Dataproc: Qwik Start - Console

GSP103

Google Cloud Self-Paced Labs

Overview

Cloud Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way. Operations that used to take hours or days take seconds or minutes instead. Create Cloud Dataproc clusters quickly and resize them at any time, so you don't have to worry about your data pipelines outgrowing your clusters.
This lab shows you how to use the Google Cloud Console to create a Google Cloud Dataproc cluster, run a simple Apache Spark job in the cluster, then modify the number of workers in the cluster.

Setup and Requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

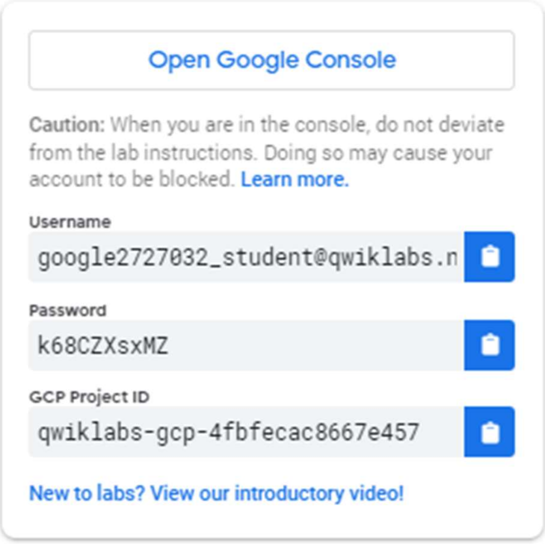
To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

Note: If you are using a Pixelbook, open an Incognito window to run this lab.

How to start your lab and sign in to the Google Cloud Console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is a panel populated with the temporary credentials that you must use for this lab.



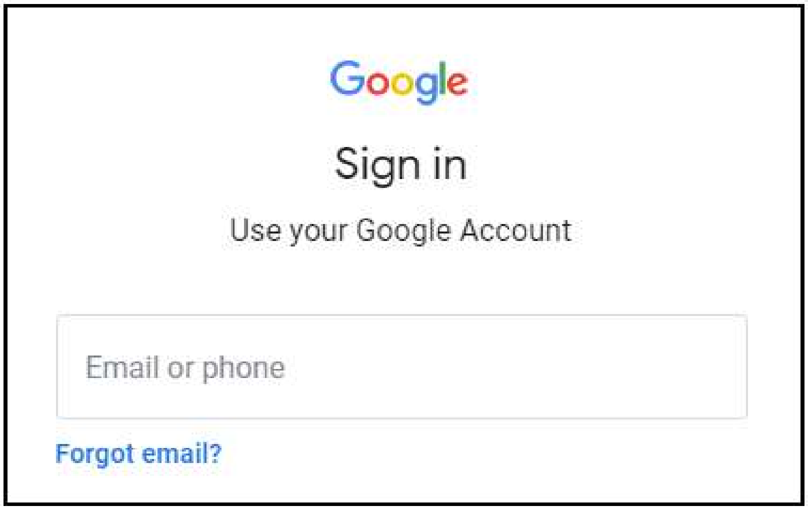
Caution: When you are in the console, do not deviate from the lab instructions. Doing so may cause your account to be blocked. Learn more.

Username: google2727032_student@qwiklabs.n

Password: k68CZXsxMZ

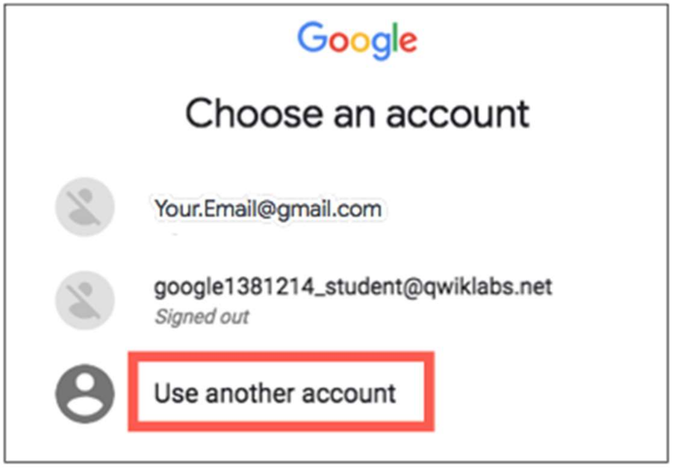
GCP Project ID: qwiklabs-gcp-4fbfecac8667e457

2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



Tip: Open the tabs in separate windows, side-by-side.

