Wordcount, Pagerank running on Spark

Kadirbek Sharau
19995


```
# Install Helm Chart for Spark
helm repo add bitnami https://charts.bitnami.com/bitnami
helm install my-spark bitnami/spark

# Check the services to get the External IP of the Spark master
kubectl get svc

# Deploy a sample application to verify that the installation is correct
kubectl run --namespace default my-spark-client --rm --tty -i --restart='Never' \
    --image docker.io/bitnami/spark:3.0.1-debian-10-r115 -- spark-submit \
    --master spark://<external-ip>:7077 --deploy-mode cluster \
    --class org.apache.spark.examples.JavaWordCount
/opt/bitnami/spark/examples/jars/spark-examples_2.12-3.0.1.jar \
    /opt/bitnami/spark/examples/src/main/resources/people.txt

# Set up a Persistent Volume and Pod
cat <<EOF >spark-pvc.yaml
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: spark-data-pvc
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 2Gi
  storageClassName: nfs
---
apiVersion: v1
kind: Pod
metadata:
  name: spark-data-pod
spec:
  volumes:
    - name: spark-data-pv
      persistentVolumeClaim:
        claimName: spark-data-pvc
```

```
  containers:
    - name: inspector
      image: bitnami/minideb
      command:
        - sleep
        - infinity
      volumeMounts:
        - mountPath: "/data"
          name: spark-data-pv
EOF

kubectl apply -f spark-pvc.yaml

# Check the status of PVC and Pods
kubectl get pvc
kubectl get pods

# Launch a Spark job using the setup
kubectl run --namespace default spark-client --rm --tty -i --restart='Never' \
    --image docker.io/bitnami/spark:3.0.1-debian-10-r115 -- spark-submit \
    --master spark://LOAD-BALANCER-External-ip-ADDRESS:7077 --deploy-mode cluster \
    --class org.apache.spark.examples.JavaWordCount /data/my.jar /data/test.txt

# If facing issues with PySpark on Bitnami Spark, modify the command to exclude problematic
arguments:
export
PYTHONPATH=/opt/bitnami/spark/python/lib/py4j-0.10.9.7-src.zip:/opt/bitnami/spark/python/:/opt/bitnami/spark/python/:$PYTHONPATH
export PYTHONSTARTUP=/opt/bitnami/spark/python/pyspark/shell.py
exec "$SPARK_HOME"/bin/spark-submit pyspark-shell-main
```

Screenshots

1

```
Updated property [core/project].
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud container clusters create spark --num-nodes=1 --machine-type=e2-highmem-2 --region=us-west1
Default change: VPC-native is the default mode during cluster creation for versions greater than 1.21.0-gke.1500. To create advanced routes based clusters, please pass the `--
no-enable-ip-alias` flag
Note: Your Pod address range (`--cluster-ipv4-cidr`) can accommodate at most 1008 node(s).
Creating cluster spark in us-west1... Cluster is being health-checked (master i
s healthy)...done.
Created [https://container.googleapis.com/v1/projects/silent-moment-426921-k7/zones/us-west1/clusters/spark].
To inspect the contents of your cluster, go to: https://console.cloud.google.com/kubernetes/workload_/gcloud/us-west1/spark?project=silent-moment-426921-k7
CRITICAL: ACTION REQUIRED: gke-gcloud-auth-plugin, which is needed for continued use of kubectl, was not found or is not executable. Install gke-gcloud-auth-plugin for use wit
h kubectl by following https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
kubeconfig entry generated for spark.
NAME   LOCATION  MASTER_VERSION      MASTER_IP       MACHINE_TYPE  NODE_VERSION        NUM_NODES  STATUS
spark  us-west1  1.29.5-gke.1091002  35.233.251.120  e2-highmem-2  1.29.5-gke.1091002  3          RUNNING
```

2

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ %        brew install helm

==> Auto-updating Homebrew...
Adjust how often this is run with HOMEBREW_AUTO_UPDATE_SECS or disable with
HOMEBREW_NO_AUTO_UPDATE. Hide these hints with HOMEBREW_NO_ENV_HINTS (see `man brew`).
==> Auto-updated Homebrew!
Updated 5 taps (homebrew/services, mongodb/brew, dart-lang/dart, homebrew/core and homebrew/cask).
==> New Formulae
cortexso         cotila           kaskade          litmusctl        nerdfetch        openbao          pug              ryelang          soapyhackrf      wcurl
==> New Casks
blip                                                      crashplan                                                 notchnook

You have 38 outdated formulae installed.

