

Run PySpark Airbnb on GCP

Google Cloud Platform

Steps to run Airbnb on GCP

- Login to GCP
- Create project (if not already created)
- Create a Kubernetes cluster on GCP
- Connect a local shell to your GCP project
- Submit your Python app to the GCP cluster
- Review your results

Assumptions

- gcloud has been installed
 - The command line interface for GCP
 - <https://cloud.google.com/sdk/docs/install>
- Review the Quickstart for GKE
 - <https://cloud.google.com/sdk/docs/install>

Login to GCP

The screenshot shows the Google Cloud Platform console interface. At the top, a blue header bar contains the Google Cloud logo, the project name 'ghTorrentPlus' with a dropdown arrow, and a search bar with the placeholder text 'Search products and resources'. Below the header, a navigation bar includes 'DASHBOARD' (highlighted), 'ACTIVITY', and 'RECOMMENDATIONS'. The main content area is divided into several sections. A 'Quick Access' section at the top contains three cards: the first with 'Google Cloud' and 'Getting Started' links, the second with 'Google Cloud' and 'Manage Resources' links, and the third with 'Google Cloud' and 'IAM Permissions' links. Below this, the 'Project info' section displays details for the 'ghTorrentPlus' project: Project name (ghTorrentPlus), Project ID (ghtorrentplus), and Project number (199861100084). It also includes a blue link 'ADD PEOPLE TO THIS PROJECT' and a button 'Go to project settings'. To the right, the 'APIs' section shows a table with the header 'Requests (requests/sec)' and a message 'No data is available for the selected'.

Google Cloud Platform ghTorrentPlus

Search products and resources

DASHBOARD ACTIVITY RECOMMENDATIONS

Quick Access

- Google Cloud
- Getting Started
- Google Cloud
- Manage Resources
- Google Cloud
- IAM Permissions

Project info

Project name
ghTorrentPlus

Project ID
ghtorrentplus

Project number
199861100084

[ADD PEOPLE TO THIS PROJECT](#)

→ Go to project settings

APIs

Requests (requests/sec)

⚠ No data is available for the selected

Create project (if not already created)

Select from project drop down menu

Select New Project

Select a project

Search projects and folders

NEW PROJECT


RECENT STARRED PREVIEW ALL

	Name	ID
✓ ☆	ghTorrentPlus ?	ghtorrentplus
☆	BQ Project ?	bq-project-1
☆	gha-project ?	gha-project-1
☆	my-spark-demo ?	my-spark-demo-273420

Create project (if not already created)

Google Cloud Platform

New Project

 You have 17 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)


[MANAGE QUOTAS](#)

Project name *
spark8795

Project ID: spark8795. It cannot be changed later. [EDIT](#)

Billing account *
Big data infrastructure (CIS 8795)

Any charges for this project will be billed to the account you select here.

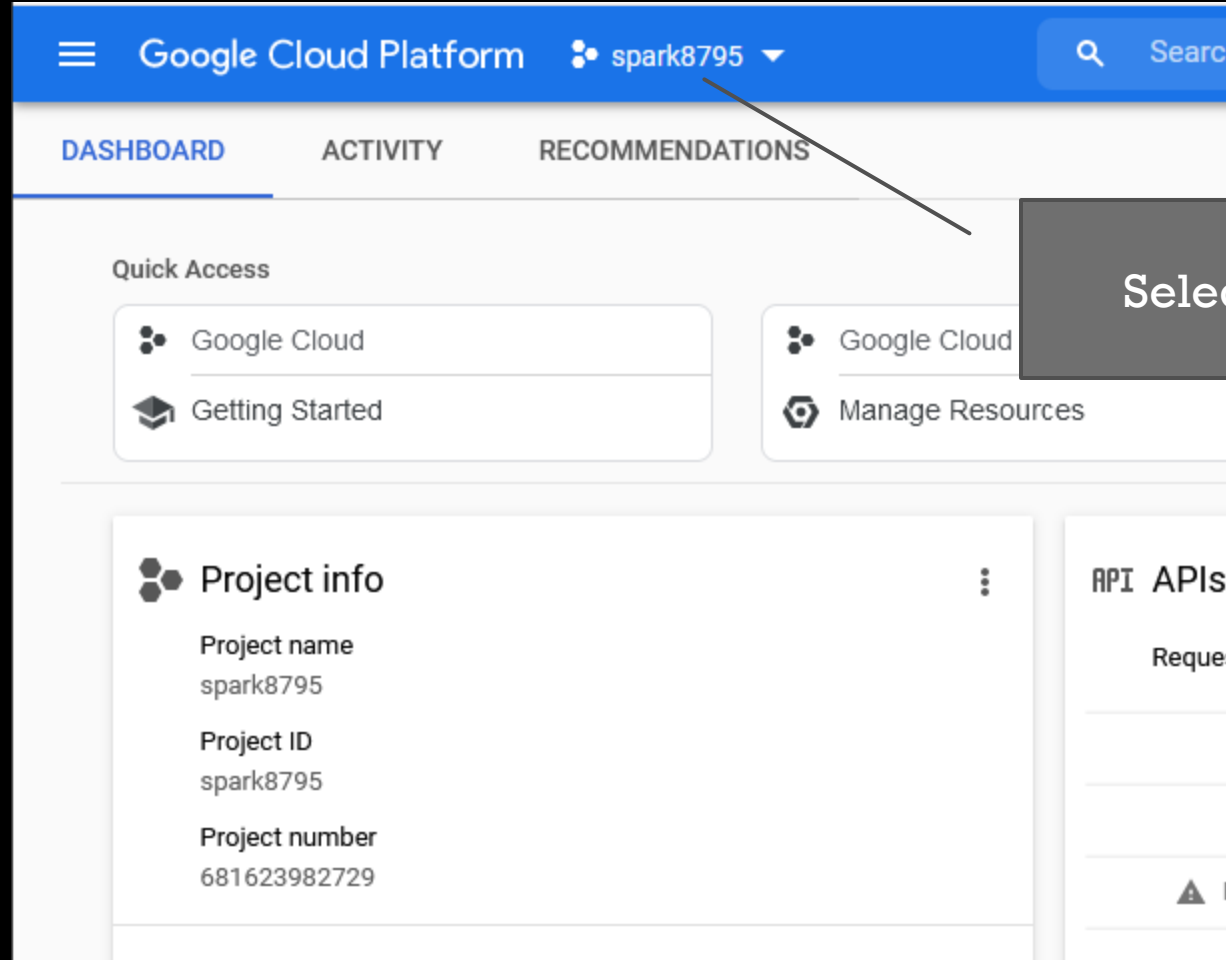
Location *
 No organization [BROWSE](#)

