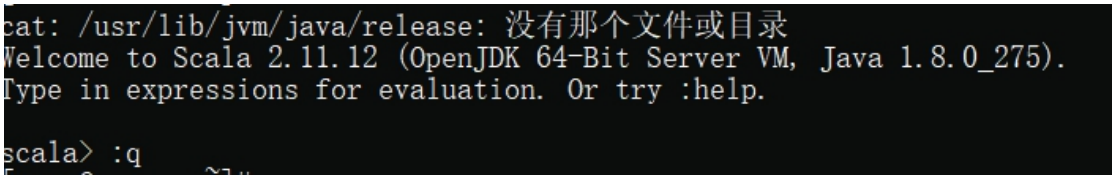
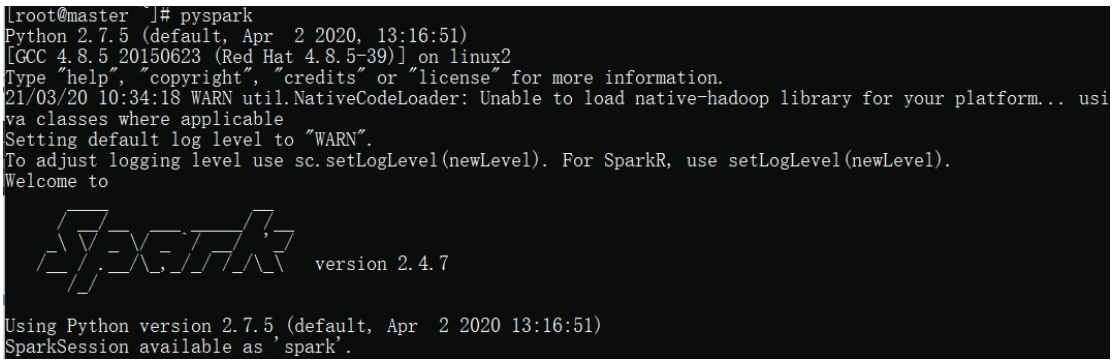
IEMS5730 HW3

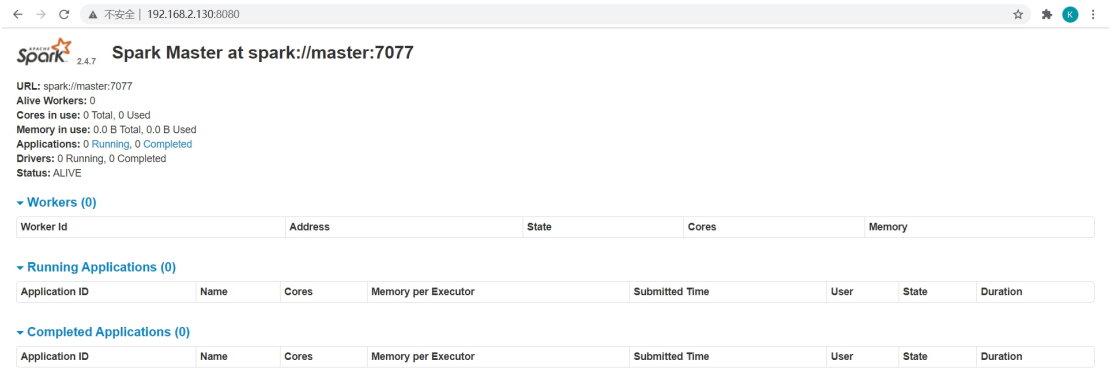
Q1 Single-node Spark



P1 Install scala (2.11.12)