==> Downloading https://ghcr.io/v2/homebrew/core/helm/manifests/3.15.3_1
######################################################################################################################################################################################## 100.
==> Fetching helm
==> Downloading https://ghcr.io/v2/homebrew/core/helm/blobs/sha256:0462f1b6824f651808aa2354b7f89743c793b37419409782632a227e820761a9
######################################################################################################################################################################################## 100.
==> Pouring helm--3.15.3_1.arm64_sonoma.bottle.tar.gz
==> Caveats
zsh completions have been installed to:
  /opt/homebrew/share/zsh/site-functions
==> Summary
🍺  /opt/homebrew/Cellar/helm/3.15.3_1: 66 files, 50.2MB
==> Running `brew cleanup helm`...
Disable this behaviour by setting HOMEBREW_NO_INSTALL_CLEANUP.
Hide these hints with HOMEBREW_NO_ENV_HINTS (see `man brew`).
```

3

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % helm repo add stable https://charts.helm.sh/stable
helm install nfs stable/nfs-server-provisioner --set persistence.enabled=true,persistence.size=5Gi

"stable" has been added to your repositories
WARNING: This chart is deprecated
Error: INSTALLATION FAILED: Kubernetes cluster unreachable: Get "https://35.233.251.120/version": getting credentials: exec: executable gke-gcloud-auth-plugin not found

It looks like you are trying to use a client-go credential plugin that is not installed.

To learn more about this feature, consult the documentation available at:
      https://kubernetes.io/docs/reference/access-authn-authz/authentication/#client-go-credential-plugins

Install gke-gcloud-auth-plugin for use with kubectl by following https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud components update

Beginning update. This process may take several minutes.
WARNING: The ARM versions of the following components are available, replacing installed x86_64 versions: [anthoscli-darwin-x86_64, gcloud-crc32c-darwin-x86_64].


Your current Google Cloud CLI version is: 462.0.1
You will be upgraded to version: 483.0.0
```

These components will be removed.		
Name	Version	Size
Google Cloud CRC32C Hash Tool	1.0.0	
anthoscli	0.2.47	

These components will be updated.		
Name	Version	Size
BigQuery Command Line Tool	2.1.6	1.7 MiB
Cloud Storage Command Line Tool	5.30	11.3 MiB
Google Cloud CLI Core Libraries	2024.06.28	18.9 MiB

4

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud components install gke-gcloud-auth-plugin


Your current Google Cloud CLI version is: 483.0.0
Installing components from version: 483.0.0

┌──────────────────────────────────────────────────┐
│     These components will be installed.          │
├──────────────────────┬─────────┬─────────────────┤
│         Name         │ Version │       Size      │
├──────────────────────┼─────────┼─────────────────┤
│ gke-gcloud-auth-plugin │   0.5.9 │ 4.0 MiB         │
└──────────────────────┴─────────┴─────────────────┘

For the latest full release notes, please visit:
  https://cloud.google.com/sdk/release_notes

Once started, canceling this operation may leave your SDK installation in an inconsistent state.

Do you want to continue (Y/n)?  y

Performing in place update...

╔══════════════════════════════════════════════════════════════╗
╠═ Downloading: gke-gcloud-auth-plugin                        ═╣
╠═ Downloading: gke-gcloud-auth-plugin                        ═╣
╠═ Installing: gke-gcloud-auth-plugin                         ═╣
╠═ Installing: gke-gcloud-auth-plugin                         ═╣
╚══════════════════════════════════════════════════════════════╝

Performing post processing steps...done.

Update done!

kadirbeksharau@Kadirbeks-MacBook-Pro ~ %
```

5

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pods

NAME                              READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0      1/1     Running   0          19m
spark-data-pod                    1/1     Running   0          100s
spark-pod                         0/1     Pending   0          13m
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % docker run -v /tmp:/tmp -it bitnami/spark -- find /opt/bitnami/spark/examples/jars/ -name spark-examples* -exec cp {} /tmp/my.ja
r \;
```

6

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud components install gke-gcloud-auth-plugin


Your current Google Cloud CLI version is: 483.0.0
Installing components from version: 483.0.0

┌──────────────────────────────────────────────────┐
│     These components will be installed.          │
├──────────────────────┬─────────┬─────────────────┤
│         Name         │ Version │       Size      │
├──────────────────────┼─────────┼─────────────────┤
│ gke-gcloud-auth-plugin │   0.5.9 │ 4.0 MiB         │
└──────────────────────┴─────────┴─────────────────┘

For the latest full release notes, please visit:
  https://cloud.google.com/sdk/release_notes

Once started, canceling this operation may leave your SDK installation in an inconsistent state.

Do you want to continue (Y/n)?  y

Performing in place update...

╔══════════════════════════════════════════════════════════════╗
╠═ Downloading: gke-gcloud-auth-plugin                        ═╣
╠═ Downloading: gke-gcloud-auth-plugin                        ═╣
╠═ Installing: gke-gcloud-auth-plugin                         ═╣
╠═ Installing: gke-gcloud-auth-plugin                         ═╣
╚══════════════════════════════════════════════════════════════╝

Performing post processing steps...done.

Update done!

kadirbeksharau@Kadirbeks-MacBook-Pro ~ % kubectl plugin list
```

7

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud components install gke-gcloud-auth-plugin


All components are up to date.
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % gcloud container clusters get-credentials [CLUSTER_NAME] --zone [ZONE] --project [PROJECT_ID]

zsh: no matches found: [CLUSTER_NAME]
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % helm install nfs stable/nfs-server-provisioner --set persistence.enabled=true,persistence.size=5Gi

WARNING: This chart is deprecated
NAME: nfs
LAST DEPLOYED: Sat Jul 13 13:19:01 2024
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
NOTES:
The NFS Provisioner service has now been installed.

A storage class named 'nfs' has now been created
and is available to provision dynamic volumes.

You can use this storageclass by creating a `PersistentVolumeClaim` with the
correct storageClassName attribute. For example:

    ---
    kind: PersistentVolumeClaim
    apiVersion: v1
    metadata:
      name: test-dynamic-volume-claim
    spec:
      storageClassName: "nfs"
      accessModes:
        - ReadWriteOnce
      resources:
        requests:
          storage: 100Mi
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % kubectl apply -f spark-pvc.yaml
```