If you see the **Choose an account** page, click **Use Another Account**.



3. In the **Sign in** page, paste the username that you copied from the Connection Details panel. Then copy and paste the password.

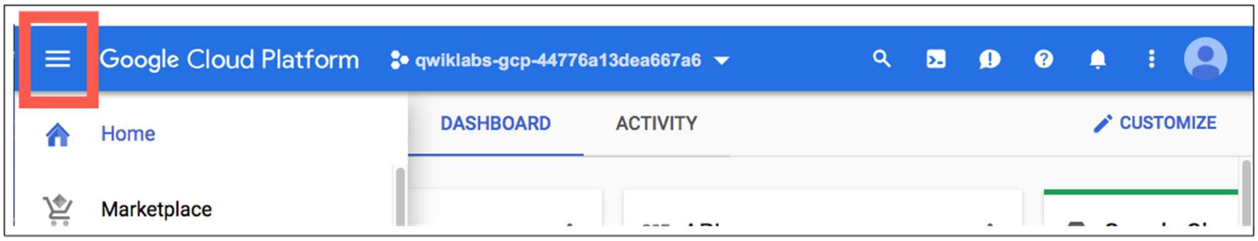
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

4. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Cloud Console opens in this tab.

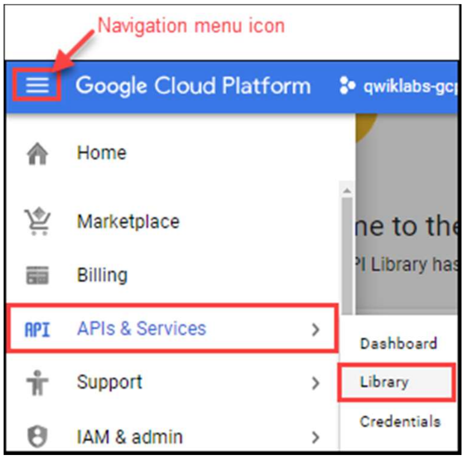
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-
left.



Confirm Cloud Dataproc API is enabled

To create a Dataproc cluster in Google Cloud, the Cloud Dataproc API must be enabled. To confirm the API is enabled:

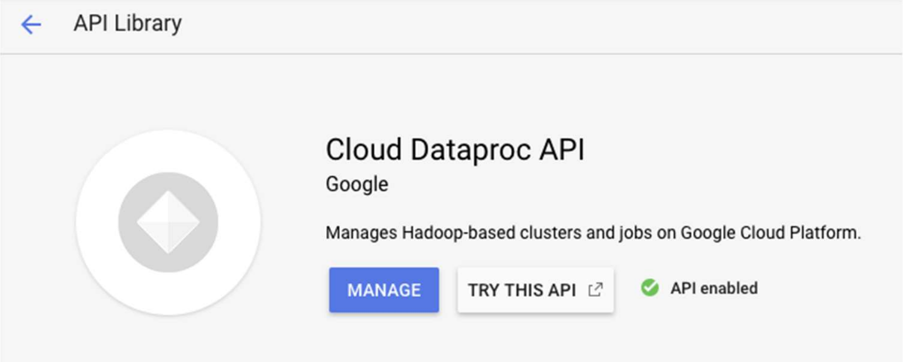
Click **Navigation menu** > **APIs & Services** > **Library**:



Type **Cloud Dataproc** in the **Search for APIs & Services** dialog. The console will display the Cloud Dataproc API in the search results.

Click on **Cloud Dataproc API** to display the status of the API. If the API is not already enabled, click the **Enable** button.