P2 Spark installation and pyspark startup



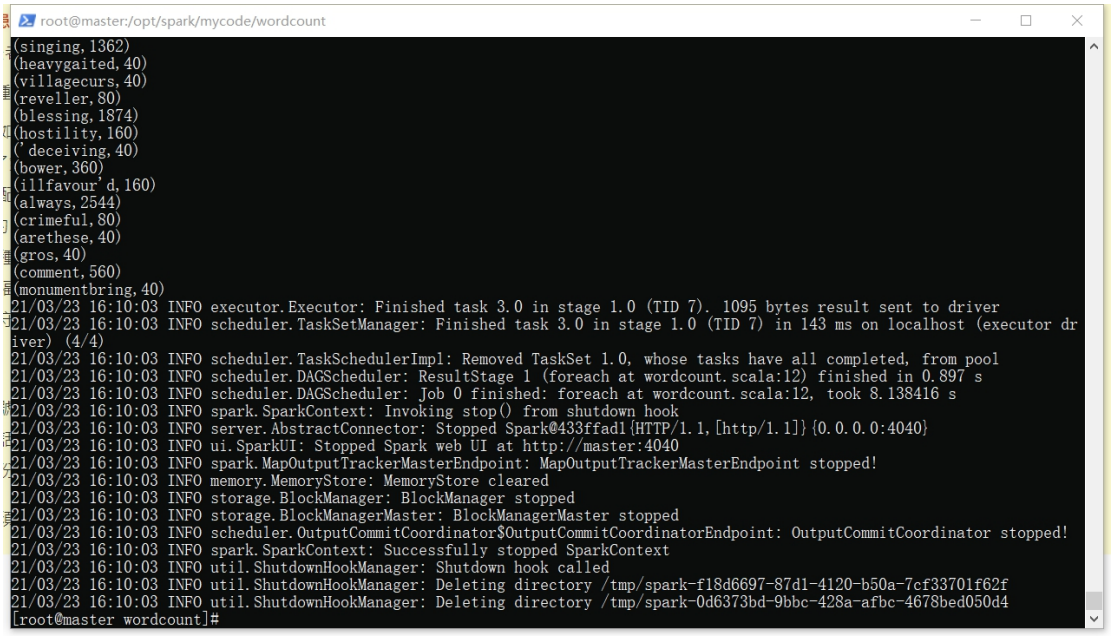
P3 Spark cluster configuration is complete

```
[root@master ~]# hadoop fs -cat /spark/output1/part-00000
(d,4291467)
(',1070168)
(s,7768732)
(e,13610507)
(p,2309813)
(x,221006)
(w,1980207)
(z,63568)
(a,7559931)
(t,7427774)
(i,6201250)
(b,1635453)
(y,1770526)
(k,1299074)
(u,3559108)
(h,4450751)
( ,18918060)
(o,7516281)
(n,6848395)
(f,1878346)
(q,146736)
(j,174870)
(v,1256332)
(r,7634095)
(g,2641335)
(l,5546644)
(m,2739274)
(c,3155424)
[root@master ~]#
```

P4 wordcount result(script in lecture notes)

```
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
export SPARK_DIST_CLASSPATH=$(hadoop classpath)
export SBT_HOME=/opt/sbt
export PATH=$PATH:$SBT_HOME/bin
```

P5 spark and sbt configuration



P6 wordcount job

① Submit and run one Spark WordCount job

Running time: 8.134s

② Submit and run three Spark WordCount jobs at the same time

Average running time: 6.890s

③ Submit and run three Spark WordCount jobs at the same time

Average running time: 6.867s

Q2 Setup and run a Spark application over Kubernetes

```
File contains no section headers.
file: file:///etc/yum.repos.d/kubernetes.repo, line: 1
'name=kubernetes Repo\n'
[root@master yum.repos.d]# ls
CentOS-Base.repo  docker-ce.repo  kubernetes.repo
[root@master yum.repos.d]# yum install gcc
已加载插件：fastestmirror


File contains no section headers.
file: file:///etc/yum.repos.d/kubernetes.repo, line: 1
'name=kubernetes Repo\n'
[root@master yum.repos.d]# yum repolist
已加载插件：fastestmirror
Determining fastest mirrors
 * base: mirrors.aliyun.com
 * extras: mirrors.aliyun.com
 * updates: mirrors.aliyun.com
base                                                                                    | 3.6 kB  0
docker-ce-stable                                                                        | 3.5 kB  0
extras                                                                                  | 2.9 kB  0
kubernetes                                                                              | 1.4 kB  0
updates                                                                                 | 2.9 kB  0
(1/7): docker-ce-stable/7/x86_64/updateinfo                                             |   55 B  0
(2/7): docker-ce-stable/7/x86_64/primary_db                                             |  58 kB  0
(3/7): extras/7/x86_64/primary_db                                                       | 227 kB  0
(4/7): kubernetes/primary                                                               |  87 kB  0
(5/7): base/7/x86_64/group_gz                                                           | 153 kB  0
(6/7): updates/7/x86_64/primary_db                                                      | 6.5 MB  0
```