Parent organization or folder

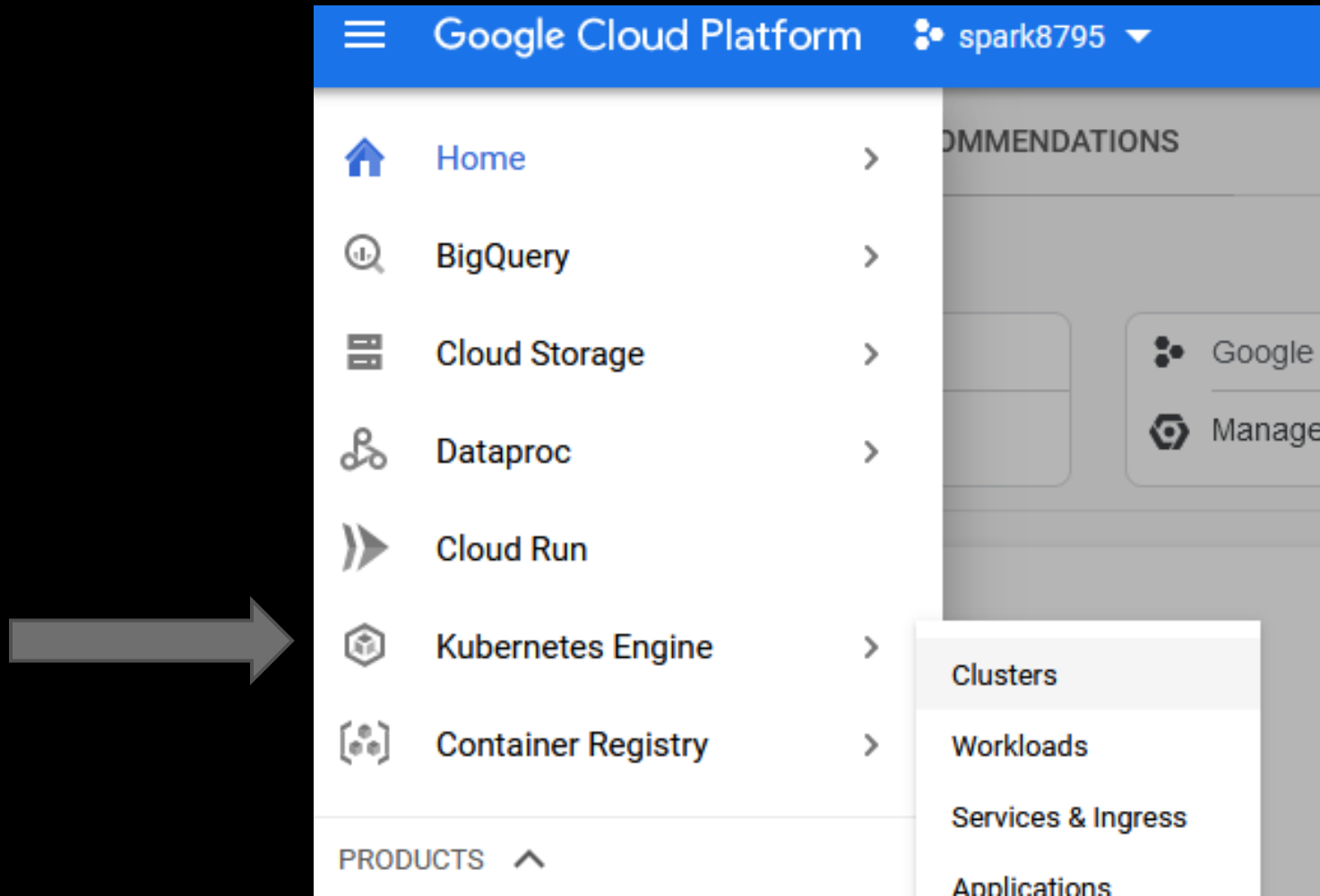
[CREATE](#) [CANCEL](#)

