Example Kubeflow Operations and Tasks
NetApp Solutions
NetApp
December 23, 2022
Table of Contents

Example Kubeflow Operations and Tasks ................................................................. 1
Example Kubeflow Operations and Tasks ................................................................. 1
Provision a Jupyter Notebook Workspace for Data Scientist or Developer Use ............. 1
Example Notebooks and Pipelines ............................................................................. 8
Example Kubeflow Operations and Tasks

This section includes examples of various operations and tasks that you may want to perform using Kubeflow.

Next: Provision a Jupyter Notebook Workspace for Data Scientist or Developer Use.

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. To provision a new Jupyter Notebook server with Kubeflow, perform the following tasks. For more information about Jupyter Notebooks within the Kubeflow context, see the official Kubeflow documentation.

1. From the Kubeflow central dashboard, click Notebook Servers in the main menu to navigate to the Jupyter Notebook server administration page.
2. Click New Server to provision a new Jupyter Notebook server.
3. Give your new server a name, choose the Docker image that you want your server to be based on, and specify the amount of CPU and RAM to be reserved by your server. If the Namespace field is blank, use the Select Namespace menu in the page header to choose a namespace. The Namespace field is then auto-populated with the chosen namespace.

In the following example, the `kubeflow-anonymous` namespace is chosen. In addition, the default values for Docker image, CPU, and RAM are accepted.

4. Specify the workspace volume details. If you choose to create a new volume, then that volume or PVC is provisioned using the default StorageClass. Because a StorageClass utilizing Trident was designated as the default StorageClass in the section `Kubeflow Deployment`, the volume or PVC is provisioned with Trident. This volume is automatically mounted as the default workspace within the Jupyter Notebook Server container. Any notebooks that a user creates on the server that are not saved to a separate data volume are automatically saved to this workspace volume. Therefore, the notebooks are persistent across reboots.
5. Add data volumes. The following example specifies an existing PVC named 'pb-fg-all' and accepts the default mount point.

6. **Optional**: Request that the desired number of GPUs be allocated to your notebook server. In the following example, one GPU is requested.

7. Click Launch to provision your new notebook server.

8. Wait for your notebook server to be fully provisioned. This can take several minutes if you have never provisioned a server using the Docker image that you specified because the image needs to be downloaded. When your server has been fully provisioned, you see a green check mark in the Status column on the Jupyter Notebook server administration page.
9. Click Connect to connect to your new server web interface.

10. Confirm that the dataset volume that was specified in step 6 is mounted on the server. Note that this volume is mounted within the default workspace by default. From the perspective of the user, this is just another folder within the workspace. The user, who is likely a data scientist and not an infrastructure expert, does not need to possess any storage expertise in order to use this volume.
11. Open a Terminal and, assuming that a new volume was requested in step 5, execute `df -h` to confirm that a new Trident-provisioned persistent volume is mounted as the default workspace.

The default workspace directory is the base directory that you are presented with when you first access the server's web interface. Therefore, any artifacts that you create by using the web interface are stored on this Trident-provisioned persistent volume.
12. Using the terminal, run `nvidia-smi` to confirm that the correct number of GPUs were allocated to the notebook server. In the following example, one GPU has been allocated to the notebook server as requested in step 7.

```
$ df -h
Filesystem Size  Used  Avail
  /                       439G   34G   382G
  /dev/shm                64M     0   64M
  /dev/sda                 252G     0   252G
  /sys/fs/cgroup          439G   34G   382G
192.168.11.111/trident_pvc_3dcfe7e5_d5a9_11e9_9b9d_00505681a82d   10G   320K    10G
  /home/jovyan            292G     0   292G
  /run/secrets/kubernetes.io/serviceaccount  10T   10T   47G
  /run/nvidia              252G   12K   252G
  /run/nvidia/persistenced/socket  51G   4.9M   51G
  /sys/firmware            252G     0   252G
  /sys/firmware            252G     0   252G
```

$ nvidia-smi
Fri Sep 13 13:52:15 2019
-----------------------------------------------------------------------------
NVIDIA-SMI 410.104  Driver Version: 410.104  CUDA Version: N/A
-----------------------------------------------------------------------------
GPU Name| Persistence-M| Fan Temp| Perf| Pwr:Usage/Cap| Bus-Id| Disp.A| Volatile Uncorr. ECC| GPU-Util| Compute M.|
-----------------------------------------------------------------------------
Tesla V100-SXM2... On| N/A| 38C| 46W / 300W| 00000000:00:00.0| Off| 0MIB / 32480MIB| 0%| Default|
-----------------------------------------------------------------------------
Processes: GPU  PID  Type  Process name  GPU Memory  Usage
-----------------------------------------------------------------------------
No running processes found
$
Example Notebooks and Pipelines

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 directly from within a Jupyter Notebook.
• Advanced NetApp data management operations 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.

Next: Apache Airflow Deployment.