P7 Kubernetes configuration

```
http://mirrors.aliyuncs.com/centos/7/os/x86_64/repodata/6d0c3a488c282fe537794b5946b01e28c7f44db79097bb06826e1c0c88bad5ef-pr
imary.sqlite.bz2: [Errno 14] curl#7 - "Failed connect to mirrors.aliyuncs.com:80; Connection refused"
正在尝试其它镜像。
(7/7): base/7/x86_64/primary_db                                                | 6.1 MB  00:00:00
kubernetes                                                                                 642/642
源标识                                源名称                                                状态
base/7/x86_64                         CentOS-7 - Base - mirrors.aliyun.com                 10,072
docker-ce-stable/7/x86_64             Docker CE Stable - x86_64                               108
extras/7/x86_64                       CentOS-7 - Extras - mirrors.aliyun.com                  460
kubernetes                            kubernetes Repo                                         642
updates/7/x86_64                      CentOS-7 - Updates - mirrors.aliyun.com               1,898
repolist: 13,180
```

P8 Kubernetes installation complete

```
[root@master yum.repos.d]# scp CentOS-Base.repo docker-ce.repo kubernetes.repo slave1:/etc/yum.repos.d/
CentOS-Base.repo                                              100% 2523     1.9MB/s   00:00
docker-ce.repo                                                100% 1919     2.6MB/s   00:00
kubernetes.repo                                               100%  199   274.0KB/s   00:00
[root@master yum.repos.d]# scp CentOS-Base.repo docker-ce.repo kubernetes.repo slave2:/etc/yum.repos.d/
CentOS-Base.repo                                              100% 2523     1.2MB/s   00:00
docker-ce.repo                                                100% 1919     2.7MB/s   00:00
kubernetes.repo                                               100%  199   376.1KB/s   00:00
[root@master yum.repos.d]# scp CentOS-Base.repo docker-ce.repo kubernetes.repo slave3:/etc/yum.repos.d/
CentOS-Base.repo                                              100% 2523     1.9MB/s   00:00
docker-ce.repo                                                100% 1919   908.6KB/s   00:00
kubernetes.repo                                               100%  199   190.6KB/s   00:00
[root@master yum.repos.d]#
```

P9 copy kubernetes to three slaves

```
root@master:~
nd "node-role.kubernetes.io/control-plane=''  (deprecated)"
[mark-control-plane] Marking the node master as control-plane by adding the taints [node-role.kubernetes.io/master:NoSch
edule]
[bootstrap-token] Using token: mnuvcg.bzakmy7sq21v74sz
[bootstrap-token] Configuring bootstrap tokens, cluster-info ConfigMap, RBAC Roles
[bootstrap-token] configured RBAC rules to allow Node Bootstrap tokens to get nodes
[bootstrap-token] configured RBAC rules to allow Node Bootstrap tokens to post CSRs in order for nodes to get long term
certificate credentials
[bootstrap-token] configured RBAC rules to allow the csrapprover controller automatically approve CSRs from a Node Boots
trap Token
[bootstrap-token] configured RBAC rules to allow certificate rotation for all node client certificates in the cluster
[bootstrap-token] Creating the "cluster-info" ConfigMap in the "kube-public" namespace
[kubelet-finalize] Updating "/etc/kubernetes/kubelet.conf" to point to a rotatable kubelet client certificate and key
[addons] Applied essential addon: CoreDNS
[addons] Applied essential addon: kube-proxy

Your Kubernetes control-plane has initialized successfully!

To start using your cluster, you need to run the following as a regular user:

  mkdir -p $HOME/.kube
  sudo cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
  sudo chown $(id -u):$(id -g) $HOME/.kube/config

Alternatively, if you are the root user, you can run:

  export KUBECONFIG=/etc/kubernetes/admin.conf

You should now deploy a pod network to the cluster.
Run "kubectl apply -f [podnetwork].yaml" with one of the options listed at:
  https://kubernetes.io/docs/concepts/cluster-administration/addons/

Then you can join any number of worker nodes by running the following on each as root:

kubeadm join 192.168.2.130:6443 --token mnuvcg.bzakmy7sq21v74sz \
    --discovery-token-ca-cert-hash sha256:9caed533cd0a1e5c73786c2dea844ebc0e6a05a2918ce8b451d1591e6ae7cd35
[root@master ~]#
```

P10 Kubeadm initialization

```
[root@master ~]# docker image ls
REPOSITORY                                TAG        IMAGE ID       CREATED         SIZE
k8s.gcr.io/kube-proxy                     v1.20.5    5384b1650507   8 days ago      118MB
k8s.gcr.io/kube-controller-manager        v1.20.5    6f0c3da8c99e   8 days ago      116MB
k8s.gcr.io/kube-scheduler                 v1.20.5    8d13f1db8bfb   8 days ago      47.3MB
k8s.gcr.io/kube-apiserver                 v1.20.5    d7e24aeb3b10   8 days ago      122MB
k8s.gcr.io/etcd                           3.4.13-0   0369cf4303ff   7 months ago    253MB
k8s.gcr.io/coredns                        1.7.0      bfe3a36ebd25   9 months ago    45.2MB
k8s.gcr.io/pause                          3.2        80d28bedfe5d   13 months ago   683kB
[root@master ~]#
```

