

Containerization and Big Data: Exercise 4-2

Brandon Hosley

July 15, 2020

4 Building an Apache Spark Cluster Using Kubernetes

4.4 Configuring Minikube

```
bhosl2@us2004lts: ~/mySparkCluster
bhosl2@us2004lts:~$ mkdir mySparkCluster
bhosl2@us2004lts:~$ cd mySparkCluster
bhosl2@us2004lts:~/mySparkCluster$ minikube start --cpus 6 --memory 14336
* minikube v1.11.0 on Ubuntu 20.04
* minikube 1.12.0 is available! Download it: https://github.com/kubernetes/minikube/releases/tag/v1.12.0
* To disable this notice, run: 'minikube config set WantUpdateNotification false'

* Automatically selected the docker driver
* Starting control plane node minikube in cluster minikube
* Creating docker container (CPUs=6, Memory=14336MB) ...
* Preparing Kubernetes v1.18.3 on Docker 19.03.2 ...
  - kubeadm.pod-network-cidr=10.244.0.0/16
* Verifying Kubernetes components...
* Enabled addons: default-storageclass, storage-provisioner
* Done! kubectl is now configured to use "minikube"
bhosl2@us2004lts:~/mySparkCluster$
```

4.5 Creating Spark Docker Images

```
bhosl2@us2004lts: ~/mySparkCluster/spark-2.4.6-bin-hadoop2.7
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$ ls
LICENSE  R          RELEASE  conf  examples  kubernetes  python  yarn
NOTICE  README.md bin      data  jars      licenses    sbin
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$ ls bin
beeline          pyspark          spark-class2.cmd  spark-submit
beeline.cmd      pyspark.cmd      spark-shell        spark-submit.cmd
docker-image-tool.sh  pyspark2.cmd    spark-shell1.cmd  spark-submit2.cmd
find-spark-home  run-example      spark-shell2.cmd  sparkR
find-spark-home.cmd  run-example.cmd  spark-sql         sparkR.cmd
load-spark-env.cmd  spark-class      spark-sql1.cmd    sparkR2.cmd
load-spark-env.sh   spark-class.cmd  spark-sql2.cmd
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$
```

```
bhosl2@us2004lts: ~/mySparkCluster/spark-2.4.6-bin-hadoop2.7
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$ docker image ls
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
spark-r              v1.0.uis           ee40aded5556       3 minutes ago      1.11GB
spark-py             v1.0.uis           9841553098a2       4 minutes ago      1.06GB
spark                v1.0.uis           b1c210a4fef4       5 minutes ago      555MB
k8s.gcr.io/kube-proxy v1.18.3            3439b7546ef29      0 weeks ago        117MB
k8s.gcr.io/kube-controller-manager v1.18.3            da26705ech4b       0 weeks ago        162MB
k8s.gcr.io/kube-apiserver v1.18.3            7e28efe97dbd       0 weeks ago        173MB
k8s.gcr.io/kube-scheduler v1.18.3            76216c34ed0c       0 weeks ago        95.3MB
kubernetesui/dashboard v2.0.0             8b32422733b3       2 months ago       222MB
openjdk              8u242-slim         21c90a4e95d3       2 months ago       284MB
k8s.gcr.io/pause      3.2                80d28bedfe5d       5 months ago       683KB
k8s.gcr.io/coredns    1.6.7              67da37a9a360       5 months ago       43.8MB
k8s.gcr.io/etcd        3.4.3-0            303ce5db0e90       8 months ago       288MB
kubernetesui/metrics-scraper v1.0.2            3b08661dc379       8 months ago       40.1MB
gcr.io/k8s-minikube/storage-provisioner v1.8.1            4689081edb10       2 years ago        80.8MB
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$
```

4.6 Running Spark App

```
bhosl2@us2004lts: ~
bhosl2@us2004lts:~$ kubectl create serviceaccount spark
serviceaccount/spark created
bhosl2@us2004lts:~$ kubectl create clusterrolebinding spark-role --clusterrole=edit --serviceaccount=default:spark --namespace=default
clusterrolebinding.rbac.authorization.k8s.io/spark-role created
bhosl2@us2004lts:~$ kubectl cluster-info
Kubernetes master is running at https://172.17.0.3:8443
KubeDNS is running at https://172.17.0.3:8443/api/v1/namespaces/kube-system/services/kube-dns:dns/proxy

To further debug and diagnose cluster problems, use 'kubectl cluster-info dump'.
bhosl2@us2004lts:~$ minikube ip
172.17.0.3
bhosl2@us2004lts:~$

bhosl2@us2004lts: ~/mySparkCluster/spark-2.4.6-bin-hadoop2.7
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$ kubectl get pods
NAME                READY   STATUS    RESTARTS   AGE
spark-pi-driver      0/1     Completed 0           2m49s
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$ kubectl logs spark-pi-driver
| grep "Pi is roughly"
Pi is roughly 3.1409757048785245
bhosl2@us2004lts:~/mySparkCluster/spark-2.4.6-bin-hadoop2.7$
```

4.7 Minikube Dashboard