If the API's enabled, you're good to go:



Cloud Dataproc API

Google

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

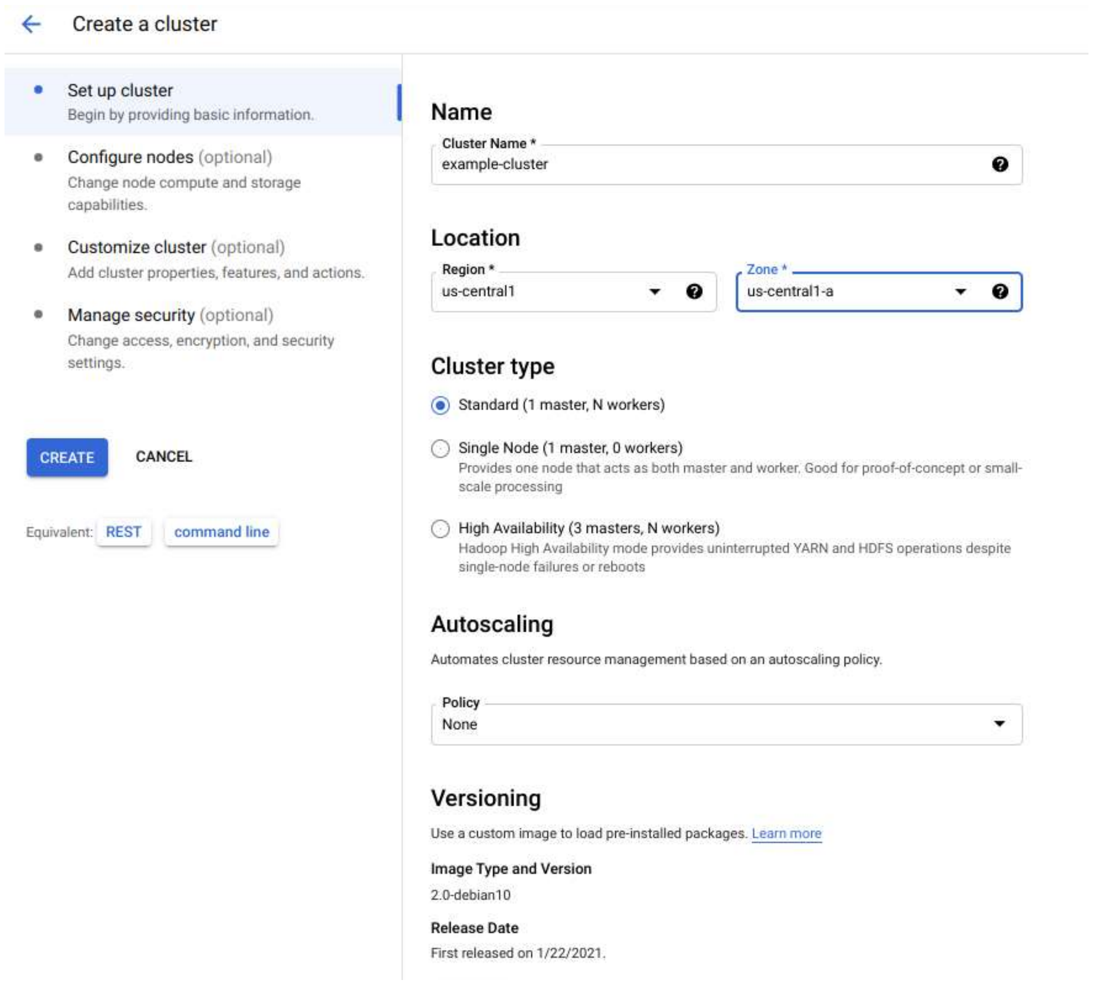
Create a cluster

In the Cloud Platform Console, select **Navigation menu** > **Dataproc** > **Clusters**, then click **Create cluster**.

Set the following fields for your cluster. Accept the default values for all other fields.