Give it a unique name

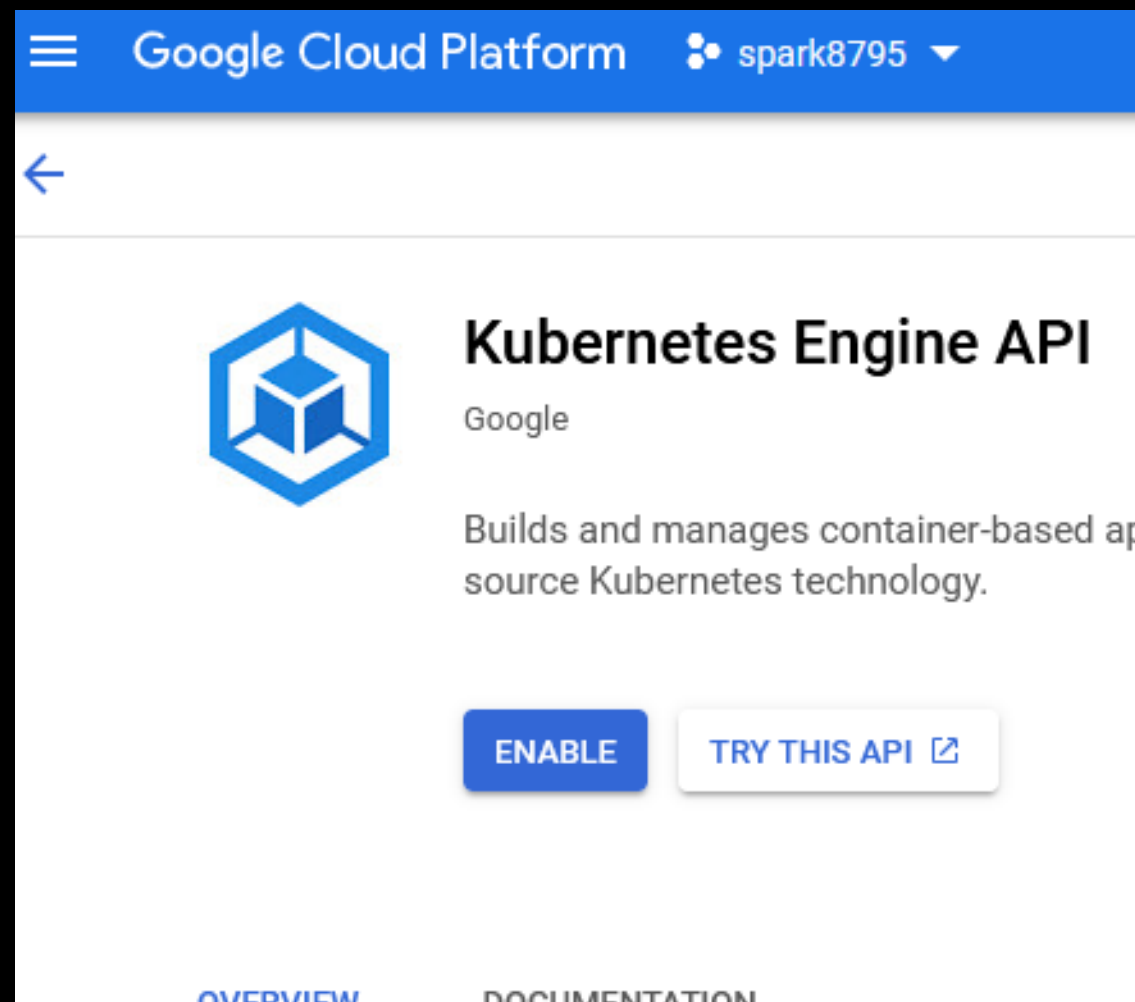
Select your new project



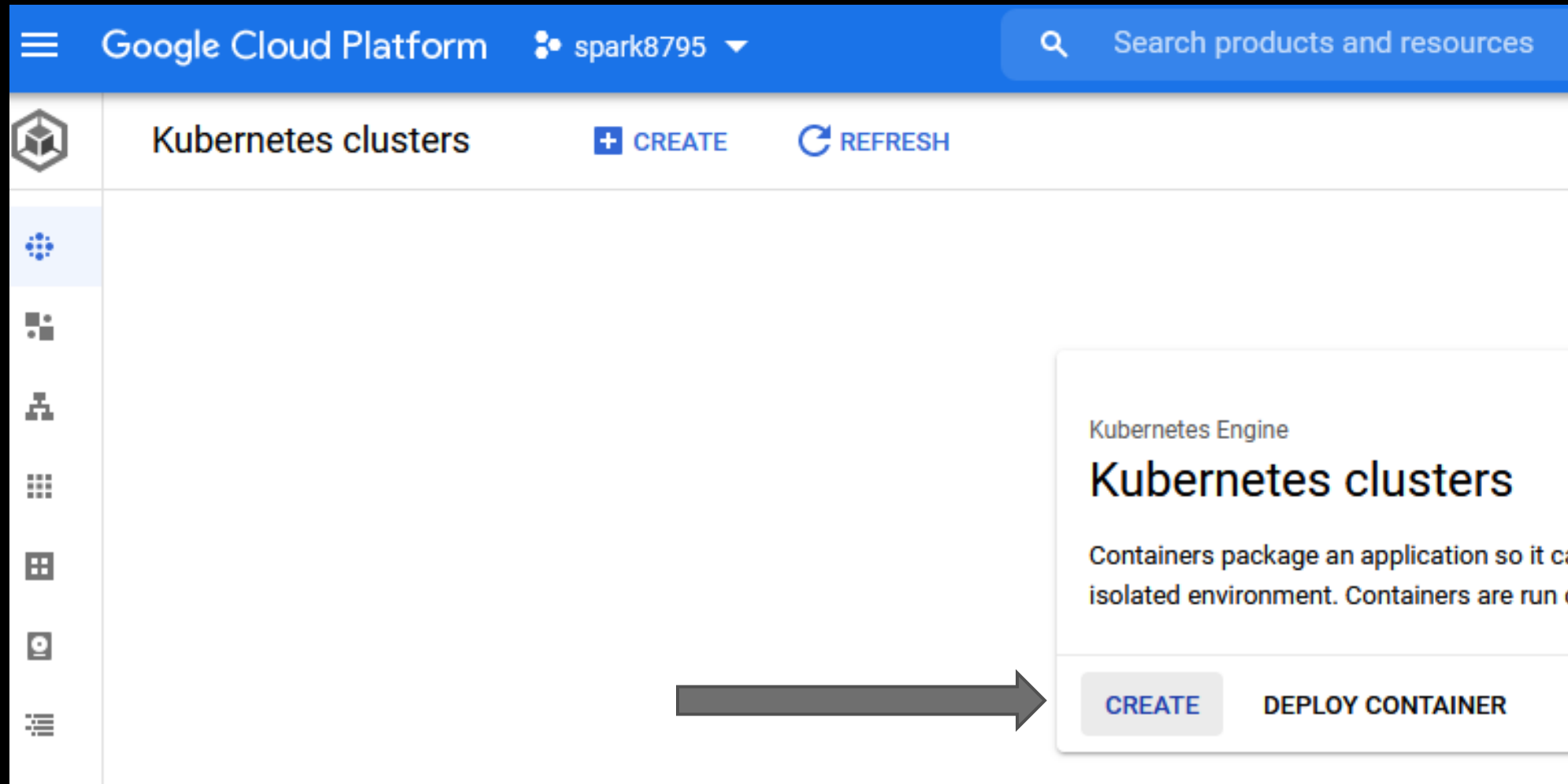
Create a Kubernetes cluster on GCP



Enable Kubernetes API (if asked)



Create cluster standard (not autopilot)



Basic cluster

- Name: my-first-cluster-1
 - Name is in k8submit_gcp.sh
- Zonal: us-east1-b
 - Others are fine, just remember your zone
- Select My first cluster
 - Customize
 - N1 standard-2 (2 vCPU, 7.5 GB memory)
- Make changes

Edit My first cluster

Cluster basics

The new cluster will be created with the name, version, and in the location you specify here. After the cluster is created, name and location can't be changed.



To experiment with an affordable cluster, try **My first cluster** in the Cluster set-up guides

Name

my-first-cluster-1



Location type

☒ Zonal

☐ Regional

Zone

us-east1-b



Cluster set-up guides



My first cluster

An affordable cluster to e

Make changes during creation

+ ADD NODE POOL

REMOVE NODE POOL

Cluster basics

The new cluster will be created with the name, version, and in the location you specify here. After the cluster is created, name and location can't be changed.

i

To experiment with an affordable cluster, try **My first cluster** in the **Cluster set-up guides**

Name
cluster-1

?

Location type
☒ Zonal
☐ Regional

Zone
us-central1-c

▼ ?

☐ Specify default node locations ?
Current default: us-central1-c

Control plane version
Choose a release channel for automatic management of your cluster's version and upgrade