8

```
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % kubectl apply -f spark-pvc.yaml

error: the path "spark-pvc.yaml" does not exist
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % ls
65                      Documents               Library                 My Games                Public                  key                     node_modules
Applications            Downloads               Movies                  Pictures                cache                   key.pub                 package-lock.json
Desktop                 Google Drive            Music                   Postman                 go                      myenv                   package.json
kadirbeksharau@Kadirbeks-MacBook-Pro ~ %
kadirbeksharau@Kadirbeks-MacBook-Pro ~ % find ~ -type f -name "spark-pvc.yaml"

find: /Users/kadirbeksharau/Library/Application Support/CallHistoryTransactions: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/CloudDocs/session/db: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/com.apple.sharedfilelist: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/Knowledge: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/com.apple.TCC: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/FileProvider: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/FaceTime: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/com.apple.avfoundation/Frecents: Operation not permitted
find: /Users/kadirbeksharau/Library/Application Support/CallHistoryDB: Operation not permitted
find: /Users/kadirbeksharau/Library/Assistant/SiriVocabulary: Operation not permitted
find: /Users/kadirbeksharau/Library/Daemon Containers: Operation not permitted
find: /Users/kadirbeksharau/Library/Autosave Information: Operation not permitted
find: /Users/kadirbeksharau/Library/IdentityServices: Operation not permitted
find: /Users/kadirbeksharau/Library/Calendars: Operation not permitted
find: /Users/kadirbeksharau/Library/Messages: Operation not permitted
find: /Users/kadirbeksharau/Library/HomeKit: Operation not permitted
find: /Users/kadirbeksharau/Library/Sharing: Operation not permitted
find: /Users/kadirbeksharau/Library/com.apple.aiml.instrumentation: Operation not permitted
find: /Users/kadirbeksharau/Library/Mail: Operation not permitted
find: /Users/kadirbeksharau/Library/Trial: Operation not permitted
find: /Users/kadirbeksharau/Library/AppleMediaServices: Operation not permitted
find: /Users/kadirbeksharau/Library/DuetExpertCenter: Operation not permitted
find: /Users/kadirbeksharau/Library/Accounts: Operation not permitted
find: /Users/kadirbeksharau/Library/Safari: Operation not permitted
find: /Users/kadirbeksharau/Library/Biome: Operation not permitted
find: /Users/kadirbeksharau/Library/IntelligencePlatform: Operation not permitted
find: /Users/kadirbeksharau/Library/Shortcuts: Operation not permitted
find: /Users/kadirbeksharau/Library/Suggestions: Operation not permitted
find: /Users/kadirbeksharau/Library/Weather: Operation not permitted
find: /Users/kadirbeksharau/Library/Group Containers/group.com.apple.stocks-news: Operation not permitted
find: /Users/kadirbeksharau/Library/Group Containers/group.com.apple.photolibraryd.private: Operation not permitted
```

9

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl apply -f spark-pvc.yaml

pod/spark-pod created
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pvc spark-pvc
kubectl get pods

Error from server (NotFound): persistentvolumeclaims "spark-pvc" not found
NAME                            READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0    1/1     Running   0          6m49s
spark-pod                       0/1     Pending   0          7s
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pvc spark-pvc
kubectl get pods

Error from server (NotFound): persistentvolumeclaims "spark-pvc" not found
NAME                            READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0    1/1     Running   0          7m19s
spark-pod                       0/1     Pending   0          37s
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pvc
kubectl get pods

NAME                               STATUS   VOLUME                                     CAPACITY   ACCESS MODES   STORAGECLASS   VOLUMEATTRIBUTESCLASS   AGE
data-nfs-nfs-server-provisioner-0   Bound    pvc-ff46bb7c-c8ae-4f2f-877c-bf683aa22dac   5Gi        RWO            standard-rwo   <unset>                 7m45s
NAME                            READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0    1/1     Running   0          7m45s
spark-pod                       0/1     Pending   0          63s
```

10

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl describe pvc spark-pvc
kubectl describe pod spark-pod

Error from server (NotFound): persistentvolumeclaims "spark-pvc" not found
Name:         spark-pod
Namespace:    default
Priority:     0
Node:         <none>
Labels:       <none>
Annotations:  cloud.google.com/cluster_autoscaler_unhelpable_since: 2024-07-13T20:25:46+0000
              cloud.google.com/cluster_autoscaler_unhelpable_until: Inf
Status:       Pending
IP:
IPs:          <none>
Containers:
  spark-container:
    Image:        nginx
    Port:         <none>
    Host Port:    <none>
    Environment:  <none>
    Mounts:
      /usr/share/nginx/html from nfs-storage (rw)
      /var/run/secrets/kubernetes.io/serviceaccount from kube-api-access-kbnjs (ro)
Conditions:
  Type           Status
  PodScheduled   False
Volumes:
  nfs-storage:
    Type:       PersistentVolumeClaim (a reference to a PersistentVolumeClaim in the same namespace)
    ClaimName:  spark-pvc
    ReadOnly:   false
  kube-api-access-kbnjs:
    Type:                     Projected (a volume that contains injected data from multiple sources)
    TokenExpirationSeconds:   3607
    ConfigMapName:            kube-root-ca.crt
    ConfigMapOptional:        <nil>
    DownwardAPI:              true
QoS Class:                    BestEffort
Node-Selectors:               <none>
Tolerations:                  node.kubernetes.io/not-ready:NoExecute op=Exists for 300s
                              node.kubernetes.io/unreachable:NoExecute op=Exists for 300s
Events:
  Type     Reason             Age                  From                Message
  ----     ------             ----                 ----                -------
```

