optional:** Clean up job artifacts. The following example commands show the deletion of the job object that was created in step 1.

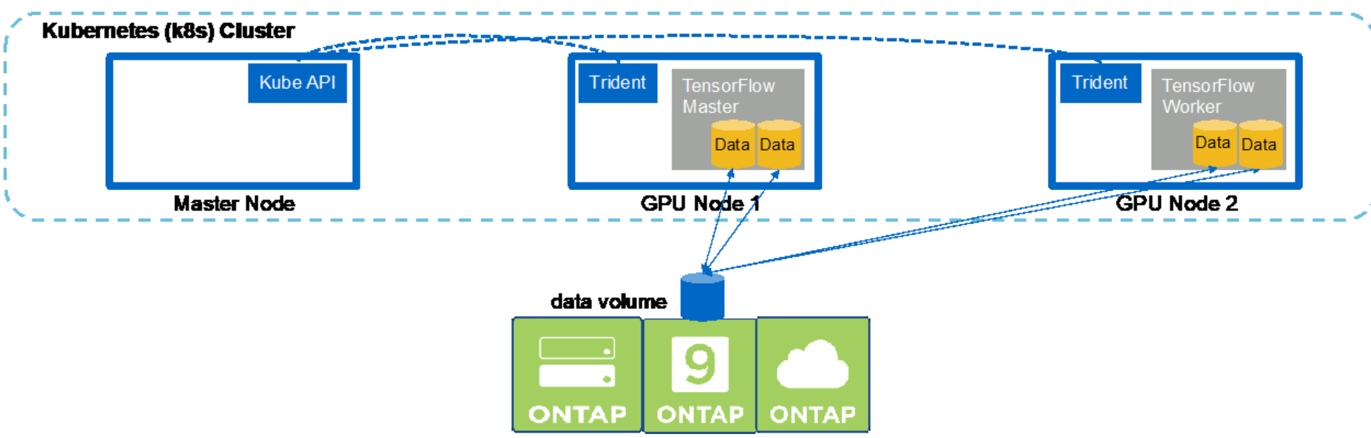
   When you delete the job object, Kubernetes automatically deletes any associated pods.

```
$ kubectl get jobs
NAME                                                COMPLETIONS   DURATION
AGE
netapp-tensorflow-single-imagenet                   1/1           5m42s
10m
$ kubectl get pods
NAME                                                      READY   STATUS
RESTARTS   AGE
netapp-tensorflow-single-imagenet-m7x92                   0/1     Completed
0          11m
$ kubectl delete job netapp-tensorflow-single-imagenet
job.batch "netapp-tensorflow-single-imagenet" deleted
$ kubectl get jobs
No resources found.
$ kubectl get pods
No resources found.
```

Execute a Synchronous Distributed AI Workload

To execute a synchronous multinode AI and ML job in your Kubernetes cluster, perform the following tasks on the deployment jump host. This process enables you to take advantage of data that is stored on a NetApp volume and to use more GPUs than a single worker node can provide. See the following figure for a depiction of a synchronous distributed AI job.

> Synchronous distributed jobs can help increase performance and training accuracy compared with asynchronous distributed jobs. A discussion of the pros and cons of synchronous jobs versus asynchronous jobs is outside the scope of this document.



1. The following example commands show the creation of one worker that participates in the synchronous distributed execution of the same TensorFlow benchmark job that was executed on a single node in the example in the section Execute a Single-Node AI Workload. In this specific example, only a single worker is deployed because the job is executed across two worker nodes.

This example worker deployment requests eight GPUs and thus can run on a single GPU worker node that features eight or more GPUs. If your GPU worker nodes feature more than eight GPUs, to maximize performance, you might want to increase this number to be equal to the number of GPUs that your worker nodes feature. For more information about Kubernetes deployments, see the official Kubernetes documentation.

A Kubernetes deployment is created in this example because this specific containerized worker would never complete on its own. Therefore, it doesn't make sense to deploy it by using the Kubernetes job construct. If your worker is designed or written to complete on its own, then it might make sense to use the job construct to deploy your worker.

The pod that is specified in this example deployment specification is given a `hostNetwork` value of `true`. This value means that the pod uses the host worker node's networking stack instead of the virtual networking stack that Kubernetes usually creates for each pod. This annotation is used in this case because the specific workload relies on Open MPI, NCCL, and Horovod to execute the workload in a synchronous distributed manner. Therefore, it requires access to the host networking stack. A discussion about Open MPI, NCCL, and Horovod is outside the scope of this document. Whether or not this `hostNetwork: true` annotation is necessary depends on the requirements of the specific workload that you are executing. For more information about the `hostNetwork` field, see the official Kubernetes documentation.

```
$ cat << EOF > ./netapp-tensorflow-multi-imagenet-worker.yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: netapp-tensorflow-multi-imagenet-worker
spec:
  replicas: 1
  selector:
    matchLabels:
      app: netapp-tensorflow-multi-imagenet-worker
  template:
    metadata:
      labels:
        app: netapp-tensorflow-multi-imagenet-worker
    spec:
      hostNetwork: true
      volumes:
      - name: dshm
        emptyDir:
          medium: Memory
      - name: testdata-iface1
        persistentVolumeClaim:
          claimName: pb-fg-all-iface1
      - name: testdata-iface2
        persistentVolumeClaim:
          claimName: pb-fg-all-iface2
      - name: results
        persistentVolumeClaim:
```

```
          claimName: tensorflow-results
      containers:
      - name: netapp-tensorflow-py2
        image: netapp/tensorflow-py2:19.03.0
        command: ["bash", "/netapp/scripts/start-slave-multi.sh",
"22122"]
        resources:
          limits:
            nvidia.com/gpu: 8
        volumeMounts:
        - mountPath: /dev/shm
          name: dshm
        - mountPath: /mnt/mount_0
          name: testdata-iface1
        - mountPath: /mnt/mount_1
          name: testdata-iface2
        - mountPath: /tmp
          name: results
        securityContext:
          privileged: true
EOF
$ kubectl create -f ./netapp-tensorflow-multi-imagenet-worker.yaml
deployment.apps/netapp-tensorflow-multi-imagenet-worker created
$ kubectl get deployments
NAME                                      DESIRED   CURRENT   UP-TO-DATE
AVAILABLE   AGE
netapp-tensorflow-multi-imagenet-worker   1         1         1
1           4s
```

2. Confirm that the worker deployment that you created in step 1 launched successfully. The following example commands confirm that a single worker pod was created for the deployment, as indicated in the deployment definition, and that this pod is currently running on one of the GPU worker nodes.

```
$ kubectl get pods -o wide
NAME                                                         READY
STATUS    RESTARTS   AGE
IP               NODE             NOMINATED NODE
netapp-tensorflow-multi-imagenet-worker-654fc7f486-v6725   1/1
Running   0          60s   10.61.218.154   10.61.218.154   <none>
$ kubectl logs netapp-tensorflow-multi-imagenet-worker-654fc7f486-v6725
22122
```

3. Create a Kubernetes job for a master that kicks off, participates in, and tracks the execution of the synchronous multinode job. The following example commands create one master that kicks off, participates in, and tracks the synchronous distributed execution of the same TensorFlow benchmark job that was executed on a single node in the example in the section Execute a Single-Node AI Workload.

This example master job requests eight GPUs and thus can run on a single GPU worker node that features eight or more GPUs. If your GPU worker nodes feature more than eight GPUs, to maximize performance, you might want to increase this number to be equal to the number of GPUs that your worker nodes feature.

The master pod that is specified in this example job definition is given a `hostNetwork` value of `true`, just as the worker pod was given a `hostNetwork` value of `true` in step 1. See step 1 for details about why this value is necessary.

```
$ cat << EOF > ./netapp-tensorflow-multi-imagenet-master.yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: netapp-tensorflow-multi-imagenet-master
spec:
  backoffLimit: 5
  template:
    spec:
      hostNetwork: true
      volumes:
      - name: dshm
        emptyDir:
          medium: Memory
      - name: testdata-iface1
        persistentVolumeClaim:
          claimName: pb-fg-all-iface1
      - name: testdata-iface2
        persistentVolumeClaim:
          claimName: pb-fg-all-iface2
      - name: results
        persistentVolumeClaim:
          claimName: tensorflow-results
      containers:
      - name: netapp-tensorflow-py2
        image: netapp/tensorflow-py2:19.03.0
        command: ["python", "/netapp/scripts/run.py", "--
dataset_dir=/mnt/mount_0/dataset/imagenet", "--port=22122", "--
num_devices=16", "--dgx_version=dgx1", "--
nodes=10.61.218.152,10.61.218.154"]
        resources:
          limits:
            nvidia.com/gpu: 8
        volumeMounts:
        - mountPath: /dev/shm
          name: dshm
        - mountPath: /mnt/mount_0
          name: testdata-iface1
        - mountPath: /mnt/mount_1
```

```
        name: testdata-iface2
      - mountPath: /tmp
        name: results
      securityContext:
        privileged: true
    restartPolicy: Never
EOF
$ kubectl create -f ./netapp-tensorflow-multi-imagenet-master.yaml
job.batch/netapp-tensorflow-multi-imagenet-master created
$ kubectl get jobs
NAME                                      COMPLETIONS   DURATION   AGE
netapp-tensorflow-multi-imagenet-master   0/1           25s        25s
```

4. Confirm that the master job that you created in step 3 is running correctly. The following example command confirms that a single master pod was created for the job, as indicated in the job definition, and that this pod is currently running on one of the GPU worker nodes. You should also see that the worker pod that you originally saw in step 1 is still running and that the master and worker pods are running on different nodes.

```
$ kubectl get pods -o wide
NAME                                                          READY
STATUS    RESTARTS   AGE
IP              NODE            NOMINATED NODE
netapp-tensorflow-multi-imagenet-master-ppwwj                 1/1
Running   0          45s   10.61.218.152   10.61.218.152   <none>
netapp-tensorflow-multi-imagenet-worker-654fc7f486-v6725   1/1
Running   0          26m   10.61.218.154   10.61.218.154   <none>
```

5. Confirm that the master job that you created in step 3 completes successfully. The following example commands confirm that the job completed successfully.

```
$ kubectl get jobs
NAME                                      COMPLETIONS   DURATION   AGE
netapp-tensorflow-multi-imagenet-master   1/1           5m50s      9m18s
$ kubectl get pods
NAME                                                          READY
STATUS      RESTARTS   AGE
netapp-tensorflow-multi-imagenet-master-ppwwj                 0/1
Completed   0          9m38s
netapp-tensorflow-multi-imagenet-worker-654fc7f486-v6725   1/1
Running     0          35m
$ kubectl logs netapp-tensorflow-multi-imagenet-master-ppwwj
[10.61.218.152:00008] WARNING: local probe returned unhandled
shell:unknown assuming bash
rm: cannot remove '/lib': Is a directory
[10.61.218.154:00033] PMIX ERROR: NO-PERMISSIONS in file gds_dstore.c at
```

```
line 702
[10.61.218.154:00033] PMIX ERROR: NO-PERMISSIONS in file gds_dstore.c at
line 711
[10.61.218.152:00008] PMIX ERROR: NO-PERMISSIONS in file gds_dstore.c at
line 702
[10.61.218.152:00008] PMIX ERROR: NO-PERMISSIONS in file gds_dstore.c at
line 711
Total images/sec = 12881.33875
================ Clean Cache !!! ==================
mpirun -allow-run-as-root -np 2 -H 10.61.218.152:1,10.61.218.154:1 -mca
pml ob1 -mca btl ^openib -mca btl_tcp_if_include enp1s0f0 -mca
plm_rsh_agent ssh -mca plm_rsh_args "-p 22122" bash -c 'sync; echo 1 >
/proc/sys/vm/drop_caches'
=========================================
mpirun -allow-run-as-root -np 16 -H 10.61.218.152:8,10.61.218.154:8
-bind-to none -map-by slot -x NCCL_DEBUG=INFO -x LD_LIBRARY_PATH -x PATH
-mca pml ob1 -mca btl ^openib -mca btl_tcp_if_include enp1s0f0 -x
NCCL_IB_HCA=mlx5 -x NCCL_NET_GDR_READ=1 -x NCCL_IB_SL=3 -x
NCCL_IB_GID_INDEX=3 -x
NCCL_SOCKET_IFNAME=enp5s0.3091,enp12s0.3092,enp132s0.3093,enp139s0.3094
-x NCCL_IB_CUDA_SUPPORT=1 -mca orte_base_help_aggregate 0 -mca
plm_rsh_agent ssh -mca plm_rsh_args "-p 22122" python
/netapp/tensorflow/benchmarks_190205/scripts/tf_cnn_benchmarks/tf_cnn_be
nchmarks.py --model=resnet50 --batch_size=256 --device=gpu
--force_gpu_compatible=True --num_intra_threads=1 --num_inter_threads=48
--variable_update=horovod --batch_group_size=20 --num_batches=500
--nodistortions --num_gpus=1 --data_format=NCHW --use_fp16=True
--use_tf_layers=False --data_name=imagenet --use_datasets=True
--data_dir=/mnt/mount_0/dataset/imagenet
--datasets_parallel_interleave_cycle_length=10
--datasets_sloppy_parallel_interleave=False --num_mounts=2
--mount_prefix=/mnt/mount_%d --datasets_prefetch_buffer_size=2000 --
datasets_use_prefetch=True --datasets_num_private_threads=4
--horovod_device=gpu >
/tmp/20190814_161609_tensorflow_horovod_rdma_resnet50_gpu_16_256_b500_im
agenet_nodistort_fp16_r10_m2_nockpt.txt 2>&1
```

6. Delete the worker deployment when you no longer need it. The following example commands show the deletion of the worker deployment object that was created in step 1.

   When you delete the worker deployment object, Kubernetes automatically deletes any associated worker pods.

```
$ kubectl get deployments
NAME                                      DESIRED   CURRENT   UP-TO-DATE
AVAILABLE   AGE
netapp-tensorflow-multi-imagenet-worker   1         1         1
1           43m
$ kubectl get pods
NAME                                                         READY
STATUS      RESTARTS   AGE
netapp-tensorflow-multi-imagenet-master-ppwwj                0/1
Completed   0          17m
netapp-tensorflow-multi-imagenet-worker-654fc7f486-v6725   1/1
Running     0          43m
$ kubectl delete deployment netapp-tensorflow-multi-imagenet-worker
deployment.extensions "netapp-tensorflow-multi-imagenet-worker" deleted
$ kubectl get deployments
No resources found.
$ kubectl get pods
NAME                                            READY   STATUS
RESTARTS   AGE
netapp-tensorflow-multi-imagenet-master-ppwwj   0/1     Completed   0
18m
```

7. **Optional:** Clean up the master job artifacts. The following example commands show the deletion of the master job object that was created in step 3.

   When you delete the master job object, Kubernetes automatically deletes any associated master pods.

```
$ kubectl get jobs
NAME                                      COMPLETIONS   DURATION   AGE
netapp-tensorflow-multi-imagenet-master   1/1           5m50s      19m
$ kubectl get pods
NAME                                            READY   STATUS
RESTARTS   AGE
netapp-tensorflow-multi-imagenet-master-ppwwj   0/1     Completed   0
19m
$ kubectl delete job netapp-tensorflow-multi-imagenet-master
job.batch "netapp-tensorflow-multi-imagenet-master" deleted
$ kubectl get jobs
No resources found.
$ kubectl get pods
No resources found.
```

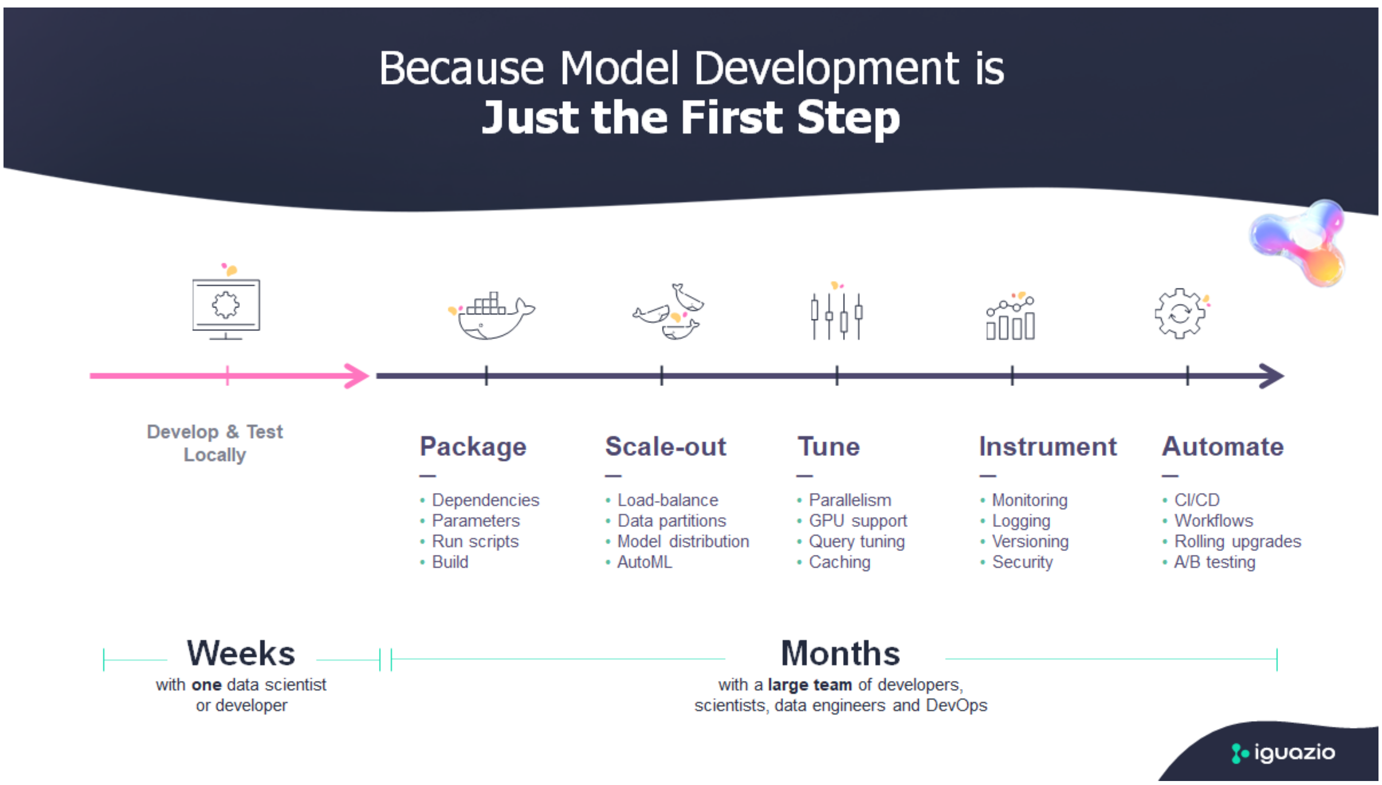
MLRun Pipeline with Iguazio

TR-4834: NetApp and Iguazio for MLRun Pipeline

Rick Huang, David Arnette, NetApp
Marcelo Litovsky, Iguazio

This document covers the details of the MLRun pipeline using NetApp ONTAP AI, NetApp AI Control Plane, NetApp Cloud Volumes software, and the Iguazio Data Science Platform. We used Nuclio serverless function, Kubernetes Persistent Volumes, NetApp Cloud Volumes, NetApp Snapshot copies, Grafana dashboard, and other services on the Iguazio platform to build an end-to-end data pipeline for the simulation of network failure detection. We integrated Iguazio and NetApp technologies to enable fast model deployment, data replication, and production monitoring capabilities on premises as well as in the cloud.

The work of a data scientist should be focused on the training and tuning of machine learning (ML) and artificial intelligence (AI) models. However, according to research by Google, data scientists spend ~80% of their time figuring out how to make their models work with enterprise applications and run at scale, as shown in the following image depicting model development in the AI/ML workflow.



To manage end-to-end AI/ML projects, a wider understanding of enterprise components is needed. Although DevOps have taken over the definition, integration, and deployment these types of components, machine learning operations target a similar flow that includes AI/ML projects. To get an idea of what an end-to-end AI/ML pipeline touches in the enterprise, see the following list of required components:

- Storage
- Networking

- Databases
- File systems
- Containers
- Continuous integration and continuous deployment (CI/CD) pipeline
- Development integrated development environment (IDE)
- Security
- Data access policies
- Hardware
- Cloud
- Virtualization
- Data science toolsets and libraries

In this paper, we demonstrate how the partnership between NetApp and Iguazio drastically simplifies the development of an end-to-end AI/ML pipeline. This simplification accelerates the time to market for all of your AI/ML applications.

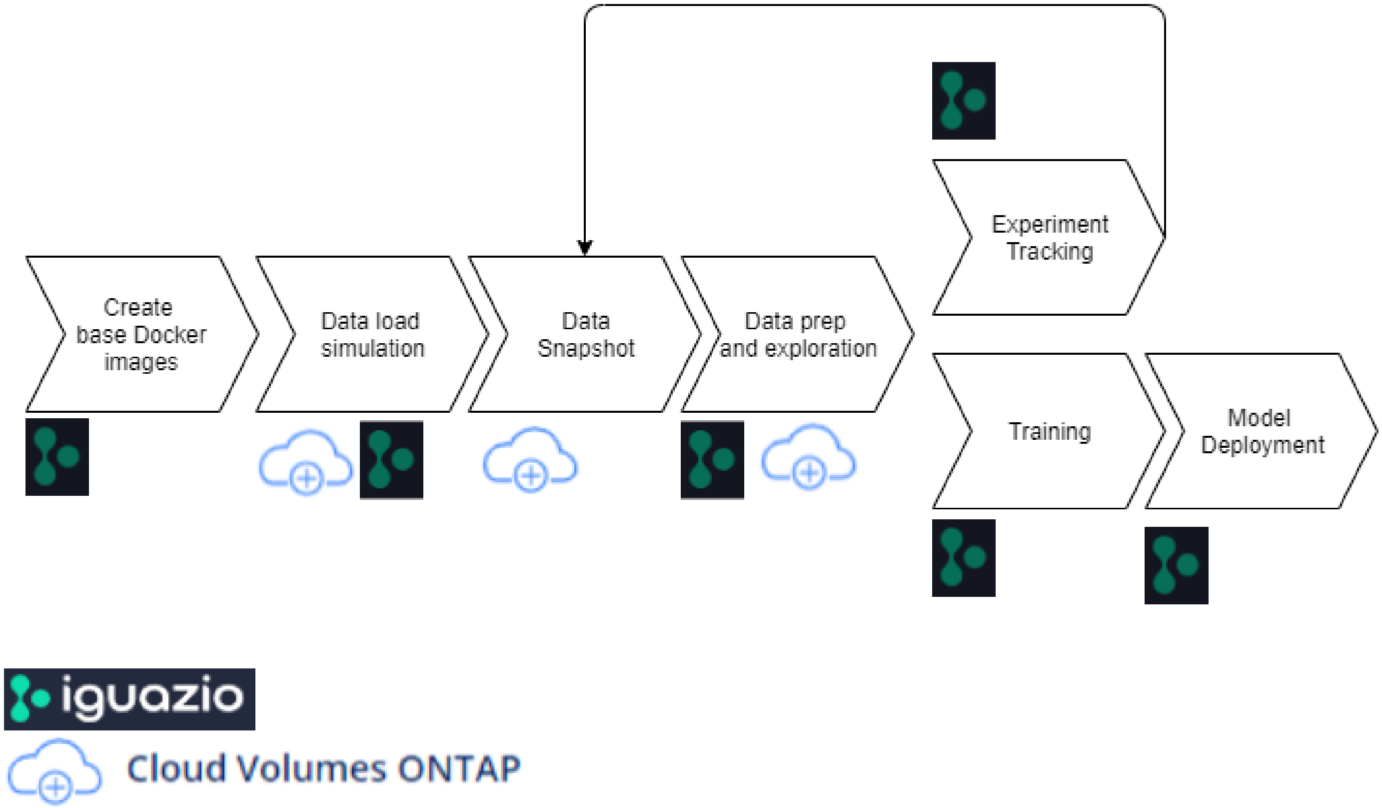
Target Audience

The world of data science touches multiple disciplines in information technology and business.

- The data scientist needs the flexibility to use their tools and libraries of choice.
- The data engineer needs to know how the data flows and where it resides.
- A DevOps engineer needs the tools to integrate new AI/ML applications into their CI/CD pipelines.
- Business users want to have access to AI/ML applications. We describe how NetApp and Iguazio help each of these roles bring value to business with our platforms.

Solution Overview

This solution follows the lifecycle of an AI/ML application. We start with the work of data scientists to define the different steps needed to prep data and train and deploy models. We follow with the work needed to create a full pipeline with the ability to track artifacts, experiment with execution, and deploy to Kubeflow. To complete the full cycle, we integrate the pipeline with NetApp Cloud Volumes to enable data versioning, as seen in the following image.



Technology Overview

This article provides an overview of the colution for MLRun pipeline using NetApp ONTAP AI, NetApp AI Control Plane, NetApp Cloud Volumes software, and the Iguazio Data Science Platform.

NetApp Overview

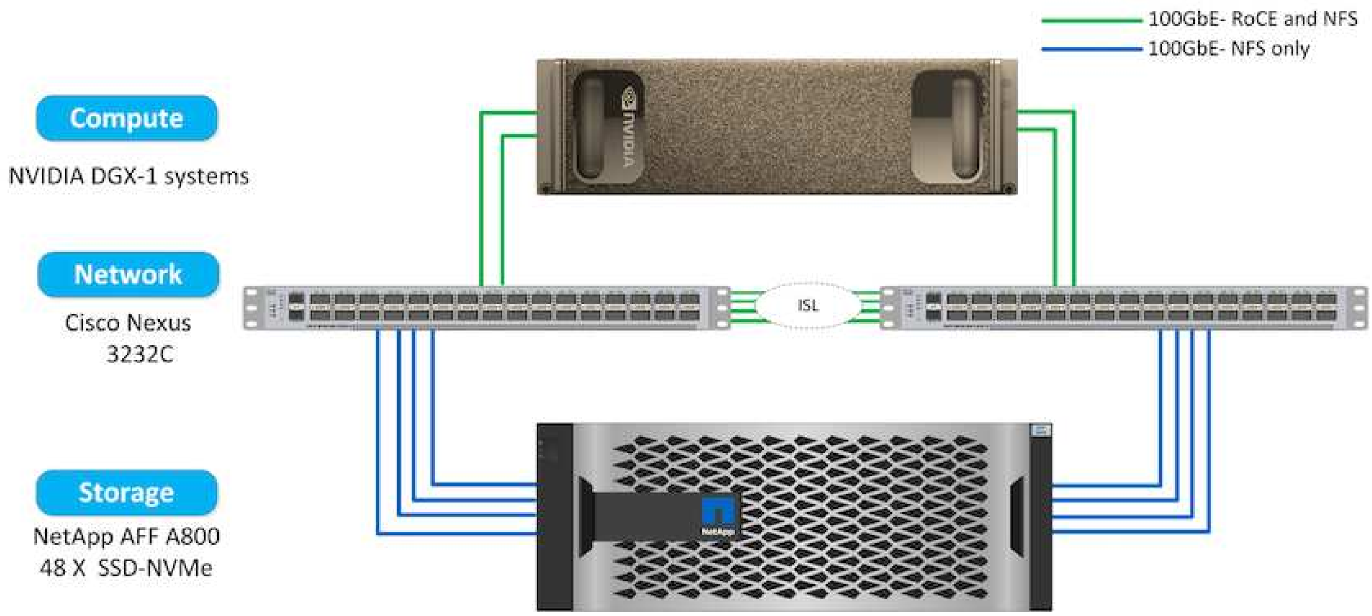
NetApp is the data authority for the hybrid cloud. NetApp provides a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, NetApp empowers global organizations to unleash the full potential of their data to expand customer touch points, foster greater innovation, and optimize their operations.

NetApp ONTAP AI

NetApp ONTAP AI, powered by NVIDIA DGX systems and NetApp cloud-connected all-flash storage, streamlines the flow of data reliably and speeds up analytics, training, and inference with your data fabric that spans from edge to core to cloud. It gives IT organizations an architecture that provides the following benefits:

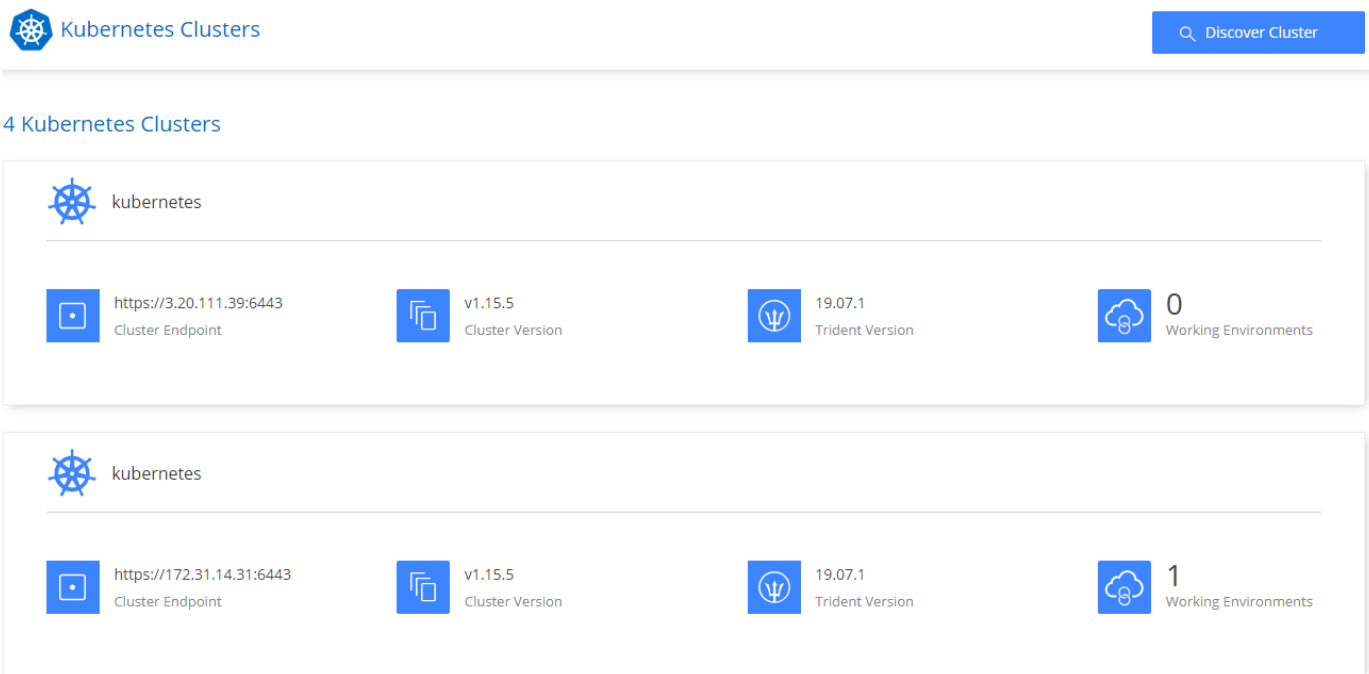
- Eliminates design complexities
- Allows independent scaling of compute and storage
- Enables customers to start small and scale seamlessly
- Offers a range of storage options for various performance and cost pointsNetApp ONTAP AI offers converged infrastructure stacks incorporating NVIDIA DGX-1, a petaflop-scale AI system, and NVIDIA Mellanox high-performance Ethernet switches to unify AI workloads, simplify deployment, and accelerate ROI. We leveraged ONTAP AI with one DGX-1 and NetApp AFF A800 storage system for this technical report. The following image shows the topology of ONTAP AI with the DGX-1 system used in this

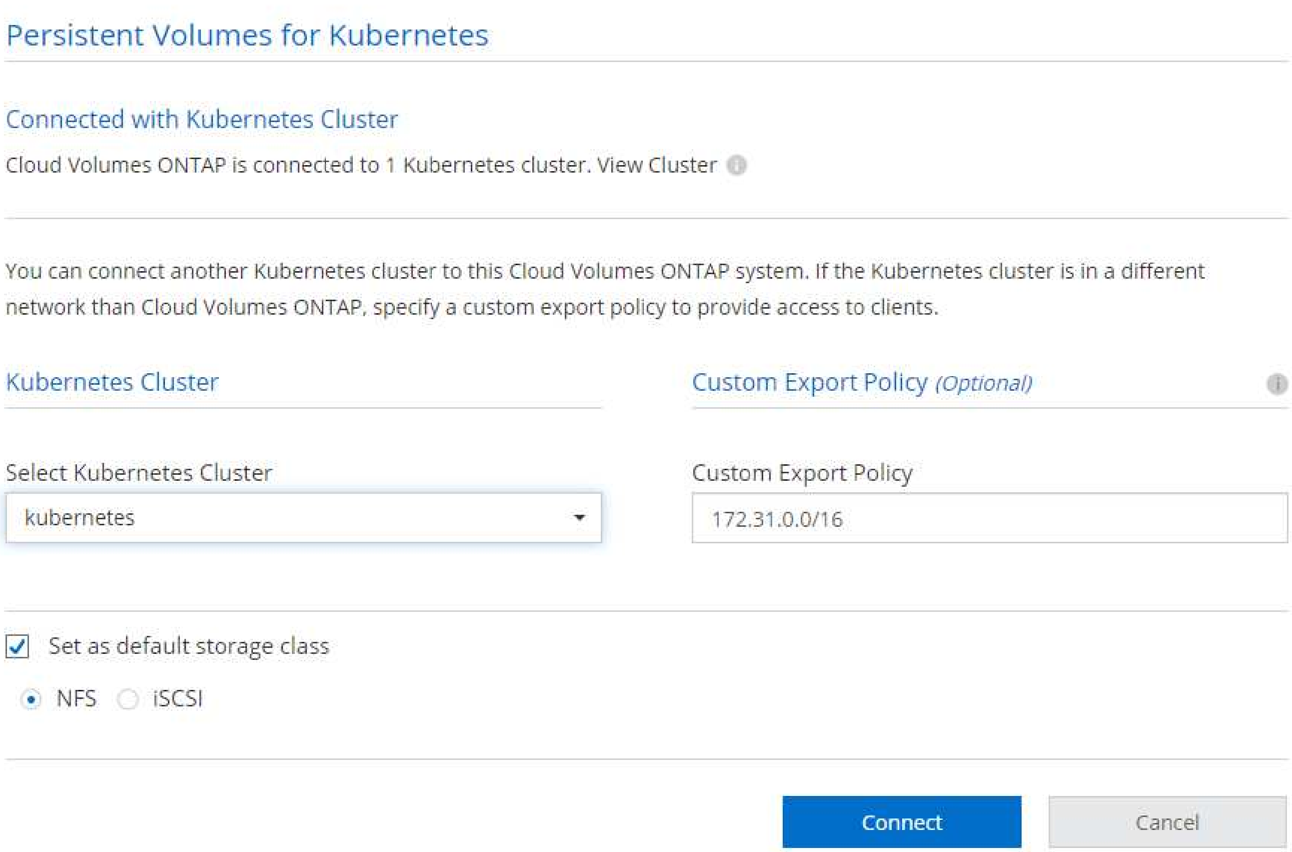
validation.