P11 Docker image information

```
[root@master ~]# ss -ntl
State      Recv-Q Send-Q              Local Address:Port                             Peer Address:Por
LISTEN     0      100                     127.0.0.1:25                                  *:*
LISTEN     0      128                     127.0.0.1:36675                               *:*
LISTEN     0      128                     127.0.0.1:10248                               *:*
LISTEN     0      128                     127.0.0.1:10249                               *:*
LISTEN     0      128                  192.168.2.130:2379                               *:*
LISTEN     0      128                     127.0.0.1:2379                                *:*
LISTEN     0      128                  192.168.2.130:2380                               *:*
LISTEN     0      128                     127.0.0.1:2381                                *:*
LISTEN     0      128                     127.0.0.1:10257                               *:*
LISTEN     0      128                     127.0.0.1:10259                               *:*
LISTEN     0      128                           *:22                                    *:*
LISTEN     0      100                       [::1]:25                                   [::]:*
LISTEN     0      128                        [::]:10250                                [::]:*
LISTEN     0      128                        [::]:6443                                 [::]:*
LISTEN     0      128                        [::]:10256                                [::]:*
LISTEN     0      128                        [::]:22                                   [::]:*
```

P12 Component and port status

```
root@master:~
State      Recv-Q Send-Q              Local Address:Port                             Peer Address:Port
LISTEN     0      100                     127.0.0.1:25                                  *:*
LISTEN     0      128                     127.0.0.1:36675                               *:*
LISTEN     0      128                     127.0.0.1:10248                               *:*
LISTEN     0      128                     127.0.0.1:10249                               *:*
LISTEN     0      128                  192.168.2.130:2379                               *:*
LISTEN     0      128                     127.0.0.1:2379                                *:*
LISTEN     0      128                  192.168.2.130:2380                               *:*
LISTEN     0      128                     127.0.0.1:2381                                *:*
LISTEN     0      128                     127.0.0.1:10257                               *:*
LISTEN     0      128                     127.0.0.1:10259                               *:*
LISTEN     0      128                           *:22                                    *:*
LISTEN     0      100                       [::1]:25                                   [::]:*
LISTEN     0      128                        [::]:10250                                [::]:*
LISTEN     0      128                        [::]:6443                                 [::]:*
LISTEN     0      128                        [::]:10256                                [::]:*
LISTEN     0      128                        [::]:22                                   [::]:*
[root@master ~]# mkdir -p $HOME/.kube
[root@master ~]# cp -i /etc/kubernetes/admin.conf $HOME/.kube/config
[root@master ~]# kubectl get cs
Warning: v1 ComponentStatus is deprecated in v1.19+
NAME                 STATUS      MESSAGE
      ERROR
scheduler            Unhealthy   Get "http://127.0.0.1:10251/healthz": dial tcp 127.0.0.1:10251: connect: connection ref
used
controller-manager   Unhealthy   Get "http://127.0.0.1:10252/healthz": dial tcp 127.0.0.1:10252: connect: connection ref
used
etcd-0               Healthy     {"health":"true"}

[root@master ~]# systemctl restart kubelet.service
[root@master ~]# kubectl get cs
Warning: v1 ComponentStatus is deprecated in v1.19+
NAME                 STATUS    MESSAGE             ERROR
scheduler            Healthy   ok
controller-manager   Healthy   ok
etcd-0               Healthy   {"health":"true"}
[root@master ~]#
```

P13 Check if the component is operating normally

```
[root@master ~]# kubectl get nodes
NAME     STATUS     ROLES                  AGE   VERSION
master   NotReady   control-plane,master   23m   v1.20.5
```

P14 View and verify whether the node information is successful

```
[root@master ~]# kubectl get  nodes
NAME     STATUS   ROLES                  AGE   VERSION
master   Ready    control-plane,master   25m   v1.20.5
```