11

12

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get storageclass

NAME                     PROVISIONER                                RECLAIMPOLICY   VOLUMEBINDINGMODE      ALLOWVOLUMEEXPANSION   AGE
nfs                      cluster.local/nfs-nfs-server-provisioner   Delete          Immediate              true                   9m18s
premium-rwo              pd.csi.storage.gke.io                      Delete          WaitForFirstConsumer   true                   55m
standard                 kubernetes.io/gce-pd                       Delete          Immediate              true                   55m
standard-rwo (default)   pd.csi.storage.gke.io                      Delete          WaitForFirstConsumer   true                   55m
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pvc spark-pvc
kubectl get pods

Error from server (NotFound): persistentvolumeclaims "spark-pvc" not found
NAME                          READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0   1/1     Running   0          10m
spark-pod                     0/1     Pending   0          4m15s
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pvc spark-pvc
kubectl get pods

Error from server (NotFound): persistentvolumeclaims "spark-pvc" not found
NAME                          READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0   1/1     Running   0          11m
spark-pod                     0/1     Pending   0          4m22s
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl apply -f spark-pvc.yaml

persistentvolumeclaim/spark-data-pvc created
pod/spark-data-pod created
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-pvc.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pods

NAME                          READY   STATUS    RESTARTS   AGE
nfs-nfs-server-provisioner-0   1/1     Running   0          19m
spark-data-pod                1/1     Running   0          100s
spark-pod                     0/1     Pending   0          13m
```

13

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % docker run -v /tmp:/tmp -it bitnami/spark -- find /opt/bitnami/spark/examples/jars/ -name "spark-examples*" -exec cp {} /tmp/my.
jar \;

Unable to find image 'bitnami/spark:latest' locally
latest: Pulling from bitnami/spark
fca6463c7c5e: Pull complete
Digest: sha256:f662a50fae302f5b2c1e93d528457ce4f1de4098a8d33eb3c8bf6991dada8f68
Status: Downloaded newer image for bitnami/spark:latest
spark 21:34:41.72 INFO  ==>
spark 21:34:41.73 INFO  ==> Welcome to the Bitnami spark container
spark 21:34:41.73 INFO  ==> Subscribe to project updates by watching https://github.com/bitnami/containers
spark 21:34:41.73 INFO  ==> Submit issues and feature requests at https://github.com/bitnami/containers/issues
spark 21:34:41.73 INFO  ==> Upgrade to Tanzu Application Catalog for production environments to access custom-configured and pre-packaged software components. Gain enhanced fe
atures, including Software Bill of Materials (SBOM), CVE scan result reports, and VEX documents. To learn more, visit https://bitnami.com/enterprise
spark 21:34:41.73 INFO  ==>

kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl cp /tmp/my.jar spark-data-pod:/data/my.jar
kubectl cp /path/to/test.txt spark-data-pod:/data/test.txt

error: /path/to/test.txt doesn't exist in local filesystem
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl cp /tmp/my.jar spark-data-pod:/data/my.jar
kubectl cp /tmp/test.txt spark-data-pod:/data/test.txt

error: /tmp/test.txt doesn't exist in local filesystem
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % echo "how much wood could a woodpecker chuck if a woodpecker could chuck wood" > /tmp/test.txt
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl cp /tmp/my.jar spark-data-pod:/data/my.jar
kubectl cp /tmp/test.txt spark-data-pod:/data/test.txt

kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-data-pod --ls -al /data
Error: unknown flag: --ls
See 'kubectl exec --help' for usage.
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-data-pod -- ls -al /data
total 1540
drwxrwsrwx 2 root root    4096 Jul 13 21:41 .
drwxr-xr-x 1 root root    4096 Jul 13 20:37 ..
-rw-r--r-- 1  501 root 1564260 Jul 13 21:41 my.jar
-rw-r--r-- 1  501 root      72 Jul 13 21:41 test.txt
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-chart.yaml
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % vim spark-chart.yaml
```