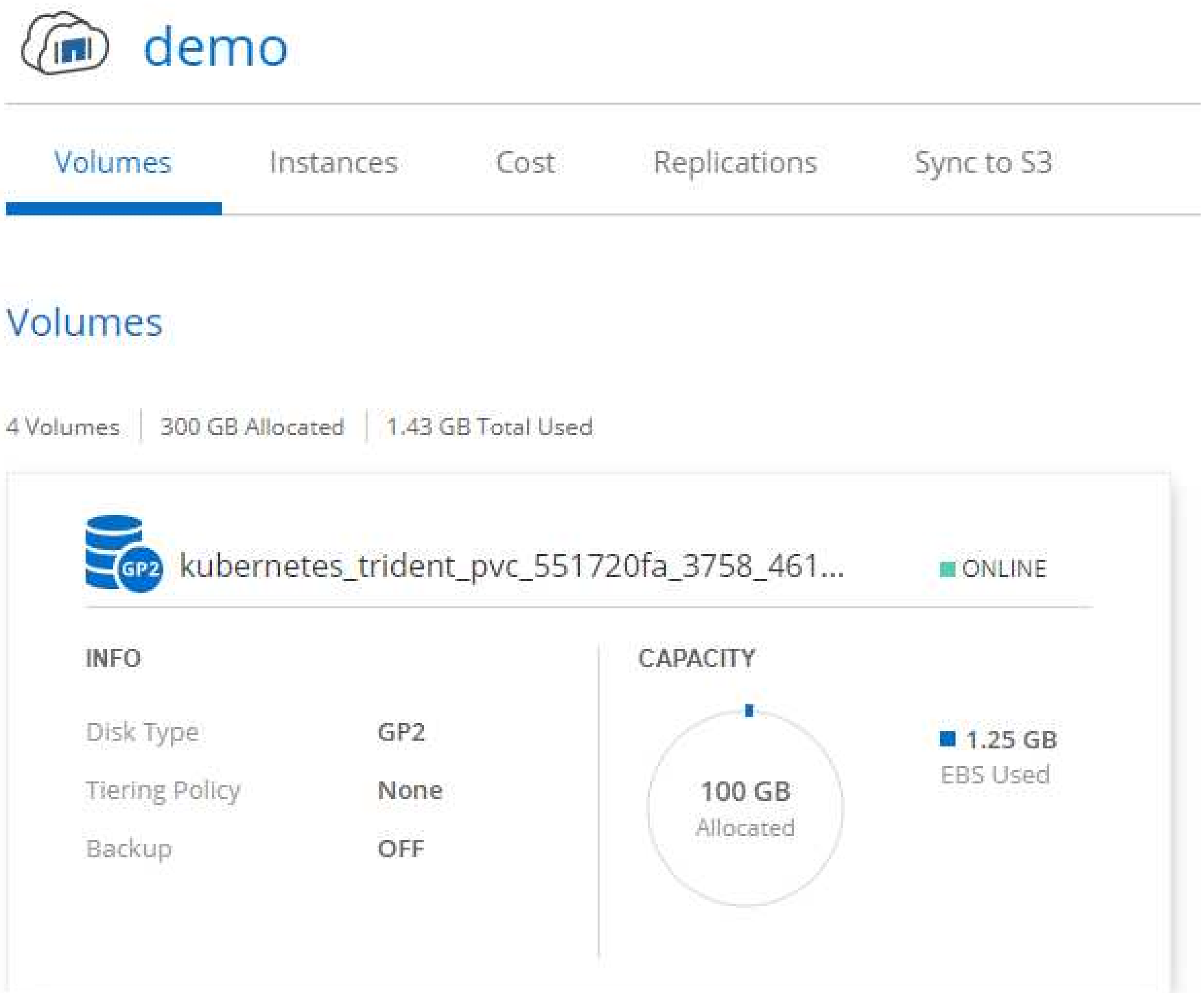


NetApp AI Control Plane

The NetApp AI Control Plane enables you to unleash AI and ML with a solution that offers extreme scalability, streamlined deployment, and nonstop data availability. The AI Control Plane solution integrates Kubernetes and Kubeflow with a data fabric enabled by NetApp. Kubernetes, the industry-standard container orchestration platform for cloud-native deployments, enables workload scalability and portability. Kubeflow is an open-source machine-learning platform that simplifies management and deployment, enabling developers to do more data science in less time. A data fabric enabled by NetApp offers uncompromising data availability and portability to make sure that your data is accessible across the pipeline, from edge to core to cloud. This technical report uses the NetApp AI Control Plane in an MLRun pipeline. The following image shows Kubernetes cluster management page where you can have different endpoints for each cluster. We connected NFS Persistent Volumes to the Kubernetes cluster, and the following images show an Persistent Volume connected to the cluster, where NetApp Trident offers persistent storage support and data management capabilities.


Kubernetes Clusters
Discover Cluster
4 Kubernetes Clusters
kubernetes
https://3.20.111.39:6443
Cluster Endpoint
v1.15.5
Cluster Version
19.07.1
Trident Version
0
Working Environments
kubernetes
https://172.31.14.31:6443
Cluster Endpoint
v1.15.5
Cluster Version
19.07.1
Trident Version
1
Working Environments

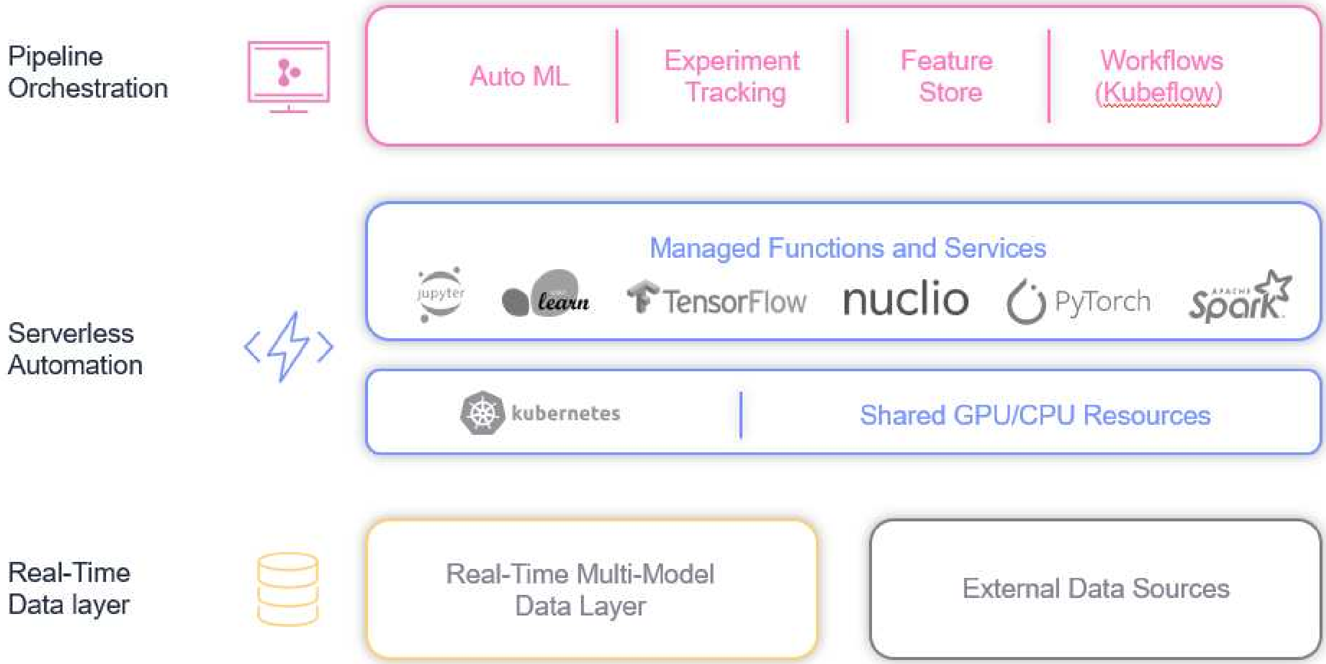

Persistent Volumes for Kubernetes
Connected with Kubernetes Cluster
Cloud Volumes ONTAP is connected to 1 Kubernetes cluster. View Cluster
You can connect another Kubernetes cluster to this Cloud Volumes ONTAP system. If the Kubernetes cluster is in a different network than Cloud Volumes ONTAP, specify a custom export policy to provide access to clients.
Kubernetes Cluster
Custom Export Policy (Optional)
Select Kubernetes Cluster
kubernetes
Custom Export Policy
172.31.0.0/16
Set as default storage class
NFS
iSCSI
Connect
Cancel



Iguazio Overview

The Iguazio Data Science Platform is a fully integrated and secure data- science platform as a service (PaaS) that simplifies development, accelerates performance, facilitates collaboration, and addresses operational challenges. This platform incorporates the following components, and the Iguazio Data Science Platform is presented in the following image:

- A data-science workbench that includes Jupyter Notebooks, integrated analytics engines, and Python packages
- Model management with experiments tracking and automated pipeline capabilities
- Managed data and ML services over a scalable Kubernetes cluster
- Nuclio, a real-time serverless functions framework
- An extremely fast and secure data layer that supports SQL, NoSQL, time-series databases, files (simple objects), and streaming
- Integration with third-party data sources such as NetApp, Amazon S3, HDFS, SQL databases, and streaming or messaging protocols
- Real-time dashboards based on Grafana



Software and Hardware Requirements

This article defines the hardware requirements that must be met in order to deploy this solution.

Network Configuration

The following is the network configuration requirement for setting up in the cloud:

- The Iguazio cluster and NetApp Cloud Volumes must be in the same virtual private cloud.
- The cloud manager must have access to port 6443 on the Iguazio app nodes.
- We used Amazon Web Services in this technical report. However, users have the option of deploying the solution in any Cloud provider.For on-premises testing in ONTAP AI with NVIDIA DGX-1, we used the Iguazio hosted DNS service for convenience.

Clients must be able to access dynamically created DNS domains. Customers can use their own DNS if desired.

Hardware Requirements

You can install Iguazio on-premises in your own cluster. We have verified the solution in NetApp ONTAP AI with an NVIDIA DGX-1 system. The following table lists the hardware used to test this solution.

Hardware	Quantity
DGX-1 systems	1
NetApp AFF A800 system	1 high-availability (HA) pair, includes 2 controllers and 48 NVMe SSDs (3.8TB or above)
Cisco Nexus 3232C network switches	2

The following table lists the software components required for on-premise testing:

Software	Version or Other Information
NetApp ONTAP data management software	9.7
Cisco NX-OS switch firmware	7.0(3)I6(1)
NVIDIA DGX OS	4.4 - Ubuntu 18.04 LTS
Docker container platform	19.03.5
Container version	20.01-tf1-py2
Machine learning framework	TensorFlow 1.15.0
Iguazio	Version 2.8+
ESX Server	6.5

This solution was fully tested with Iguazio version 2.5 and NetApp Cloud Volumes ONTAP for AWS. The Iguazio cluster and NetApp software are both running on AWS.

Software	Version or Type
Iguazio	Version 2.8+
App node	M5.4xlarge
Data node	I3.4xlarge

Network Device Failure Prediction Use Case Summary

This use case is based on an Iguazio customer in the telecommunications space in Asia. With 100K enterprise customers and 125k network outage events per year, there was a critical need to predict and take proactive action to prevent network failures from affecting customers. This solution provided them with the following benefits:

- Predictive analytics for network failures
- Integration with a ticketing system
- Taking proactive action to prevent network failuresAs a result of this implementation of Iguazio, 60% of failures were proactively prevented.

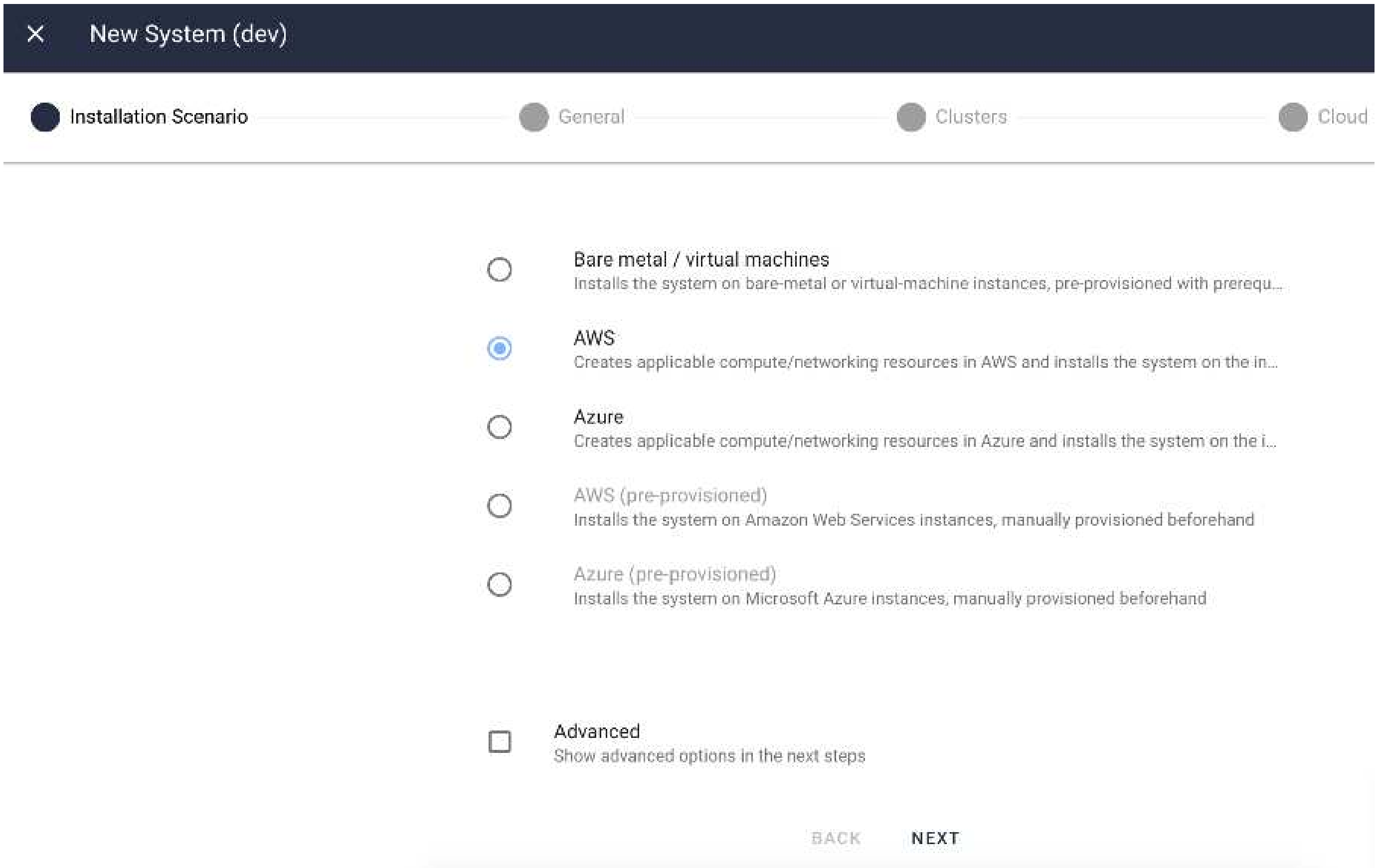
Setup Overview

Iguazio can be installed on-premises or on a cloud provider.

Iguazio Installation

Provisioning can be done as a service and managed by Iguazio or by the customer. In both cases, Iguazio provides a deployment application (Provazio) to deploy and manage clusters.

For on-premises installation, please refer to NVA-1121 for compute, network, and storage setup. On-premises deployment of Iguazio is provided by Iguazio without additional cost to the customer. See this page for DNS and SMTP server configurations. The Provazio installation page is shown as follows.



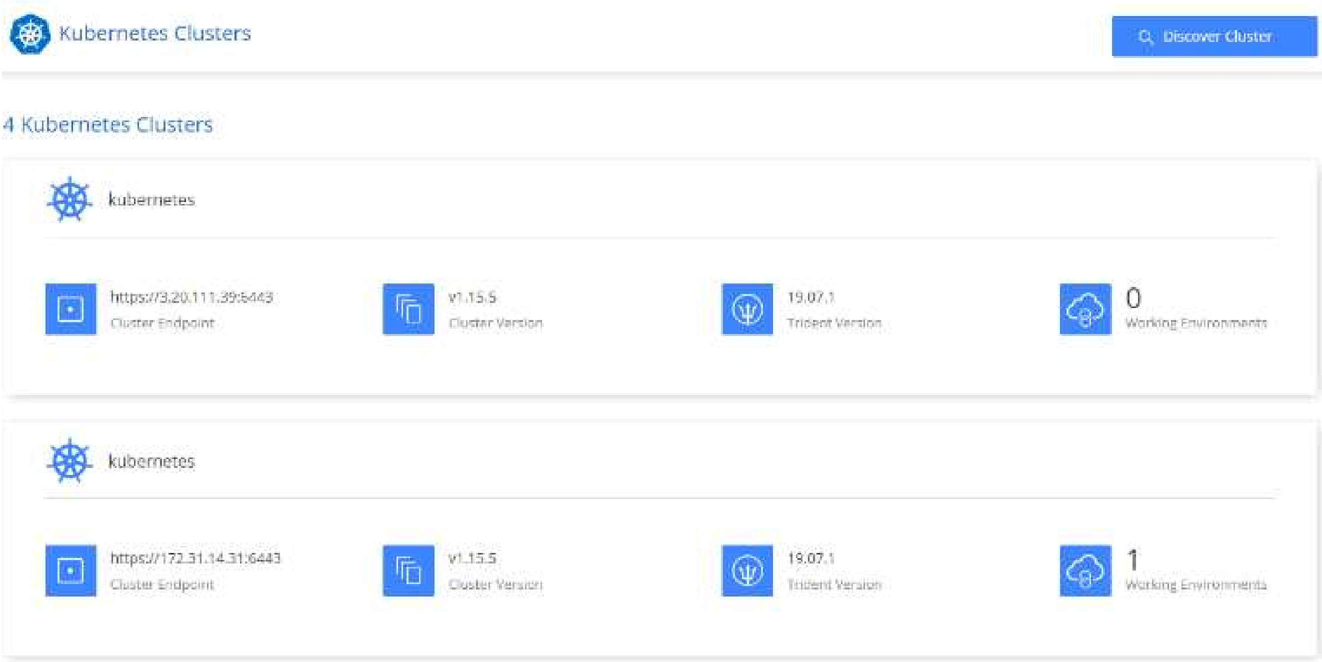
Configuring Kubernetes Cluster

This section is divided into two parts for cloud and on-premises deployment respectively.

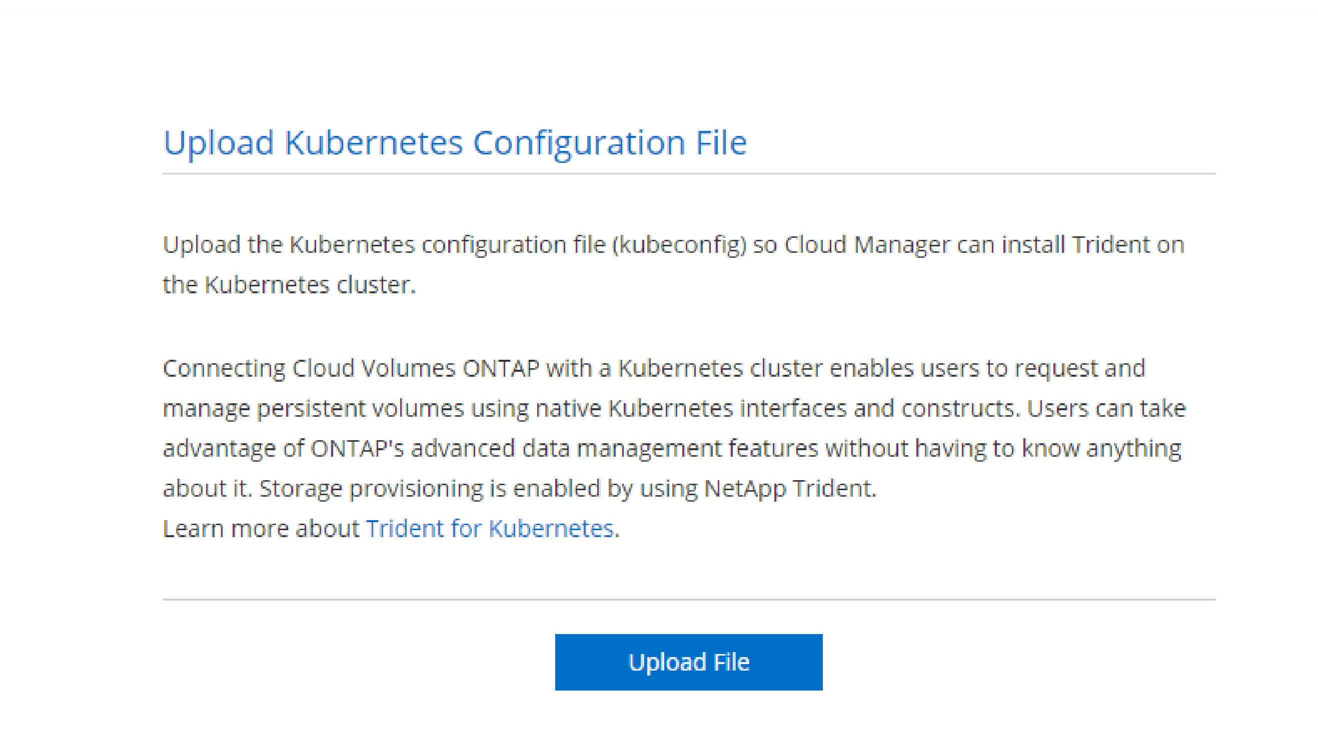
Cloud Deployment Kubernetes Configuration

Through NetApp Cloud Manager, you can define the connection to the Iguazio Kubernetes cluster. Trident requires access to multiple resources in the cluster to make the volume available.

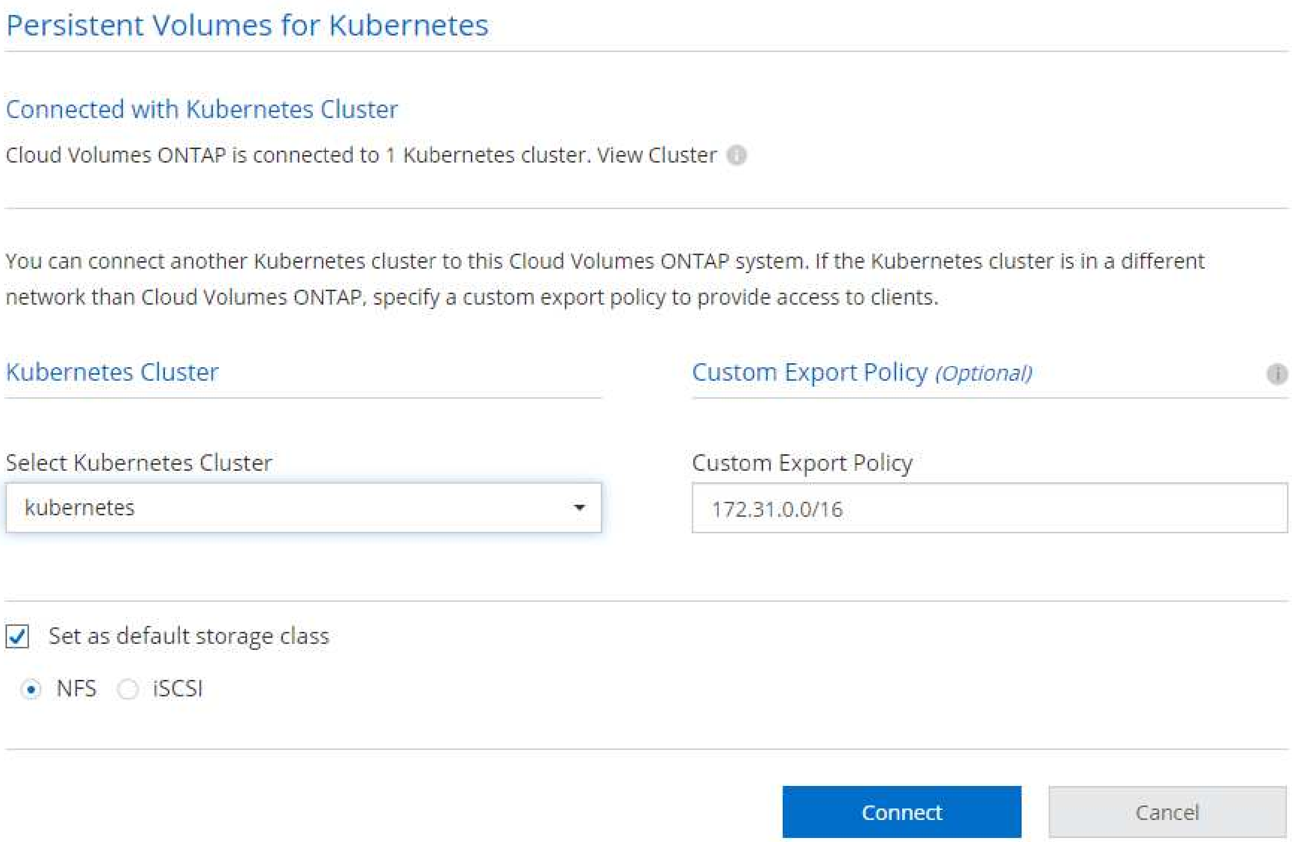
1. To enable access, obtain the Kubernetes config file from one the Iguazio nodes. The file is located under `/home/Iguazio/.kube/config.` Download this file to your desktop.
2. Go to Discover Cluster to configure.



3. Upload the Kubernetes config file. See the following image.



4. Deploy Trident and associate a volume with the cluster. See the following image on defining and assigning a Persistent Volume to the Iguazio cluster.This process creates a Persistent Volume (PV) in Iguazio’s Kubernetes cluster. Before you can use it, you must define a Persistent Volume Claim (PVC).



On-Premises Deployment Kubernetes Configuration

For on-premises installation of NetApp Trident, see TR-4798 for details. After configuring your Kubernetes cluster and installing NetApp Trident, you can connect Trident to the Iguazio cluster to enable NetApp data management capabilities, such as taking Snapshot copies of your data and model.

Define Persistent Volume Claim

This article demonstrates how to define a persistent volume claim on a Jupyter notebook.

1. Save the following YAML to a file to create a PVC of type Basic.

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: basic
spec:
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 100Gi
  storageClassName: netapp-file
```

2. Apply the YAML file to your Iguazio Kubernetes cluster.

```
Kubectl -n default-tenant apply -f <your yaml file>
```

Attach NetApp Volume to the Jupyter Notebook

Iguazio offers several managed services to provide data scientists with a full end-to-end stack for development and deployment of AI/ML applications. You can read more about these components at the Iguazio Overview of Application Services and Tools.

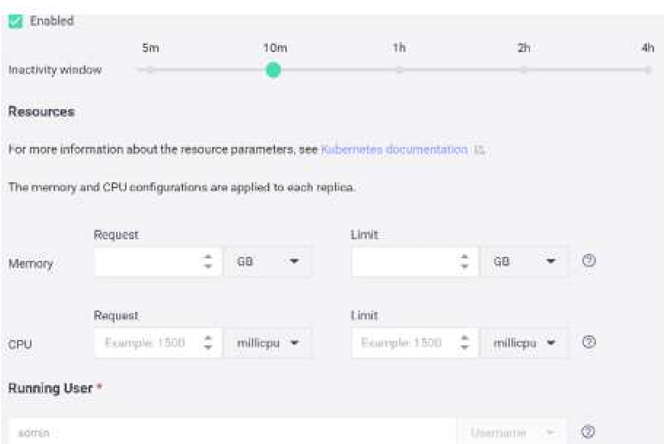
One of the managed services is Jupyter Notebook. Each developer gets its own deployment of a notebook container with the resources they need for development. To give them access to the NetApp Cloud Volume, you can assign the volume to their container and resource allocation, running user, and environment variable settings for Persistent Volume Claims is presented in the following image.

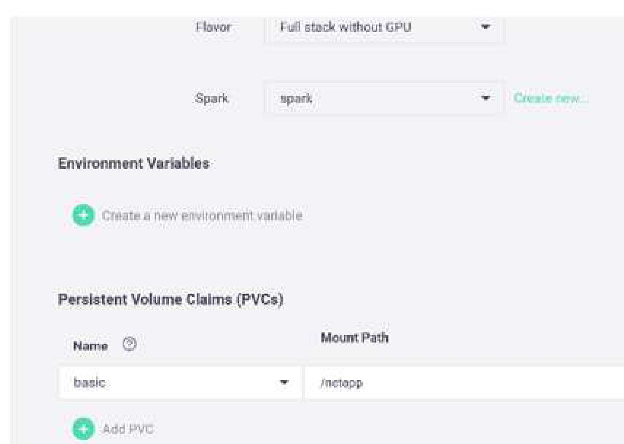
For an on-premises configuration, you can refer to TR-4798 on the Trident setup to enable NetApp ONTAP data management capabilities, such as taking Snapshot copies of your data or model for versioning control. Add the following line in your Trident back- end config file to make Snapshot directories visible:

```
{
    …
    "defaults": {
        "snapshotDir": "true"
    }
}
```

You must create a Trident back- end config file in JSON format, and then run the following Trident command to reference it:

```
tridentctl create backend -f <backend-file>
```





Deploying the Application

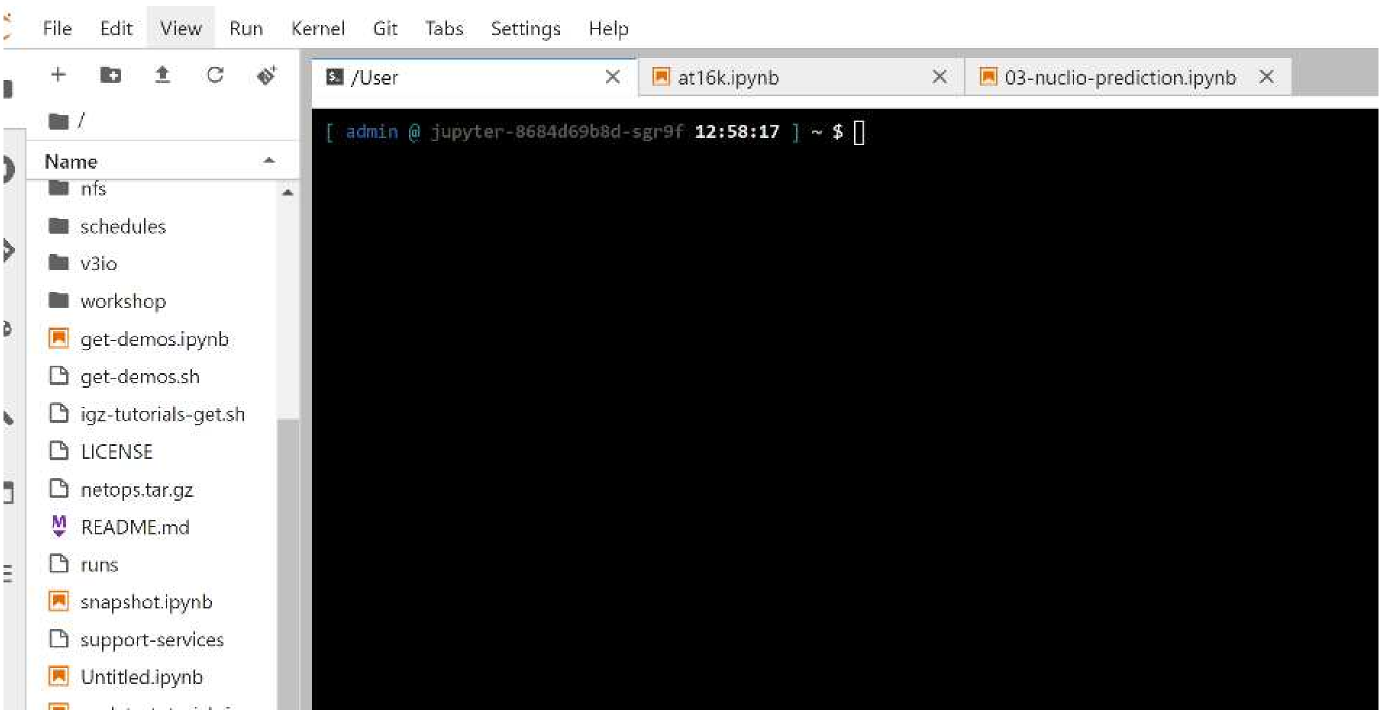
The following sections describe how to install and deploy the application.

Get Code from GitHub

Now that the NetApp Cloud Volume or NetApp Trident volume is available to the Iguazio cluster and the developer environment, you can start reviewing the application.

Users have their own workspace (directory). On every notebook, the path to the user directory is `/User`. The Iguazio platform manages the directory. If you follow the instructions above, the NetApp Cloud volume is available in the `/netapp` directory.

Get the code from GitHub using a Jupyter terminal.



At the Jupyter terminal prompt, clone the project.