P15 After installing flannel, the status becomes ready

```
[root@master ~]# scp /usr/lib/systemd/system/docker.service slave1:/usr/lib/systemd/system/docker.service
docker.service                                                                100% 1795     2.0MB/s   00:00
[root@master ~]# scp /etc/sysconfig/kubelet slave1:/etc/sysconfig/
kubelet                                                                       100%   42    59.0KB/s   00:00
[root@master ~]# scp /usr/lib/systemd/system/docker.service slave2:/usr/lib/systemd/system/docker.service
docker.service                                                                100% 1795     1.8MB/s   00:00
[root@master ~]# scp /etc/sysconfig/kubelet slave2:/etc/sysconfig/
kubelet                                                                       100%   42    52.6KB/s   00:00
[root@master ~]# scp /usr/lib/systemd/system/docker.service slave3:/usr/lib/systemd/system/docker.service
docker.service                                                                100% 1795   717.8KB/s   00:00
[root@master ~]# scp /etc/sysconfig/kubelet slave3:/etc/sysconfig/
kubelet                                                                       100%   42    26.0KB/s   00:00
[root@master ~]#
```

P16 Transfer the configuration file to the three slaves

```
[root@master ~]# kubectl get nodes
NAME     STATUS   ROLES                  AGE     VERSION
master   Ready    control-plane,master   28m     v1.20.5
slave1   Ready    <none>                 9m19s   v1.20.5
slave2   Ready    <none>                 4m45s   v1.20.5
slave3   Ready    <none>                 38s     v1.20.5
[root@master ~]# kubectl get pods -n kube-system -o wide
NAME                              READY   STATUS    RESTARTS   AGE     IP              NODE     NOMINATED NODE   READINE
S GATES
coredns-74ff55c5b-n5s4k           1/1     Running   0          28m     10.244.0.3      master   <none>           <none>
coredns-74ff55c5b-t6cm5           1/1     Running   0          28m     10.244.0.2      master   <none>           <none>
etcd-master                       1/1     Running   0          28m     192.168.2.130   master   <none>           <none>
kube-apiserver-master             1/1     Running   0          28m     192.168.2.130   master   <none>           <none>
kube-controller-manager-master    1/1     Running   0          24m     192.168.2.130   master   <none>           <none>
kube-flannel-ds-5zpxh             1/1     Running   0          4m57s   192.168.2.132   slave2   <none>           <none>
kube-flannel-ds-d8625             1/1     Running   0          23m     192.168.2.130   master   <none>           <none>
kube-flannel-ds-m72wm             1/1     Running   0          50s     192.168.2.133   slave3   <none>           <none>
kube-flannel-ds-rjqrt             1/1     Running   0          9m31s   192.168.2.131   slave1   <none>           <none>
kube-proxy-6v48d                  1/1     Running   0          28m     192.168.2.130   master   <none>           <none>
kube-proxy-jjn9b                  1/1     Running   0          4m57s   192.168.2.132   slave2   <none>           <none>
kube-proxy-rjltk                  1/1     Running   0          9m31s   192.168.2.131   slave1   <none>           <none>
kube-proxy-xwg86                  1/1     Running   0          50s     192.168.2.133   slave3   <none>           <none>
kube-scheduler-master             1/1     Running   0          24m     192.168.2.130   master   <none>           <none>
```

P17 The kubernetes cluster is built

```
[root@master ~]# kubectl create -f /opt/spark/namespace-spark-cluster.yaml
namespace/spark-cluster created
[root@master ~]# kubectl get namespaces
NAME              STATUS   AGE
default           Active   17h
kube-node-lease   Active   17h
kube-public       Active   17h
kube-system       Active   17h
spark-cluster     Active   9s
```

P18 Add spark cluster in kubernetes

```
[root@master ~]# kubectl get pods |grep spark-master
spark-master-controller-96rmd   1/1     Running   0          2m19s
```

P19 Spark-master is running

```
[root@master ~]# kubectl create -f /opt/spark/spark-master-service.yaml
service/spark-master created
[root@master ~]# kubectl get svc |grep spark-master
spark-master   ClusterIP   10.98.247.207   <none>        7077/TCP,8080/TCP   25s
```