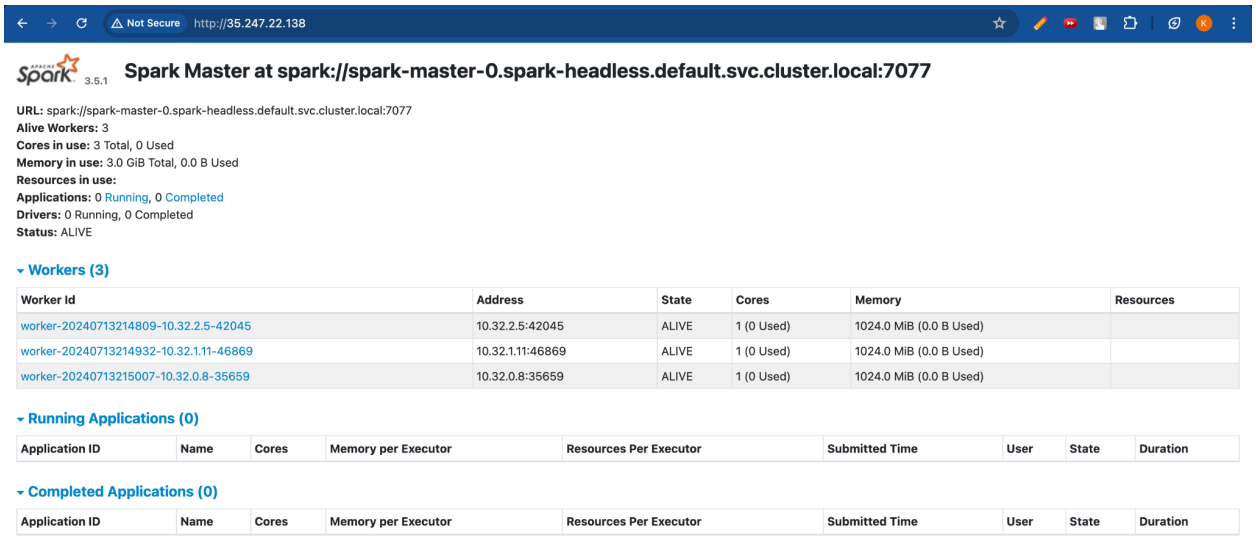
14

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % helm install spark bitnami/spark -f spark-chart.yaml

NAME: spark
LAST DEPLOYED: Sat Jul 13 14:46:59 2024
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
NOTES:
CHART NAME: spark
CHART VERSION: 9.2.5
APP VERSION: 3.5.1

** Please be patient while the chart is being deployed **

1. Get the Spark master WebUI URL by running these commands:

    NOTE: It may take a few minutes for the LoadBalancer IP to be available.
    You can watch the status of by running 'kubectl get --namespace default svc -w spark-master-svc'

  export SERVICE_IP=$(kubectl get --namespace default svc spark-master-svc -o jsonpath="{.status.loadBalancer.ingress[0]['ip', 'hostname'] }")
  echo http://$SERVICE_IP:80

2. Submit an application to the cluster:

  To submit an application to the cluster the spark-submit script must be used. That script can be
  obtained at https://github.com/apache/spark/tree/master/bin. Also you can use kubectl run.

  Run the commands below to obtain the master IP and submit your application.

  export EXAMPLE_JAR=$(kubectl exec -ti --namespace default spark-worker-0 -- find examples/jars/ -name 'spark-example*\.jar' | tr -d '\r')
  export SUBMIT_IP=$(kubectl get --namespace default svc spark-master-svc -o jsonpath="{.status.loadBalancer.ingress[0]['ip', 'hostname'] }")

  kubectl run --namespace default spark-client --rm --tty -i --restart='Never' \
    --image docker.io/bitnami/spark:3.5.1-debian-12-r8 \
    -- spark-submit --master spark://$SUBMIT_IP:7077 \
    --deploy-mode cluster \
    --class org.apache.spark.examples.SparkPi \
    $EXAMPLE_JAR 1000

** IMPORTANT: When submit an application the --master parameter should be set to the service IP, if not, the application will not resolve the master. **
```

15

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get svc -l "app.kubernetes.io/instance=spark,app.kubernetes.io/name=spark"
NAME               TYPE           CLUSTER-IP      EXTERNAL-IP     PORT(S)                       AGE
spark-headless     ClusterIP      None            <none>          <none>                        3m46s
spark-master-svc   LoadBalancer   34.118.230.89   35.247.22.138   7077:31878/TCP,80:32697/TCP   3m46s
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata %
```