```
cd /User
git clone .
```

You should now see the `netops- netapp` folder on the file tree in Jupyter workspace.

Configure Working Environment

Copy the `Notebook set_env-Example.ipynb` as `set_env.ipynb`. Open and edit `set_env.ipynb`. This notebook sets variables for credentials, file locations, and execution drivers.

If you follow the instructions above, the following steps are the only changes to make:

1. Obtain this value from the Iguazio services dashboard: `docker_registry`

   Example: `docker-registry.default-tenant.app.clusterq.iguaziodev.com:80`

2. Change `admin` to your Iguazio username:

```
IGZ_CONTAINER_PATH = '/users/admin'
```

The following are the ONTAP system connection details. Include the volume name that was generated when Trident was installed. The following setting is for an on-premises ONTAP cluster:

```
ontapClusterMgmtHostname = '0.0.0.0'
ontapClusterAdminUsername = 'USER'
ontapClusterAdminPassword = 'PASSWORD'
sourceVolumeName = 'SOURCE VOLUME'
```

The following setting is for Cloud Volumes ONTAP:

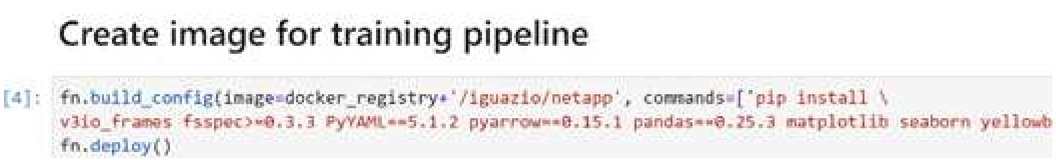
```
MANAGER=ontapClusterMgmtHostname
svm='svm'
email='email'
password=ontapClusterAdminPassword
weid="weid"
volume=sourceVolumeName
```

Create Base Docker Images

Everything you need to build an ML pipeline is included in the Iguazio platform. The developer can define the specifications of the Docker images required to run the pipeline and execute the image creation from Jupyter Notebook. Open the notebook `create- images.ipynb` and Run All Cells.

This notebook creates two images that we use in the pipeline.

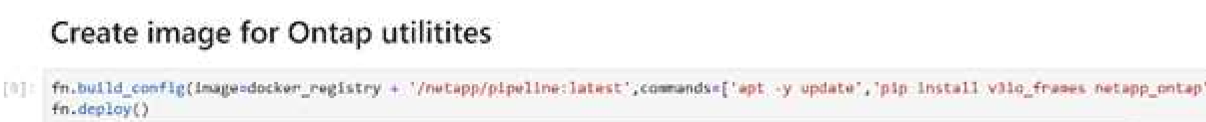
- `iguazio/netapp`. Used to handle ML tasks.



Create image for training pipeline

```
[4]: fn.build_config(image=docker_registry+'/iguazio/netapp', commands=['pip install \
v3io_frames fsspec>=0.3.3 PyYAML==5.1.2 pyarrow==0.15.1 pandas==0.25.3 matplotlib seaborn yellowb
fn.deploy()
```

- `netapp/pipeline`. Contains utilities to handle NetApp Snapshot copies.



Create image for Ontap utilitites

```
fn.build_config(image=docker_registry + '/netapp/pipeline:latest',commands=['apt -y update','pip install v3io_frames netapp_ontap'
fn.deploy()
```

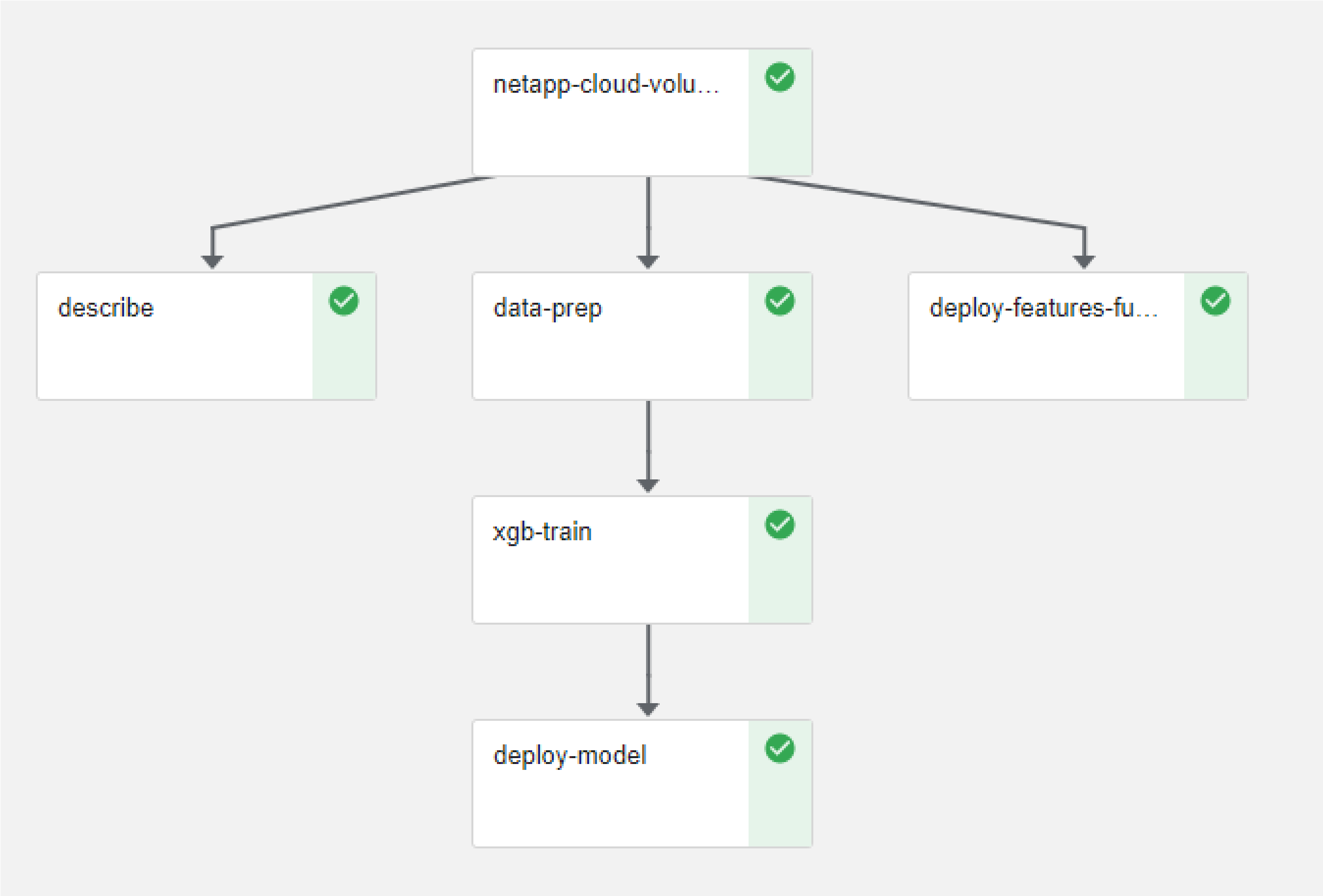
Review Individual Jupyter Notebooks

The following table lists the libraries and frameworks we used to build this task. All these components have been fully integrated with Iguazio's role- based access and security controls.

Libraries/Framework	Description
MLRun	An managed by Iguazio to enable the assembly, execution, and monitoring of an ML/AI pipeline.
Nuclio	A serverless functions framework integrated with Iguazio. Also available as an open-source project managed by Iguazio.
Kubeflow	A Kubernetes-based framework to deploy the pipeline. This is also an open-source project to which Iguazio contributes. It is integrated with Iguazio for added security and integration with the rest of the infrastructure.
Docker	A Docker registry run as a service in the Iguazio platform. You can also change this to connect to your registry.
NetApp Cloud Volumes	Cloud Volumes running on AWS give us access to large amounts of data and the ability to take Snapshot copies to version the datasets used for training.
Trident	Trident is an open-source project managed by NetApp. It facilitates the integration with storage and compute resources in Kubernetes.

We used several notebooks to construct the ML pipeline. Each notebook can be tested individually before being brought together in the pipeline. We cover each notebook individually following the deployment flow of this demonstration application.

The desired result is a pipeline that trains a model based on a Snapshot copy of the data and deploys the model for inference. A block diagram of a completed MLRun pipeline is shown in the following image.



Deploy Data Generation Function

This section describes how we used Nuclio serverless functions to generate network device data. The use case is adapted from an Iguazio client that deployed the pipeline and used Iguazio services to monitor and predict network device failures.

We simulated data coming from network devices. Executing the Jupyter notebook `data- generator.ipynb` creates a serverless function that runs every 10 minutes and generates a Parquet file with new data. To deploy the function, run all the cells in this notebook. See the [Nuclio website](#) to review any unfamiliar components in this notebook.

A cell with the following comment is ignored when generating the function. Every cell in the notebook is assumed to be part of the function. Import the Nuclio module to enable `%nuclio magic`.

```
# nuclio: ignore
import nuclio
```

In the spec for the function, we defined the environment in which the function executes, how it is triggered, and the resources it consumes.

```
spec = nuclio.ConfigSpec(config={"spec.triggers.inference.kind":"cron",

"spec.triggers.inference.attributes.interval" :"10m",
                                "spec.readinessTimeoutSeconds" : 60,
                                "spec.minReplicas" : 1},……
```

The `init_context` function is invoked by the Nuclio framework upon initialization of the function.

```
def init_context(context):
    ….
```

Any code not in a function is invoked when the function initializes. When you invoke it, a handler function is executed. You can change the name of the handler and specify it in the function spec.

```
def handler(context, event):
            …
```

You can test the function from the notebook prior to deployment.

```
%%time
# nuclio: ignore
init_context(context)
event = nuclio.Event(body='')
output = handler(context, event)
output
```

The function can be deployed from the notebook or it can be deployed from a CI/CD pipeline (adapting this code).

```
addr = nuclio.deploy_file(name='generator',project='netops',spec=spec,
tag='v1.1')
```

Pipeline Notebooks

These notebooks are not meant to be executed individually for this setup. This is just a review of each notebook. We invoked them as part of the pipeline. To execute them individually, review the MLRun documentation to execute them as Kubernetes jobs.

snap_cv.ipynb

This notebook handles the Cloud Volume Snapshot copies at the beginning of the pipeline. It passes the name of the volume to the pipeline context. This notebook invokes a shell script to handle the Snapshot copy. While running in the pipeline, the execution context contains variables to help locate all files needed for execution.

While writing this code, the developer does not have to worry about the file location in the container that executes it. As described later, this application is deployed with all its dependencies, and it is the definition of the pipeline parameters that provides the execution context.

```
command = os.path.join(context.get_param('APP_DIR'),"snap_cv.sh")
```

The created Snapshot copy location is placed in the MLRun context to be consumed by steps in the pipeline.

```
context.log_result('snapVolumeDetails',snap_path)
```

The next three notebooks are run in parallel.

data-prep.ipynb

Raw metrics must be turned into features to enable model training. This notebook reads the raw metrics from the Snapshot directory and writes the features for model training to the NetApp volume.

When running in the context of the pipeline, the input `DATA_DIR` contains the Snapshot copy location.

```
metrics_table = os.path.join(str(mlruncontext.get_input('DATA_DIR',
os.getenv('DATA_DIR','/netpp'))),
                             mlruncontext.get_param('metrics_table',
os.getenv('metrics_table','netops_metrics_parquet')))
```

describe.ipynb

To visualize the incoming metrics, we deploy a pipeline step that provides plots and graphs that are available through the Kubeflow and MLRun UIs. Each execution has its own version of this visualization tool.

```
ax.set_title("features correlation")
plt.savefig(os.path.join(base_path, "plots/corr.png"))
context.log_artifact(PlotArtifact("correlation",  body=plt.gcf()),
local_path="plots/corr.html")
```

deploy-feature-function.ipynb

We continuously monitor the metrics looking for anomalies. This notebook creates a serverless function that generates the features need to run prediction on incoming metrics. This notebook invokes the creation of the function. The function code is in the notebook `data- prep.ipynb`. Notice that we use the same notebook as a step in the pipeline for this purpose.

training.ipynb

After we create the features, we trigger the model training. The output of this step is the model to be used for inferencing. We also collect statistics to keep track of each execution (experiment).

For example, the following command enters the accuracy score into the context for that experiment. This value is visible in Kubeflow and MLRun.

```
context.log_result('accuracy',score)
```

deploy-inference-function.ipynb

The last step in the pipeline is to deploy the model as a serverless function for continuous inferencing. This notebook invokes the creation of the serverless function defined in `nuclio-inference- function.ipynb`.

Review and Build Pipeline

The combination of running all the notebooks in a pipeline enables the continuous run of experiments to reassess the accuracy of the model against new metrics. First, open the `pipeline.ipynb` notebook. We take you through details that show how NetApp and Iguazio simplify the deployment of this ML pipeline.

We use MLRun to provide context and handle resource allocation to each step of the pipeline. The MLRun API service runs in the Iguazio platform and is the point of interaction with Kubernetes resources. Each developer cannot directly request resources; the API handles the requests and enables access controls.

```
# MLRun API connection definition
mlconf.dbpath = 'http://mlrun-api:8080'
```

The pipeline can work with NetApp Cloud Volumes and on-premises volumes. We built this demonstration to use Cloud Volumes, but you can see in the code the option to run on-premises.

```
# Initialize the NetApp snap fucntion once for all functions in a notebook
if [ NETAPP_CLOUD_VOLUME ]:
    snapfn =
code_to_function('snap',project='NetApp',kind='job',filename="snap_cv.ipyn
b").apply(mount_v3io())
    snap_params = {
    "metrics_table" : metrics_table,
    "NETAPP_MOUNT_PATH" : NETAPP_MOUNT_PATH,
    'MANAGER' : MANAGER,
    'svm' : svm,
    'email': email,
    'password': password ,
    'weid': weid,
    'volume': volume,
    "APP_DIR" : APP_DIR
       }
else:
    snapfn =
code_to_function('snap',project='NetApp',kind='job',filename="snapshot.ipy
nb").apply(mount_v3io())
….
snapfn.spec.image = docker_registry + '/netapp/pipeline:latest'
snapfn.spec.volume_mounts =
[snapfn.spec.volume_mounts[0],netapp_volume_mounts]
      snapfn.spec.volumes = [ snapfn.spec.volumes[0],netapp_volumes]
```

The first action needed to turn a Jupyter notebook into a Kubeflow step is to turn the code into a function. A function has all the specifications required to run that notebook. As you scroll down the notebook, you can see that we define a function for every step in the pipeline.

Part of the Notebook	**Description**
<code_to_function> (part of the MLRun module)	Name of the function: Project name. used to organize all project artifacts. This is visible in the MLRun UI. Kind. In this case, a Kubernetes job. This could be Dask, mpi, sparkk8s, and more. See the MLRun documentation for more details. File. The name of the notebook. This can also be a location in Git (HTTP).
image	The name of the Docker image we are using for this step. We created this earlier with the create-image.ipynb notebook.
volume_mounts & volumes	Details to mount the NetApp Cloud Volume at run time.

We also define parameters for the steps.

```
params={   "FEATURES_TABLE":FEATURES_TABLE,
           "SAVE_TO" : SAVE_TO,
           "metrics_table" : metrics_table,
           'FROM_TSDB': 0,
           'PREDICTIONS_TABLE': PREDICTIONS_TABLE,
           'TRAIN_ON_LAST': '1d',
           'TRAIN_SIZE':0.7,
           'NUMBER_OF_SHARDS' : 4,
           'MODEL_FILENAME' : 'netops.v3.model.pickle',
           'APP_DIR' : APP_DIR,
           'FUNCTION_NAME' : 'netops-inference',
           'PROJECT_NAME' : 'netops',
           'NETAPP_SIM' : NETAPP_SIM,
           'NETAPP_MOUNT_PATH': NETAPP_MOUNT_PATH,
           'NETAPP_PVC_CLAIM' : NETAPP_PVC_CLAIM,
           'IGZ_CONTAINER_PATH' : IGZ_CONTAINER_PATH,
           'IGZ_MOUNT_PATH' : IGZ_MOUNT_PATH
            }
```

After you have the function definition for all steps, you can construct the pipeline. We use the `kfp` module to make this definition. The difference between using MLRun and building on your own is the simplification and shortening of the coding.

The functions we defined are turned into step components using the `as_step` function of MLRun.

Snapshot Step Definition

Initiate a Snapshot function, output, and mount v3io as source:

```
snap = snapfn.as_step(NewTask(handler='handler',params=snap_params),
name='NetApp_Cloud_Volume_Snapshot',outputs=['snapVolumeDetails','training
_parquet_file']).apply(mount_v3io())
```

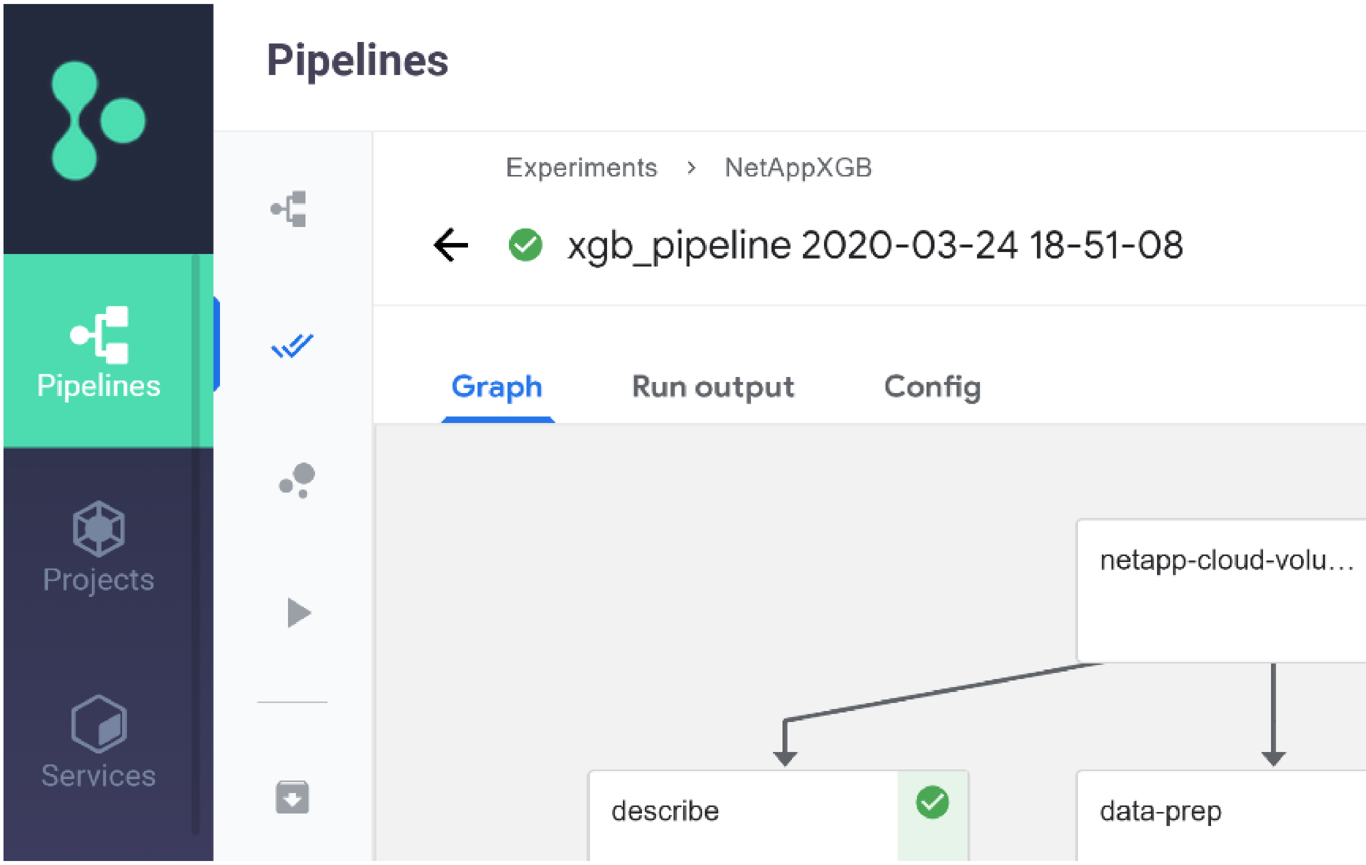
Parameters	Details
NewTask	NewTask is the definition of the function run.
(MLRun module)	Handler. Name of the Python function to invoke. We used the name handler in the notebook, but it is not required. params. The parameters we passed to the execution. Inside our code, we use context.get_param (‘PARAMETER’) to get the values.

Parameters	Details
as_step	Name. Name of the Kubeflow pipeline step. outputs. These are the values that the step adds to the dictionary on completion. Take a look at the snap_cv.ipynb notebook. mount_v3io(). This configures the step to mount /User for the user executing the pipeline.

```
prep = data_prep.as_step(name='data-prep',
handler='handler',params=params,
                          inputs = {'DATA_DIR':
snap.outputs['snapVolumeDetails']} ,

