# Deploy the Client for Triton Inference Server (Automated Deployment)

## NetApp Solutions

Dorian Henderson, Kevin Hoke
January 14, 2021

# Table of Contents

# Deploy the Client for Triton Inference Server (Automated Deployment)

To deploy the client for the Triton Inference Server, complete the following steps:

1. Open a VI editor, create a deployment for the Triton client, and call the file `triton_client.yaml`.

```
---
apiVersion: apps/v1
kind: Deployment
metadata:
  labels:
    app: triton-client
  name: triton-client
  namespace: triton
spec:
  replicas: 1
  selector:
    matchLabels:
      app: triton-client
      version: v1
  template:
    metadata:
      labels:
        app: triton-client
        version: v1
    spec:
      containers:
      - image: nvcr.io/nvidia/tritonserver:20.07- v1- py3-clientsdk
        imagePullPolicy: IfNotPresent
        name: triton-client
        resources:
          limits:
            cpu: "2"
            memory: 4Gi
          requests:
            cpu: "2"
            memory: 4Gi
```

2. Deploy the client.

```
kubectl apply -f triton_client.yaml
```

Next: Collect Inference Metrics from Triton Inference Server