Field	Value
Name	example-cluster
Region	us-central1
Zone	us-central1-a

Note: A Zone is a special multi-region namespace that is capable of deploying instances into all Google Compute zones globally. You can also specify distinct regions, such as `us-east1` or `europe-west1`, to isolate resources (including VM instances and Cloud Storage) and metadata storage locations utilized by Cloud Dataproc within the user-specified region.



Create a cluster

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customize cluster (optional)
  Add cluster properties, features, and actions.
- Manage security (optional)
  Change access, encryption, and security settings.

Name

Cluster Name *: example-cluster

Location

Region *: us-central1

Zone *: us-central1-a

Cluster type

● Standard (1 master, N workers)

○ Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

○ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling

Automates cluster resource management based on an autoscaling policy.

Policy: None

Versioning

Use a custom image to load pre-installed packages. Learn more

Image Type and Version
2.0-debian10

Release Date
First released on 1/22/2021.

Click **Create** to create the cluster.

Your new cluster will appear in the Clusters list. It may take a few minutes to create, the cluster Status shows as **Provisioning** until the cluster is ready to use, then changes to **Running**.

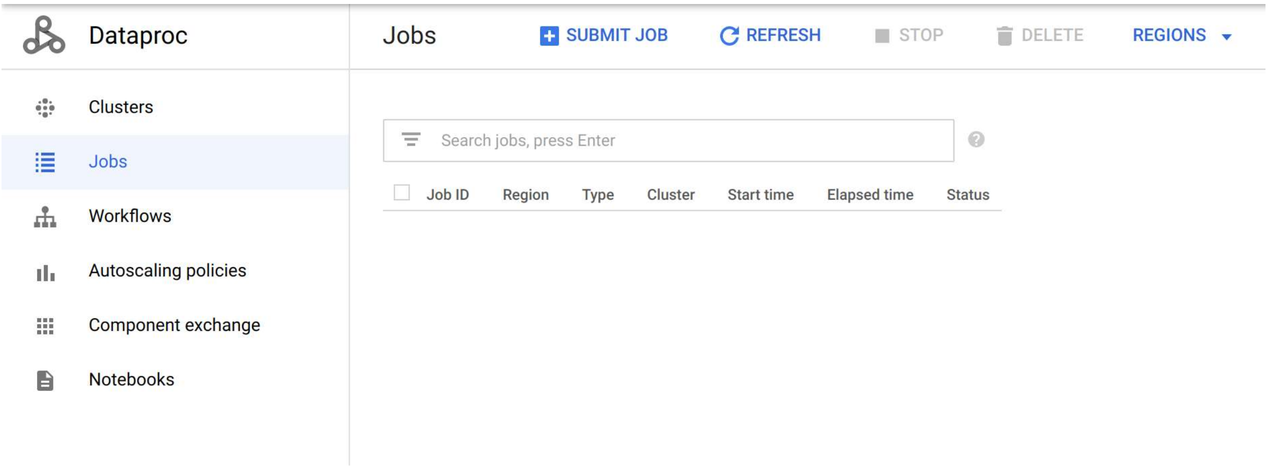
Test Completed Task

Click **Check my progress** to verify your performed task.