out_path=artifacts_path).apply(mount_v3io()).after(snap)
```

Parameters	Details
inputs	You can pass to a step the outputs of a previous step. In this case, snap.outputs['snapVolumeDetails'] is the name of the Snapshot copy we created on the snap step.
out_path	A location to place artifacts generating using the MLRun module log_artifacts.

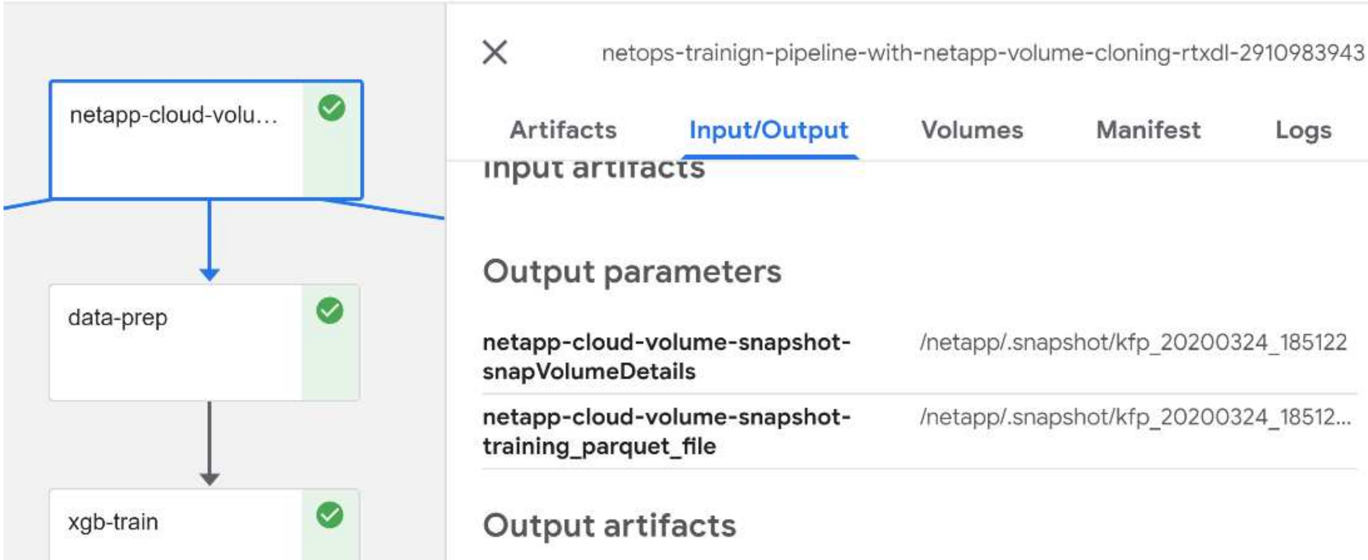
You can run `pipeline.ipynb` from top to bottom. You can then go to the Pipelines tab from the Iguazio dashboard to monitor progress as seen in the Iguazio dashboard Pipelines tab.



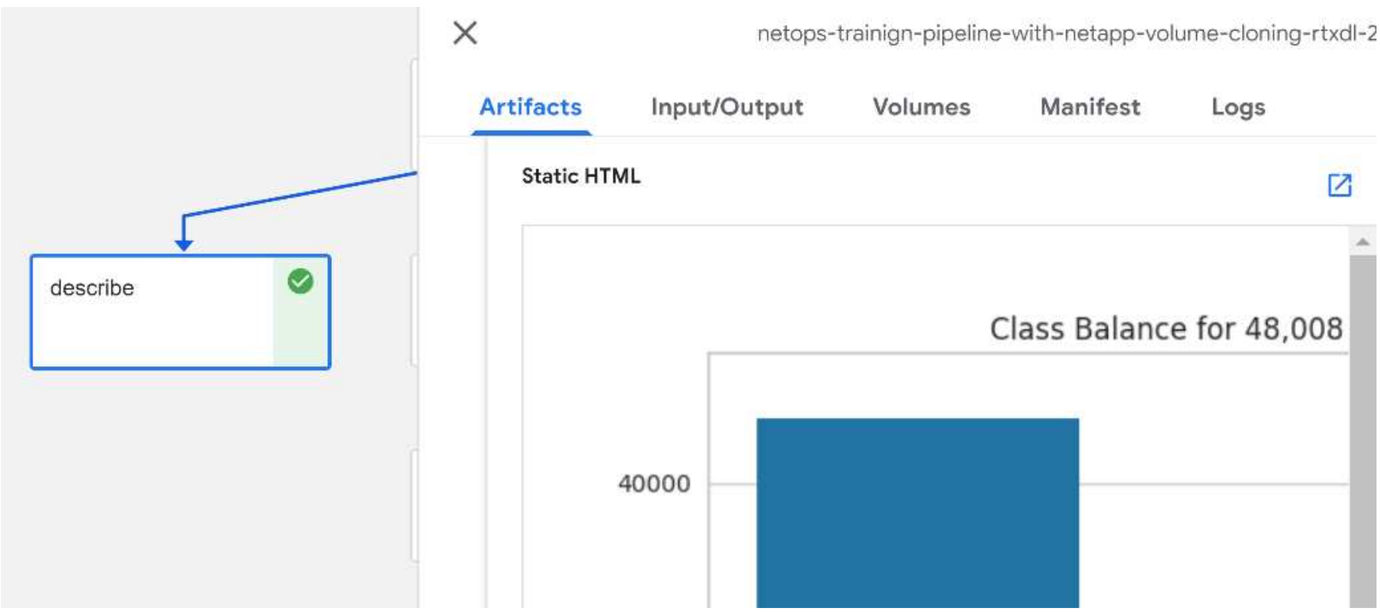
Because we logged the accuracy of training step in every run, we have a record of accuracy for each experiment, as seen in the record of training accuracy.

	Run name	Status	Duration	Pipeline Version	Recurring ...	Start time	accuracy
☐	xgb_pipeline 2020-03-24 18-51-...	✓	0:08:43	[View pipeline]	-	3/24/2020, 2:51:09 PM	0.985
☐	xgb_pipeline 2020-03-19 13-31-...	✓	0:08:14	[View pipeline]	-	3/19/2020, 9:31:19 AM	0.980
☐	xgb_pipeline 2020-03-18 12-56-...	✓	0:08:11	[View pipeline]	-	3/18/2020, 8:56:08 AM	0.990
☐	xgb_pipeline 2020-03-17 19-49-...	✓	0:08:03	[View pipeline]	-	3/17/2020, 3:49:31 PM	0.985
☐	xgb_pipeline 2020-03-17 18-34-...	✓	0:05:54	[View pipeline]	-	3/17/2020, 2:34:56 PM	0.980
☐	xgb_pipeline 2020-03-17 17-34-...	✓	0:04:48	[View pipeline]	-	3/17/2020, 1:34:16 PM	0.982
☐	xgb_pipeline 2020-03-17 17-01-...	✓	0:05:25	[View pipeline]	-	3/17/2020, 1:01:58 PM	0.987
☐	xgb_pipeline 2020-03-16 16-47-...	✓	0:06:08	[View pipeline]	-	3/16/2020, 12:47:19 ...	0.983
☐	xgb_pipeline 2020-03-16 13-57-...	✓	0:05:18	[View pipeline]	-	3/16/2020, 9:57:03 AM	0.980

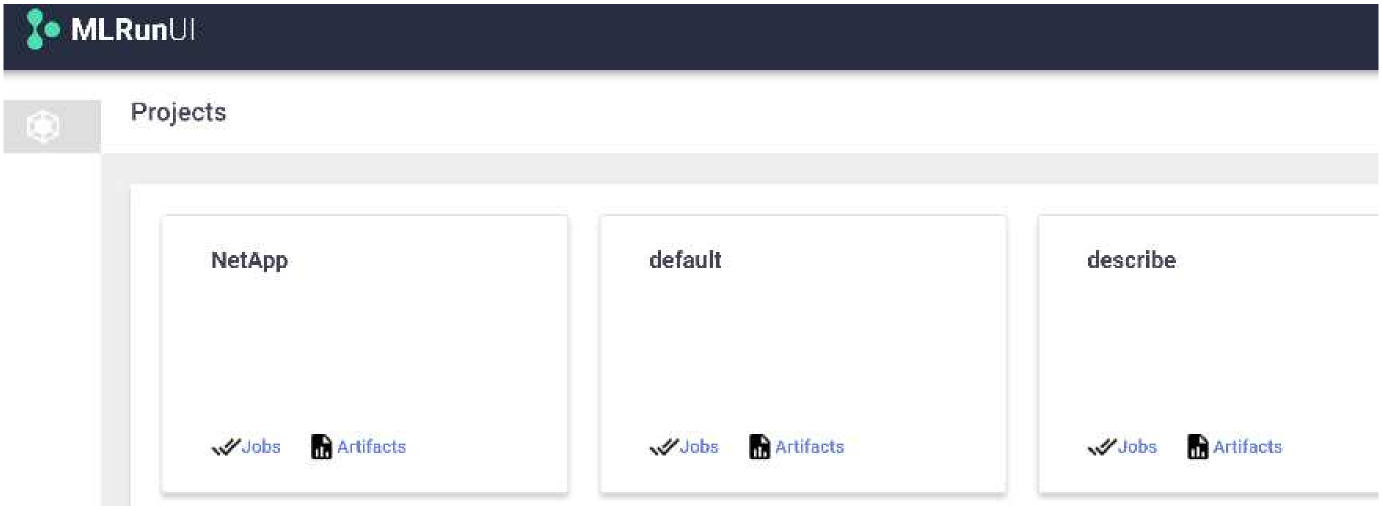
If you select the Snapshot step, you can see the name of the Snapshot copy that was used to run this experiment.



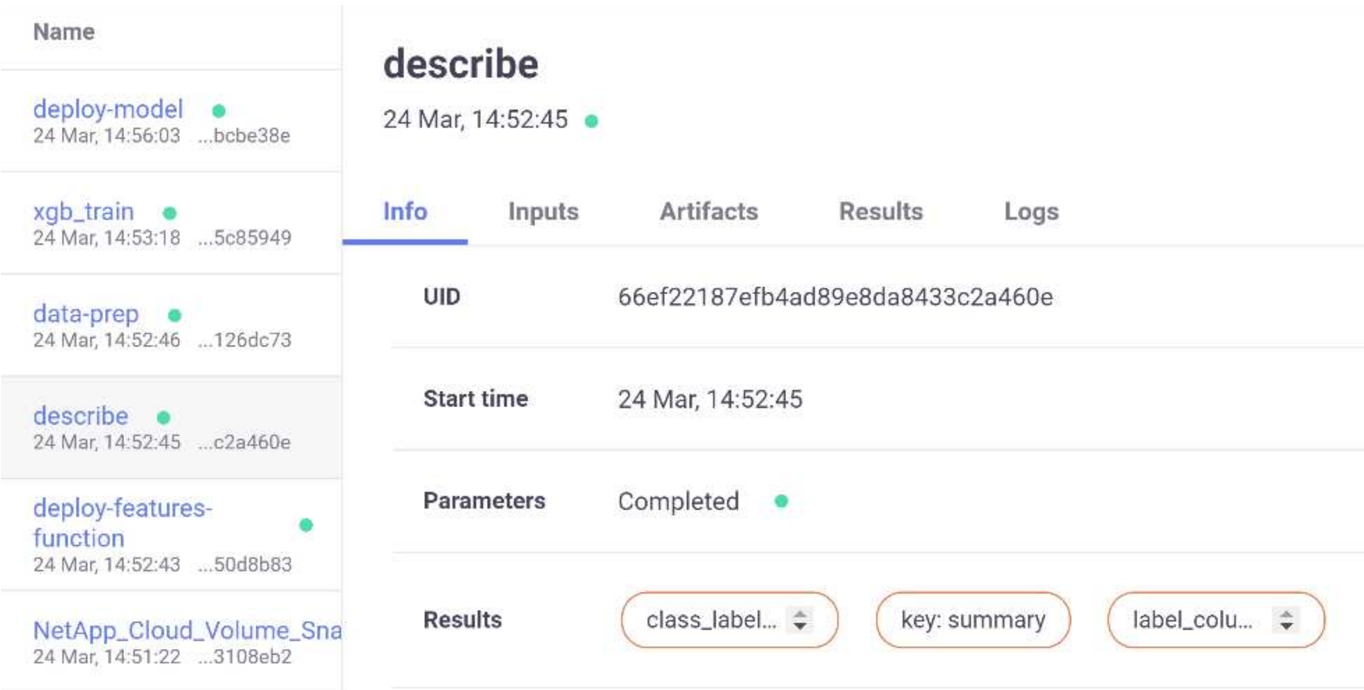
The described step has visual artifacts to explore the metrics we used. You can expand to view the full plot as seen in the following image.



The MLRun API database also tracks inputs, outputs, and artifacts for each run organized by project. An example of inputs, outputs, and artifacts for each run can be seen in the following image.



For each job, we store additional details.



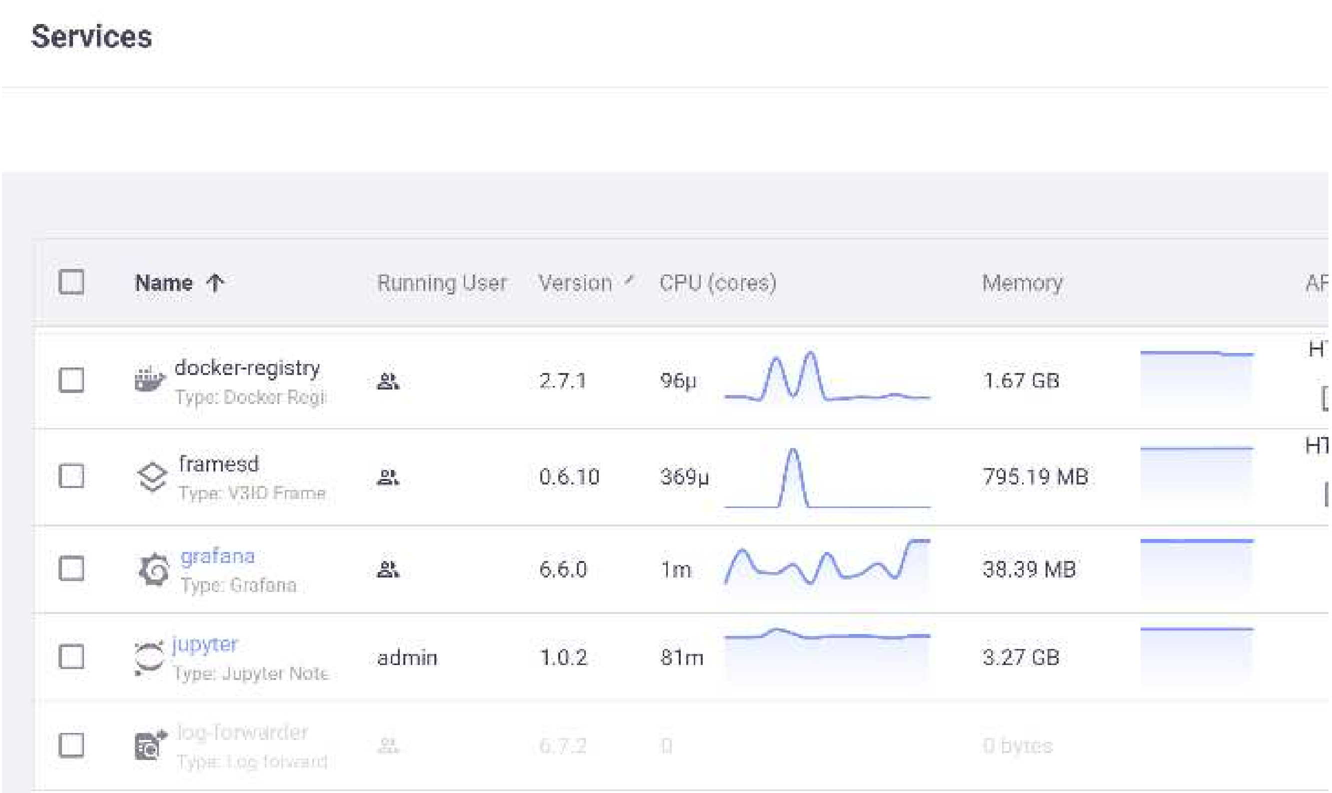
There is more information about MLRun than we can cover in this document. Al artifacts, including the definition of the steps and functions, can be saved to the API database, versioned, and invoked individually or as a full project. Projects can also be saved and pushed to Git for later use. We encourage you to learn more at the MLRun GitHub site.

Deploy Grafana Dashboard

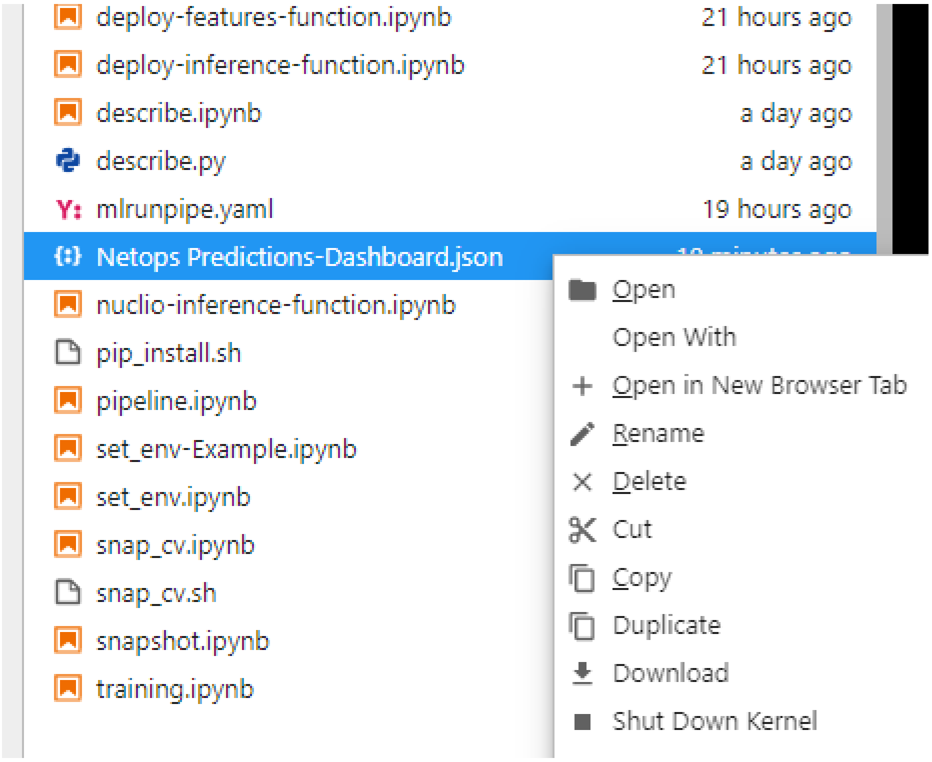
After everything is deployed, we run inferences on new data. The models predict failure on network device equipment. The results of the prediction are stored in an Iguazio TimeSeries table. You can visualize the results with Grafana in the platform integrated with Iguazio's security and data access policy.

You can deploy the dashboard by importing the provided JSON file into the Grafana interfaces in the cluster.

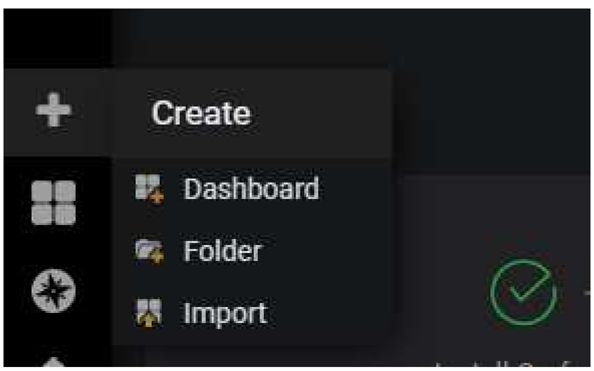
1. To verify that the Grafana service is running, look under Services.



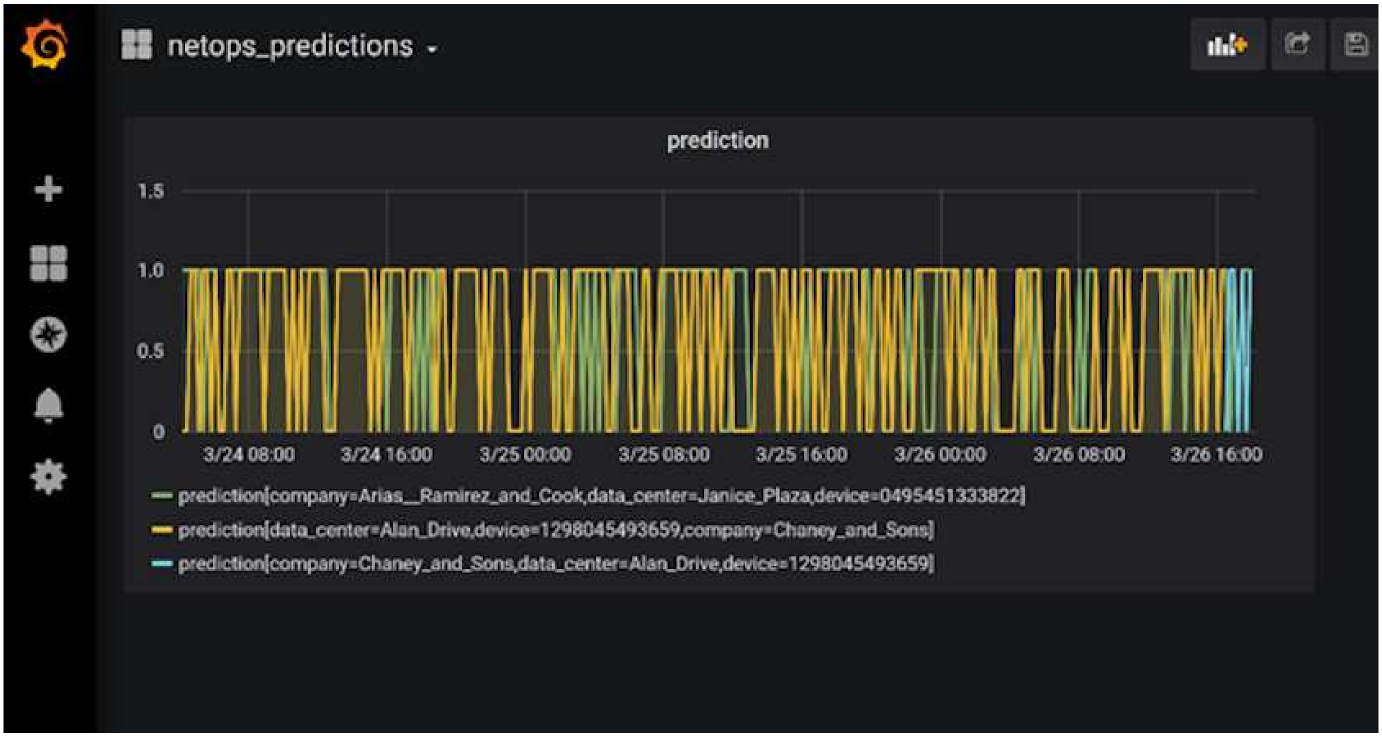
2. If it is not present, deploy an instance from the Services section:
    a. Click New Service.
    b. Select Grafana from the list.
    c. Accept the defaults.
    d. Click Next Step.
    e. Enter your user ID.
    f. Click Save Service.
    g. Click Apply Changes at the top.
3. To deploy the dashboard, download the file `NetopsPredictions-Dashboard.json` through the Jupyter interface.



4. Open Grafana from the Services section and import the dashboard.



5. Click Upload `*.json` File and select the file that you downloaded earlier (`NetopsPredictions-Dashboard.json`). The dashboard displays after the upload is completed.



Deploy Cleanup Function

When you generate a lot of data, it is important to keep things clean and organized. To do so, deploy the cleanup function with the `cleanup.ipynb` notebook.

Benefits

NetApp and Iguazio speed up and simplify the deployment of AI and ML applications by building in essential frameworks, such as Kubeflow, Apache Spark, and TensorFlow, along with orchestration tools like Docker and Kubernetes. By unifying the end-to-end data pipeline, NetApp and Iguazio reduce the latency and complexity inherent in many advanced computing workloads, effectively bridging the gap between development and operations. Data scientists can run queries on large datasets and securely share data and algorithmic models with authorized users during the training phase. After the containerized models are ready for production, you can easily move them from development environments to operational environments.

Conclusion

When building your own AI/ML pipelines, configuring the integration, management, security, and accessibility of the components in an architecture is a challenging task. Giving developers access and control of their environment presents another set of challenges.

The combination of NetApp and Iguazio brings these technologies together as managed services to accelerate technology adoption and improve the time to market for new AI/ML applications.

TR-4915: Data movement with E-Series and BeeGFS for AI and analytics workflows

Cody Harryman and Ryan Rodine, NetApp

TR-4915 describes how to move data from any data repository into a BeeGFS file system backed by NetApp E-Series SAN storage. For artificial intelligence (AI) and machine learning (ML) applications, customers might routinely need to move large data sets exceeding many petabytes of data into their BeeGFS clusters for model development. This document explores how to accomplish this by using NetApp XCP and NetApp BlueXP Copy and Sync tools.

TR-4915: Data movement with E-Series and BeeGFS for AI and analytics workflows

Vector Database Solution with NetApp

Karthikeyan Nagalingam and Rodrigo Nascimento, NetApp

This document provides a thorough exploration of the deployment and management of vector databases, such as Milvus, and pgvecto an open-source PostgreSQL extension, using NetApp's storage solutions. It details the infrastructure guidelines for using NetApp ONTAP and StorageGRID object storage and validates the application of Milvus database in AWS FSX for NetApp ONTAP. The document elucidates NetApp's file-object duality and its utility for vector databases and applications that support vector embeddings. It emphasizes the capabilities of SnapCenter, NetApp's enterprise management product, in offering backup and restore functionalities for vector databases, ensuring data integrity and availability. The document further delves into NetApp's hybrid cloud solution, discussing its role in data replication and protection across on-premises and cloud environments. It includes insights into the performance validation of vector databases on NetApp ONTAP, and concludes with two practical use cases on generative AI : RAG with LLM and the NetApp's internal ChatAI. This document serves as a comprehensive guide for leveraging NetApp's storage solutions for managing vector databases.

The Reference Architecture focus on the following:

1. Introduction
2. Solution Overview
3. Vector Database
4. Technology Requirement
5. Deployment Procedure
6. Solution Verification Overview
    - Milvus cluster setup with Kubernetes in on-premises
    - Milvus with Amazon FSxN for NetApp ONTAP – file and object duality
    - Vector database protection using NetApp SnapCenter.
    - Disaster Recovery using NetApp SnapMirror
    - Performance validation
7. Vector Database with Instaclustr using PostgreSQL: pgvector
8. Vector Database Use Cases

Introduction

This section provide an introduction to vector database solution for NetApp.

Introduction

Vector databases effectively address the challenges that are designed to handle the complexities of semantic search in Large Language Models (LLMs) and generative Artificial Intelligence (AI). Unlike traditional data management systems, vector databases are capable of processing and searching through various types of data, including images, videos, text, audio, and other forms of unstructured data, by using the content of the data itself rather than labels or tags.

The limitations of Relational Database Management Systems (RDBMS) are well-documented, particularly their struggles with high-dimensional data representations and unstructured data common in AI applications. RDBMS often necessitate a time-consuming and error-prone process of flattening data into more manageable structures, leading to delays and inefficiencies in searches. Vector databases, however, are designed to circumvent these issues, offering a more efficient and accurate solution for managing and searching through complex and high-dimensional data, thus facilitating the advancement of AI applications.

This document serves as a comprehensive guide for customers who are currently using or planning to use vector databases, detailing the best practices for utilizing vector databases on platforms such as NetApp ONTAP, NetApp StorageGRID, Amazon FSxN for NetApp ONTAP, and SnapCenter. The content provided herein covers a range of topics:

- Infrastructure guidelines for vector databases, like Milvus, provided by NetApp storage through NetApp ONTAP and StorageGRID object storage.
- Validation of the Milvus database in AWS FSX for NetApp ONTAP through file and object store.
- Delves into NetApp's file-object duality, demonstrating its utility for data in vector databases as well as other applications.
- How NetApp's Data Protection Management product, SnapCenter, offers backup and restore functionalities for vector database data.
- How NetApp's Hybrid Cloud offers data replication and protection across on-premises and cloud environments.
- Provides insights into the performance validation of vector databases like Milvus and pgvector on NetApp ONTAP.
- Two specific use cases: Retrieval Augmented Generation (RAG) with Large Language Models(LLM) and the NetApp IT team's ChatAI, thereby offering practical examples of the concepts and practices outlined.

Solution Overview

This section provides an overview for the NetApp vector database solution.

Solution overview

This solution showcases the distinctive benefits and capabilities that NetApp brings to the table to tackle the challenges faced by vector database customers. By leveraging NetApp ONTAP, StorageGRID, NetApp's cloud solutions, and SnapCenter, customers can add significant value to their business operations. These tools not only address existing issues but also enhance efficiency and productivity, thereby contributing to overall business growth.

Why NetApp?

- NetApp's offerings, such as ONTAP and StorageGRID, allow for the separation of storage and compute, enabling optimal resource utilization based on specific requirements. This flexibility empowers customers to independently scale their storage using NetApp storage solutions.
- By leveraging NetApp's storage controllers, customers can efficiently serve data to their vector database using NFS and S3 protocols. These protocols facilitate customer data storage and manage the vector database index, eliminating the need for multiple copies of data accessed through file and object methods.
- NetApp ONTAP provides native support for NAS and Object storage across leading cloud service providers like AWS, Azure, and Google Cloud. This wide compatibility ensures seamless integration, enabling customer data mobility, global accessibility, disaster recovery, dynamic scalability, and high performance.
- With NetApp's robust data management capabilities, customers can rest assured knowing that their data is well-protected against potential risks and threats. NetApp prioritizes data security, offering peace of mind to customers regarding the safety and integrity of their valuable information.

Vector Database

This section covers the definition and use of a vector database in NetApp AI solutions.

Vector Database

A vector database is a specialized type of database designed to handle, index, and search unstructured data using embeddings from machine learning models. Instead of organizing data in a traditional tabular format, it arranges data as high-dimensional vectors, also known as vector embeddings. This unique structure allows the database to handle complex, multi-dimensional data more efficiently and accurately.

One of the key capabilities of a vector database is its use of generative AI to perform analytics. This includes similarity searches, where the database identifies data points that are like a given input, and anomaly detection, where it can spot data points that deviate significantly from the norm.

Furthermore, vector databases are well-suited to handle temporal data, or time-stamped data. This type of data provides information about 'what' happened and when it happened, in sequence and in relation to all other events within a given IT system. This ability to handle and analyze temporal data makes vector databases particularly useful for applications that require an understanding of events over time.

Advantages of vector database for ML and AI:

- High-dimensional Search: Vector databases excel in managing and retrieving high-dimensional data, which is often generated in AI and ML applications.
- Scalability: They can efficiently scale to handle large volumes of data, supporting the growth and expansion of AI and ML projects.
- Flexibility: Vector databases offer a high degree of flexibility, allowing for the accommodation of diverse data types and structures.

- Performance: They provide high-performance data management and retrieval, critical for the speed and efficiency of AI and ML operations.
- Customizable Indexing: Vector databases offer customizable indexing options, enabling optimized data organization and retrieval based on specific needs.

Vector databases and use cases.

This section provides varies vector databases and their use case details.

Faiss and ScaNN

They are libraries that serve as crucial tools in the realm of vector search. These libraries provide functionality that is instrumental in managing and searching through vector data, making them invaluable resources in this specialized area of data management.

Elasticsearch

It's a widely used search and analytics engine, has recently incorporated vector search capabilities. This new feature enhances its functionality, enabling it to handle and search through vector data more effectively.

Pinecone

It is a robust vector database with a unique set of features. It supports both dense and sparse vectors in its indexing functionality, which enhances its flexibility and adaptability. One of its key strengths lies in its ability to combine traditional search methods with AI-based dense vector search, creating a hybrid search approach that leverages the best of both worlds.

Primarily cloud-based, Pinecone is designed for machine learning applications and integrates well with a variety of platforms, including GCP, AWS, Open AI, GPT-3, GPT-3.5, GPT-4, Catgut Plus, Elasticsearch, Haystack, and more. It's important to note that Pinecone is a closed-source platform and is available as a Software as a Service (SaaS) offering.

Given its advanced capabilities, Pinecone is particularly well-suited for the cybersecurity industry, where its high-dimensional search and hybrid search capabilities can be leveraged effectively to detect and respond to threats.

Chroma

It's a vector database that has a Core-API with four primary functions, one of which includes an in-memory document-vector store. It also utilizes the Face Transformers library to vectorize documents, enhancing its functionality and versatility.
Chroma is designed to operate both in the cloud and on-premises, offering flexibility based on user needs. Particularly, it excels in audio-related applications, making it an excellent choice for audio-based search engines, music recommendation systems, and other audio-related use cases.

Weaviate

It's a versatile vector database that allows users to vectorize their content using either its built-in modules or custom modules, providing flexibility based on specific needs. It offers both fully managed and self-hosted solutions, catering to a variety of deployment preferences.

One of Weaviate's key features is its ability to store both vectors and objects, enhancing its data handling capabilities. It is widely used for a range of applications, including semantic search and data classification in ERP systems. In the e-commerce sector, it powers search and recommendation engines. Weaviate is also used for image search, anomaly detection, automated data harmonization, and cybersecurity threat analysis, showcasing its versatility across multiple domains.

Redis

Redis is a high-performing vector database known for its fast in-memory storage, offering low latency for read-write operations. This makes it an excellent choice for recommendation systems, search engines, and data analytics applications that require quick data access.

Redis supports various data structures for vectors, including lists, sets, and sorted sets. It also provides vector operations such as calculating distances between vectors or finding intersections and unions. These features are particularly useful for similarity search, clustering, and content-based recommendation systems.

In terms of scalability and availability, Redis excels in handling high throughput workloads and offers data replication. It also integrates well with other data types, including traditional relational databases (RDBMS). Redis includes a Publish/Subscribe (Pub/Sub) feature for real-time updates, which is beneficial for managing real-time vectors. Moreover, Redis is lightweight and simple to use, making it a user-friendly solution for managing vector data.

Milvus

It’s a versatile vector database that offers an API like a document store, much like MongoDB. It stands out due to its support for a wide variety of data types, making it a popular choice in the data science and machine learning fields.

One of Milvus’ unique features is its multi-vectorization capability, which allows users to specify at runtime the type of vector to use for the search. Furthermore, it utilizes Knowwhere, a library that sits atop other libraries like Faiss, to manage communication between queries and the vector search algorithms.

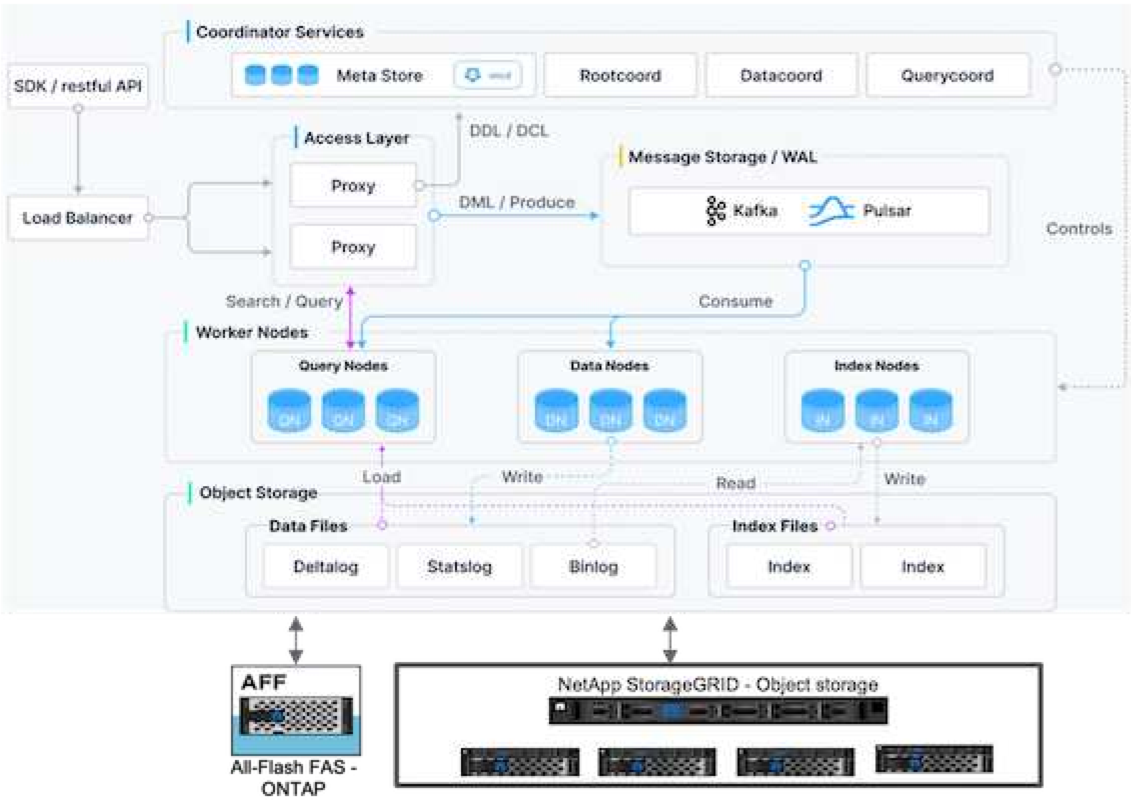
Milvus also offers seamless integration with machine learning workflows, thanks to its compatibility with PyTorch and TensorFlow. This makes it an excellent tool for a range of applications, including e-commerce, image and video analysis, object recognition, image similarity search, and content-based image retrieval. In the realm of natural language processing, Milvus is used for document clustering, semantic search, and question-answering systems.

For this solution, we picked milvus for the solution validation. For performance, we used both milvus and postgres(pgvecto.rs).

Why we chose milvus for this solution?

- Open-Source: Milvus is an open-source vector database, encouraging community-driven development and improvements.
- AI Integration: It leverages embedding similarity search and AI applications to enhance vector database functionality.
- Large Volume Handling: Milvus has the capacity to store, index, and manage over a billion embedding vectors generated by Deep Neural Networks (DNN) and Machine Learning (ML) models.
- User-Friendly: It is easy to use, with setup taking less than a minute. Milvus also offers SDKs for different programming languages.
- Speed: It offers blazing fast retrieval speeds, up to 10 times faster than some alternatives.
- Scalability and Availability: Milvus is highly scalable, with options to scale up and out as needed.
- Feature-Rich: It supports different data types, attribute filtering, User-Defined Function (UDF) support, configurable consistency levels, and travel time, making it a versatile tool for various applications.

Milvus architecture overview



This section provides higher lever components and services are used in Milvus architecture.
* Access layer – It's composed of a group of stateless proxies and serves as the front layer of the system and endpoint to users.
* Coordinator service – it assigns the tasks to the worker nodes and act as a system's brain. It has three coordinator types: root coord,data coord and query coord.
* Worker nodes : It follows the instruction from coordinator service and execute user triggered DML/DDL commands.it has three types of worker nodes such as query node, data node and index node.
* Storage: it's responsible for data persistence. It comprises meta storage, log broker, and object storage. NetApp storage such as ONTAP and StorageGRID provides object storage and File based storage to Milvus for both customer data and vector database data.

Technology Requirement

This section provides an overview of the requirements for the NetApp vector database solution.

Technology Requirement

The hardware and software configurations outlined below were utilized for the majority of the validations performed in this document, with the exception of performance. These configurations serve as a guideline to help you set up your environment. However, please note that the specific components may vary depending on individual customer requirements.

Hardware requirements

Hardware	Details
NetApp AFF Storage array HA Pair	* A800 * ONTAP 9.14.1 * 48 x 3.49TB SSD-NVM * Two Flexible group volumes: metadata and data. * Metadata NFS volume has 12 x Persistent Volumes with 250GB. * Data is a ONTAP NAS S3 volume
6 x FUJITSU PRIMERGY RX2540 M4	* 64 CPUs * Intel® Xeon® Gold 6142 CPU @ 2.60GHz * 256 GM Physical Memory * 1 x 100GbE network port
Networking	100 GbE
StorageGRID	* 1 x SG100, 3xSGF6024 * 3 x 24 x 7.68TB

Software requirements

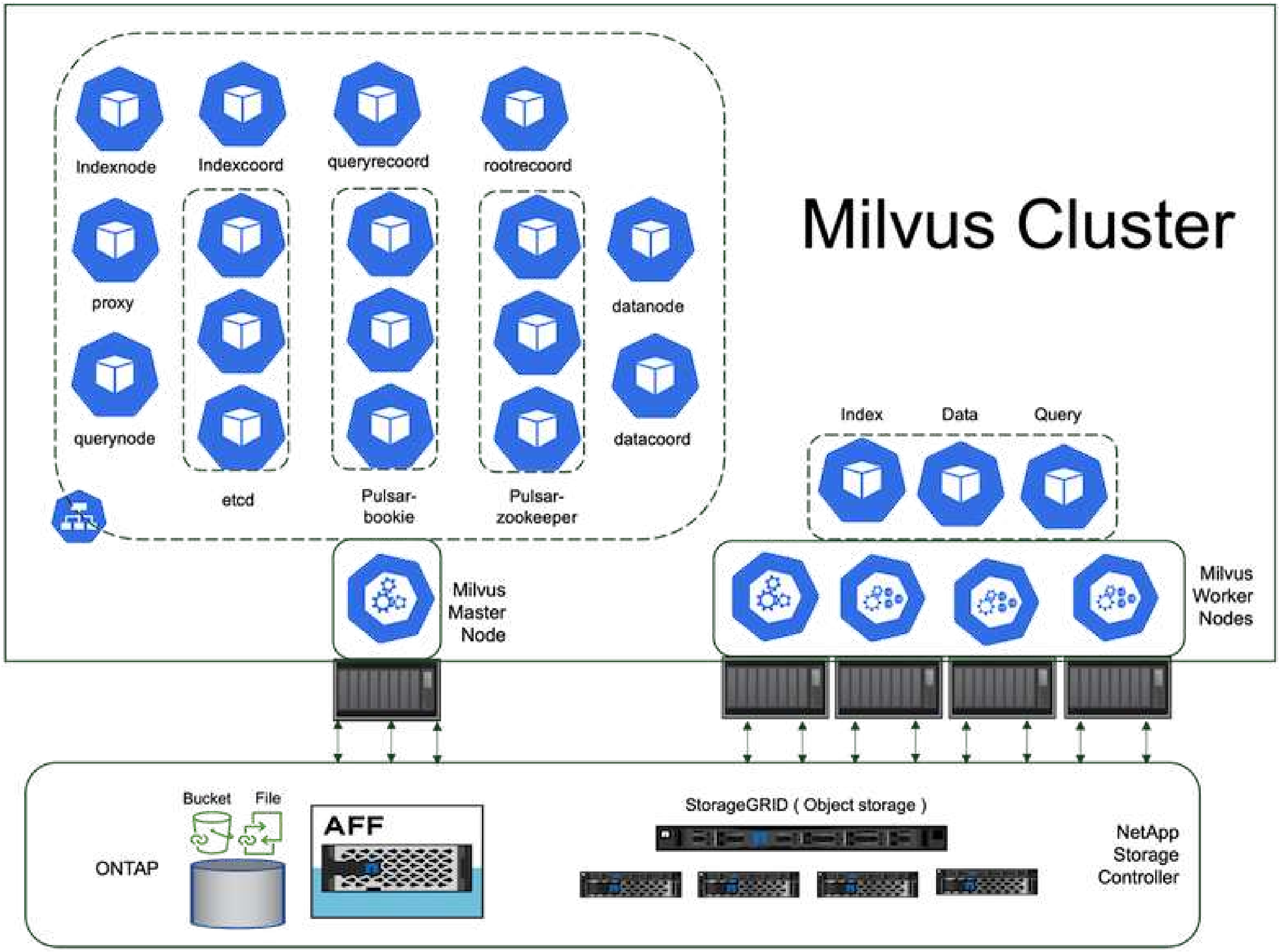
Software	Details
Milvus cluster	* CHART - milvus-4.1.11. * APP Version – 2.3.4 * Dependent bundles such as bookkeeper, zookeeper, pulsar, etcd, proxy, querynode, worker
Kubernetes	* 5 node K8s cluster * 1 Master node and 4 Worker nodes * Version – 1.7.2
Python	*3.10.12.

Deployment Procedure

This section discusses the deployment procedure for the vector database solution for NetApp.

Deployment procedure

In this deployment section, we used milvus vector database with Kubernetes for the lab setup as below.



The netapp storage provides the storage for the cluster to keep customers data and milvus cluster data.

NetApp storage setup – ONTAP

- Storage system initialization
- Storage virtual machine (SVM) creation
- Assignment of logical network interfaces
- NFS, S3 configuration and licensing

Please follow the steps below for NFS (Network File System):

1. Create a FlexGroup volume for NFSv4. In our set up for this validation, we have used 48 SSDs, 1 SSD dedicated for the controller's root volume and 47 SSDs spread across for NFSv4]].Verify that the NFS export policy for the FlexGroup volume has read/write permissions for the Kubernetes (K8s) nodes network. If these permissions are not in place, grant read/write (rw) permissions for the K8s nodes network.
2. On all K8s nodes, create a folder and mount the FlexGroup volume onto this folder through a Logical Interface (LIF) on each K8s nodes.

Please follow the steps below for NAS S3 (Network Attached Storage Simple Storage Service):

1. Create a FlexGroup volume for NFS.
2. Set up an object-store-server with HTTP enabled and the admin status set to 'up' using the "vserver object-

store-server create" command. You have the option to enable HTTPS and set a custom listener port.

3. Create an object-store-server user using the "vserver object-store-server user create -user <username>" command.
4. To obtain the access key and secret key, you can run the following command: "set diag; vserver object-store-server user show -user <username>". However, moving forward, these keys will be supplied during the user creation process or can be retrieved using REST API calls.
5. Establish an object-store-server group using the user created in step 2 and grant access. In this example, we have provided "FullAccess".
6. Create a NAS bucket by setting its type to "nas" and supplying the path to the NFSv3 volume. It's also possible to utilize an S3 bucket for this purpose.

NetApp storage setup – StorageGRID

1. Install the storageGRID software.
2. Create a tenant and bucket.
3. Create user with required permission.

Please check more details in https://docs.netapp.com/us-en/storagegrid-116/primer/index.html

Solution Overview

We have conducted a comprehensive solution validation focused on five key areas, the details of which are outlined below. Each section delves into the challenges faced by customers, the solutions provided by NetApp, and the subsequent benefits to the customer.

1. Milvus cluster setup with Kubernetes in on-premises
Customer challenges to scale independently on storage and compute, effective infrastructure management and data management. In this section, we detail the process of installing a Milvus cluster on Kubernetes, utilizing a NetApp storage controller for both cluster data and customer data.
2. Milvus with Amazon FSxN for NetApp ONTAP – file and object duality
In this section, Why we need to deploy vector database in cloud as well as steps to deploy vector database (milvus standalone) in Amazon FSxN for NetApp ONTAP within docker containers.
3. Vector database protection using NetApp SnapCenter.
In this section, we delve into how SnapCenter safeguards the vector database data and Milvus data residing in ONTAP. For this example, we utilized a NAS bucket (milvusdbvol1) derived from an NFS ONTAP volume (vol1) for customer data, and a separate NFS volume (vectordbpv) for Milvus cluster configuration data.
4. Disaster Recovery using NetApp SnapMirror
In this section, we discuss about the importance of Disaster recovery(DR) for vector database and how netapp disaster recovery product snapmirror provides DR solution to vector database.
5. Performance validation
In this section, we aim to delve into the performance validation of vector databases, such as Milvus and pgvecto.rs, focusing on their storage performance characteristics such as I/O profile and netapp storage controller behavious in support of RAG and inference workloads within the LLM Lifecycle. We will evaluate and identify any performance differentiators when these databases are combined with the ONTAP storage solution. Our analysis will be based on key performance indicators, such as the number of queries processed per second(QPS).

Milvus Cluster Setup with Kubernetes in on-premises

This section discusses the milvus cluster setup for the vector database solution for NetApp.

Milvus cluster setup with Kubernetes in on-premises

Customer challenges to scale independently on storage and compute, effective infrastructure management and data management,
Kubernetes and vector databases together form a powerful, scalable solution for managing large data operations. Kubernetes optimizes resources and manages containers, while vector databases efficiently handle high-dimensional data and similarity searches. This combination enables swift processing of complex queries on large datasets and seamlessly scales with growing data volumes, making it ideal for big data applications and AI workloads.

1. In this section, we detail the process of installing a Milvus cluster on Kubernetes, utilizing a NetApp storage controller for both cluster data and customer data.
2. To install a Milvus cluster, Persistent Volumes (PVs) are required for storing data from various Milvus cluster components. These components include etcd (three instances), pulsar-bookie-journal (three instances), pulsar-bookie-ledgers (three instances), and pulsar-zookeeper-data (three instances).

> In milvus cluster, we can use either pulsar or kafka for the underlying engine supporting Milvus cluster's reliable storage and publication/subscription of message streams. For Kafka with NFS,NetApp has made improvements in ONTAP 9.12.1 and later, and these enhancements, along with NFSv4.1 and Linux changes that are included in RHEL 8.7 or 9.1 and higher, resolve the "silly rename" issue that can occur when running Kafka over NFS. if you interested in more in-depth information on the topic of running kafka with netapp NFS solution, please check - this link.

3. We created a single NFS volume from NetApp ONTAP and established 12 persistent volumes, each with 250GB of storage. The storage size can vary depending on the cluster size; for instance, we have another cluster where each PV has 50GB. Please refer below to one of the PV YAML files for more details; we had 12 such files in total. In each file, the storageClassName is set to 'default', and the storage and path are unique to each PV.