Cluster settings

✓ Cluster name: my-first-cluster-2

✓ Cluster zone: us-central1-b

✓ Version: Rapid release channel instead of default version

✓ Boot disk size: 32GB instead of 100GB boot disk size

✓ Autoscaling: Disabled

✓ Cloud Operations for GKE: Disabled

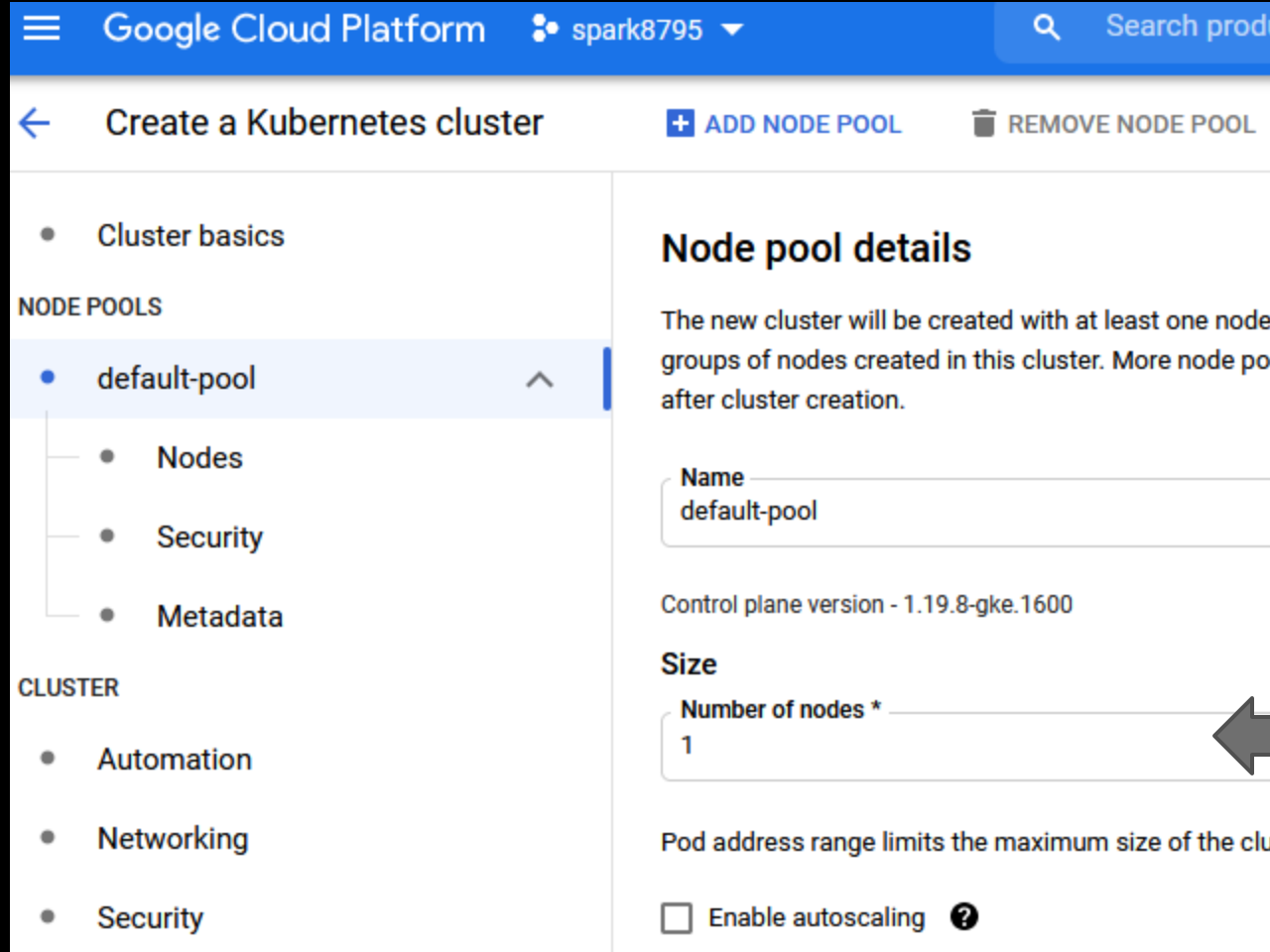
Different from recommendations

- Machine type: n1-standard-2 instead of e2-medium
Recommendation: g1-small to save money

MAKE CHANGES

CREATE CLUSTER NOW

Reduce the nodes to 1 (default-pool)



Google Cloud Platform spark8795

Create a Kubernetes cluster

+ ADD NODE POOL REMOVE NODE POOL

Cluster basics

NODE POOLS

- default-pool

Nodes

Security

Metadata

CLUSTER

- Automation
- Networking
- Security

Node pool details

The new cluster will be created with at least one node groups of nodes created in this cluster. More node pool after cluster creation.

Name
default-pool

Control plane version - 1.19.8-gke.1600

Size

Number of nodes *
1

Pod address range limits the maximum size of the cluster

☐ Enable autoscaling ?

Eventually, it's ready

Google Cloud Platform

spark8795

Search products and resources


Kubernetes clusters

+ CREATE

+ DEPLOY

REFRESH

DELETE



Introducing Autopilot mode

An optimized cluster with a hands-off experience. When you create a cluster in Autopilot mode, Google provisions and manages the entire

[Compare cluster modes](#)

- ✓ Get a production-ready cluster based on your workload requirements
- ✓ Eliminate the overhead of node management
- ✓ Pay per Pod, only for the resources that you use
- ✓ Increase security with Google best practices built-in
- ✓ Gain higher workload availability