Submit a job

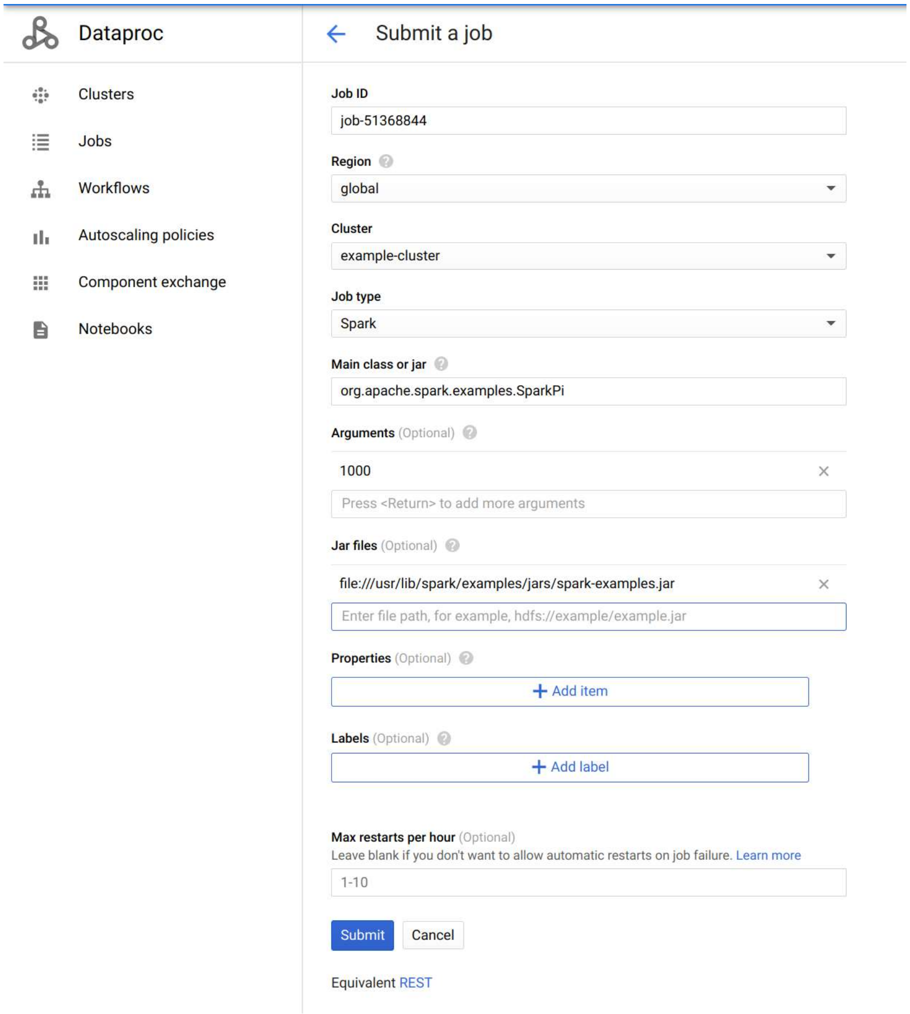
To run a sample Spark job:

Click **Jobs** in the left pane to switch to Dataproc's jobs view, then click **Submit job**:



Set the following fields to update Job. Accept the default values for all other fields.

Field	Value
Cluster	example-cluster
Job type	Spark
Main class or jar	org.apache.spark.examples.SparkPi
Arguments	1000 (This sets the number of tasks.)
Jar file	file:///usr/lib/spark/examples/jars/spark-examples.jar



Job ID: job-51368844

Region: global

Cluster: example-cluster

Job type: Spark

Main class or jar: org.apache.spark.examples.SparkPi

Arguments (Optional): 1000

Press <Return> to add more arguments

Jar files (Optional): file:///usr/lib/spark/examples/jars/spark-examples.jar

Enter file path, for example, hdfs://example/example.jar

Properties (Optional)

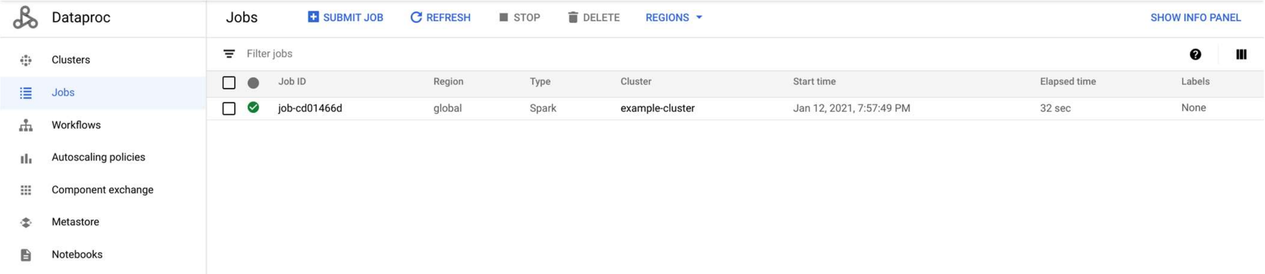
Labels (Optional)

Max restarts per hour (Optional)
Leave blank if you don't want to allow automatic restarts on job failure. Learn more

1-10

Click **Submit**.