```
root@node2:~# cat sai_nfs_to_default_pv1.yaml
apiVersion: v1
kind: PersistentVolume
metadata:
  name: karthik-pv1
spec:
  capacity:
    storage: 250Gi
  volumeMode: Filesystem
  accessModes:
  - ReadWriteOnce
  persistentVolumeReclaimPolicy: Retain
  storageClassName: default
  local:
    path: /vectordbsc/milvus/milvus1
  nodeAffinity:
    required:
      nodeSelectorTerms:
      - matchExpressions:
        - key: kubernetes.io/hostname
          operator: In
          values:
          - node2
          - node3
          - node4
          - node5
          - node6
root@node2:~#
```

4. Execute the 'kubectl apply' command for each PV YAML file to create the Persistent Volumes, and then verify their creation using ‘kubectl get pv’

```
root@node2:~# for i in $( seq 1 12 ); do kubectl apply -f
sai_nfs_to_default_pv$i.yaml; done
persistentvolume/karthik-pv1 created
persistentvolume/karthik-pv2 created
persistentvolume/karthik-pv3 created
persistentvolume/karthik-pv4 created
persistentvolume/karthik-pv5 created
persistentvolume/karthik-pv6 created
persistentvolume/karthik-pv7 created
persistentvolume/karthik-pv8 created
persistentvolume/karthik-pv9 created
persistentvolume/karthik-pv10 created
persistentvolume/karthik-pv11 created
persistentvolume/karthik-pv12 created
root@node2:~#
```

5. For storing customer data, Milvus supports object storage solutions such as MinIO, Azure Blob, and S3. In this guide, we utilize S3. The following steps apply to both ONTAP S3 and StorageGRID object store. We use Helm to deploy the Milvus cluster. Download the configuration file, values.yaml, from the Milvus download location. Please refer to the appendix for the values.yaml file we used in this document.
6. Ensure that the 'storageClass' is set to 'default' in each section, including those for the log, etcd, zookeeper, and bookkeeper.
7. In the MinIO section, disable MinIO.
8. Create a NAS bucket from ONTAP or StorageGRID object storage and include them in an External S3 with the object storage credentials.

```
###################################
# External S3
# - these configs are only used when `externalS3.enabled` is true
###################################
externalS3:
  enabled: true
  host: "192.168.150.167"
  port: "80"
  accessKey: "24G4C1316APP2BIPDE5S"
  secretKey: "Zd28p43rgZaU44PX_ftT279z9nt4jBSro97j87Bx"
  useSSL: false
  bucketName: "milvusdbvol1"
  rootPath: ""
  useIAM: false
  cloudProvider: "aws"
  iamEndpoint: ""
  region: ""
  useVirtualHost: false
```

9. Before creating the Milvus cluster, ensure that the PersistentVolumeClaim (PVC) does not have any pre-existing resources.

```
root@node2:~# kubectl get pvc
No resources found in default namespace.
root@node2:~#
```

10. Utilize Helm and the values.yaml configuration file to install and start the Milvus cluster.

```
root@node2:~# helm upgrade --install my-release milvus/milvus --set
global.storageClass=default  -f values.yaml
Release "my-release" does not exist. Installing it now.
NAME: my-release
LAST DEPLOYED: Thu Mar 14 15:00:07 2024
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
root@node2:~#
```

11. Verify the status of the PersistentVolumeClaims (PVCs).

```
root@node2:~# kubectl get pvc
NAME                                                                    STATUS
VOLUME          CAPACITY   ACCESS MODES   STORAGECLASS   AGE
data-my-release-etcd-0                                                  Bound
karthik-pv8     250Gi      RWO            default        3s
data-my-release-etcd-1                                                  Bound
karthik-pv5     250Gi      RWO            default        2s
data-my-release-etcd-2                                                  Bound
karthik-pv4     250Gi      RWO            default        3s
my-release-pulsar-bookie-journal-my-release-pulsar-bookie-0             Bound
karthik-pv10    250Gi      RWO            default        3s
my-release-pulsar-bookie-journal-my-release-pulsar-bookie-1             Bound
karthik-pv3     250Gi      RWO            default        3s
my-release-pulsar-bookie-journal-my-release-pulsar-bookie-2             Bound
karthik-pv1     250Gi      RWO            default        3s
my-release-pulsar-bookie-ledgers-my-release-pulsar-bookie-0             Bound
karthik-pv2     250Gi      RWO            default        3s
my-release-pulsar-bookie-ledgers-my-release-pulsar-bookie-1             Bound
karthik-pv9     250Gi      RWO            default        3s
my-release-pulsar-bookie-ledgers-my-release-pulsar-bookie-2             Bound
karthik-pv11    250Gi      RWO            default        3s
my-release-pulsar-zookeeper-data-my-release-pulsar-zookeeper-0          Bound
karthik-pv7     250Gi      RWO            default        3s
root@node2:~#
```

12. Check the status of the pods.

```
root@node2:~# kubectl get pods -o wide
NAME                                                READY   STATUS
RESTARTS        AGE     IP               NODE     NOMINATED NODE
READINESS GATES
<content removed to save page space>
```

Please make sure the pods status are ‘running’ and working as expected

13. Test data writing and reading in Milvus and NetApp object storage.
    - Write data using the "prepare_data_netapp_new.py" Python program.

```
root@node2:~# date;python3 prepare_data_netapp_new.py ;date
Thu Apr  4 04:15:35 PM UTC 2024
=== start connecting to Milvus     ===
=== Milvus host: localhost         ===
Does collection hello_milvus_ntapnew_update2_sc exist in Milvus:
False
=== Drop collection - hello_milvus_ntapnew_update2_sc ===
=== Drop collection - hello_milvus_ntapnew_update2_sc2 ===
=== Create collection `hello_milvus_ntapnew_update2_sc` ===
=== Start inserting entities       ===
Number of entities in hello_milvus_ntapnew_update2_sc: 3000
Thu Apr  4 04:18:01 PM UTC 2024
root@node2:~#
```

◦ Read the data using the "verify_data_netapp.py" Python file.

```
root@node2:~# python3 verify_data_netapp.py
=== start connecting to Milvus     ===
=== Milvus host: localhost         ===

Does collection hello_milvus_ntapnew_update2_sc exist in Milvus: True
{'auto_id': False, 'description': 'hello_milvus_ntapnew_update2_sc',
'fields': [{'name': 'pk', 'description': '', 'type': <DataType.INT64:
5>, 'is_primary': True, 'auto_id': False}, {'name': 'random',
'description': '', 'type': <DataType.DOUBLE: 11>}, {'name': 'var',
'description': '', 'type': <DataType.VARCHAR: 21>, 'params':
{'max_length': 65535}}, {'name': 'embeddings', 'description': '',
'type': <DataType.FLOAT_VECTOR: 101>, 'params': {'dim': 16}}]}
Number of entities in Milvus: hello_milvus_ntapnew_update2_sc : 3000

=== Start Creating index IVF_FLAT  ===

=== Start loading                  ===

=== Start searching based on vector similarity ===

hit: id: 2998, distance: 0.0, entity: {'random': 0.9728033590489911},
random field: 0.9728033590489911
hit: id: 2600, distance: 0.602496862411499, entity: {'random':
0.3098157043984633}, random field: 0.3098157043984633
hit: id: 1831, distance: 0.6797959804534912, entity: {'random':
0.6331477114129169}, random field: 0.6331477114129169
hit: id: 2999, distance: 0.0, entity: {'random':
0.02316334456872482}, random field: 0.02316334456872482
hit: id: 2524, distance: 0.5918987989425659, entity: {'random':
```

```
0.285283165889066}, random field: 0.285283165889066
hit: id: 264, distance: 0.7254047393798828, entity: {'random':
0.3329096143562196}, random field: 0.3329096143562196
search latency = 0.4533s

=== Start querying with `random > 0.5` ===

query result:
-{'random': 0.6378742006852851, 'embeddings': [0.20963514,
0.39746657, 0.12019053, 0.6947492, 0.9535575, 0.5454552, 0.82360446,
0.21096309, 0.52323616, 0.8035404, 0.77824664, 0.80369574, 0.4914803,
0.8265614, 0.6145269, 0.80234545], 'pk': 0}
search latency = 0.4476s

=== Start hybrid searching with `random > 0.5` ===

hit: id: 2998, distance: 0.0, entity: {'random': 0.9728033590489911},
random field: 0.9728033590489911
hit: id: 1831, distance: 0.6797959804534912, entity: {'random':
0.6331477114129169}, random field: 0.6331477114129169
hit: id: 678, distance: 0.7351570129394531, entity: {'random':
0.5195484662306603}, random field: 0.5195484662306603
hit: id: 2644, distance: 0.8620758056640625, entity: {'random':
0.9785952878381153}, random field: 0.9785952878381153
hit: id: 1960, distance: 0.9083120226860046, entity: {'random':
0.6376039340439571}, random field: 0.6376039340439571
hit: id: 106, distance: 0.9792704582214355, entity: {'random':
0.9679994241326673}, random field: 0.9679994241326673
search latency = 0.1232s
Does collection hello_milvus_ntapnew_update2_sc2 exist in Milvus:
True
{'auto_id': True, 'description': 'hello_milvus_ntapnew_update2_sc2',
'fields': [{'name': 'pk', 'description': '', 'type': <DataType.INT64:
5>, 'is_primary': True, 'auto_id': True}, {'name': 'random',
'description': '', 'type': <DataType.DOUBLE: 11>}, {'name': 'var',
'description': '', 'type': <DataType.VARCHAR: 21>, 'params':
{'max_length': 65535}}, {'name': 'embeddings', 'description': '',
'type': <DataType.FLOAT_VECTOR: 101>, 'params': {'dim': 16}}]}
```

Based on the above validation, the integration of Kubernetes with a vector database, as demonstrated through the deployment of a Milvus cluster on Kubernetes using a NetApp storage controller, offers customers a robust, scalable, and efficient solution for managing large-scale data operations. This setup provides customers with the ability to handle high-dimensional data and execute complex queries rapidly and efficiently, making it an ideal solution for big data applications and AI workloads. The use of Persistent Volumes (PVs) for various cluster components, along with the creation of a single NFS volume from NetApp ONTAP, ensures optimal resource utilization and data management. The process of verifying the status of PersistentVolumeClaims (PVCs) and pods, as well as testing data writing and reading, provides customers with the assurance of reliable and consistent data operations. The use of ONTAP or StorageGRID object storage for customer data further enhances data accessibility and security. Overall, this setup empowers customers with a resilient and high-performing data management solution that can seamlessly scale with their growing data needs.

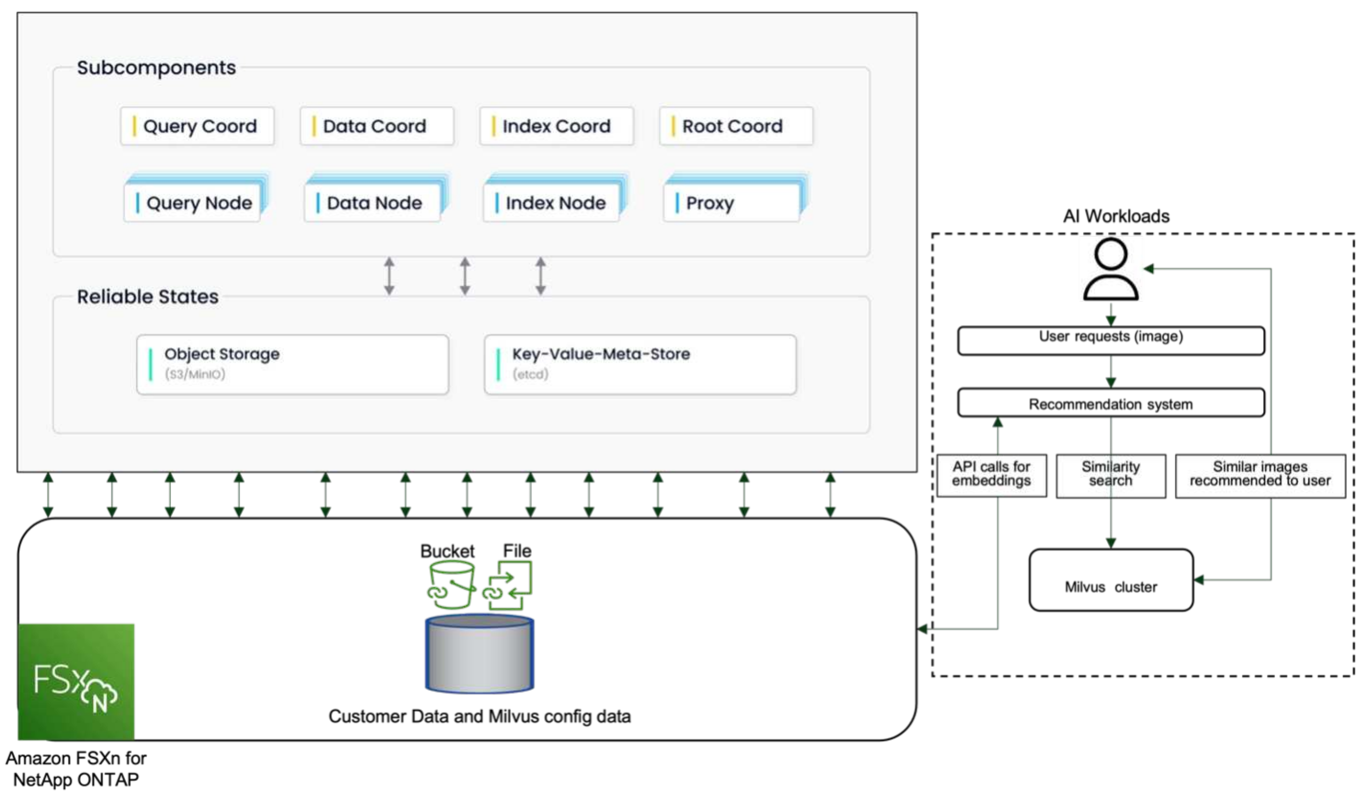
Milvus with Amazon FSxN for NetApp ONTAP - file and object duality

This section discusses the milvus cluster setup with Amazon FSxN for the vector database solution for NetApp.

Milvus with Amazon FSxN for NetApp ONTAP – file and object duality

In this section, Why we need to deploy vector database in cloud as well as steps to deploy vector database (milvus standalone) in Amazon FSxN for NetApp ONTAP within docker containers.

Deploying a vector database in the cloud provides several significant benefits, particularly for applications that require handling high-dimensional data and executing similarity searches. First, cloud-based deployment offers scalability, allowing for the easy adjustment of resources to match the growing data volumes and query loads. This ensures that the database can efficiently handle increased demand while maintaining high performance. Second, cloud deployment provides high availability and disaster recovery, as data can be replicated across different geographical locations, minimizing the risk of data loss, and ensuring continuous service even during unexpected events. Third, it provides cost-effectiveness, as you only pay for the resources you use, and can scale up or down based on demand, avoiding the need for substantial upfront investment in hardware. Finally, deploying a vector database in the cloud can enhance collaboration, as data can be accessed and shared from anywhere, facilitating team-based work and data-driven decision making.
Please check the architecture of the milvus standalone with Amazon FSxN for NetApp ONTAP used in this validation.



1. Create an Amazon FSxN for NetApp ONTAP instance and note down the details of the VPC, VPC security groups, and subnet. This information will be required when creating an EC2 instance. You can find more

details here - https://us-east-1.console.aws.amazon.com/fsx/home?region=us-east-1#file-system-create

2. Create an EC2 instance, ensuring that the VPC, Security Groups, and subnet match those of the Amazon FSxN for NetApp ONTAP instance.
3. Install nfs-common using the command 'apt-get install nfs-common' and update the package information using 'sudo apt-get update'.
4. Create a mount folder and mount the Amazon FSxN for NetApp ONTAP on it.

```
ubuntu@ip-172-31-29-98:~$ mkdir /home/ubuntu/milvusvectordb
ubuntu@ip-172-31-29-98:~$ sudo mount 172.31.255.228:/vol1
/home/ubuntu/milvusvectordb
ubuntu@ip-172-31-29-98:~$ df -h /home/ubuntu/milvusvectordb
Filesystem            Size  Used Avail Use% Mounted on
172.31.255.228:/vol1  973G  126G  848G  13% /home/ubuntu/milvusvectordb
ubuntu@ip-172-31-29-98:~$
```

5. Install Docker and Docker Compose using 'apt-get install'.
6. Set up a Milvus cluster based on the docker-compose.yaml file, which can be downloaded from the Milvus website.

```
root@ip-172-31-22-245:~# wget https://github.com/milvus-
io/milvus/releases/download/v2.0.2/milvus-standalone-docker-compose.yml
-O docker-compose.yml
--2024-04-01 14:52:23--  https://github.com/milvus-
io/milvus/releases/download/v2.0.2/milvus-standalone-docker-compose.yml
<removed some output to save page space>
```

7. In the 'volumes' section of the docker-compose.yml file, map the NetApp NFS mount point to the corresponding Milvus container path, specifically in etcd, minio, and standalone.Check Appendix D: docker-compose.yml for details about changes in yml
8. Verify the mounted folders and files.

```
ubuntu@ip-172-31-29-98:~/milvusvectordb$ ls -ltrh
/home/ubuntu/milvusvectordb
total 8.0K
-rw-r--r-- 1 root root 1.8K Apr  2 16:35 s3_access.py
drwxrwxrwx 2 root root 4.0K Apr  4 20:19 volumes
ubuntu@ip-172-31-29-98:~/milvusvectordb$ ls -ltrh
/home/ubuntu/milvusvectordb/volumes/
total 0
ubuntu@ip-172-31-29-98:~/milvusvectordb$ cd
ubuntu@ip-172-31-29-98:~$ ls
docker-compose.yml  docker-compose.yml~  milvus.yaml  milvusvectordb
vectordbvol1
ubuntu@ip-172-31-29-98:~$
```

9. Run 'docker-compose up -d' from the directory containing the docker-compose.yml file.
10. Check the status of the Milvus container.

```
ubuntu@ip-172-31-29-98:~$ sudo docker-compose ps
      Name                         Command                        State
Ports
------------------------------------------------------------------------
------------------------------------------------------------------------
----------
milvus-etcd         etcd -advertise-client-url ...   Up (healthy)
2379/tcp, 2380/tcp
milvus-minio        /usr/bin/docker-entrypoint ...   Up (healthy)
0.0.0.0:9000->9000/tcp,:::9000->9000/tcp, 0.0.0.0:9001-
>9001/tcp,:::9001->9001/tcp
milvus-standalone   /tini -- milvus run standalone   Up (healthy)
0.0.0.0:19530->19530/tcp,:::19530->19530/tcp, 0.0.0.0:9091-
>9091/tcp,:::9091->9091/tcp
ubuntu@ip-172-31-29-98:~$
ubuntu@ip-172-31-29-98:~$ ls -ltrh /home/ubuntu/milvusvectordb/volumes/
total 12K
drwxr-xr-x 3 root root 4.0K Apr  4 20:21 etcd
drwxr-xr-x 4 root root 4.0K Apr  4 20:21 minio
drwxr-xr-x 5 root root 4.0K Apr  4 20:21 milvus
ubuntu@ip-172-31-29-98:~$
```

11. To validate the read and write functionality of vector database and it’s data in Amazon FSxN for NetApp ONTAP, we used the Python Milvus SDK and a sample program from PyMilvus. Install the necessary packages using 'apt-get install python3-numpy python3-pip' and install PyMilvus using 'pip3 install pymilvus'.
12. Validate data writing and reading operations from Amazon FSxN for NetApp ONTAP in the vector

database.

```
root@ip-172-31-29-98:~/pymilvus/examples# python3
prepare_data_netapp_new.py
=== start connecting to Milvus     ===
=== Milvus host: localhost         ===
Does collection hello_milvus_ntapnew_sc exist in Milvus: True
=== Drop collection - hello_milvus_ntapnew_sc ===
=== Drop collection - hello_milvus_ntapnew_sc2 ===
=== Create collection `hello_milvus_ntapnew_sc` ===
=== Start inserting entities       ===
Number of entities in hello_milvus_ntapnew_sc: 9000
root@ip-172-31-29-98:~/pymilvus/examples# find
/home/ubuntu/milvusvectordb/
…
<removed content to save page space >
…
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/103/4487898457
91411923/b3def25f-c117-4fba-8256-96cb7557cd6c
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/103/4487898457
91411923/b3def25f-c117-4fba-8256-96cb7557cd6c/part.1
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/103/4487898457
91411923/xl.meta
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/0
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/0/448789845791
411924
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/0/448789845791
411924/xl.meta
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/1
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/1/448789845791
411925
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/1/448789845791
411925/xl.meta
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/100
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/100/4487898457
```

```
91411920
/home/ubuntu/milvusvectordb/volumes/minio/a-bucket/files/insert_log
/448789845791611912/448789845791611913/448789845791611939/100/4487898457
91411920/xl.meta
```

13. Check the reading operation using the verify_data_netapp.py script.

```
root@ip-172-31-29-98:~/pymilvus/examples# python3 verify_data_netapp.py
=== start connecting to Milvus     ===

=== Milvus host: localhost         ===

Does collection hello_milvus_ntapnew_sc exist in Milvus: True
{'auto_id': False, 'description': 'hello_milvus_ntapnew_sc', 'fields':
[{'name': 'pk', 'description': '', 'type': <DataType.INT64: 5>,
'is_primary': True, 'auto_id': False}, {'name': 'random', 'description':
'', 'type': <DataType.DOUBLE: 11>}, {'name': 'var', 'description': '',
'type': <DataType.VARCHAR: 21>, 'params': {'max_length': 65535}},
{'name': 'embeddings', 'description': '', 'type': <DataType.
FLOAT_VECTOR: 101>, 'params': {'dim': 8}}], 'enable_dynamic_field':
False}
Number of entities in Milvus: hello_milvus_ntapnew_sc : 9000

=== Start Creating index IVF_FLAT  ===


=== Start loading                  ===


=== Start searching based on vector similarity ===

hit: id: 2248, distance: 0.0, entity: {'random': 0.2777646777746381},
random field: 0.2777646777746381
hit: id: 4837, distance: 0.07805602252483368, entity: {'random':
0.6451650959930306}, random field: 0.6451650959930306
hit: id: 7172, distance: 0.07954417169094086, entity: {'random':
0.6141351712303128}, random field: 0.6141351712303128
hit: id: 2249, distance: 0.0, entity: {'random': 0.7434908973629817},
random field: 0.7434908973629817
hit: id: 830, distance: 0.05628090724349022, entity: {'random':
0.8544487225667627}, random field: 0.8544487225667627
hit: id: 8562, distance: 0.07971227169036865, entity: {'random':
0.4464554280115878}, random field: 0.4464554280115878
search latency = 0.1266s

=== Start querying with `random > 0.5` ===
```

```
query result:
-{'random': 0.6378742006852851, 'embeddings': [0.3017092, 0.74452263,
0.8009826, 0.4927033, 0.12762444, 0.29869467, 0.52859956, 0.23734547],
'pk': 0}
search latency = 0.3294s

=== Start hybrid searching with `random > 0.5` ===

hit: id: 4837, distance: 0.07805602252483368, entity: {'random':
0.6451650959930306}, random field: 0.6451650959930306
hit: id: 7172, distance: 0.07954417169094086, entity: {'random':
0.6141351712303128}, random field: 0.6141351712303128
hit: id: 515, distance: 0.09590047597885132, entity: {'random':
0.8013175797590888}, random field: 0.8013175797590888
hit: id: 2249, distance: 0.0, entity: {'random': 0.7434908973629817},
random field: 0.7434908973629817
hit: id: 830, distance: 0.05628090724349022, entity: {'random':
0.8544487225667627}, random field: 0.8544487225667627
hit: id: 1627, distance: 0.08096684515476227, entity: {'random':
0.9302397069516164}, random field: 0.9302397069516164
search latency = 0.2674s
Does collection hello_milvus_ntapnew_sc2 exist in Milvus: True
{'auto_id': True, 'description': 'hello_milvus_ntapnew_sc2', 'fields':
[{'name': 'pk', 'description': '', 'type': <DataType.INT64: 5>,
'is_primary': True, 'auto_id': True}, {'name': 'random', 'description':
'', 'type': <DataType.DOUBLE: 11>}, {'name': 'var', 'description': '',
'type': <DataType.VARCHAR: 21>, 'params': {'max_length': 65535}},
{'name': 'embeddings', 'description': '', 'type': <DataType.
FLOAT_VECTOR: 101>, 'params': {'dim': 8}}], 'enable_dynamic_field':
False}
```

14. If the customer wants to access (read) NFS data tested in the vector database via the S3 protocol for AI workloads, this can be validated using a straightforward Python program. An example of this could be a similarity search of images from another application as mentioned in the picture that is in the beginning of this section.

```
root@ip-172-31-29-98:~/pymilvus/examples# sudo python3
/home/ubuntu/milvusvectordb/s3_access.py -i 172.31.255.228 --bucket
milvusnasvol --access-key PY6UF318996I86NBYNDD --secret-key
hoPctr9aD88c1j0SkIYZ2uPa03vlbqKA0c5feK6F
OBJECTS in the bucket milvusnasvol are :
***************************************
…
<output content removed to save page space>
…
```

```
bucket/files/insert_log/448789845791611912/448789845791611913/4487898457
91611920/0/448789845791411917/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/1/448789845791411918/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/100/448789845791411913/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/101/448789845791411914/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/102/448789845791411915/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/103/448789845791411916/1c48ab6e-
1546-4503-9084-28c629216c33/part.1
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611920/103/448789845791411916/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/0/448789845791411924/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/1/448789845791411925/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/100/448789845791411920/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/101/448789845791411921/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/102/448789845791411922/xl.meta
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/103/448789845791411923/b3def25f-
c117-4fba-8256-96cb7557cd6c/part.1
volumes/minio/a-bucket/files/insert_log/448789845791611912
/448789845791611913/448789845791611939/103/448789845791411923/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791211880
/448789845791211881/448789845791411889/100/1/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791211880
/448789845791211881/448789845791411889/100/448789845791411912/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791611912
/448789845791611913/448789845791611920/100/1/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791611912
/448789845791611913/448789845791611920/100/448789845791411919/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791611912
/448789845791611913/448789845791611939/100/1/xl.meta
volumes/minio/a-bucket/files/stats_log/448789845791611912
/448789845791611913/448789845791611939/100/448789845791411926/xl.meta
***************************************
root@ip-172-31-29-98:~/pymilvus/examples#
```

This section effectively demonstrates how customers can deploy and operate a standalone Milvus setup within Docker containers, utilizing Amazon's NetApp FSxN for NetApp ONTAP data storage. This setup allows customers to leverage the power of vector databases for handling high-dimensional data and executing complex queries, all within the scalable and efficient environment of Docker containers. By creating an Amazon FSxN for NetApp ONTAP instance and matching EC2 instance, customers can ensure optimal resource utilization and data management. The successful validation of data writing and reading operations from FSxN in the vector database provides customers with the assurance of reliable and consistent data operations. Additionally, the ability to list (read) data from AI workloads via the S3 protocol offers enhanced data accessibility. This comprehensive process, therefore, provides customers with a robust and efficient solution for managing their large-scale data operations, leveraging the capabilities of Amazon's FSxN for NetApp ONTAP.

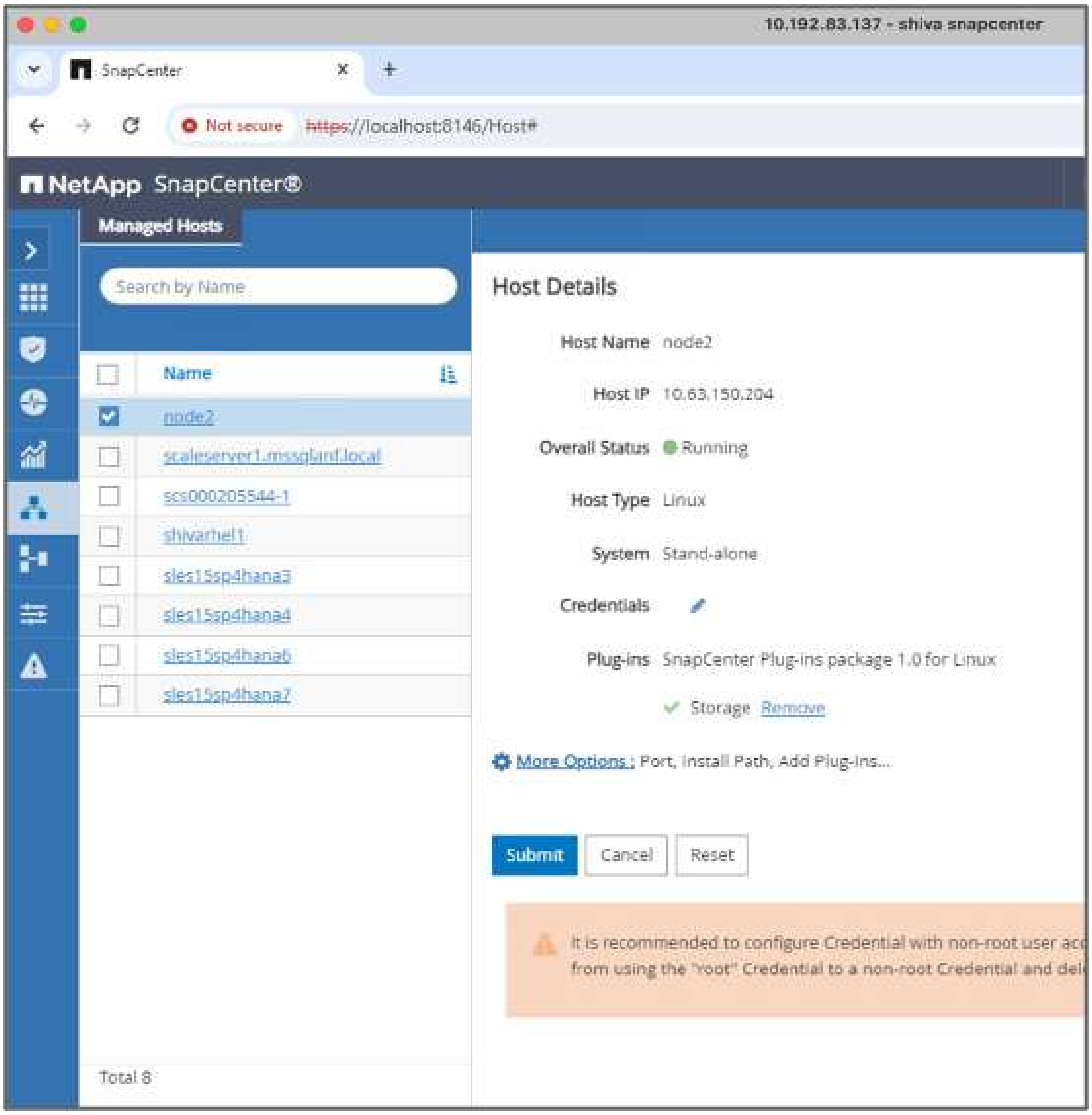
Vector Database Protection using SnapCenter

This section describes how to provide data protection for the vector database using NetApp SnapCenter.

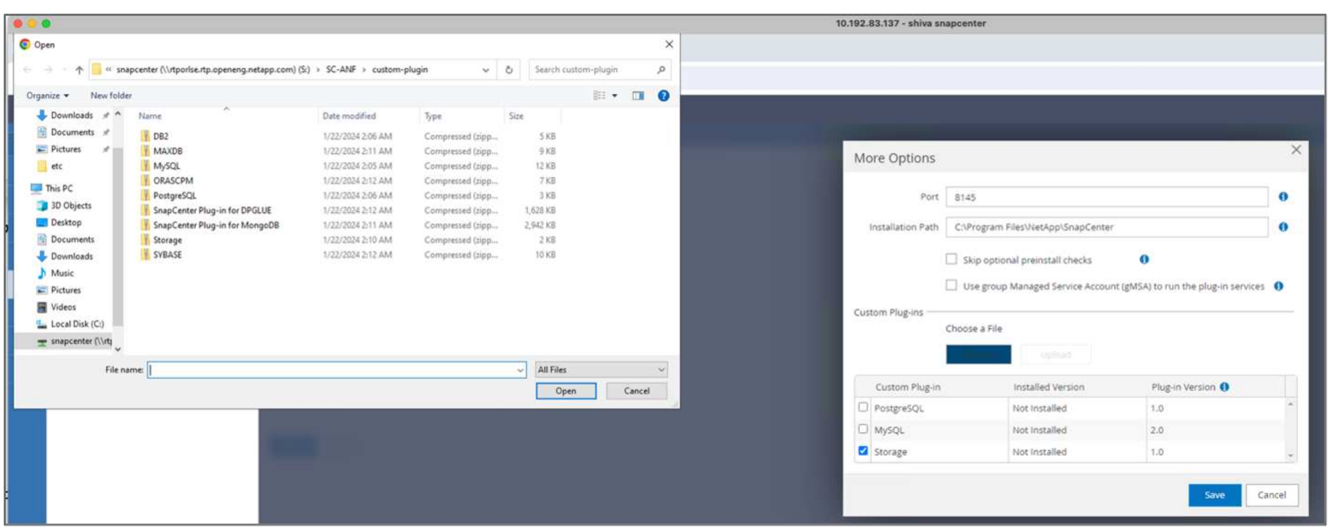
Vector database protection using NetApp SnapCenter.

For example, in the film production industry, customers often possess critical embedded data such as video and audio files. Loss of this data, due to issues like hard drive failures, can have a significant impact on their operations, potentially jeopardizing multimillion-dollar ventures. We have encountered instances where invaluable content was lost, causing substantial disruption and financial loss. Ensuring the security and integrity of this essential data is therefore of paramount importance in this industry.
In this section, we delve into how SnapCenter safeguards the vector database data and Milvus data residing in ONTAP. For this example, we utilized a NAS bucket (milvusdbvol1) derived from an NFS ONTAP volume (vol1) for customer data, and a separate NFS volume (vectordbpv) for Milvus cluster configuration data. please check the here for the snapcenter backup workflow

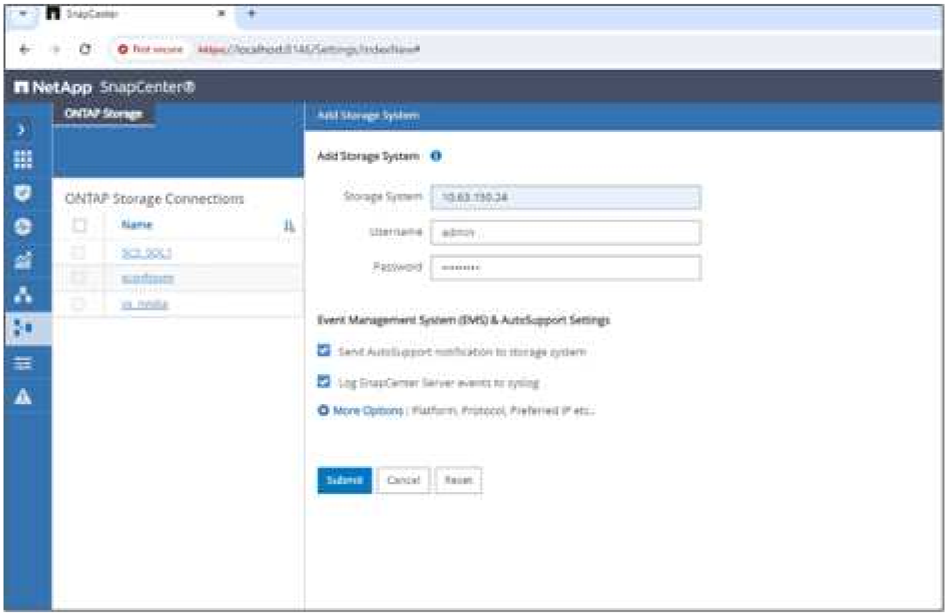
1. Set up the host that will be used to execute SnapCenter commands.



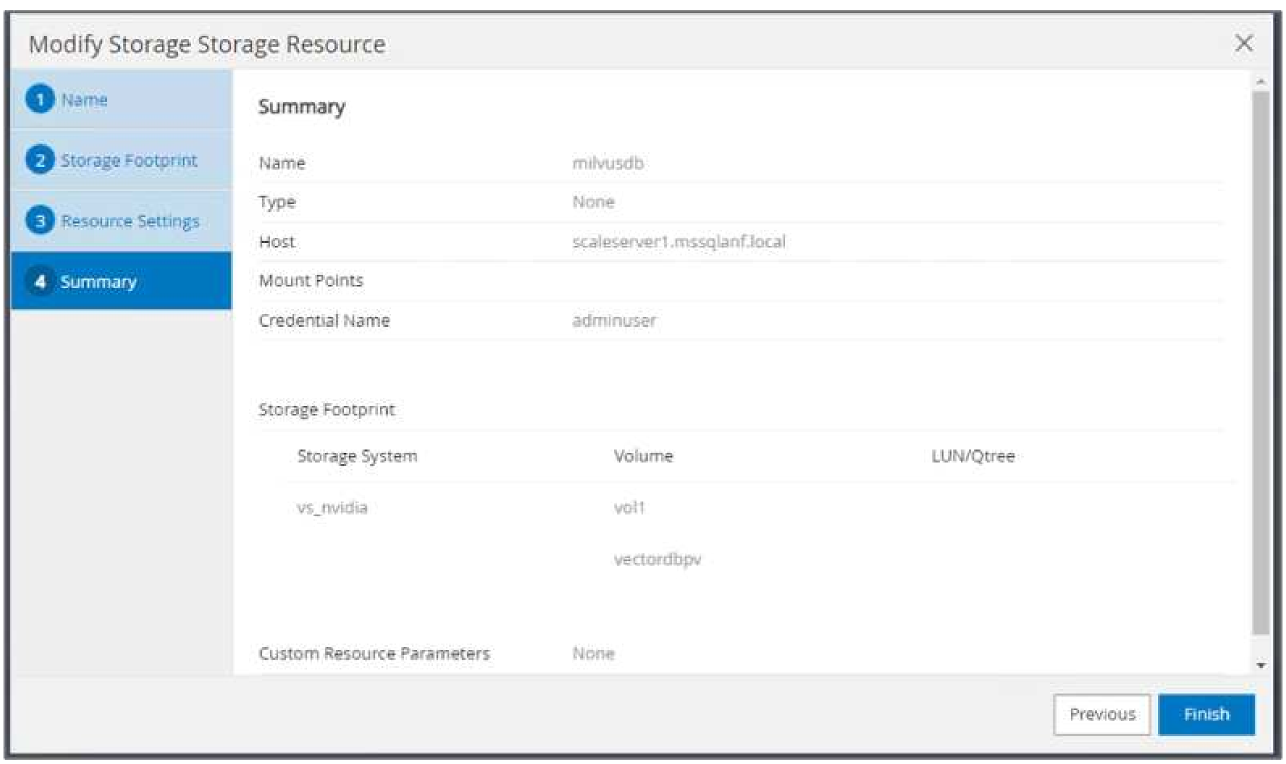
2. Install and configure the storage plugin. From the added host, select "More Options". Navigate to and select the downloaded storage plugin from the NetApp Automation Store. Install the plugin and save the configuration.



3. Set up the storage system and volume: Add the storage system under "Storage System" and select the SVM (Storage Virtual Machine). In this example, we've chosen "vs_nvidia".



4. Establish a resource for the vector database, incorporating a backup policy and a custom snapshot name.
   - Enable Consistency Group Backup with default values and enable SnapCenter without filesystem consistency.
   - In the Storage Footprint section, select the volumes associated with the vector database customer data and Milvus cluster data. In our example, these are "vol1" and "vectordbpv".
   - Create policy for vector database protection and protect vector database resource using the policy.



5. Insert data into the S3 NAS bucket using a Python script. In our case, we modified the backup script provided by Milvus, namely 'prepare_data_netapp.py', and executed the 'sync' command to flush the data

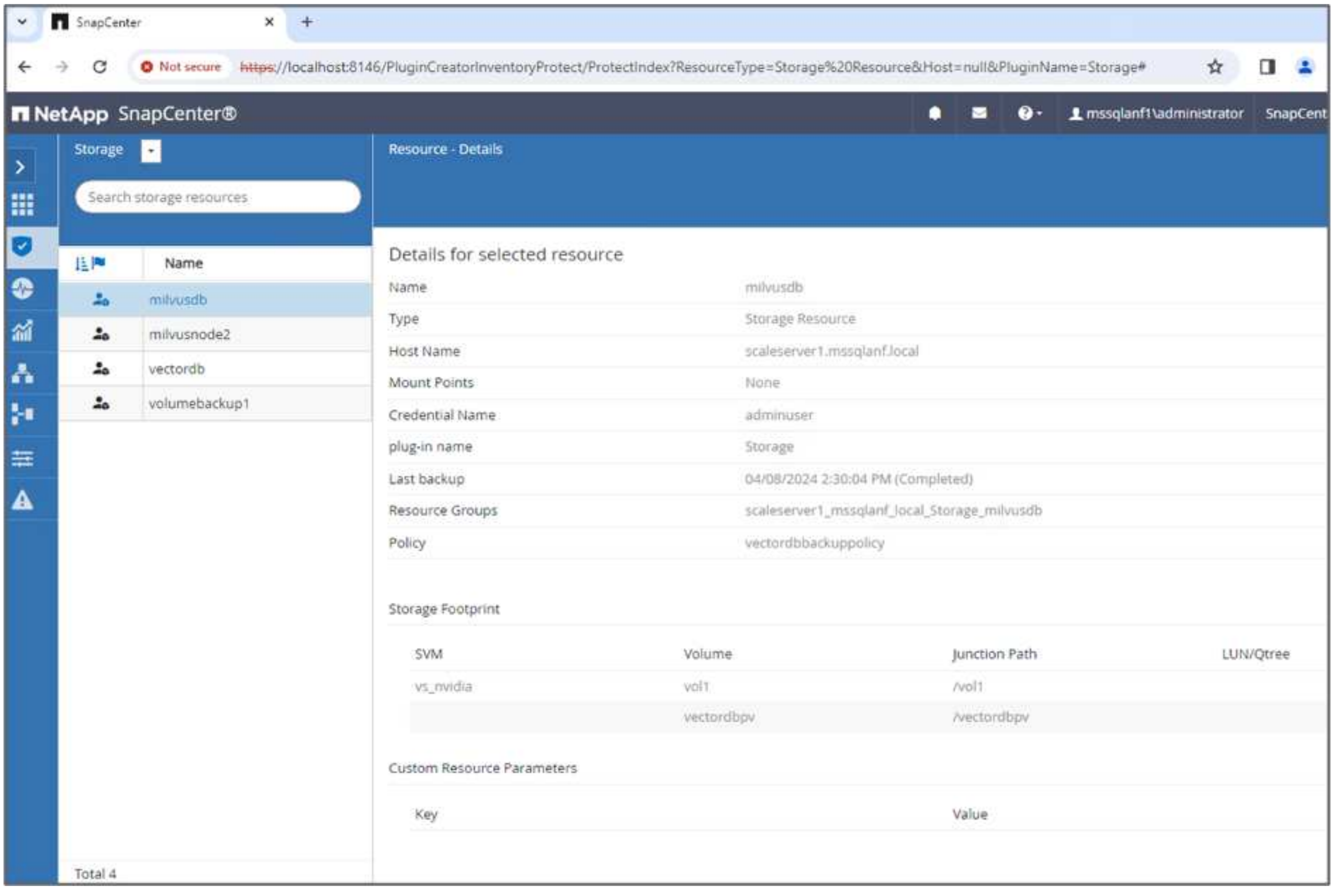
from the operating system.