TRY THE DEMO

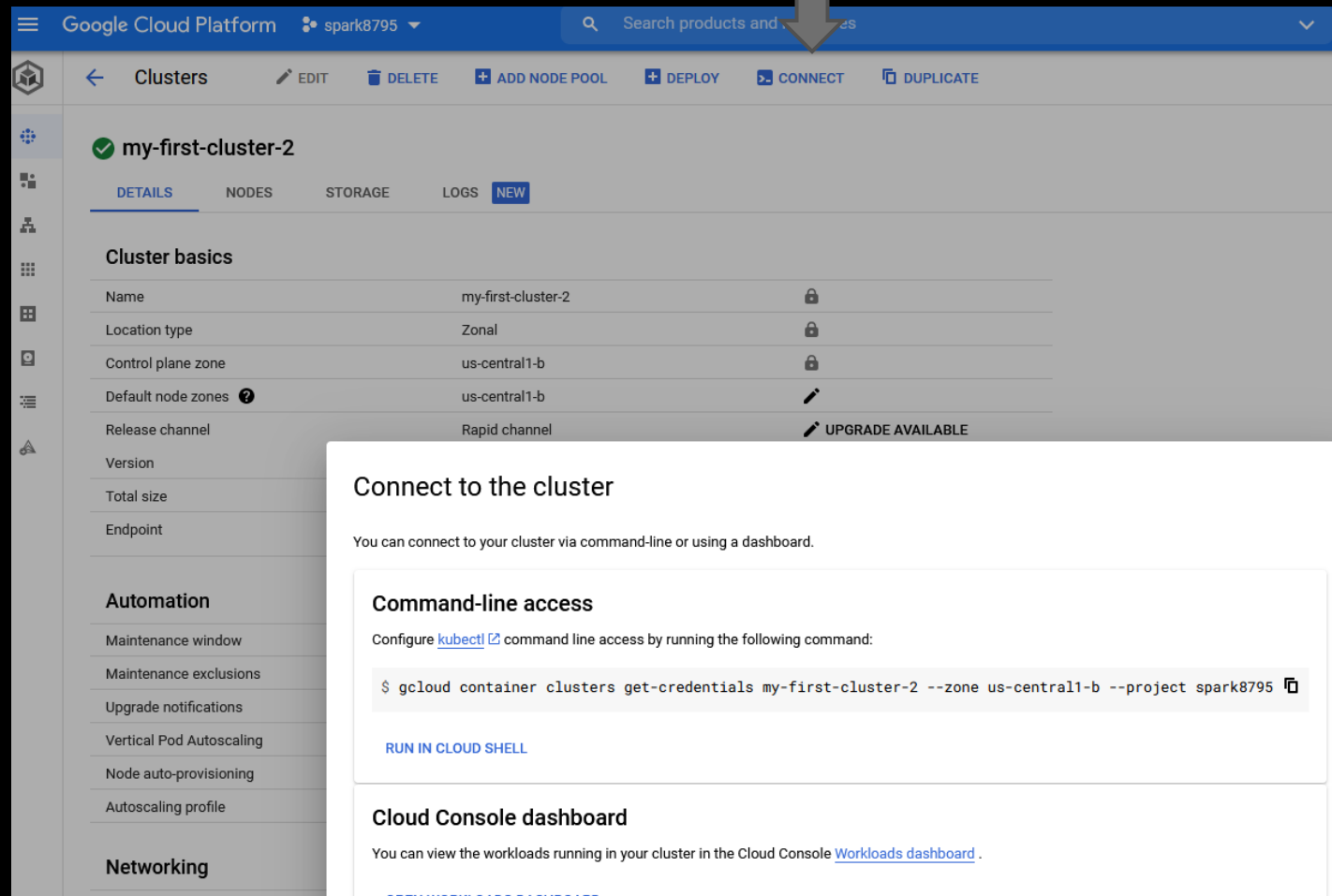
LEARN MORE

Filter

Enter property name or value

<input type="checkbox"/>	<input checked="" type="radio"/>	Name ↑	Location	Number of nodes	Total vCPUs	Total memory	Notifications	Labels
<input type="checkbox"/>	<input checked="" type="radio"/>	my-first-cluster-2	us-central1-b	1	2	7.5 GB		—

Connect to the cluster copy command to local shell



The screenshot shows the Google Cloud Platform interface for a Kubernetes cluster named 'my-first-cluster-2'. The 'CONNECT' button in the top navigation bar is highlighted, and a modal dialog titled 'Connect to the cluster' is open. The modal provides instructions on how to connect to the cluster via command-line or dashboard. A large grey arrow points from the 'CONNECT' button to the modal, and another large grey arrow points from the 'Command-line access' section to the terminal at the bottom of the image.

Cluster basics

Property	Value	Actions
Name	my-first-cluster-2	🔒
Location type	Zonal	🔒
Control plane zone	us-central1-b	🔒
Default node zones	us-central1-b	✎
Release channel	Rapid channel	✎ UPGRADE AVAILABLE
Version		
Total size		
Endpoint		

Automation

- Maintenance window
- Maintenance exclusions
- Upgrade notifications
- Vertical Pod Autoscaling
- Node auto-provisioning
- Autoscaling profile

Networking

Connect to the cluster

You can connect to your cluster via command-line or using a dashboard.

Command-line access

Configure [kubectl](#) command line access by running the following command:

```
$ gcloud container clusters get-credentials my-first-cluster-2 --zone us-central1-b --project spark8795
```

[RUN IN CLOUD SHELL](#)

Cloud Console dashboard

You can view the workloads running in your cluster in the Cloud Console [Workloads dashboard](#).

```
wnr@Lincoln:/mnt/c/Users/robin$ gcloud container clusters get-credentials my-first-cluster-2 --zone us-central1-b --project spark8795
Fetching cluster endpoint and auth data.
kubeconfig entry generated for my-first-cluster-2.
```

Connect a local shell to your GCP project

- First
 - `gcloud auth login`
 - Then, select the link to login

```
(base) wnr@Coolidge:robin$ gcloud auth login  
Go to the following link in your browser:
```

- After login, the browser will provide you with a key to copy
 - paste it into your shell, where it says Enter verification code

```
You are now logged in as [robinson.wn@gmail.com].  
Your current project is [None]. You can change this setting by running:  
$ gcloud config set project PROJECT_ID
```


Point kubectl to your GCP project

- `kubectl config get-contexts`
`gcloud config set project spark8795`
`gcloud config list`
`kubectl config use-context`