P20 start spark service

```
[root@master docker]# sudo docker build -t wordcount-app .
Sending build context to Docker daemon  3.072kB
[WARNING]: Empty continuation line found in:
    RUN echo "==> Install curl helper tool..."&& apt-get update && DEBIAN_FRONTEND=noninteractive apt-get install -y --f
orce-yes curl && echo "===> install from Typesafe repo (contains old versions but they have all dependencies we need lat
er on)"&& curl -sSL http://apt.typesafe.com/repo-deb-build-0002.deb -o repo-deb.deb && dpkg -i repo-deb.deb && apt-get u
pdate && echo "===> install Scala"&& DEBIAN_FRONTEND=noninteractive apt-get install -y --force-yes libjansi-java && curl
 -sSL $SCALA_TARBALL -o scala.deb && dpkg -i scala.deb && echo "===> clean up..."&& rm -f *.deb && apt-get remove -y --a
uto-remove curl && apt-get clean && rm -rf /var/lib/apt/lists/*
[WARNING]: Empty continuation lines will become errors in a future release.
Step 1/8 : FROM centos:latest
latest: Pulling from library/centos
7a0437f04f83: Pull complete
Digest: sha256:5528e8b1b1719d34604c87e11dcd1c0a20bedf46e83b5632cdeac91b8c04efc1
Status: Downloaded newer image for centos:latest
 ---> 300e315adb2f
Step 2/8 : MAINTAINER brianlk7(1k120@ie.cuhk.edu.hk)
 ---> Running in e6bc9e0f68d7
Removing intermediate container e6bc9e0f68d7
 ---> e67041981b17
Step 3/8 : ENV SCALA_VERSION 2.11.12
 ---> Running in 881de0bd6ca6
Removing intermediate container 881de0bd6ca6
 ---> 75f64637d8e4
Step 4/8 : ENV SCALA_TARBALL http://www.scala-lang.org/files/archive/scala-$SCALA_VERSION.deb
 ---> Running in 88cb6e380c5d
Removing intermediate container 88cb6e380c5d
 ---> ec29d7e6fc46
Step 5/8 : RUN echo "==> Install curl helper tool..."&& apt-get update && DEBIAN_FRONTEND=noninteractive apt-get install
 -y --force-yes curl && echo "===> install from Typesafe repo (contains old versions but they have all dependencies we n
eed later on)"&& curl -sSL http://apt.typesafe.com/repo-deb-build-0002.deb -o repo-deb.deb && dpkg -i repo-deb.deb && ap
t-get update && echo "===> install Scala"&& DEBIAN_FRONTEND=noninteractive apt-get install -y --force-yes libjansi-java
&& curl -sSL $SCALA_TARBALL -o scala.deb && dpkg -i scala.deb && echo "===> clean up..."&& rm -f *.deb && apt-get remove
 -y --auto-remove curl && apt-get clean && rm -rf /var/lib/apt/lists/*
 ---> Running in 79da19bdd58c
```

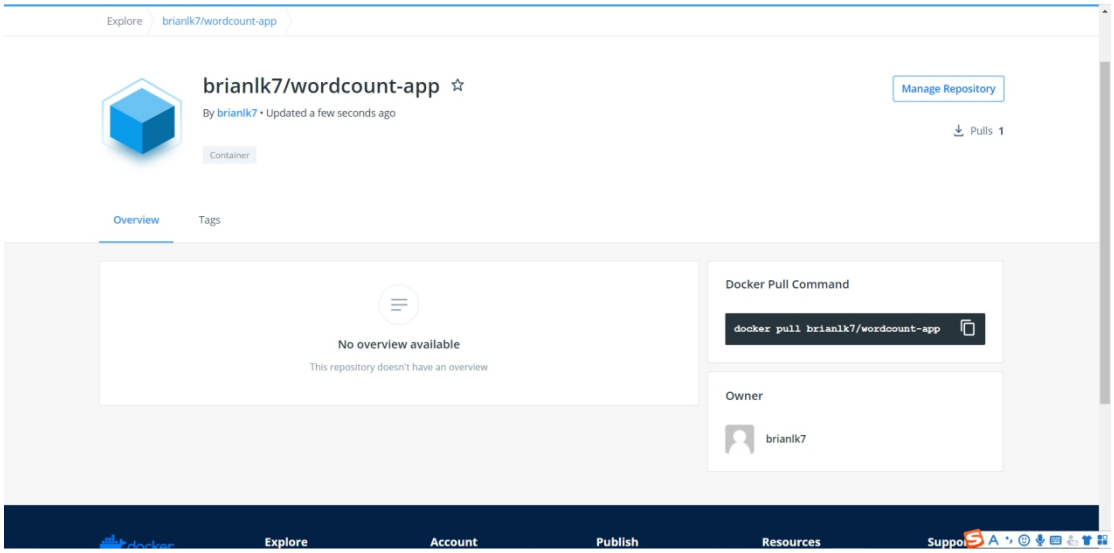
P21 docker image building

```
[root@master scala-docker-app]# docker images
REPOSITORY                                TAG                 IMAGE ID            CREATED             SIZE
wordcount-app                             latest              10eb8fc75457        6 minutes ago       1.01GB
<none>                                    <none>              ec29d7e6fc46        35 minutes ago      209MB
k8s.gcr.io/kube-proxy                     v1.20.5             5384b1650507        10 days ago         118MB
k8s.gcr.io/kube-controller-manager        v1.20.5             6f0c3da8c99e        10 days ago         116MB
k8s.gcr.io/kube-apiserver                 v1.20.5             d7e24aeb3b10        10 days ago         122MB
k8s.gcr.io/kube-scheduler                 v1.20.5             8d13f1db8bfb        10 days ago         47.3MB
quay.io/coreos/flannel                    v0.13.1-rc2         dee1cac4dd20        7 weeks ago         64.3MB
k8s.gcr.io/etcd                           3.4.13-0            0369cf4303ff        7 months ago        253MB
k8s.gcr.io/coredns                        1.7.0               bfe3a36ebd25        9 months ago        45.2MB
k8s.gcr.io/pause                          3.2                 80d28bedfe5d        13 month 截图(Alt + A) 83kB
```

