# Table of Contents

Kubeflow ................................................................. 1
  Kubeflow Deployment .................................................. 1
  Example Kubeflow Operations and Tasks .......................... 2
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.

   ```bash
   # Designate the default StorageClass
   kubectl config set-context my-context --namespace=default --cluster=default --context=default
   kubectl get storageclasses ontap-nas-flexgroup
   kubectl get storageclasses
   ```

   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.
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 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.

```yaml
# ml.pipeline
# name = "MLST Classification Pipeline"
# description = "Train a simple ML for classification"

def mlst_pipeline():
    mlst_train_task = mlst_train.op()
    mlst_train_task.apply(
        kgp.unpromount.prev('mlst-data', 'mlst-data-wol', '/mnt/data/'))

    mlst_text_task = mlst_text.op()
    mlst_text_task.apply(
        kgp.unpromount.prev('mlst-data', 'mlst-data-wol', '/mnt/data/'))

    volume_snapshot_name = "mlst-pytorch-snapshot"
    dataset_snapshot = dlx(terminate_op)
    name="dataset.snapshot",
    image="pytorch:1.9.9",
    command="{}")

    python3 -m pop install netapp-dataops-k8s &&
    echo "" | kubectl exec -it volume-snapshot mlst-pytorch-snapshot
        --overwrite=true --snapshot-name="mlst-data" --snapshot-name="%s"
    netapp-dataops k8s CLI to create volume-snapshot

    mlst_test_task = mlst_test.op()
    mlst_test_task.after(mlst_train.task)
    dataset_snapshot.after(mlst_train.task)
```

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