Point kubectl to your GCP project (from docker-desktop to GCP project)

```
wnr@Lincoln:/mnt/c/Users/robin$ kubectl config get-contexts
CURRENT  NAME                                     CLUSTER                                AUTHINFO
*        docker-desktop                         docker-desktop                        docker-desktop
gke_gha-demo_us-east1-b_gha-cluster        gke_gha-demo_us-east1-b_gha-cluster  gke_gha-demo_us-east1-b_gha-cluster
gke_gha-project-1_us-east1-b_gha-cluster    gke_gha-project-1_us-east1-b_gha-cluster
gke_gha-project-1_us-east1-c_gha-cluster    gke_gha-project-1_us-east1-c_gha-cluster
gke_gha-project-1_us-east1-d_gha-cluster    gke_gha-project-1_us-east1-d_gha-cluster
gke_gha-project-288318_us-east1-b_gha-cluster
gke_spark8795_us-central1-b_my-first-cluster-2  gke_spark8795_us-central1-b_my-first-cluster-2

wnr@Lincoln:/mnt/c/Users/robin$ kubectl config current-context
docker-desktop
wnr@Lincoln:/mnt/c/Users/robin$ kubectl config use-context gke_spark8795_us-central1-b_my-first-cluster-2
Switched to context "gke_spark8795_us-central1-b_my-first-cluster-2".
wnr@Lincoln:/mnt/c/Users/robin$ kubectl config current-context
gke_spark8795_us-central1-b_my-first-cluster-2
wnr@Lincoln:/mnt/c/Users/robin$ kubectl config get-contexts
CURRENT  NAME                                     CLUSTER                                AUTHINFO
*        docker-desktop                         docker-desktop                        docker-desktop
gke_gha-demo_us-east1-b_gha-cluster        gke_gha-demo_us-east1-b_gha-cluster  gke_gha-demo_us-east1-b_gha-cluster
gke_gha-project-1_us-east1-b_gha-cluster    gke_gha-project-1_us-east1-b_gha-cluster
gke_gha-project-1_us-east1-c_gha-cluster    gke_gha-project-1_us-east1-c_gha-cluster
gke_gha-project-1_us-east1-d_gha-cluster    gke_gha-project-1_us-east1-d_gha-cluster
gke_gha-project-288318_us-east1-b_gha-cluster
gke_spark8795_us-central1-b_my-first-cluster-2  gke_spark8795_us-central1-b_my-first-cluster-2
```

Create service account

- `kubectl create serviceaccount spark`
- `kubectl create clusterrolebinding spark-role --clusterrole=edit --serviceaccount=default:spark --namespace=default`

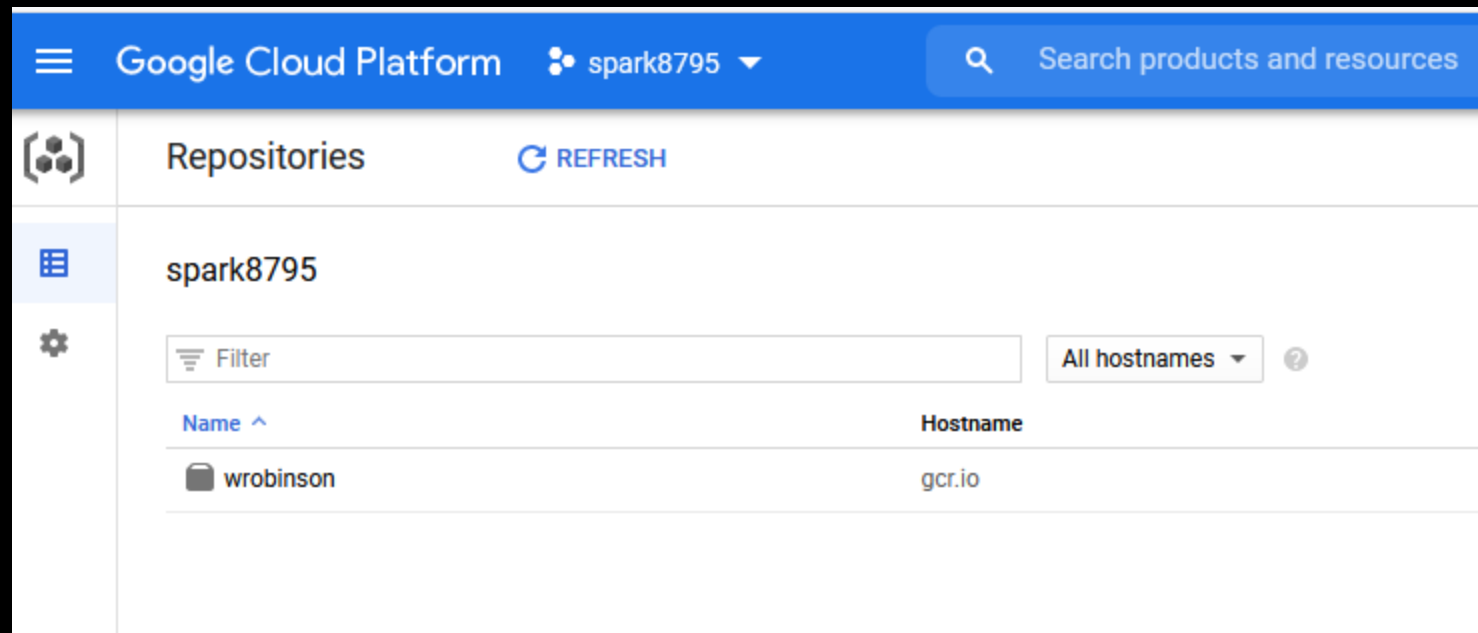
```
wnr@Coolidge:airbnb$ kubectl create serviceaccount spark
serviceaccount/spark created
wnr@Coolidge:airbnb$ kubectl create clusterrolebinding spark-role --clusterrole=edit --serviceaccount=default:spark --namespace=default
clusterrolebinding.rbac.authorization.k8s.io/spark-role created
```

Docker build and push image to cloud

- On WSL2
 - `sudo hwclock -s`
 - Synchronize time/clock
- `gcloud auth configure-docker`
- `docker build -f lnode.dockerfile -t wrobinson/airbnb:1.0 -t gcr.io/spark8795/wrobinson/airbnb:1.0 .`
- `docker push gcr.io/spark8795/wrobinson/airbnb:1.0`

Open container registry

- Search for **container registry** in GCP
 - Confirm container is there...



