AutoScaling

오토스케일링 구현을 위한 Metric 서버 설치와 HPA 구성후 테스트 실습

1. 실습 디렉토리 이동

```
cd ~/k8s_course/lab4/yaml
```

2. Metic-server 설치

```
kubectl create -f metric-server.yaml
```

3. Deployment 편집

```
kubectl edit deployment metrics-server -n kube-system
```

```
/kubelet-use-node-status-port

- --kubelet-use-node-status-port
- --metric-resolution=15s
- --kubelet-insecure-tls
```

맨 아랫줄을 추가해줍니다.(- --kubelet-insecure-tls)

4. 동작 확인

```
kubectl top node
```

5. 실습에 필요한 Deployment와 Service 생성

```
kubectl create -f php-apache.yaml
```

6. HPA 생성

```
kubectl autoscale deployment php-apache \
--cpu-percent=50 --min=1 --max=10
```

7. 생성한 HPA 확인

```
kubectl get hpa
```

8. 부하 생성(다른 터미널을 추가하여 진행)

```
kubectl run -i --tty load-generator --rm \
--image=busybox --restart=Never -- \
/bin/sh -c "while sleep 0.01; do \
wget -q -O- http://php-apache; done"
```

9. 1분정도 뒤 Pod 증가 확인

```
kubectl get hpa
kubectl get deployment php-apache
```

리소스

10. 부하 중지(부하 생성 시 터미널에서 진행)

```
Ctrl + C
```

11. 약 5~7분 뒤 Pod 감소 확인

```
kubectl get hpa
kubectl get deployment php-apache
```

12. clear

```
kubectl delete pod,hpa,deploy --all
```

참고 : https://kubernetes.io/ko/docs/tasks/run-application/horizontal-pod-autoscale-walkthrough/