```
root@node2:~# python3 prepare_data_netapp.py

=== start connecting to Milvus     ===


=== Milvus host: localhost         ===

Does collection hello_milvus_netapp_sc_test exist in Milvus: False

=== Create collection `hello_milvus_netapp_sc_test` ===


=== Start inserting entities       ===

Number of entities in hello_milvus_netapp_sc_test: 3000

=== Create collection `hello_milvus_netapp_sc_test2` ===

Number of entities in hello_milvus_netapp_sc_test2: 6000
root@node2:~# for i in 2 3 4 5 6   ; do ssh node$i "hostname; sync; echo
'sync executed';" ; done
node2
sync executed
node3
sync executed
node4
sync executed
node5
sync executed
node6
sync executed
root@node2:~#
```

6. Verify the data in the S3 NAS bucket. In our example, the files with the timestamp '2024-04-08 21:22' were created by the 'prepare_data_netapp.py' script.

```
root@node2:~# aws s3 ls --profile ontaps3  s3://milvusdbvol1/
--recursive | grep '2024-04-08'

<output content removed to save page space>
2024-04-08 21:18:14       5656
stats_log/448950615991000809/448950615991000810/448950615991001854/100/1
2024-04-08 21:18:12       5654
stats_log/448950615991000809/448950615991000810/448950615991001854/100/4
48950615990800869
2024-04-08 21:18:17       5656
stats_log/448950615991000809/448950615991000810/448950615991001872/100/1
2024-04-08 21:18:15       5654
stats_log/448950615991000809/448950615991000810/448950615991001872/100/4
48950615990800876
2024-04-08 21:22:46       5625
stats_log/448950615991003377/448950615991003378/448950615991003385/100/1
2024-04-08 21:22:45       5623
stats_log/448950615991003377/448950615991003378/448950615991003385/100/4
48950615990800899
2024-04-08 21:22:49       5656
stats_log/448950615991003408/448950615991003409/448950615991003416/100/1
2024-04-08 21:22:47       5654
stats_log/448950615991003408/448950615991003409/448950615991003416/100/4
48950615990800906
2024-04-08 21:22:52       5656
stats_log/448950615991003408/448950615991003409/448950615991003434/100/1
2024-04-08 21:22:50       5654
stats_log/448950615991003408/448950615991003409/448950615991003434/100/4
48950615990800913
root@node2:~#
```

7. Initiate a backup using the Consistency Group (CG) snapshot from the 'milvusdb' resource

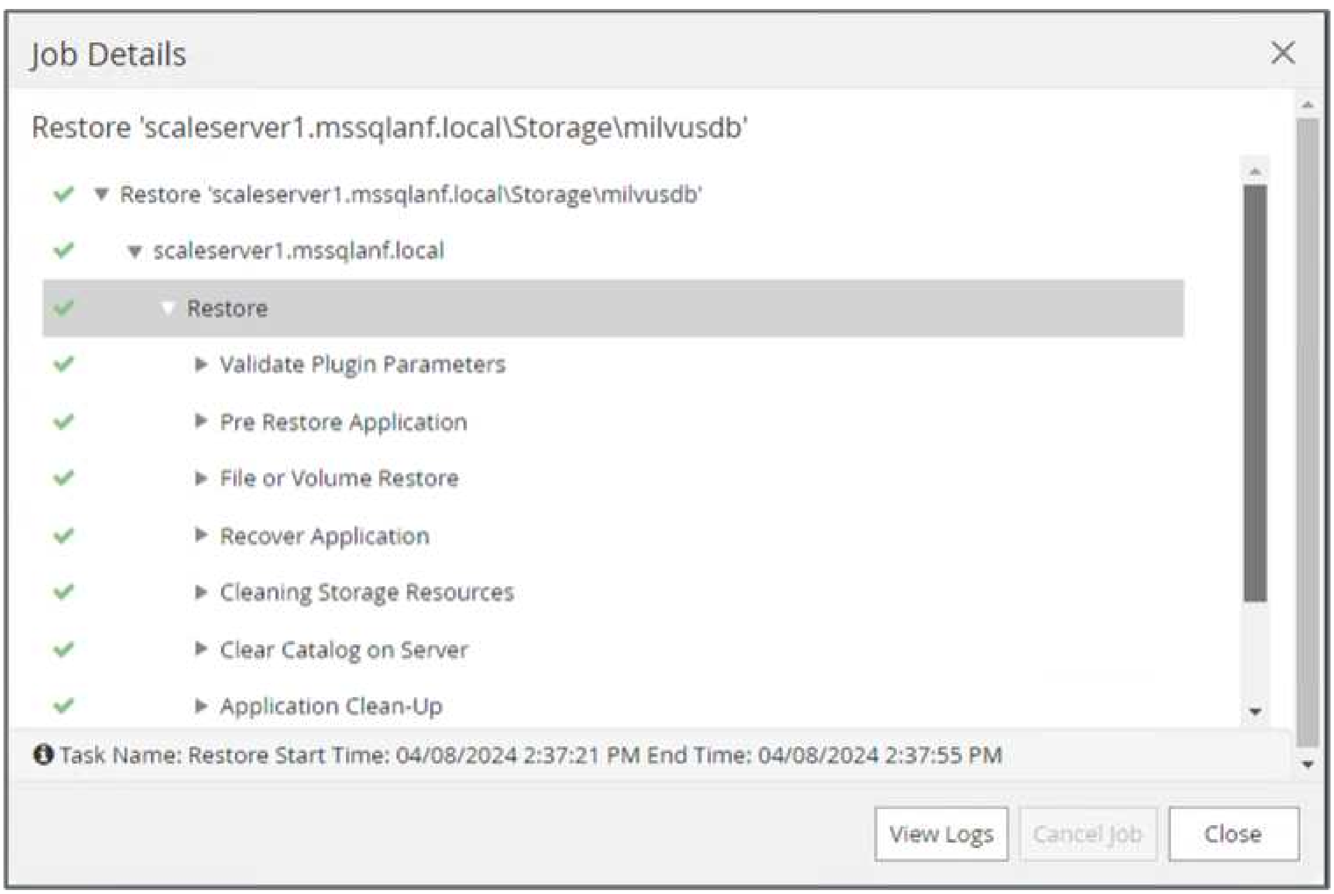


8. To test the backup functionality, we either added a new table after the backup process or removed some data from the NFS (S3 NAS bucket).

   For this test, imagine a scenario where someone created a new, unnecessary, or inappropriate collection after the backup. In such a case, we would need to revert the vector database to its state before the new collection was added. For instance, new collections such as 'hello_milvus_netapp_sc_testnew' and 'hello_milvus_netapp_sc_testnew2' have been inserted.

```
root@node2:~# python3 prepare_data_netapp.py

=== start connecting to Milvus     ===


=== Milvus host: localhost         ===

Does collection hello_milvus_netapp_sc_testnew exist in Milvus: False

=== Create collection `hello_milvus_netapp_sc_testnew` ===


=== Start inserting entities       ===

Number of entities in hello_milvus_netapp_sc_testnew: 3000

=== Create collection `hello_milvus_netapp_sc_testnew2` ===

Number of entities in hello_milvus_netapp_sc_testnew2: 6000
root@node2:~#
```

9. Execute a full restore of the S3 NAS bucket from the previous snapshot.



10. Use a Python script to verify the data from the 'hello_milvus_netapp_sc_test' and 'hello_milvus_netapp_sc_test2' collections.

```
root@node2:~# python3 verify_data_netapp.py

=== start connecting to Milvus     ===


=== Milvus host: localhost         ===

Does collection hello_milvus_netapp_sc_test exist in Milvus: True
{'auto_id': False, 'description': 'hello_milvus_netapp_sc_test',
'fields': [{'name': 'pk', 'description': '', 'type': <DataType.INT64: 5
>, 'is_primary': True, 'auto_id': False}, {'name': 'random',
'description': '', 'type': <DataType.DOUBLE: 11>}, {'name': 'var',
'description': '', 'type': <DataType.VARCHAR: 21>, 'params':
{'max_length': 65535}}, {'name': 'embeddings', 'description': '',
'type': <DataType.FLOAT_VECTOR: 101>, 'params': {'dim': 8}}]}
Number of entities in Milvus: hello_milvus_netapp_sc_test : 3000

=== Start Creating index IVF_FLAT  ===


=== Start loading                  ===


=== Start searching based on vector similarity ===

hit: id: 2998, distance: 0.0, entity: {'random': 0.9728033590489911},
random field: 0.9728033590489911
hit: id: 1262, distance: 0.08883658051490784, entity: {'random':
0.2978858685751561}, random field: 0.2978858685751561
hit: id: 1265, distance: 0.09590047597885132, entity: {'random':
0.3042039939240304}, random field: 0.3042039939240304
hit: id: 2999, distance: 0.0, entity: {'random': 0.02316334456872482},
random field: 0.02316334456872482
hit: id: 1580, distance: 0.05628091096878052, entity: {'random':
0.3855988746044062}, random field: 0.3855988746044062
hit: id: 2377, distance: 0.08096685260534286, entity: {'random':
0.8745922204004368}, random field: 0.8745922204004368
search latency = 0.2832s

=== Start querying with `random > 0.5` ===

query result:
-{'random': 0.6378742006852851, 'embeddings': [0.20963514, 0.39746657,
```

```
0.12019053, 0.6947492, 0.9535575, 0.5454552, 0.82360446, 0.21096309],
'pk': 0}
search latency = 0.2257s

=== Start hybrid searching with `random > 0.5` ===

hit: id: 2998, distance: 0.0, entity: {'random': 0.9728033590489911},
random field: 0.9728033590489911
hit: id: 747, distance: 0.14606499671936035, entity: {'random':
0.5648774800635661}, random field: 0.5648774800635661
hit: id: 2527, distance: 0.1530652642250061, entity: {'random':
0.8928974315571507}, random field: 0.8928974315571507
hit: id: 2377, distance: 0.08096685260534286, entity: {'random':
0.8745922204004368}, random field: 0.8745922204004368
hit: id: 2034, distance: 0.20354536175727844, entity: {'random':
0.5526117606328499}, random field: 0.5526117606328499
hit: id: 958, distance: 0.21908017992973328, entity: {'random':
0.6647383716417955}, random field: 0.6647383716417955
search latency = 0.5480s
Does collection hello_milvus_netapp_sc_test2 exist in Milvus: True
{'auto_id': True, 'description': 'hello_milvus_netapp_sc_test2',
'fields': [{'name': 'pk', 'description': '', 'type': <DataType.INT64: 5
>, 'is_primary': True, 'auto_id': True}, {'name': 'random',
'description': '', 'type': <DataType.DOUBLE: 11>}, {'name': 'var',
'description': '', 'type': <DataType.VARCHAR: 21>, 'params':
{'max_length': 65535}}, {'name': 'embeddings', 'description': '',
'type': <DataType.FLOAT_VECTOR: 101>, 'params': {'dim': 8}}]}
Number of entities in Milvus: hello_milvus_netapp_sc_test2 : 6000

=== Start Creating index IVF_FLAT  ===


=== Start loading                  ===



=== Start searching based on vector similarity ===

hit: id: 448950615990642008, distance: 0.07805602252483368, entity:
{'random': 0.5326684390871348}, random field: 0.5326684390871348
hit: id: 448950615990645009, distance: 0.07805602252483368, entity:
{'random': 0.5326684390871348}, random field: 0.5326684390871348
hit: id: 448950615990640618, distance: 0.13562293350696564, entity:
{'random': 0.7864676926688837}, random field: 0.7864676926688837
hit: id: 448950615990642314, distance: 0.10414951294660568, entity:
{'random': 0.2209597460821181}, random field: 0.2209597460821181
hit: id: 448950615990645315, distance: 0.10414951294660568, entity:
```

```
{'random': 0.2209597460821181}, random field: 0.2209597460821181
hit: id: 448950615990640004, distance: 0.11571306735277176, entity:
{'random': 0.7765521996186631}, random field: 0.7765521996186631
search latency = 0.2381s

=== Start querying with `random > 0.5` ===

query result:
-{'embeddings': [0.15983285, 0.72214717, 0.7414838, 0.44471496,
0.50356466, 0.8750043, 0.316556, 0.7871702], 'pk': 448950615990639798,
'random': 0.7820620141382767}
search latency = 0.3106s

=== Start hybrid searching with `random > 0.5` ===

hit: id: 448950615990642008, distance: 0.07805602252483368, entity:
{'random': 0.5326684390871348}, random field: 0.5326684390871348
hit: id: 448950615990645009, distance: 0.07805602252483368, entity:
{'random': 0.5326684390871348}, random field: 0.5326684390871348
hit: id: 448950615990640618, distance: 0.13562293350696564, entity:
{'random': 0.7864676926688837}, random field: 0.7864676926688837
hit: id: 448950615990640004, distance: 0.11571306735277176, entity:
{'random': 0.7765521996186631}, random field: 0.7765521996186631
hit: id: 448950615990643005, distance: 0.11571306735277176, entity:
{'random': 0.7765521996186631}, random field: 0.7765521996186631
hit: id: 448950615990640402, distance: 0.13665105402469635, entity:
{'random': 0.9742541034109935}, random field: 0.9742541034109935
search latency = 0.4906s
root@node2:~#
```

11. Verify that the unnecessary or inappropriate collection is no longer present in the database.