Run

- `./k8submit_gcp.sh`
- `kubectl logs -f -l app=airbnb --tail=-1`

Run

```
wnr@Lincoln:/mnt/c/Users/robin/OneDrive/Profile/Documents/dev/8795/airbnb$ ./k8submit_gcp.sh
/opt/spark/bin/spark-submit --master k8s://https://34.67.4.150:443 --deploy-mode cluster --driver-memory 1g --
amicAllocation.enabled=false --conf spark.executor.heartbeatInterval=20s --conf spark.shuffle.io.retryWait=60
=4 --conf spark.kubernetes.driver.limit.cores=3500m --conf spark.kubernetes.executor.request.cores=3500m --co
1.0 --conf spark.kubernetes.container.image.pullPolicy=Always --name airbnb --conf spark.kubernetes.authentic
abel.app=airbnb --conf spark.kubernetes.driver.label.app=airbnb --conf spark.kubernetes.driverEnv.PYSPARK_APP
/work-dir/airbnb.zip --conf spark.kubernetes.executorEnv.PYTHONPATH=/opt/spark/work-dir/airbnb.zip local:///c
py_files=/opt/spark/work-dir/airbnb.zip
```

```
21/04/07 11:40:32 WARN Utils: Your hostname, Lincoln resolves to a loopback address: 127.0.1.1; using 172.21.
21/04/07 11:40:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark-3.0.1-bin-hadoop2.7/j
fer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/04/07 11:40:35 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/04/07 11:40:35 INFO SparkKubernetesClientFactory: Auto-configuring K8S client using current context from u
21/04/07 11:40:36 INFO KerberosConfDriverFeatureStep: You have not specified a krb5.conf file locally or via
er image.
21/04/07 11:40:39 INFO LoggingPodStatusWatcherImpl: State changed, new state:
  pod name: airbnb-383ddb78acf12db-driver
  namespace: default
  labels: app -> airbnb, spark-app-selector -> spark-bd98ee96f5c849698d80d2663c98f92e, spark-role -> d
  pod uid: 628de121-fa1d-4db0-9dcc-51cd90cac129
  creation time: 2021-04-07T15:39:25Z
  service account name: spark
  volumes: spark-local-dir-1, spark-conf-volume, spark-token-6fjjv
  node name: N/A
```

```
wnr@Lincoln:/mnt/c/Users/robin$ kubectl logs -f -l app=airbnb
Error from server (BadRequest): container "spark-kubernetes-driver" in pod
wnr@Lincoln:/mnt/c/Users/robin$ kubectl logs -f -l app=airbnb
WARNING: Please consider reporting this to the maintainers of org.apache.sp
WARNING: Use --illegal-access=warn to enable warnings of further illegal re
WARNING: All illegal access operations will be denied in a future release
21/04/07 15:39:32 WARN NativeCodeLoader: Unable to load native-hadoop libra
run_env executing in /opt/spark/work-dir
  the python sys path is: ['/opt/spark/work-dir', '/opt/spark/python/lib/pys
pt/spark/work-dir/airbnb.zip', '/usr/lib/python37.zip', '/usr/lib/python3.7
ckages']
file/folder in this directory: sleep.py
file/folder in this directory: run_env.py
file/folder in this directory: airbnb.zip
Loading module specified in environment variable PYSPARK_APP_MODULE.
Cannot find logging.ini file
Imported module airbnb. Calling main
21/04/07 15:39:34 INFO airbnb: Start of airbnb
21/04/07 15:39:34 DEBUG airbnb: Reading data
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.proper
21/04/07 15:39:36 INFO SparkContext: Running Spark version 3.0.1
21/04/07 15:39:36 INFO ResourceUtils: =====
21/04/07 15:39:36 INFO ResourceUtils: Resources for spark.driver:

21/04/07 15:39:36 INFO ResourceUtils: =====
21/04/07 15:39:36 INFO SparkContext: Submitted application: airbnb
21/04/07 15:39:36 INFO SecurityManager: Changing view acls to: spark,wnr
21/04/07 15:39:36 INFO SecurityManager: Changing modify acls to: spark,wnr
21/04/07 15:39:36 INFO SecurityManager: Changing view acls groups to:
21/04/07 15:39:36 INFO SecurityManager: Changing modify acls groups to:
21/04/07 15:39:36 INFO SecurityManager: SecurityManager: authentication di
```

Useful Kubernetes commands

- Cheat sheet
 - <https://kubernetes.io/docs/reference/kubectl/cheatsheet/>
- Examples
 - `kubectl get all`
 - `kubectl get pods -l app=airbnb`
 - `kubectl logs -f -l app=airbnb --tail=-1`
 - `kubectl delete pods -l app=airbnb`
 - `kubectl delete pod/PODNAME`
 - `kubectl get pods --all-namespaces`
 - `kubectl exec <CONTAINER ID> -ti /bin/bash`
 - `kubectl apply -f FILENAME.yaml`

Debug

- If your deployment is Pending for a long time, you may have an issue with insufficient resources
 - To view pods:
 - `kubectl get pods`
 - To view a pods description:
 - `kubectl describe pod POD_NAME_FROM_GET_PODS_ABOVE`
 - If the above indicates insufficient resources in log, then delete old pods:
 - `kubectl delete pods -l app=airbnb`
 - Typically, delete pods after each run