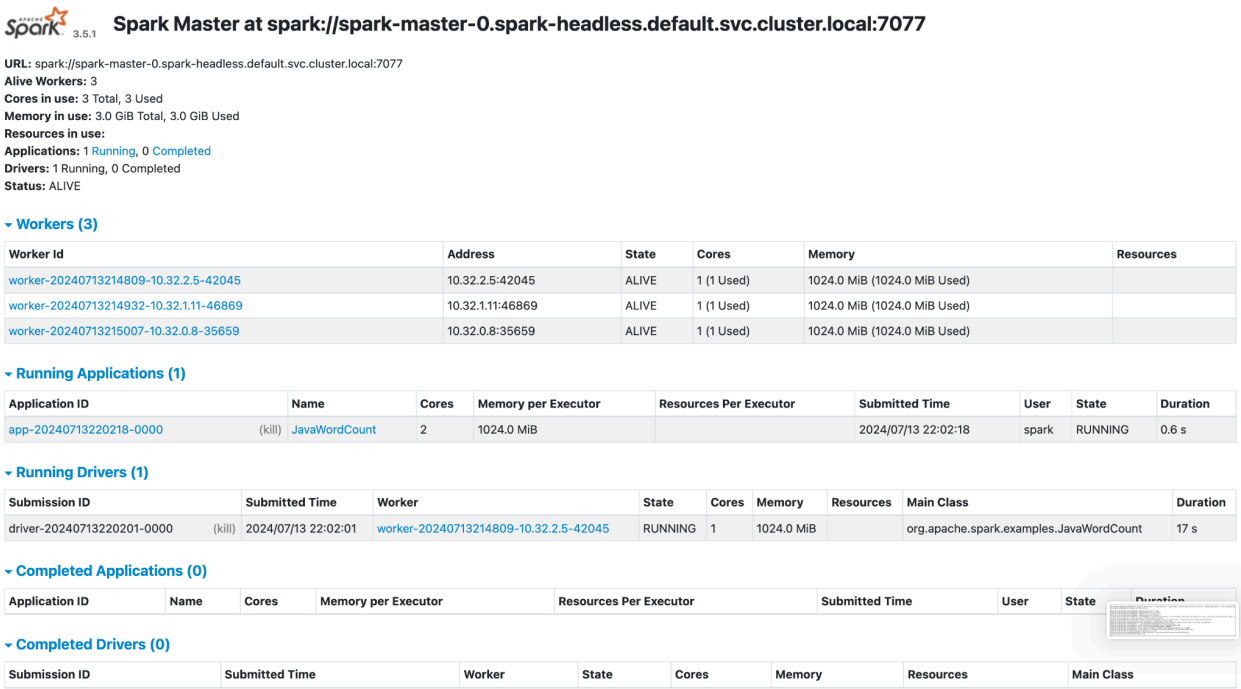
16



17

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-master-0 -- spark-submit --master spark://35.247.22.138:7077 --deploy-mode cluster --class org.apache.spark.examples.JavaWordCount /data/my.jar /data/test.txt

24/07/13 22:01:58 INFO SecurityManager: Changing view acls to: spark
24/07/13 22:01:58 INFO SecurityManager: Changing modify acls to: spark
24/07/13 22:01:58 INFO SecurityManager: Changing view acls groups to:
24/07/13 22:01:58 INFO SecurityManager: Changing modify acls groups to:
24/07/13 22:01:58 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: spark; groups with view permissions: EMPTY; users with modify permissions: spark; groups with modify permissions: EMPTY
24/07/13 22:01:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/07/13 22:02:00 INFO Utils: Successfully started service 'driverClient' on port 46205.
24/07/13 22:02:01 INFO TransportClientFactory: Successfully created connection to /35.247.22.138:7077 after 291 ms (0 ms spent in bootstraps)
24/07/13 22:02:01 INFO ClientEndpoint: ... waiting before polling master for driver state
24/07/13 22:02:02 INFO ClientEndpoint: Driver successfully submitted as driver-20240713220201-0000
24/07/13 22:02:06 INFO ClientEndpoint: State of driver-20240713220201-0000 is RUNNING
24/07/13 22:02:06 INFO ClientEndpoint: Driver running on 10.32.2.5:42045 (worker-20240713214809-10.32.2.5-42045)
24/07/13 22:02:06 INFO ClientEndpoint: spark-submit not configured to wait for completion, exiting spark-submit JVM.
24/07/13 22:02:06 INFO ShutdownHookManager: Shutdown hook called
24/07/13 22:02:06 INFO ShutdownHookManager: Deleting directory /tmp/spark-cb833191-ccd0-4672-a266-6123d916e4bc
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata %
```

18



Spark 3.5.1

Spark Master at spark://spark-master-0.spark-headless.default.svc.cluster.local:7077

URL: spark://spark-master-0.spark-headless.default.svc.cluster.local:7077
Alive Workers: 3
Cores in use: 3 Total, 3 Used
Memory in use: 3.0 GiB Total, 3.0 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 1 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240713214809-10.32.2.5-42045	10.32.2.5:42045	ALIVE	1 (1 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240713214932-10.32.1.11-46869	10.32.1.11:46869	ALIVE	1 (1 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240713215007-10.32.0.8-35659	10.32.0.8:35659	ALIVE	1 (1 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240713220218-0000 (kill)	JavaWordCount	2	1024.0 MiB		2024/07/13 22:02:18	spark	RUNNING	0.6 s

Running Drivers (1)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class	Duration
driver-20240713220201-0000 (kill)	2024/07/13 22:02:01	worker-20240713214809-10.32.2.5-42045	RUNNING	1	1024.0 MiB		org.apache.spark.examples.JavaWordCount	17 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Drivers (0)

Submission ID	Submitted Time	Worker	State	Cores	Memory	Resources	Main Class

19

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl get pods -o wide
NAME                          READY   STATUS    RESTARTS   AGE    IP           NODE                                      NOMINATED NODE   READINESS GATES
nfs-nfs-server-provisioner-0  1/1     Running   0          105m   10.32.0.5    gke-spark-default-pool-dfc02541-f7jg      <none>           <none>
spark-data-pod                1/1     Running   0          87m    10.32.0.6    gke-spark-default-pool-dfc02541-f7jg      <none>           <none>
spark-master-0                1/1     Running   0          17m    10.32.0.7    gke-spark-default-pool-dfc02541-f7jg      <none>           <none>
spark-pod                     0/1     Pending   0          98m    <none>       <none>                                    <none>           <none>
spark-worker-0                1/1     Running   0          17m    10.32.2.5    gke-spark-default-pool-fd8b4837-xmdg      <none>           <none>
spark-worker-1                1/1     Running   0          15m    10.32.1.11   gke-spark-default-pool-5f42b2f5-lw29      <none>           <none>
spark-worker-2                1/1     Running   0          14m    10.32.0.8    gke-spark-default-pool-dfc02541-f7jg      <none>           <none>
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata %
```