```
root@node2:~# python3 verify_data_netapp.py

=== start connecting to Milvus     ===


=== Milvus host: localhost         ===

Does collection hello_milvus_netapp_sc_testnew exist in Milvus: False
Traceback (most recent call last):
  File "/root/verify_data_netapp.py", line 37, in <module>
    recover_collection = Collection(recover_collection_name)
  File "/usr/local/lib/python3.10/dist-
packages/pymilvus/orm/collection.py", line 137, in __init__
    raise SchemaNotReadyException(
pymilvus.exceptions.SchemaNotReadyException: <SchemaNotReadyException:
(code=1, message=Collection 'hello_milvus_netapp_sc_testnew' not exist,
or you can pass in schema to create one.)>
root@node2:~#
```

In conclusion, the use of NetApp's SnapCenter to safeguard vector database data and Milvus data residing in ONTAP offers significant benefits to customers, particularly in industries where data integrity is paramount, such as film production. SnapCenter's ability to create consistent backups and perform full data restores ensures that critical data, such as embedded video and audio files, are protected against loss due to hard drive failures or other issues. This not only prevents operational disruption but also safeguards against substantial financial loss.

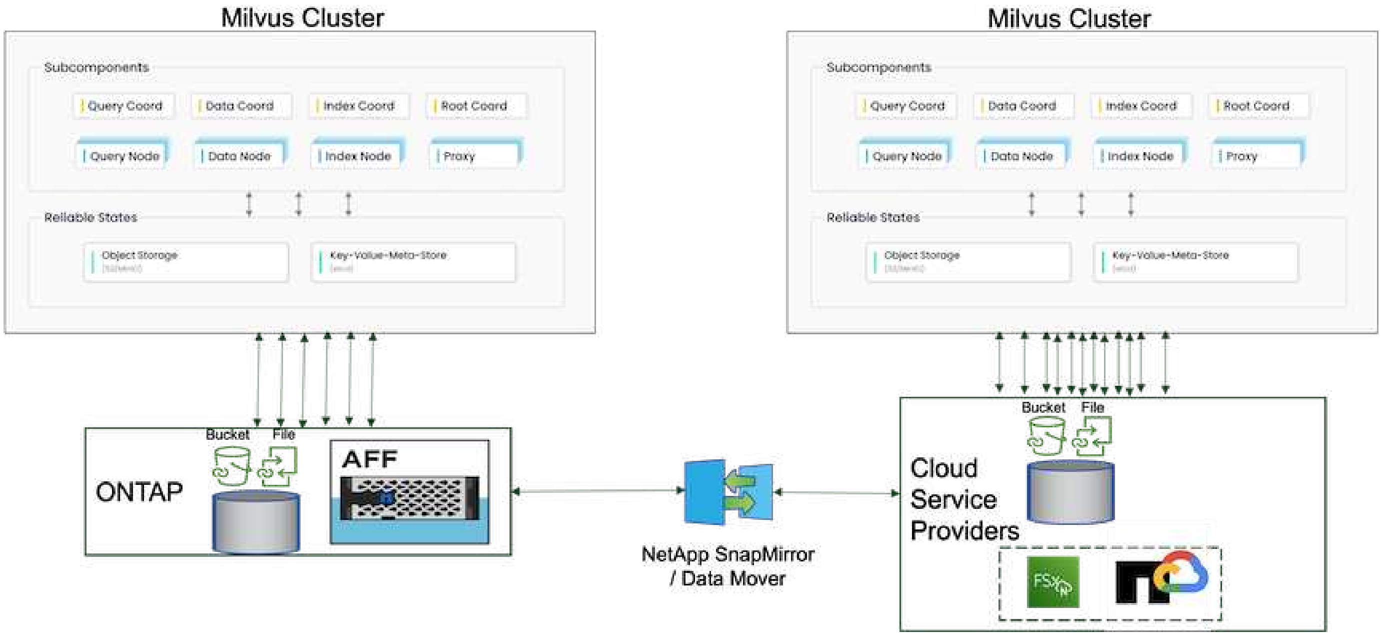
In this section, we demonstrated how SnapCenter can be configured to protect data residing in ONTAP, including the setup of hosts, installation and configuration of storage plugins, and the creation of a resource for the vector database with a custom snapshot name. We also showcased how to perform a backup using the Consistency Group snapshot and verify the data in the S3 NAS bucket.

Furthermore, we simulated a scenario where an unnecessary or inappropriate collection was created after the backup. In such cases, SnapCenter's ability to perform a full restore from a previous snapshot ensures that the vector database can be reverted to its state before the addition of the new collection, thus maintaining the integrity of the database. This capability to restore data to a specific point in time is invaluable for customers, providing them with the assurance that their data is not only secure but also correctly maintained. Thus, NetApp's SnapCenter product offers customers a robust and reliable solution for data protection and management.

Disaster Recovery using NetApp SnapMirror

This section discusses DR (disaster recovery) with SnapMirror for the vector database solution for NetApp.

Disaster Recovery using NetApp SnapMirror



Disaster recovery is crucial for maintaining the integrity and availability of a vector database, especially given its role in managing high-dimensional data and executing complex similarity searches. A well-planned and implemented disaster recovery strategy ensures that data is not lost or compromised in the event of unforeseen incidents, such as hardware failures, natural disasters, or cyber-attacks. This is particularly significant for applications relying on vector databases, where the loss or corruption of data could lead to significant operational disruptions and financial losses. Moreover, a robust disaster recovery plan also ensures business continuity by minimizing downtime and allowing for the quick restoration of services. This is achieved through NetApp data replication product SnapMirrror across different geographical locations, regular backups, and failover mechanisms. Therefore, disaster recovery is not just a protective measure, but a critical component of responsible and efficient vector database management.

NetApp's SnapMirror provides data replication from one NetApp ONTAP storage controller to another, primarily used for disaster recovery (DR) and hybrid solutions. In the context of a vector database, this tool facilitates the smooth transition of data between on-premises and cloud environments. This transition is achieved without necessitating any data conversions or application refactoring, thereby enhancing the efficiency and flexibility of data management across multiple platforms.

NetApp Hybrid solution in a vector database scenario can bring about more advantages:

1. Scalability: NetApp's hybrid cloud solution offers the ability to scale your resources as per your requirements. You can utilize on-premises resources for regular, predictable workloads and cloud resources such as Amazon FSxN for NetApp ONTAP and Google Cloud NetApp Volume (GCNV) for peak times or unexpected loads.
2. Cost Efficiency: NetApp's hybrid cloud model allows you to optimize your costs by using on-premises resources for regular workloads and only paying for cloud resources when you need them. This pay-as-you-go model can be quite cost-effective with a NetApp instaclustr service offering. For on-prem and major cloud service providers, instaclustr provids support and consultation.
3. Flexibility: NetApp's hybrid cloud gives you the flexibility to choose where to process your data. For example, you might choose to perform complex vector operations on-premises where you have more powerful hardware, and less intensive operations in the cloud.
4. Business Continuity: In the event of a disaster, having your data in a NetApp hybrid cloud can ensure business continuity. You can quickly switch to the cloud if your on-premises resources are affected. We can leverage NetApp SnapMirror to move the data from on-prem to cloud and vice versa.

5. Innovation: NetApp's hybrid cloud solutions can also enable faster innovation by providing access to cutting-edge cloud services and technologies. NetApp innovations in cloud such as Amazon FSxN for NetApp ONTAP, Azure NetApp Files and Google Cloud NetApp Volumes are cloud service providers innovative products and preferred NAS.

Vector Database Performance Validation

This section highlights the performance validation that was performed on the vector database.

Performance validation

Performance validation plays a critical role in both vector databases and storage systems, serving as a key factor in ensuring optimal operation and efficient resource utilization. Vector databases, known for handling high-dimensional data and executing similarity searches, need to maintain high performance levels to process complex queries swiftly and accurately. Performance validation helps identify bottlenecks, fine-tune configurations, and ensure the system can handle expected loads without degradation in service. Similarly, in storage systems, performance validation is essential to ensure data is stored and retrieved efficiently, without latency issues or bottlenecks that could impact overall system performance. It also aids in making informed decisions about necessary upgrades or changes in storage infrastructure. Therefore, performance validation is a crucial aspect of system management, contributing significantly to maintaining high service quality, operational efficiency, and overall system reliability.

In this section, we aim to delve into the performance validation of vector databases, such as Milvus and pgvecto.rs, focusing on their storage performance characteristics such as I/O profile and netapp storage controller behavious in support of RAG and inference workloads within the LLM Lifecycle. We will evaluate and identify any performance differentiators when these databases are combined with the ONTAP storage solution. Our analysis will be based on key performance indicators, such as the number of queries processed per second(QPS).

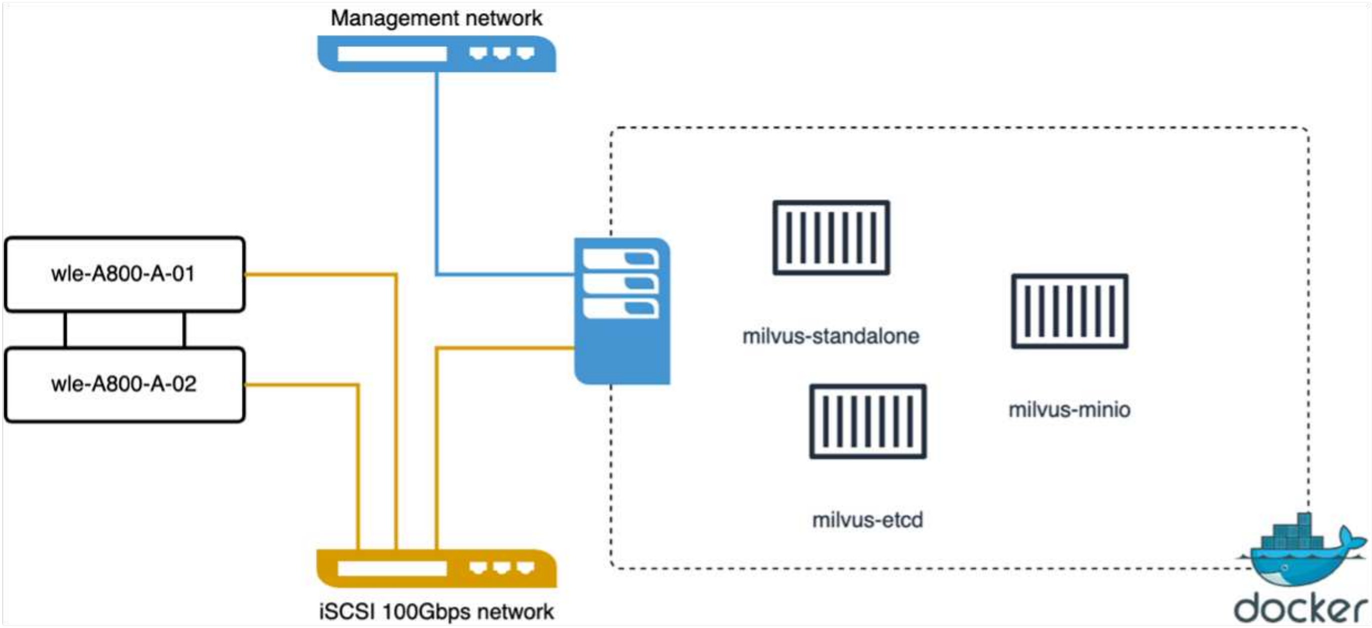
Please check the methodology used for milvus and progress below.

Details	Milvus (Standalone and Cluster)	Postgres(pgvecto.rs) #
version	2.3.2	0.2.0
Filesystem	XFS on iSCSI LUNs	
Workload Generator	VectorDB-Bench – v0.0.5	
Datasets	LAION Dataset * 10Million Embeddings * 768 Dimensions * ~300GB dataset size	
Storage controller	AFF 800 * Version – 9.14.1 * 4 x 100GbE – for milvus and 2x 100GbE for postgres * iscsi	

VectorDB-Bench with Milvus standalone cluster

we did the following performance validation on milvus standalone cluster with vectorDB-Bench.

The network and server connectivity of the milvus standalone cluster is below.



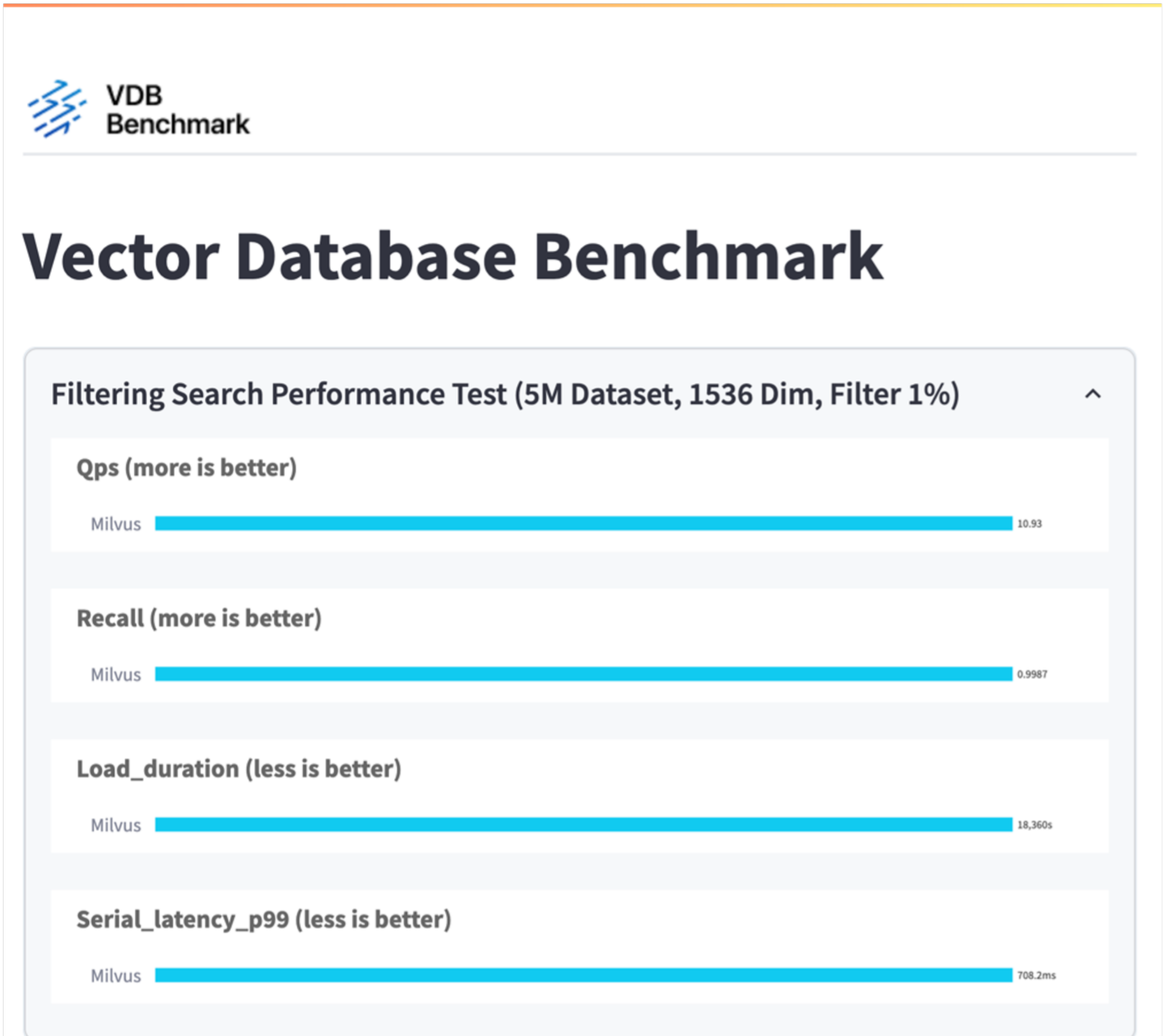
In this section, we share our observations and results from testing the Milvus standalone database.
. We selected DiskANN as the index type for these tests.
. Ingesting, optimizing, and creating indexes for a dataset of approximately 100GB took around 5 hours. For most of this duration, the Milvus server, equipped with 20 cores (which equates to 40 vcpus when Hyper-Threading is enabled), was operating at its maximum CPU capacity of 100%.We found that DiskANN is particularly important for large datasets that exceed the system memory size.
. In the query phase, we observed a Queries per Second (QPS) rate of 10.93 with a recall of 0.9987. The 99th percentile latency for queries was measured at 708.2 milliseconds.

From the storage perspective, the database issued about 1,000 ops/sec during the ingest, post-insert optimization, and index creation phases. In the query phase, it demanded 32,000 ops/sec.

The following section presents the storage performance metrics.

Workload Phase	Metric	Value
Data Ingestion and Post insert optimization	IOPS	< 1,000
	Latency	< 400 usecs
	Workload	Read/Write mix, mostly writes
	IO size	64KB
Query	IOPS	Peak at 32,000
	Latency	< 400 usecs
	Workload	100% cached read
	IO size	Mainly 8KB

The vectorDB-bench result is below.

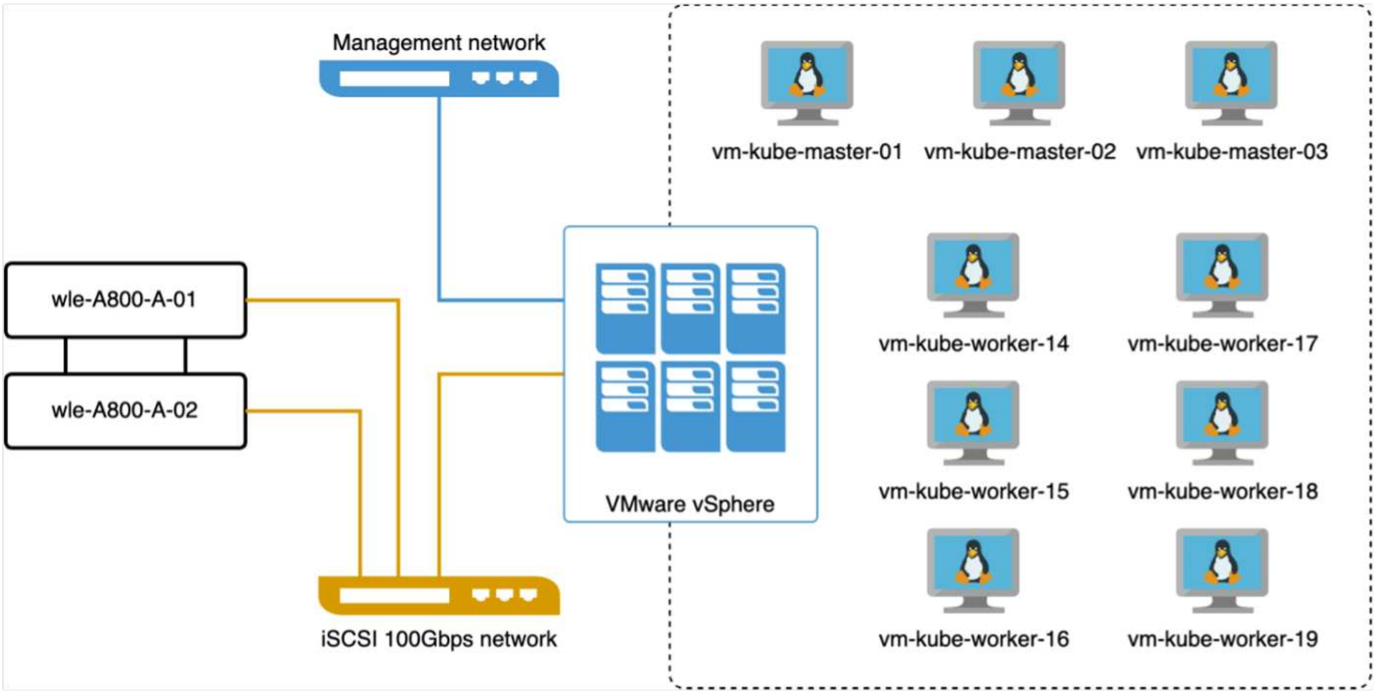


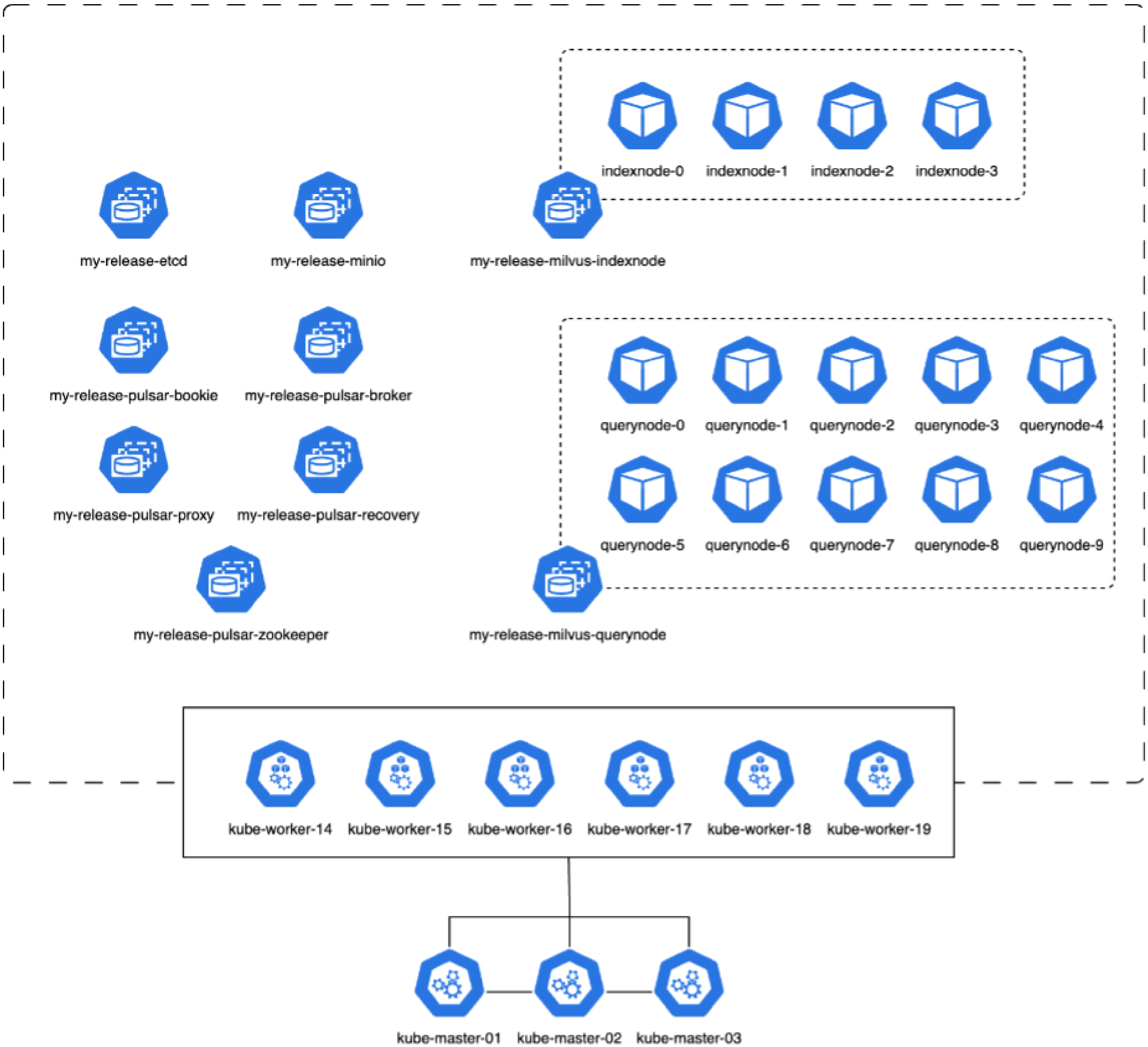
From the performance validation of the standalone Milvus instance, it’s evident that the current setup is insufficient to support a dataset of 5 million vectors with a dimensionality of 1536. we’ve determined that the storage possesses adequate resources and does not constitute a bottleneck in the system.

VectorDB-Bench with milvus cluster

In this section, we discuss the deployment of a Milvus cluster within a Kubernetes environment. This Kubernetes setup was constructed atop a VMware vSphere deployment, which hosted the Kubernetes master and worker nodes.

The details of the VMware vSphere and Kubernetes deployments are presented in the following sections.


Management network
wle-A800-A-01
wle-A800-A-02
VMware vSphere
iSCSI 100Gbps network
vm-kube-master-01
vm-kube-master-02
vm-kube-master-03
vm-kube-worker-14
vm-kube-worker-17
vm-kube-worker-15
vm-kube-worker-18
vm-kube-worker-16
vm-kube-worker-19



In this section, we present our observations and results from testing the Milvus database.
* The index type used was DiskANN.
* The table below provides a comparison between the standalone and cluster deployments when working with 5 million vectors at a dimensionality of 1536. We observed that the time taken for data ingestion and post-insert optimization was lower in the cluster deployment. The 99th percentile latency for queries was reduced by six times in the cluster deployment compared to the standalone setup.
* Although the Queries per Second (QPS) rate was higher in the cluster deployment, it was not at the desired level.

Metric	Milvus Standalone	Milvus Cluster	Difference
QPS @ Recall	10.93 @ 0.9987	18.42 @ 0.9952	+40%
p99 Latency (less is better)	708.2 ms	117.6 ms	-83%
Load Duration time (less is better)	18,360 secs	12,730 secs	-30%

The images below provide a view of various storage metrics, including storage cluster latency and total IOPS

(Input/Output Operations Per Second).

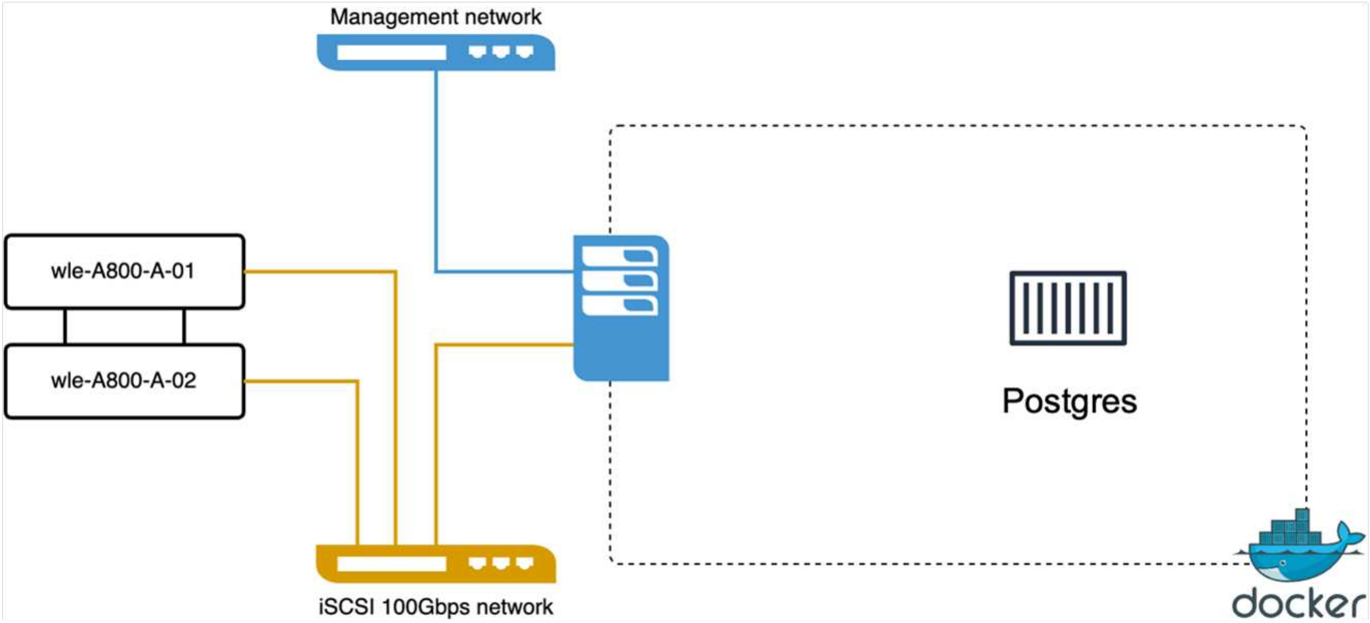
The following section presents the key storage performance metrics.

Workload Phase	Metric	Value
Data Ingestion and Post insert optimization	IOPS	< 1,000
	Latency	< 400 usecs
	Workload	Read/Write mix, mostly writes
	IO size	64KB
Query	IOPS	Peak at 147,000
	Latency	< 400 usecs
	Workload	100% cached read
	IO size	Mainly 8KB

Based on the performance validation of both the standalone Milvus and the Milvus cluster, we present the details of the storage I/O profile.
* We observed that the I/O profile remains consistent across both standalone and cluster deployments.
* The observed difference in peak IOPS can be attributed to the larger number of clients in the cluster deployment.

vectorDB-Bench with Postgres (pgvecto.rs)

We conducted the following actions on PostgreSQL(pgvecto.rs) using VectorDB-Bench:
The details regarding the network and server connectivity of PostgreSQL (specifically, pgvecto.rs) are as follows:



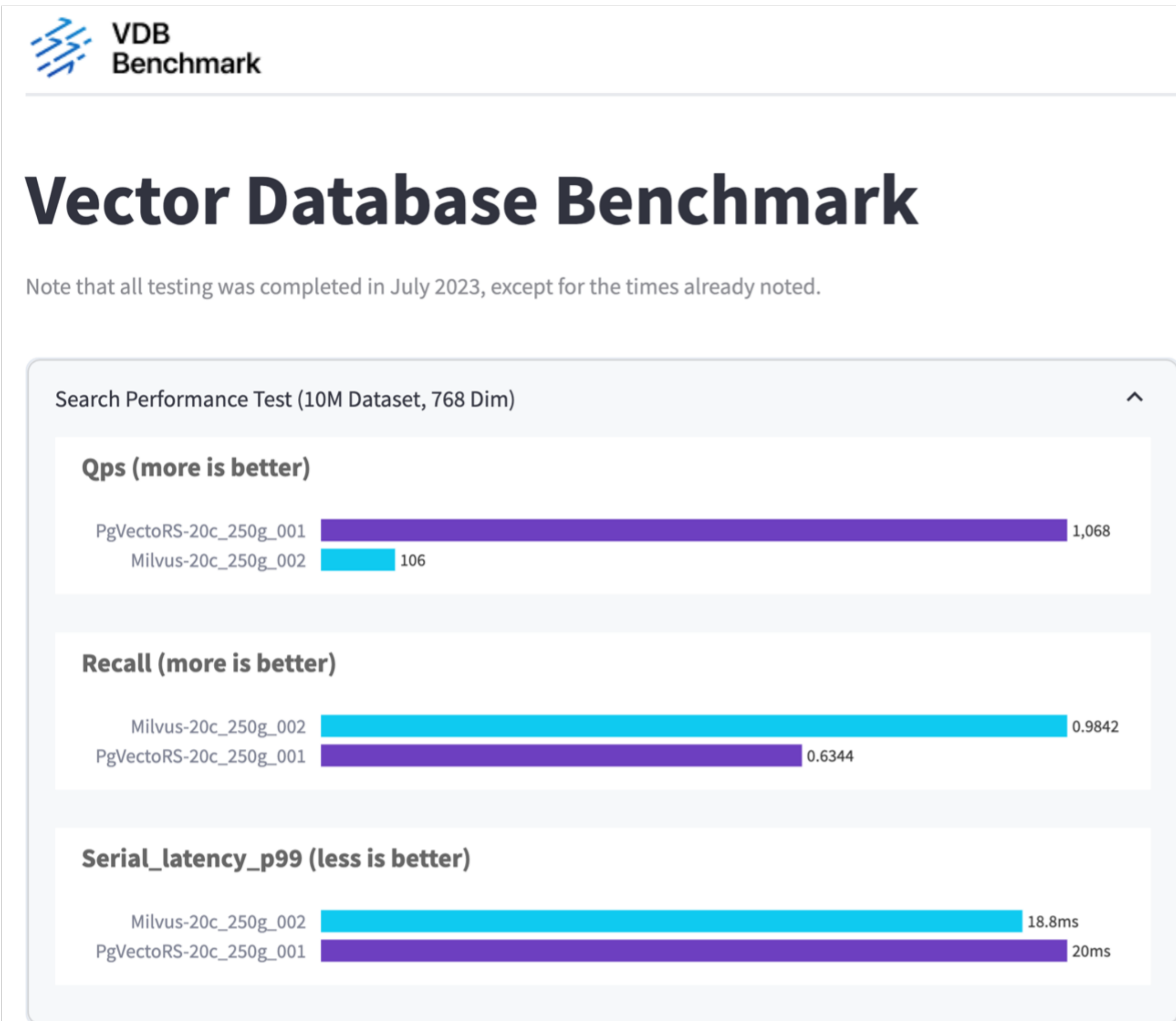
In this section, we share our observations and results from testing the PostgreSQL database, specifically using pgvecto.rs.
* We selected HNSW as the index type for these tests because at the time of testing, DiskANN wasn't available for pgvecto.rs.
* During the data ingestion phase, we loaded the Cohere dataset, which consists of 10 million vectors at a dimensionality of 768. This process took approximately 4.5 hours.
* In the query phase, we observed a Queries per Second (QPS) rate of 1,068 with a recall of 0.6344. The 99th percentile latency for queries was measured at 20 milliseconds. Throughout most of the runtime, the client CPU was operating at 100% capacity.

The images below provide a view of various storage metrics, including storage cluster latency total IOPS (Input/Output Operations Per Second).


```
The following section presents the key storage performance metrics.
```

Workload Phase	Metric	Milvus Standalone	Milvus Cluster
Data Ingestion and Post-Optimization	IOPS	< 1,000	< 1,000
	Latency	< 400 usecs	< 400 usecs
	Workload Mix	Read/Write mix, mostly writes	Read/Write mix, mostly writes
	IO Size	64 KB	64 KB
Query	IOPS	Peak at 32,000	Peak at 147,000
	Latency	< 400 usecs	< 400 usecs
	Workload Mix	100% cache reads	100% cache reads
	IO Size	Mainly 8KB	Mainly 8KB

Performance comparison between milvus and postgres on vector DB Bench



Based on our performance validation of Milvus and PostgreSQL using VectorDBBench, we observed the following:

- Index Type: HNSW

- Dataset: Cohere with 10 million vectors at 768 dimensions

We found that pgvecto.rs achieved a Queries per Second (QPS) rate of 1,068 with a recall of 0.6344, while Milvus achieved a QPS rate of 106 with a recall of 0.9842.

If high precision in your queries is a priority, Milvus outperforms pgvecto.rs as it retrieves a higher proportion of relevant items per query. However, if the number of queries per second is a more crucial factor, pgvecto.rs exceeds Milvus. It's important to note, though, that the quality of the data retrieved via pgvecto.rs is lower, with around 37% of the search results being irrelevant items.

Observation based on our performance validations:

Based on our performance validations, we have made the following observations:

In Milvus, the I/O profile closely resembles an OLTP workload, such as that seen with Oracle SLOB. The benchmark consists of three phases: Data Ingestion, Post-Optimization, and Query. The initial stages are primarily characterized by 64KB write operations, while the query phase predominantly involves 8KB reads. We expect ONTAP to handle the Milvus I/O load proficiently.

The PostgreSQL I/O profile does not present a challenging storage workload. Given the in-memory implementation currently in progress, we didn't observe any disk I/O during the query phase.

DiskANN emerges as a crucial technology for storage differentiation. It enables the efficient scaling of vector DB search beyond the system memory boundary. However, it's unlikely to establish storage performance differentiation with in-memory vector DB indices such as HNSW.

It's also worth noting that storage does not play a critical role during the query phase when the index type is HSNW, which is the most important operating phase for vector databases supporting RAG applications. The implication here is that the storage performance does not significantly impact the overall performance of these applications.

Vector Database with Instaclustr using PostgreSQL: pgvector

This section discusses the specifics of how instaclustr product integrates with postgreSQL on pgvector fuctionality in the vector database solution for NetApp.

Vector Database with Instaclustr using PostgreSQL: pgvector

In this section, we delve into the specifics of how instaclustr product integrates with postgreSQL on pgvector fuctionality. We have an example of "How To Improve Your LLM Accuracy and Performance With PGVector and PostgreSQL®: Introduction to Embeddings and the Role of PGVector". Please check the blog to get more information.

Vector Database Use Cases

This section provides an overview of the use cases for the NetApp vector database solution.

Vector Database Use Cases

In this section, we discuss about two use cases such as Retrieval Augmented Generation with Large Language Models and NetApp IT chatbot.

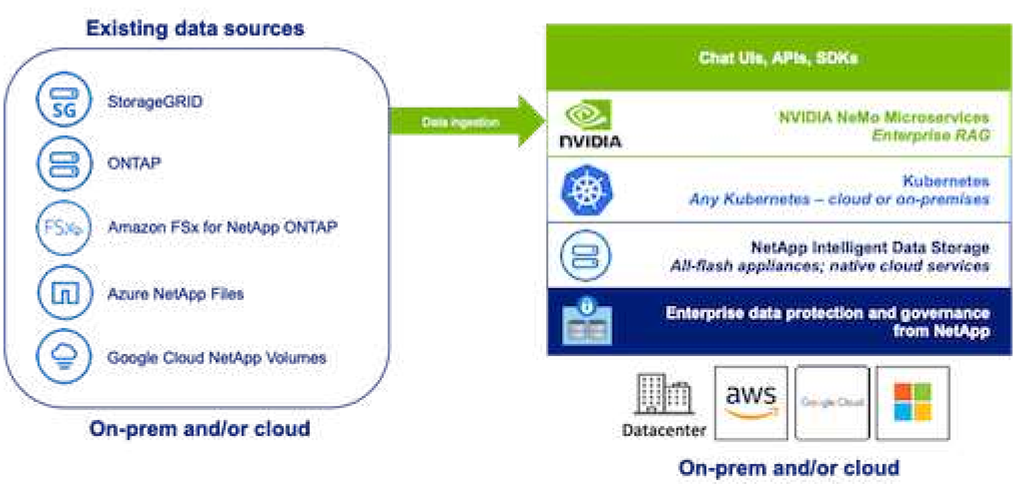
Retrieval Augmented Generation (RAG) with Large Language Models (LLMs)

```
Retrieval-augmented generation, or RAG, is a technique for enhancing the
accuracy and reliability of Large Language Models, or LLMs, by augmenting
prompts with facts fetched from external sources. In a traditional RAG
deployment, vector embeddings are generated from an existing dataset and
then stored in a vector database, often referred to as a knowledgebase.
Whenever a user submits a prompt to the LLM, a vector embedding
representation of the prompt is generated, and the vector database is
searched using that embedding as the search query. This search operation
returns similar vectors from the knowledgebase, which are then fed to the
LLM as context alongside the original user prompt. In this way, an LLM can
be augmented with additional information that was not part of its original
training dataset.
```

The NVIDIA Enterprise RAG LLM Operator is a useful tool for implementing RAG in the enterprise. This operator can be used to deploy a full RAG pipeline. The RAG pipeline can be customized to utilize either Milvus or pgvecto as the vector database for storing knowledgebase embeddings. Refer to the documentation for details.

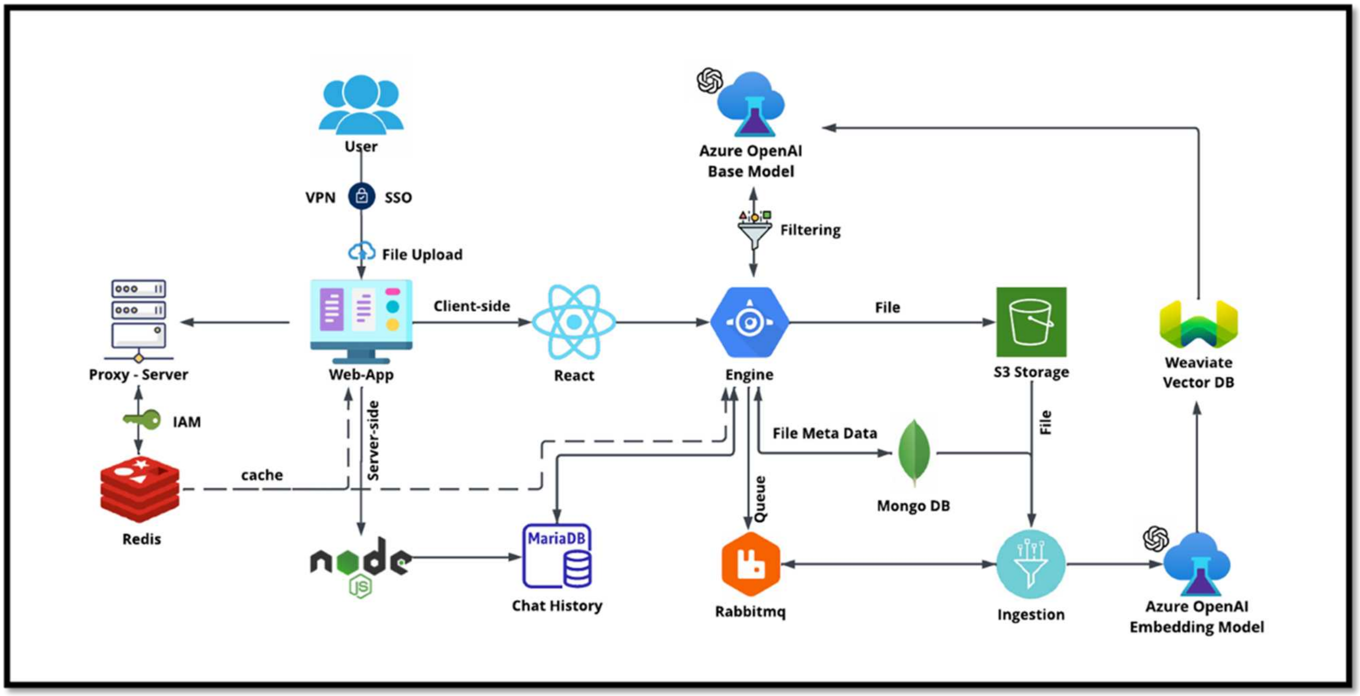
```
NetApp has validated an enterprise RAG architecture powered by the NVIDIA
Enterprise RAG LLM Operator alongside NetApp storage. Refer to our blog
post for more information and to see a demo. Figure 1 provides an overview
of this architecture.
```

Figure 1) Enterprise RAG powered by NVIDIA NeMo Microservices and NetApp



NetApp IT chatbot use case

NetApp's chatbot serves as another real-time use case for the vector database. In this instance, the NetApp Private OpenAI Sandbox provides an effective, secure, and efficient platform for managing queries from NetApp's internal users. By incorporating stringent security protocols, efficient data management systems, and sophisticated AI processing capabilities, it guarantees high-quality, precise responses to users based on their roles and responsibilities in the organization via SSO authentication. This architecture highlights the potential of merging advanced technologies to create user-focused, intelligent systems.



The use case can be divided into four primary sections.

User Authentication and Verification:

- User queries first go through the NetApp Single Sign-On (SSO) process to confirm the user's identity.
- After successful authentication, the system checks the VPN connection to ensure a secure data transmission.

Data Transmission and Processing:

- Once the VPN is validated, the data is sent to MariaDB through the NetAIChat or NetAICreate web applications. MariaDB is a fast and efficient database system used to manage and store user data.
- MariaDB then sends the information to the NetApp Azure instance, which connects the user data to the AI processing unit.

Interaction with OpenAI and Content Filtering:

- The Azure instance sends the user's questions to a content filtering system. This system cleans up the query and prepares it for processing.
- The cleaned-up input is then sent to the Azure OpenAI base model, which generates a response based on the input.

Response Generation and Moderation:

- The response from the base model is first checked to ensure it is accurate and meets content standards.
- After passing the check, the response is sent back to the user. This process ensures that the user receives a clear, accurate, and appropriate answer to their query.

Conclusion

This section concludes the vector database solution for NetApp.

Conclusion

In conclusion, this document provides a comprehensive overview of deploying and managing vector databases, such as Milvus and pgvector, on NetApp storage solutions. We discussed the infrastructure guidelines for leveraging NetApp ONTAP and StorageGRID object storage and validated the Milvus database in AWS FSX for NetApp ONTAP through file and object store.

We explored NetApp's file-object duality, demonstrating its utility not only for data in vector databases but also for other applications. We also highlighted how SnapCenter, NetApp's enterprise management product, offers backup, restore, and clone functionalities for vector database data, ensuring data integrity and availability.

The document also delves into how NetApp's Hybrid Cloud solution offers data replication and protection across on-premises and cloud environments, providing a seamless and secure data management experience. We provided insights into the performance validation of vector databases like Milvus and pgvecto on NetApp ONTAP, offering valuable information on their efficiency and scalability.

Finally, we discussed two generative AI use cases: RAG with LLM and the NetApp's internal ChatAI. These practical examples underscore the real-world applications and benefits of the concepts and practices outlined in this document. Overall, this document serves as a comprehensive guide for anyone looking to leverage NetApp's powerful storage solutions for managing vector databases.

Acknowledgments

The author like to heartfelt thanks to the below contributors, others who provided their feedback and comments to make this paper valuable to NetApp customers and NetApp fields.

1. Sathish Thyagarajan, Technical Marketing Engineer, ONTAP AI & Analytics, NetApp
2. Mike Oglesby, Technical Marketing Engineer, NetApp
3. AJ Mahajan, Senior Director, NetApp
4. Joe Scott, Manager, Workload Performance Engineering, NetApp
5. Puneet Dhawan, Senior Director, Product Management Fsx, NetApp
6. Yuval Kalderon, Senior Product Manager, FSx Product Team, NetApp

Where to find additional information

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

- Milvus documentation - https://milvus.io/docs/overview.md
- Milvus standalone documentation - https://milvus.io/docs/v2.0.x/install_standalone-docker.md
- NetApp Product Documentation
  https://www.netapp.com/support-and-training/documentation/
- instaclustr - instalclustr documentation

Version history

Version	Date	Document version history
Version 1.0	April 2024	Initial release

Appendix A: Values.yaml

This section provides sample YAML code for the values used in the NetApp vector database solution.

Appendix A: Values.yaml

```
root@node2:~# cat  values.yaml
## Enable or disable Milvus Cluster mode
cluster:
  enabled: true

image:
  all:
    repository: milvusdb/milvus
    tag: v2.3.4
    pullPolicy: IfNotPresent
    ## Optionally specify an array of imagePullSecrets.
    ## Secrets must be manually created in the namespace.
    ## ref: https://kubernetes.io/docs/tasks/configure-pod-container/pull-
image-private-registry/
    ##
    # pullSecrets:
    #   - myRegistryKeySecretName
  tools:
    repository: milvusdb/milvus-config-tool
    tag: v0.1.2
    pullPolicy: IfNotPresent

# Global node selector
# If set, this will apply to all milvus components
# Individual components can be set to a different node selector
nodeSelector: {}

# Global tolerations
# If set, this will apply to all milvus components
# Individual components can be set to a different tolerations
tolerations: []

# Global affinity
# If set, this will apply to all milvus components
# Individual components can be set to a different affinity
```