P22 docker images output

```
[root@master scala-docker-app]# docker push brianlk7/wordcount-app
Using default tag: latest
The push refers to repository [docker.io/brianlk7/wordcount-app]
48b7d04926f9: Pushed
d0c1fce89d41: Pushed
626c2c867535: Mounted from williamyeh/java7
8d4d1ab5ff74: Mounted from williamyeh/java7
latest: digest: sha256:9fc2f53603f58d3b36f07b7ab7ea0e72093f447f951827fb3cf3ca0eab041484 size: 1162
```

P23 Push docker image to docker.hub

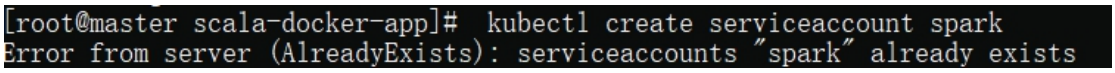


P24 URL: https://hub.docker.com/r/brianlk7/wordcount-app

```
[root@master ~]# /opt/spark/bin/spark-submit \ --master k8s://https://192.168.2.130:6443 \ --deploy-mode cluster \ --name WordC
ount \ --class /opt/spark/mycode/wordcount/wordcount.scala \ --conf spark.app.name=WordCount \ --conf spark.kubernetes.authenti
cate.diver.serviceAccountName=spark \ --conf spark.kubernetes.container.image=http://docker.io/myrepo/spark:v2.4.7 \ local:///o
pt/spark/mycode/wordcount/target/scala-2.11/wordcount-project_2.11-1.0.jar
```

P25 submit spark job to kubernetes (d)

P26 wordcount result



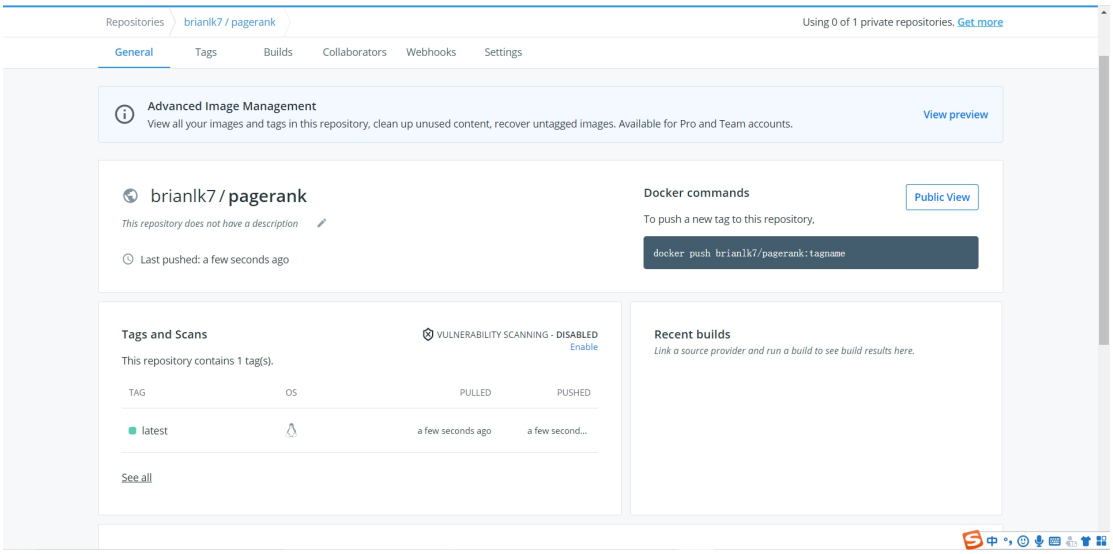
P27 service account setup (c)

Q3 Working with RDDs in Spark

(a) The program and result is attached in the zip

I choose to iterate 50 times, each iteration 2-3 minutes, the total time is about two hours.