20

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-worker-0 -- bash

I have no name!@spark-worker-0:/opt/bitnami/spark$
```

21

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-worker-0 -- bash

I have no name!@spark-worker-0:/opt/bitnami/spark$ cd /opt/bitnami/spark/work
I have no name!@spark-worker-0:/opt/bitnami/spark/work$ ls -l
total 4
drwxr-sr-x 2 1001 1001 4096 Jul 13 22:02 driver-20240713220201-0000
I have no name!@spark-worker-0:/opt/bitnami/spark/work$
```

22

```
total 4
drwxr-sr-x 2 1001 1001 4096 Jul 13 22:02 driver-20240713220201-0000
I have no name!@spark-worker-0:/opt/bitnami/spark/work$ cd driver-20240713220201-0000
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$ cat stdout
if: 1
a: 2
how: 1
could: 2
wood: 2
woodpecker: 2
much: 1
chuck: 2
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$
```

23

```
kadirbeksharu@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-worker-0 -- bash

I have no name!@spark-worker-0:/opt/bitnami/spark$ cd /opt/bitnami/spark/work
I have no name!@spark-worker-0:/opt/bitnami/spark/work$ ls -l
total 4
drwxr-sr-x 2 1001 1001 4096 Jul 13 22:02 driver-20240713220201-0000
I have no name!@spark-worker-0:/opt/bitnami/spark/work$ cd driver-20240713220201-0000
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$ cat stdout
if: 1
a: 2
how: 1
could: 2
wood: 2
woodpecker: 2
much: 1
chuck: 2
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$ ^C
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$ ^C
I have no name!@spark-worker-0:/opt/bitnami/spark/work/driver-20240713220201-0000$ cd ..
I have no name!@spark-worker-0:/opt/bitnami/spark/work$ cd ..
I have no name!@spark-worker-0:/opt/bitnami/spark$ pyspark
Error: pyspark does not support any application options.

Usage: ./bin/pyspark [options]

Options:
  --master MASTER_URL         spark://host:port, mesos://host:port, yarn,
                              k8s://https://host:port, or local (Default: local[*]).
  --deploy-mode DEPLOY_MODE   Whether to launch the driver program locally ("client") or
                              on one of the worker machines inside the cluster ("cluster")
                              (Default: client).
  --class CLASS_NAME          Your application's main class (for Java / Scala apps).
  --name NAME                 A name of your application.
  --jars JARS                 Comma-separated list of jars to include on the driver
                              and executor classpaths.
  --packages                  Comma-separated list of maven coordinates of jars to include
                              on the driver and executor classpaths. Will search the local
                              maven repo, then maven central and any additional remote
                              repositories given by --repositories. The format for the
                              coordinates should be groupId:artifactId:version.
  --exclude-packages          Comma-separated list of groupId:artifactId, to exclude while
                              resolving the dependencies provided in --packages to avoid
```

24

```
I have no name!@spark-worker-0:/opt/bitnami/spark$ export PYTHONPATH=/opt/bitnami/spark/python/lib/py4j-0.10.9.7-src.zip:/opt/bitnami/spark/python/:/opt/bitnami/spark/python/:
$PYTHONPATH
I have no name!@spark-worker-0:/opt/bitnami/spark$ export PYTHONSTARTUP=/opt/bitnami/spark/python/pyspark/shell.py
I have no name!@spark-worker-0:/opt/bitnami/spark$ exec "$SPARK_HOME"/bin/spark-submit pyspark-shell-main
Python 3.11.9 (main, Jun 22 2024, 04:32:05) [GCC 12.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/07/13 22:15:50 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/

Using Python version 3.11.9 (main, Jun 22 2024 04:32:05)
Spark context Web UI available at http://spark-worker-0.spark-headless.default.svc.cluster.local:4040
Spark context available as 'sc' (master = local[*], app id = local-1720908953482).
SparkSession available as 'spark'.
>>>
```

