Kubernetes provides some powerful tools for doing autoscaling. A horizontal pod autoscaler (HPA) is one way to implement pod scaling based upon a variety of metrics. This lesson demonstrates how to implement a simple horizontal pod autoscaler within a Kubernetes cluster.

Be sure to check out the full HPA documentation: https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/

Here is the GitHub repo for the source code used in the demo. Check the `example-solution` branch for the demo's final state: https://github.com/linuxacademy/cicd-pipeline-train-schedule-autoscaling

Here are the commands used to set up the autoscaler in the demo:

```
cd ~/
git clone https://github.com/kubernetes-incubator/metrics-server.git
cd metrics-server/
kubectl create -f deploy/1.8+/
kubectl get --raw /apis/metrics.k8s.io/
cd ~/
git clone <your fork of the sample code>
cd cicd-pipeline-train-schedule-autoscaling/
vi train-schedule-kube.yml
kubectl apply -f train-schedule-kube.yml
```

You can find the changes made to `train-schedule-kube.yml` in the `example-solution` branch of the GitHub repo.

Once you have deployed the app and the HPA, you can generate cpu load to test it by spinning up a busybox shell:

```
kubectl run -i --tty load-generator --image=busybox /bin/sh
```

Then run this in the busybox shell to create load:

```
while true; do wget -q -O- http://<kubernetes node private ip>:8080/generate-cpu-load; done
```