```yaml
affinity: {}

# Global labels and annotations
# If set, this will apply to all milvus components
labels: {}
annotations: {}

# Extra configs for milvus.yaml
# If set, this config will merge into milvus.yaml
# Please follow the config structure in the milvus.yaml
# at https://github.com/milvus-io/milvus/blob/master/configs/milvus.yaml
# Note: this config will be the top priority which will override the
config
# in the image and helm chart.
extraConfigFiles:
  user.yaml: |+
    #    For example enable rest http for milvus proxy
    #    proxy:
    #      http:
    #        enabled: true
    ##  Enable tlsMode and set the tls cert and key
    #  tls:
    #    serverPemPath: /etc/milvus/certs/tls.crt
    #    serverKeyPath: /etc/milvus/certs/tls.key
    #   common:
    #     security:
    #       tlsMode: 1

## Expose the Milvus service to be accessed from outside the cluster
(LoadBalancer service).
## or access it from within the cluster (ClusterIP service). Set the
service type and the port to serve it.
## ref: http://kubernetes.io/docs/user-guide/services/
##
service:
  type: ClusterIP
  port: 19530
  portName: milvus
  nodePort: ""
  annotations: {}
  labels: {}

  ## List of IP addresses at which the Milvus service is available
  ## Ref: https://kubernetes.io/docs/user-guide/services/#external-ips
  ##
  externalIPs: []
```

```yaml
  #   - externalIp1

  # LoadBalancerSourcesRange is a list of allowed CIDR values, which are
combined with ServicePort to
  # set allowed inbound rules on the security group assigned to the master
load balancer
  loadBalancerSourceRanges:
  - 0.0.0.0/0
  # Optionally assign a known public LB IP
  # loadBalancerIP: 1.2.3.4

ingress:
  enabled: false
  annotations:
    # Annotation example: set nginx ingress type
    # kubernetes.io/ingress.class: nginx
    nginx.ingress.kubernetes.io/backend-protocol: GRPC
    nginx.ingress.kubernetes.io/listen-ports-ssl: '[19530]'
    nginx.ingress.kubernetes.io/proxy-body-size: 4m
    nginx.ingress.kubernetes.io/ssl-redirect: "true"
  labels: {}
  rules:
    - host: "milvus-example.local"
      path: "/"
      pathType: "Prefix"
    # - host: "milvus-example2.local"
    #   path: "/otherpath"
    #   pathType: "Prefix"
  tls: []
  #  - secretName: chart-example-tls
  #    hosts:
  #      - milvus-example.local

serviceAccount:
  create: false
  name:
  annotations:
  labels:

metrics:
  enabled: true

  serviceMonitor:
    # Set this to `true` to create ServiceMonitor for Prometheus operator
    enabled: false
    interval: "30s"
```

```
    scrapeTimeout: "10s"
    # Additional labels that can be used so ServiceMonitor will be
discovered by Prometheus
    additionalLabels: {}

livenessProbe:
  enabled: true
  initialDelaySeconds: 90
  periodSeconds: 30
  timeoutSeconds: 5
  successThreshold: 1
  failureThreshold: 5

readinessProbe:
  enabled: true
  initialDelaySeconds: 90
  periodSeconds: 10
  timeoutSeconds: 5
  successThreshold: 1
  failureThreshold: 5

log:
  level: "info"
  file:
    maxSize: 300    # MB
    maxAge: 10    # day
    maxBackups: 20
  format: "text"    # text/json

  persistence:
    mountPath: "/milvus/logs"
    ## If true, create/use a Persistent Volume Claim
    ## If false, use emptyDir
    ##
    enabled: false
    annotations:
      helm.sh/resource-policy: keep
    persistentVolumeClaim:
      existingClaim: ""
      ## Milvus Logs Persistent Volume Storage Class
      ## If defined, storageClassName: <storageClass>
      ## If set to "-", storageClassName: "", which disables dynamic
provisioning
      ## If undefined (the default) or set to null, no storageClassName
spec is
      ##   set, choosing the default provisioner.
```

```yaml
      ## ReadWriteMany access mode required for milvus cluster.
      ##
      storageClass: default
      accessModes: ReadWriteMany
      size: 10Gi
      subPath: ""

## Heaptrack traces all memory allocations and annotates these events with
stack traces.
## See more: https://github.com/KDE/heaptrack
## Enable heaptrack in production is not recommended.
heaptrack:
  image:
    repository: milvusdb/heaptrack
    tag: v0.1.0
    pullPolicy: IfNotPresent

standalone:
  replicas: 1  # Run standalone mode with replication disabled
  resources: {}
  # Set local storage size in resources
  # limits:
  #    ephemeral-storage: 100Gi
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  disk:
    enabled: true
    size:
      enabled: false  # Enable local storage size limit
  profiling:
    enabled: false  # Enable live profiling

  ## Default message queue for milvus standalone
  ## Supported value: rocksmq, natsmq, pulsar and kafka
  messageQueue: rocksmq
  persistence:
    mountPath: "/var/lib/milvus"
    ## If true, alertmanager will create/use a Persistent Volume Claim
    ## If false, use emptyDir
    ##
    enabled: true
    annotations:
```

```yaml
      helm.sh/resource-policy: keep
    persistentVolumeClaim:
      existingClaim: ""
      ## Milvus Persistent Volume Storage Class
      ## If defined, storageClassName: <storageClass>
      ## If set to "-", storageClassName: "", which disables dynamic provisioning
      ## If undefined (the default) or set to null, no storageClassName spec is
      ##   set, choosing the default provisioner.
      ##
      storageClass:
      accessModes: ReadWriteOnce
      size: 50Gi
      subPath: ""

proxy:
  enabled: true
  # You can set the number of replicas to -1 to remove the replicas field in case you want to use HPA
  replicas: 1
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
  http:
    enabled: true  # whether to enable http rest server
    debugMode:
      enabled: false
  # Mount a TLS secret into proxy pod
  tls:
    enabled: false
## when enabling proxy.tls, all items below should be uncommented and the key and crt values should be populated.
#    enabled: true
#    secretName: milvus-tls
## expecting base64 encoded values here: i.e. $(cat tls.crt | base64 -w 0) and $(cat tls.key | base64 -w 0)
#    key: LS0tLS1CRUdJTiBQU--REDUCT
#    crt: LS0tLS1CRUdJTiBDR--REDUCT
#  volumes:
```

```yaml
#   - secret:
#       secretName: milvus-tls
#     name: milvus-tls
#   volumeMounts:
#   - mountPath: /etc/milvus/certs/
#     name: milvus-tls

rootCoordinator:
  enabled: true
  # You can set the number of replicas greater than 1, only if enable
active standby
  replicas: 1  # Run Root Coordinator mode with replication disabled
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
  activeStandby:
    enabled: false  # Enable active-standby when you set multiple replicas
for root coordinator

  service:
    port: 53100
    annotations: {}
    labels: {}
    clusterIP: ""

queryCoordinator:
  enabled: true
  # You can set the number of replicas greater than 1, only if enable
active standby
  replicas: 1  # Run Query Coordinator mode with replication disabled
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
  activeStandby:
```

```
    enabled: false  # Enable active-standby when you set multiple replicas
for query coordinator

  service:
    port: 19531
    annotations: {}
    labels: {}
    clusterIP: ""

queryNode:
  enabled: true
  # You can set the number of replicas to -1 to remove the replicas field
in case you want to use HPA
  replicas: 1
  resources: {}
  # Set local storage size in resources
  # limits:
  #    ephemeral-storage: 100Gi
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  disk:
    enabled: true  # Enable querynode load disk index, and search on disk
index
    size:
      enabled: false  # Enable local storage size limit
  profiling:
    enabled: false  # Enable live profiling

indexCoordinator:
  enabled: true
  # You can set the number of replicas greater than 1, only if enable
active standby
  replicas: 1   # Run Index Coordinator mode with replication disabled
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
```

```
  activeStandby:
    enabled: false  # Enable active-standby when you set multiple replicas
for index coordinator

  service:
    port: 31000
    annotations: {}
    labels: {}
    clusterIP: ""

indexNode:
  enabled: true
  # You can set the number of replicas to -1 to remove the replicas field
in case you want to use HPA
  replicas: 1
  resources: {}
  # Set local storage size in resources
  # limits:
  #    ephemeral-storage: 100Gi
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
  disk:
    enabled: true  # Enable index node build disk vector index
    size:
      enabled: false  # Enable local storage size limit

dataCoordinator:
  enabled: true
  # You can set the number of replicas greater than 1, only if enable
active standby
  replicas: 1           # Run Data Coordinator mode with replication
disabled
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
```

```yaml
    enabled: false  # Enable live profiling
  activeStandby:
    enabled: false  # Enable active-standby when you set multiple replicas
for data coordinator

  service:
    port: 13333
    annotations: {}
    labels: {}
    clusterIP: ""

dataNode:
  enabled: true
  # You can set the number of replicas to -1 to remove the replicas field
in case you want to use HPA
  replicas: 1
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling

## mixCoordinator contains all coord
## If you want to use mixcoord, enable this and disable all of other
coords
mixCoordinator:
  enabled: false
  # You can set the number of replicas greater than 1, only if enable
active standby
  replicas: 1           # Run Mixture Coordinator mode with replication
disabled
  resources: {}
  nodeSelector: {}
  affinity: {}
  tolerations: []
  extraEnv: []
  heaptrack:
    enabled: false
  profiling:
    enabled: false  # Enable live profiling
  activeStandby:
    enabled: false  # Enable active-standby when you set multiple replicas
```

```yaml
for Mixture coordinator

  service:
    annotations: {}
    labels: {}
    clusterIP: ""

attu:
  enabled: false
  name: attu
  image:
    repository: zilliz/attu
    tag: v2.2.8
    pullPolicy: IfNotPresent
  service:
    annotations: {}
    labels: {}
    type: ClusterIP
    port: 3000
    # loadBalancerIP: ""
  resources: {}
  podLabels: {}
  ingress:
    enabled: false
    annotations: {}
    # Annotation example: set nginx ingress type
    # kubernetes.io/ingress.class: nginx
    labels: {}
    hosts:
      - milvus-attu.local
    tls: []
    #  - secretName: chart-attu-tls
    #    hosts:
    #      - milvus-attu.local


## Configuration values for the minio dependency
## ref: https://github.com/minio/charts/blob/master/README.md
##

minio:
  enabled: false
  name: minio
  mode: distributed
  image:
    tag: "RELEASE.2023-03-20T20-16-18Z"
```

```yaml
    pullPolicy: IfNotPresent
  accessKey: minioadmin
  secretKey: minioadmin
  existingSecret: ""
  bucketName: "milvus-bucket"
  rootPath: file
  useIAM: false
  iamEndpoint: ""
  region: ""
  useVirtualHost: false
  podDisruptionBudget:
    enabled: false
  resources:
    requests:
      memory: 2Gi

  gcsgateway:
    enabled: false
    replicas: 1
    gcsKeyJson: "/etc/credentials/gcs_key.json"
    projectId: ""

  service:
    type: ClusterIP
    port: 9000

  persistence:
    enabled: true
    existingClaim: ""
    storageClass:
    accessMode: ReadWriteOnce
    size: 500Gi

  livenessProbe:
    enabled: true
    initialDelaySeconds: 5
    periodSeconds: 5
    timeoutSeconds: 5
    successThreshold: 1
    failureThreshold: 5

  readinessProbe:
    enabled: true
    initialDelaySeconds: 5
    periodSeconds: 5
    timeoutSeconds: 1
```

```yaml
    successThreshold: 1
    failureThreshold: 5

  startupProbe:
    enabled: true
    initialDelaySeconds: 0
    periodSeconds: 10
    timeoutSeconds: 5
    successThreshold: 1
    failureThreshold: 60

## Configuration values for the etcd dependency
## ref: https://artifacthub.io/packages/helm/bitnami/etcd
##

etcd:
  enabled: true
  name: etcd
  replicaCount: 3
  pdb:
    create: false
  image:
    repository: "milvusdb/etcd"
    tag: "3.5.5-r2"
    pullPolicy: IfNotPresent

  service:
    type: ClusterIP
    port: 2379
    peerPort: 2380

  auth:
    rbac:
      enabled: false

  persistence:
    enabled: true
    storageClass: default
    accessMode: ReadWriteOnce
    size: 10Gi

  ## Change default timeout periods to mitigate zoobie probe process
  livenessProbe:
    enabled: true
    timeoutSeconds: 10
```

```yaml
  readinessProbe:
    enabled: true
    periodSeconds: 20
    timeoutSeconds: 10

  ## Enable auto compaction
  ## compaction by every 1000 revision
  ##
  autoCompactionMode: revision
  autoCompactionRetention: "1000"

  ## Increase default quota to 4G
  ##
  extraEnvVars:
  - name: ETCD_QUOTA_BACKEND_BYTES
    value: "4294967296"
  - name: ETCD_HEARTBEAT_INTERVAL
    value: "500"
  - name: ETCD_ELECTION_TIMEOUT
    value: "2500"

## Configuration values for the pulsar dependency
## ref: https://github.com/apache/pulsar-helm-chart
##

pulsar:
  enabled: true
  name: pulsar

  fullnameOverride: ""
  persistence: true

  maxMessageSize: "5242880"  # 5 * 1024 * 1024 Bytes, Maximum size of each
message in pulsar.

  rbac:
    enabled: false
    psp: false
    limit_to_namespace: true

  affinity:
    anti_affinity: false

## enableAntiAffinity: no

  components:
    zookeeper: true
```

```yaml
    bookkeeper: true
    # bookkeeper - autorecovery
    autorecovery: true
    broker: true
    functions: false
    proxy: true
    toolset: false
    pulsar_manager: false

  monitoring:
    prometheus: false
    grafana: false
    node_exporter: false
    alert_manager: false

  images:
    broker:
      repository: apachepulsar/pulsar
      pullPolicy: IfNotPresent
      tag: 2.8.2
    autorecovery:
      repository: apachepulsar/pulsar
      tag: 2.8.2
      pullPolicy: IfNotPresent
    zookeeper:
      repository: apachepulsar/pulsar
      pullPolicy: IfNotPresent
      tag: 2.8.2
    bookie:
      repository: apachepulsar/pulsar
      pullPolicy: IfNotPresent
      tag: 2.8.2
    proxy:
      repository: apachepulsar/pulsar
      pullPolicy: IfNotPresent
      tag: 2.8.2
    pulsar_manager:
      repository: apachepulsar/pulsar-manager
      pullPolicy: IfNotPresent
      tag: v0.1.0

  zookeeper:
    volumes:
      persistence: true
      data:
        name: data
```

```yaml
        size: 20Gi    #SSD Required
        storageClassName: default
    resources:
      requests:
        memory: 1024Mi
        cpu: 0.3
    configData:
      PULSAR_MEM: >
        -Xms1024m
        -Xmx1024m
      PULSAR_GC: >
         -Dcom.sun.management.jmxremote
         -Djute.maxbuffer=10485760
         -XX:+ParallelRefProcEnabled
         -XX:+UnlockExperimentalVMOptions
         -XX:+DoEscapeAnalysis
         -XX:+DisableExplicitGC
         -XX:+PerfDisableSharedMem
         -Dzookeeper.forceSync=no
    pdb:
      usePolicy: false

  bookkeeper:
    replicaCount: 3
    volumes:
      persistence: true
      journal:
        name: journal
        size: 100Gi
        storageClassName: default
      ledgers:
        name: ledgers
        size: 200Gi
        storageClassName: default
    resources:
      requests:
        memory: 2048Mi
        cpu: 1
    configData:
      PULSAR_MEM: >
        -Xms4096m
        -Xmx4096m
        -XX:MaxDirectMemorySize=8192m
      PULSAR_GC: >
        -Dio.netty.leakDetectionLevel=disabled
        -Dio.netty.recycler.linkCapacity=1024
```

```
        -XX:+UseG1GC -XX:MaxGCPauseMillis=10
        -XX:+ParallelRefProcEnabled
        -XX:+UnlockExperimentalVMOptions
        -XX:+DoEscapeAnalysis
        -XX:ParallelGCThreads=32
        -XX:ConcGCThreads=32
        -XX:G1NewSizePercent=50
        -XX:+DisableExplicitGC
        -XX:-ResizePLAB
        -XX:+ExitOnOutOfMemoryError
        -XX:+PerfDisableSharedMem
        -XX:+PrintGCDetails
      nettyMaxFrameSizeBytes: "104867840"
    pdb:
      usePolicy: false

  broker:
    component: broker
    podMonitor:
      enabled: false
    replicaCount: 1
    resources:
      requests:
        memory: 4096Mi
        cpu: 1.5
    configData:
      PULSAR_MEM: >
        -Xms4096m
        -Xmx4096m
        -XX:MaxDirectMemorySize=8192m
      PULSAR_GC: >
        -Dio.netty.leakDetectionLevel=disabled
        -Dio.netty.recycler.linkCapacity=1024
        -XX:+ParallelRefProcEnabled
        -XX:+UnlockExperimentalVMOptions
        -XX:+DoEscapeAnalysis
        -XX:ParallelGCThreads=32
        -XX:ConcGCThreads=32
        -XX:G1NewSizePercent=50
        -XX:+DisableExplicitGC
        -XX:-ResizePLAB
        -XX:+ExitOnOutOfMemoryError
      maxMessageSize: "104857600"
      defaultRetentionTimeInMinutes: "10080"
      defaultRetentionSizeInMB: "-1"
      backlogQuotaDefaultLimitGB: "8"
```

```yaml
      ttlDurationDefaultInSeconds: "259200"
      subscriptionExpirationTimeMinutes: "3"
      backlogQuotaDefaultRetentionPolicy: producer_exception
    pdb:
      usePolicy: false

  autorecovery:
    resources:
      requests:
        memory: 512Mi
        cpu: 1

  proxy:
    replicaCount: 1
    podMonitor:
      enabled: false
    resources:
      requests:
        memory: 2048Mi
        cpu: 1
    service:
      type: ClusterIP
    ports:
      pulsar: 6650
    configData:
      PULSAR_MEM: >
        -Xms2048m -Xmx2048m
      PULSAR_GC: >
        -XX:MaxDirectMemorySize=2048m
      httpNumThreads: "100"
    pdb:
      usePolicy: false

  pulsar_manager:
    service:
      type: ClusterIP

  pulsar_metadata:
    component: pulsar-init
    image:
      # the image used for running `pulsar-cluster-initialize` job
      repository: apachepulsar/pulsar
      tag: 2.8.2


## Configuration values for the kafka dependency
```

```yaml
## ref: https://artifacthub.io/packages/helm/bitnami/kafka
##

kafka:
  enabled: false
  name: kafka
  replicaCount: 3
  image:
    repository: bitnami/kafka
    tag: 3.1.0-debian-10-r52
  ## Increase graceful termination for kafka graceful shutdown
  terminationGracePeriodSeconds: "90"
  pdb:
    create: false

  ## Enable startup probe to prevent pod restart during recovering
  startupProbe:
    enabled: true

  ## Kafka Java Heap size
  heapOpts: "-Xmx4096m -Xms4096m"
  maxMessageBytes: _10485760
  defaultReplicationFactor: 3
  offsetsTopicReplicationFactor: 3
  ## Only enable time based log retention
  logRetentionHours: 168
  logRetentionBytes: _-1
  extraEnvVars:
  - name: KAFKA_CFG_MAX_PARTITION_FETCH_BYTES
    value: "5242880"
  - name: KAFKA_CFG_MAX_REQUEST_SIZE
    value: "5242880"
  - name: KAFKA_CFG_REPLICA_FETCH_MAX_BYTES
    value: "10485760"
  - name: KAFKA_CFG_FETCH_MESSAGE_MAX_BYTES
    value: "5242880"
  - name: KAFKA_CFG_LOG_ROLL_HOURS
    value: "24"

  persistence:
    enabled: true
    storageClass:
    accessMode: ReadWriteOnce
    size: 300Gi

  metrics:
```

```
    ## Prometheus Kafka exporter: exposes complimentary metrics to JMX
exporter
    kafka:
      enabled: false
      image:
        repository: bitnami/kafka-exporter
        tag: 1.4.2-debian-10-r182

    ## Prometheus JMX exporter: exposes the majority of Kafkas metrics
    jmx:
      enabled: false
      image:
        repository: bitnami/jmx-exporter
        tag: 0.16.1-debian-10-r245

    ## To enable serviceMonitor, you must enable either kafka exporter or
jmx exporter.
    ## And you can enable them both
    serviceMonitor:
      enabled: false

  service:
    type: ClusterIP
    ports:
      client: 9092

  zookeeper:
    enabled: true
    replicaCount: 3

###################################
# External S3
# - these configs are only used when `externalS3.enabled` is true
###################################
externalS3:
  enabled: true
  host: "192.168.150.167"
  port: "80"
  accessKey: "24G4C1316APP2BIPDE5S"
  secretKey: "Zd28p43rgZaU44PX_ftT279z9nt4jBSro97j87Bx"
  useSSL: false
  bucketName: "milvusdbvol1"
  rootPath: ""
  useIAM: false
  cloudProvider: "aws"
  iamEndpoint: ""
```

```yaml
  region: ""
  useVirtualHost: false

###################################
# GCS Gateway
# - these configs are only used when `minio.gcsgateway.enabled` is true
###################################
externalGcs:
  bucketName: ""

###################################
# External etcd
# - these configs are only used when `externalEtcd.enabled` is true
###################################
externalEtcd:
  enabled: false
  ## the endpoints of the external etcd
  ##
  endpoints:
    - localhost:2379

###################################
# External pulsar
# - these configs are only used when `externalPulsar.enabled` is true
###################################
externalPulsar:
  enabled: false
  host: localhost
  port: 6650
  maxMessageSize: "5242880"  # 5 * 1024 * 1024 Bytes, Maximum size of each message in pulsar.
  tenant: public
  namespace: default
  authPlugin: ""
  authParams: ""

###################################
# External kafka
# - these configs are only used when `externalKafka.enabled` is true
###################################
externalKafka:
  enabled: false
  brokerList: localhost:9092
  securityProtocol: SASL_SSL
  sasl:
    mechanisms: PLAIN
```

```
    username: ""
    password: ""
root@node2:~#
```

Appendix B: prepare_data_netapp_new.py

This section provides a sample Python script used to prepare data for the vector database.

Appendix B: prepare_data_netapp_new.py

```
root@node2:~# cat prepare_data_netapp_new.py
# hello_milvus.py demonstrates the basic operations of PyMilvus, a Python
SDK of Milvus.
# 1. connect to Milvus
# 2. create collection
# 3. insert data
# 4. create index
# 5. search, query, and hybrid search on entities
# 6. delete entities by PK
# 7. drop collection
import time
import os
import numpy as np
from pymilvus import (
    connections,
    utility,
    FieldSchema, CollectionSchema, DataType,
    Collection,
)

fmt = "\n=== {:30} ===\n"
search_latency_fmt = "search latency = {:.4f}s"
#num_entities, dim = 3000, 8
num_entities, dim = 3000, 16

###########################################################################
#######
# 1. connect to Milvus
# Add a new connection alias `default` for Milvus server in
`localhost:19530`
# Actually the "default" alias is a buildin in PyMilvus.
# If the address of Milvus is the same as `localhost:19530`, you can omit
all
# parameters and call the method as: `connections.connect()`.
```

```
#
# Note: the `using` parameter of the following methods is default to
"default".
print(fmt.format("start connecting to Milvus"))

host = os.environ.get('MILVUS_HOST')
if host == None:
    host = "localhost"
print(fmt.format(f"Milvus host: {host}"))
#connections.connect("default", host=host, port="19530")
connections.connect("default", host=host, port="27017")

has = utility.has_collection("hello_milvus_ntapnew_update2_sc")
print(f"Does collection hello_milvus_ntapnew_update2_sc exist in Milvus:
{has}")

#drop the collection
print(fmt.format(f"Drop collection - hello_milvus_ntapnew_update2_sc"))
utility.drop_collection("hello_milvus_ntapnew_update2_sc")
#drop the collection
print(fmt.format(f"Drop collection - hello_milvus_ntapnew_update2_sc2"))
utility.drop_collection("hello_milvus_ntapnew_update2_sc2")

##########################################################################
#######
# 2. create collection
# We're going to create a collection with 3 fields.
# +-+------------+------------+------------------
+------------------------------+
# | | field name | field type | other attributes |       field description
|
# +-+------------+------------+------------------
+------------------------------+
# |1|    "pk"    |    Int64   |  is_primary=True |      "primary field"
|
# | |            |            |   auto_id=False  |
|
# +-+------------+------------+------------------
+------------------------------+
# |2|  "random"  |    Double  |                  |      "a double field"
|
# +-+------------+------------+------------------
+------------------------------+
# |3|"embeddings"| FloatVector|     dim=8        |  "float vector with dim
8"    |
# +-+------------+------------+------------------
```

```
+-------------------------------+
fields = [
    FieldSchema(name="pk", dtype=DataType.INT64, is_primary=True, auto_id
=False),
    FieldSchema(name="random", dtype=DataType.DOUBLE),
    FieldSchema(name="var", dtype=DataType.VARCHAR, max_length=65535),
    FieldSchema(name="embeddings", dtype=DataType.FLOAT_VECTOR, dim=dim)
]

schema = CollectionSchema(fields, "hello_milvus_ntapnew_update2_sc")

print(fmt.format("Create collection `hello_milvus_ntapnew_update2_sc`"))
hello_milvus_ntapnew_update2_sc = Collection
("hello_milvus_ntapnew_update2_sc", schema, consistency_level="Strong")

##########################################################################
######
# 3. insert data
# We are going to insert 3000 rows of data into
`hello_milvus_ntapnew_update2_sc`
# Data to be inserted must be organized in fields.
#
# The insert() method returns:
# - either automatically generated primary keys by Milvus if auto_id=True
in the schema;
# - or the existing primary key field from the entities if auto_id=False
in the schema.

print(fmt.format("Start inserting entities"))
rng = np.random.default_rng(seed=19530)
entities = [
    # provide the pk field because `auto_id` is set to False
    [i for i in range(num_entities)],
    rng.random(num_entities).tolist(),  # field random, only supports list
    [str(i) for i in range(num_entities)],
    rng.random((num_entities, dim)),    # field embeddings, supports
numpy.ndarray and list
]

insert_result = hello_milvus_ntapnew_update2_sc.insert(entities)
hello_milvus_ntapnew_update2_sc.flush()
print(f"Number of entities in hello_milvus_ntapnew_update2_sc:
{hello_milvus_ntapnew_update2_sc.num_entities}")  # check the num_entites

# create another collection
fields2 = [
```

```
    FieldSchema(name="pk", dtype=DataType.INT64, is_primary=True, auto_id
=True),
    FieldSchema(name="random", dtype=DataType.DOUBLE),
    FieldSchema(name="var", dtype=DataType.VARCHAR, max_length=65535),
    FieldSchema(name="embeddings", dtype=DataType.FLOAT_VECTOR, dim=dim)
]


schema2 = CollectionSchema(fields2, "hello_milvus_ntapnew_update2_sc2")

print(fmt.format("Create collection `hello_milvus_ntapnew_update2_sc2`"))
hello_milvus_ntapnew_update2_sc2 = Collection
("hello_milvus_ntapnew_update2_sc2", schema2, consistency_level="Strong")

entities2 = [
    rng.random(num_entities).tolist(),  # field random, only supports list
    [str(i) for i in range(num_entities)],
    rng.random((num_entities, dim)),    # field embeddings, supports
numpy.ndarray and list
]

insert_result2 = hello_milvus_ntapnew_update2_sc2.insert(entities2)
hello_milvus_ntapnew_update2_sc2.flush()
insert_result2 = hello_milvus_ntapnew_update2_sc2.insert(entities2)
hello_milvus_ntapnew_update2_sc2.flush()

# index_params = {"index_type": "IVF_FLAT", "params": {"nlist": 128},
"metric_type": "L2"}
# hello_milvus_ntapnew_update2_sc.create_index("embeddings", index_params)
#
hello_milvus_ntapnew_update2_sc2.create_index(field_name="var",index_name=
"scalar_index")

# index_params2 = {"index_type": "Trie"}
# hello_milvus_ntapnew_update2_sc2.create_index("var", index_params2)

print(f"Number of entities in hello_milvus_ntapnew_update2_sc2:
{hello_milvus_ntapnew_update2_sc2.num_entities}")  # check the num_entites

root@node2:~#
```

Appendix C: verify_data_netapp.py

This section contains a sample Python script that can be used to validate the vector database in the NetApp vector database solution.

Appendix C: verify_data_netapp.py

```
root@node2:~# cat verify_data_netapp.py
import time
import os
import numpy as np
from pymilvus import (
    connections,
    utility,
    FieldSchema, CollectionSchema, DataType,
    Collection,
)

fmt = "\n=== {:30} ===\n"
search_latency_fmt = "search latency = {:.4f}s"
num_entities, dim = 3000, 16
rng = np.random.default_rng(seed=19530)
entities = [
    # provide the pk field because `auto_id` is set to False
    [i for i in range(num_entities)],
    rng.random(num_entities).tolist(),  # field random, only supports list
    rng.random((num_entities, dim)),    # field embeddings, supports
numpy.ndarray and list
]

###########################################################################
######
# 1. get recovered collection hello_milvus_ntapnew_update2_sc
print(fmt.format("start connecting to Milvus"))
host = os.environ.get('MILVUS_HOST')
if host == None:
    host = "localhost"
print(fmt.format(f"Milvus host: {host}"))
#connections.connect("default", host=host, port="19530")
connections.connect("default", host=host, port="27017")

recover_collections = ["hello_milvus_ntapnew_update2_sc",
"hello_milvus_ntapnew_update2_sc2"]

for recover_collection_name in recover_collections:
    has = utility.has_collection(recover_collection_name)
    print(f"Does collection {recover_collection_name} exist in Milvus:
{has}")
    recover_collection = Collection(recover_collection_name)
    print(recover_collection.schema)
    recover_collection.flush()
```

```
    print(f"Number of entities in Milvus: {recover_collection_name} :
{recover_collection.num_entities}")  # check the num_entites


#############################################################################
######
    # 4. create index
    # We are going to create an IVF_FLAT index for
hello_milvus_ntapnew_update2_sc collection.
    # create_index() can only be applied to `FloatVector` and
`BinaryVector` fields.
    print(fmt.format("Start Creating index IVF_FLAT"))
    index = {
        "index_type": "IVF_FLAT",
        "metric_type": "L2",
        "params": {"nlist": 128},
    }

    recover_collection.create_index("embeddings", index)


#############################################################################
######
    # 5. search, query, and hybrid search
    # After data were inserted into Milvus and indexed, you can perform:
    # - search based on vector similarity
    # - query based on scalar filtering(boolean, int, etc.)
    # - hybrid search based on vector similarity and scalar filtering.
    #

    # Before conducting a search or a query, you need to load the data in
`hello_milvus` into memory.
    print(fmt.format("Start loading"))
    recover_collection.load()

    #
-----------------------------------------------------------------------------
---
    # search based on vector similarity
    print(fmt.format("Start searching based on vector similarity"))
    vectors_to_search = entities[-1][-2:]
    search_params = {
        "metric_type": "L2",
        "params": {"nprobe": 10},
    }
```

```
    start_time = time.time()
    result = recover_collection.search(vectors_to_search, "embeddings",
search_params, limit=3, output_fields=["random"])
    end_time = time.time()

    for hits in result:
        for hit in hits:
            print(f"hit: {hit}, random field: {hit.entity.get('random')}")
    print(search_latency_fmt.format(end_time - start_time))

    #
---------------------------------------------------------------------------
---
    # query based on scalar filtering(boolean, int, etc.)
    print(fmt.format("Start querying with `random > 0.5`"))

    start_time = time.time()
    result = recover_collection.query(expr="random > 0.5", output_fields=
["random", "embeddings"])
    end_time = time.time()

    print(f"query result:\n-{result[0]}")
    print(search_latency_fmt.format(end_time - start_time))

    #
---------------------------------------------------------------------------
---
    # hybrid search
    print(fmt.format("Start hybrid searching with `random > 0.5`"))

    start_time = time.time()
    result = recover_collection.search(vectors_to_search, "embeddings",
search_params, limit=3, expr="random > 0.5", output_fields=["random"])
    end_time = time.time()

    for hits in result:
        for hit in hits:
            print(f"hit: {hit}, random field: {hit.entity.get('random')}")
    print(search_latency_fmt.format(end_time - start_time))


###########################################################################
#####
    # 7. drop collection
    # Finally, drop the hello_milvus, hello_milvus_ntapnew_update2_sc
collection
```

```
    #print(fmt.format(f"Drop collection {recover_collection_name}"))
    #utility.drop_collection(recover_collection_name)

root@node2:~#
```

Appendix D: docker-compose.yml

This section includes sample YAML code for the vector database solution for NetApp.

Appendix D: docker-compose.yml

```
version: '3.5'

services:
  etcd:
    container_name: milvus-etcd
    image:: quay.io/coreos/etcd:v3.5.5
    environment:
      - ETCD_AUTO_COMPACTION_MODE=revision
      - ETCD_AUTO_COMPACTION_RETENTION=1000
      - ETCD_QUOTA_BACKEND_BYTES=4294967296
      - ETCD_SNAPSHOT_COUNT=50000
    volumes:
      - /home/ubuntu/milvusvectordb/volumes/etcd:/etcd
    command: etcd -advertise-client-urls=http://127.0.0.1:2379 -listen
-client-urls http://0.0.0.0:2379 --data-dir /etcd
    healthcheck:
      test: ["CMD", "etcdctl", "endpoint", "health"]
      interval: 30s
      timeout: 20s
      retries: 3

  minio:
    container_name: milvus-minio
    image:: minio/minio:RELEASE.2023-03-20T20-16-18Z
    environment:
      MINIO_ACCESS_KEY: minioadmin
      MINIO_SECRET_KEY: minioadmin
    ports:
      - "9001:9001"
      - "9000:9000"
    volumes:
      - /home/ubuntu/milvusvectordb/volumes/minio:/minio_data
    command: minio server /minio_data --console-address ":9001"
    healthcheck:
```

```
      test: ["CMD", "curl", "-f",
"http://localhost:9000/minio/health/live"]
      interval: 30s
      timeout: 20s
      retries: 3

  standalone:
    container_name: milvus-standalone
    image:: milvusdb/milvus:v2.4.0-rc.1
    command: ["milvus", "run", "standalone"]
    security_opt:
    - seccomp:unconfined
    environment:
      ETCD_ENDPOINTS: etcd:2379
      MINIO_ADDRESS: minio:9000
    volumes:
      - /home/ubuntu/milvusvectordb/volumes/milvus:/var/lib/milvus
    healthcheck:
      test: ["CMD", "curl", "-f", "http://localhost:9091/healthz"]
      interval: 30s
      start_period: 90s
      timeout: 20s
      retries: 3
    ports:
      - "19530:19530"
      - "9091:9091"
    depends_on:
      - "etcd"
      - "minio"

networks:
  default:
    name: milvus
```

Copyright information

Copyright © 2024 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.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, 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) (FEB 2014).

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.