25

```
kadirbeksharau@Kadirbeks-MacBook-Pro bigdata % kubectl exec -it spark-worker-0 -- bash

I have no name!@spark-worker-0:/opt/bitnami/spark$ cd spark
bash: cd: spark: No such file or directory
I have no name!@spark-worker-0:/opt/bitnami/spark$ ls
LICENSE  NOTICE  R  README.md  RELEASE  bin  conf  conf.default  data  examples  jars  kubernetes  licenses  logs  python  sbin  tmp  venv  work  yarn
I have no name!@spark-worker-0:/opt/bitnami/spark$ cd opt/bitnami
bash: cd: opt/bitnami: No such file or directory
I have no name!@spark-worker-0:/opt/bitnami/spark$ cd /opt/bitnami/spark/examples/src/main/python
I have no name!@spark-worker-0:/opt/bitnami/spark/examples/src/main/python$ ls
__init__.py  avro_inputformat.py  logistic_regression.py  mllib       parquet_inputformat.py  sort.py  status_api_demo.py  transitive_closure.py
als.py       kmeans.py            ml                      pagerank.py  pi.py                   sql      streaming           wordcount.py
I have no name!@spark-worker-0:/opt/bitnami/spark/examples/src/main/python$ spark-submit pagerank.py /opt 2
WARN: This is a naive implementation of PageRank and is given as an example!
Please refer to PageRank implementation provided by graphx
24/07/13 22:19:18 INFO SparkContext: Running Spark version 3.5.1
24/07/13 22:19:18 INFO SparkContext: OS info Linux, 6.1.85+, amd64
24/07/13 22:19:18 INFO SparkContext: Java version 17.0.11
24/07/13 22:19:18 INFO ResourceUtils: ==============================================================
24/07/13 22:19:18 INFO ResourceUtils: No custom resources configured for spark.driver.
24/07/13 22:19:18 INFO ResourceUtils: ==============================================================
24/07/13 22:19:18 INFO SparkContext: Submitted application: PythonPageRank
24/07/13 22:19:18 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, a
mount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/07/13 22:19:18 INFO ResourceProfile: Limiting resource is cpu
24/07/13 22:19:18 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/07/13 22:19:18 INFO SecurityManager: Changing view acls to: spark
24/07/13 22:19:18 INFO SecurityManager: Changing modify acls to: spark
24/07/13 22:19:18 INFO SecurityManager: Changing view acls groups to:
24/07/13 22:19:18 INFO SecurityManager: Changing modify acls groups to:
24/07/13 22:19:18 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: spark; groups with view permissions: EMPTY; us
ers with modify permissions: spark; groups with modify permissions: EMPTY
24/07/13 22:19:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/07/13 22:19:20 INFO Utils: Successfully started service 'sparkDriver' on port 34813.
24/07/13 22:19:20 INFO SparkEnv: Registering MapOutputTracker
24/07/13 22:19:20 INFO SparkEnv: Registering BlockManagerMaster
24/07/13 22:19:20 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/07/13 22:19:20 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/07/13 22:19:20 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/07/13 22:19:20 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-1c8cd8a7-68e2-41b9-bd9c-1d5be58a2b81
24/07/13 22:19:20 INFO MemoryStore: MemoryStore started with capacity 413.9 MiB
24/07/13 22:19:20 INFO SparkEnv: Registering OutputCommitCoordinator
24/07/13 22:19:21 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/07/13 22:19:21 INFO Utils: Successfully started service 'SparkUI' on port 4040.
```

26

```
If provided paths are partition directories, please set "basePath" in the options of the data source to specify the root directory of the table. If there are multiple root dir
ectories, please load them separately and then union them.
        at scala.Predef$.assert(Predef.scala:223)
        at org.apache.spark.sql.execution.datasources.PartitioningUtils$.parsePartitions(PartitioningUtils.scala:178)
        at org.apache.spark.sql.execution.datasources.PartitioningUtils$.parsePartitions(PartitioningUtils.scala:110)
        at org.apache.spark.sql.execution.datasources.PartitioningAwareFileIndex.inferPartitioning(PartitioningAwareFileIndex.scala:201)
        at org.apache.spark.sql.execution.datasources.InMemoryFileIndex.partitionSpec(InMemoryFileIndex.scala:75)
        at org.apache.spark.sql.execution.datasources.PartitioningAwareFileIndex.partitionSchema(PartitioningAwareFileIndex.scala:51)
        at org.apache.spark.sql.execution.datasources.DataSource.getOrInferFileFormatSchema(DataSource.scala:167)
        at org.apache.spark.sql.execution.datasources.DataSource.resolveRelation(DataSource.scala:407)
        at org.apache.spark.sql.DataFrameReader.loadV1Source(DataFrameReader.scala:229)
        at org.apache.spark.sql.DataFrameReader.$anonfun$load$2(DataFrameReader.scala:211)
        at scala.Option.getOrElse(Option.scala:189)
        at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:211)
        at org.apache.spark.sql.DataFrameReader.text(DataFrameReader.scala:646)
        at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:77)
        at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.base/java.lang.reflect.Method.invoke(Method.java:568)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:374)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.ClientServerConnection.waitForCommands(ClientServerConnection.java:182)
        at py4j.ClientServerConnection.run(ClientServerConnection.java:106)
        at java.base/java.lang.Thread.run(Thread.java:840)

24/07/13 22:19:53 INFO SparkContext: Invoking stop() from shutdown hook
24/07/13 22:19:53 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/07/13 22:19:53 INFO SparkUI: Stopped Spark web UI at http://spark-worker-0.spark-headless.default.svc.cluster.local:4040
24/07/13 22:19:53 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/07/13 22:19:53 INFO MemoryStore: MemoryStore cleared
24/07/13 22:19:53 INFO BlockManager: BlockManager stopped
24/07/13 22:19:53 INFO BlockManagerMaster: BlockManagerMaster stopped
24/07/13 22:19:53 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/07/13 22:19:53 INFO SparkContext: Successfully stopped SparkContext
24/07/13 22:19:53 INFO ShutdownHookManager: Shutdown hook called
24/07/13 22:19:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-98964e84-9201-4655-8ca7-320b375a3216
24/07/13 22:19:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-eb6ea915-37d7-4704-9e0c-b31c582e33a0
24/07/13 22:19:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-eb6ea915-37d7-4704-9e0c-b31c582e33a0/pyspark-03a515c1-78a3-403f-8f79-6df932674f2d
I have no name!@spark-worker-0:/opt/bitnami/spark/examples/src/main/python$
```