(b)



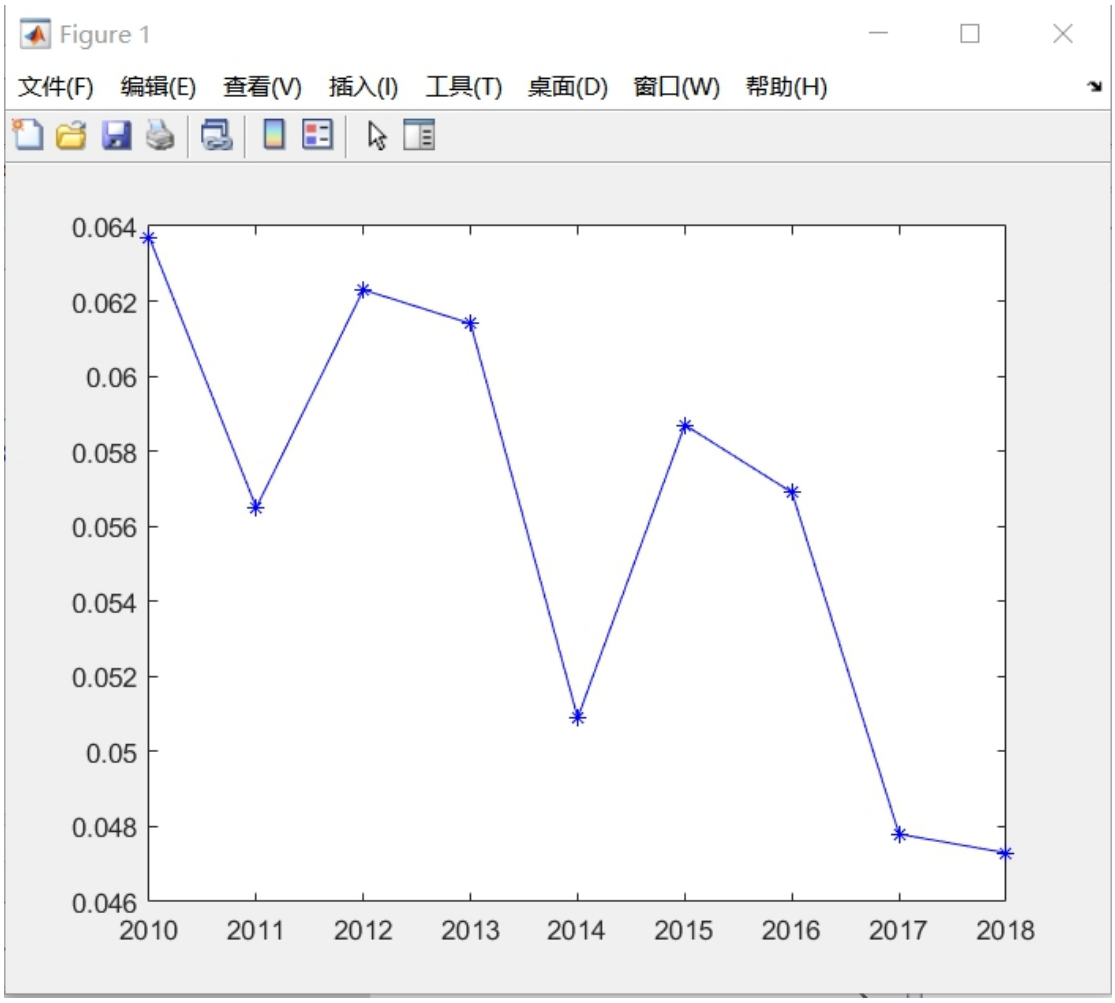
P28 URL:registry.hub.docker.com/repository/docker/brianlk7/pagerank

In this part, the running time is obviously reduced, the average time is about 50 minutes. And the result stay the same.

Q4 Using Spark SQL

The result of (a) and (b) are attached in the zip (in the folder sparkSQL)

For question(c), I list the result in the textfile(GUN). And we can see the changing of the percentage of gun offense in the picture below. We can know from the figure that the gun percentage overall slightly decreased. The Obama’ s executive actions on gun control has got some positive effect.



P29 Gun percentage