How the job calculates Pi: The Spark job estimates a value of Pi using the [Monte Carlo method](). It generates x,y points on a coordinate plane that models a circle enclosed by a unit square. The input argument (1000) determines the number of x,y pairs to generate; the more pairs generated, the greater the accuracy of the estimation. This estimation leverages Cloud Dataproc worker nodes to parallelize the computation. For more information, see [Estimating Pi using the Monte Carlo Method]() and see [JavaSparkPi.java on GitHub]().
Your job should appear in the **Jobs** list, which shows your project's jobs with its cluster, type, and current status. Job status displays as **Running**, and then **Succeeded** after it completes.



□	●	Job ID	Region	Type	Cluster	Start time	Elapsed time	Labels
□	✓	job-cd01466d	global	Spark	example-cluster	Jan 12, 2021, 7:57:49 PM	32 sec	None

Test Completed Task

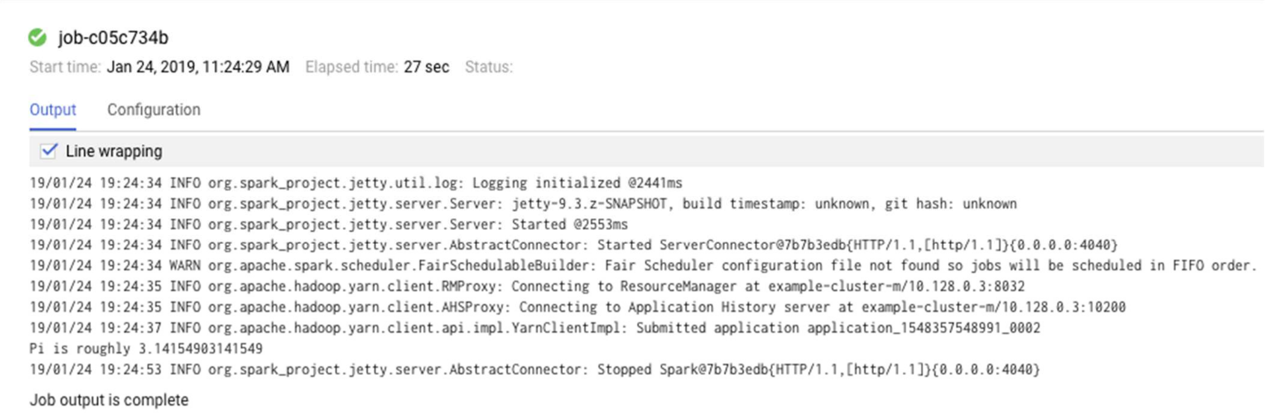
Click **Check my progress** to verify your performed task.

View the job output

To see your completed job's output:

Click the job ID in the **Jobs** list.

Check **Line wrapping** or scroll all the way to the right to see the calculated value of Pi. Your output, with **Line wrapping** checked, should look something like this:



job-c05c734b

Start time: Jan 24, 2019, 11:24:29 AM Elapsed time: 27 sec Status:

☑ Line wrapping

```
19/01/24 19:24:34 INFO org.spark_project.jetty.util.log: Logging initialized @2441ms
19/01/24 19:24:34 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
19/01/24 19:24:34 INFO org.spark_project.jetty.server.Server: Started @2553ms
19/01/24 19:24:34 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@7b7b3edb{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
19/01/24 19:24:34 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order.
19/01/24 19:24:35 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at example-cluster-m/10.128.0.3:8032
19/01/24 19:24:35 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at example-cluster-m/10.128.0.3:10200
19/01/24 19:24:37 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1548357548991_0002
Pi is roughly 3.14154903141549
19/01/24 19:24:53 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@7b7b3edb{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```

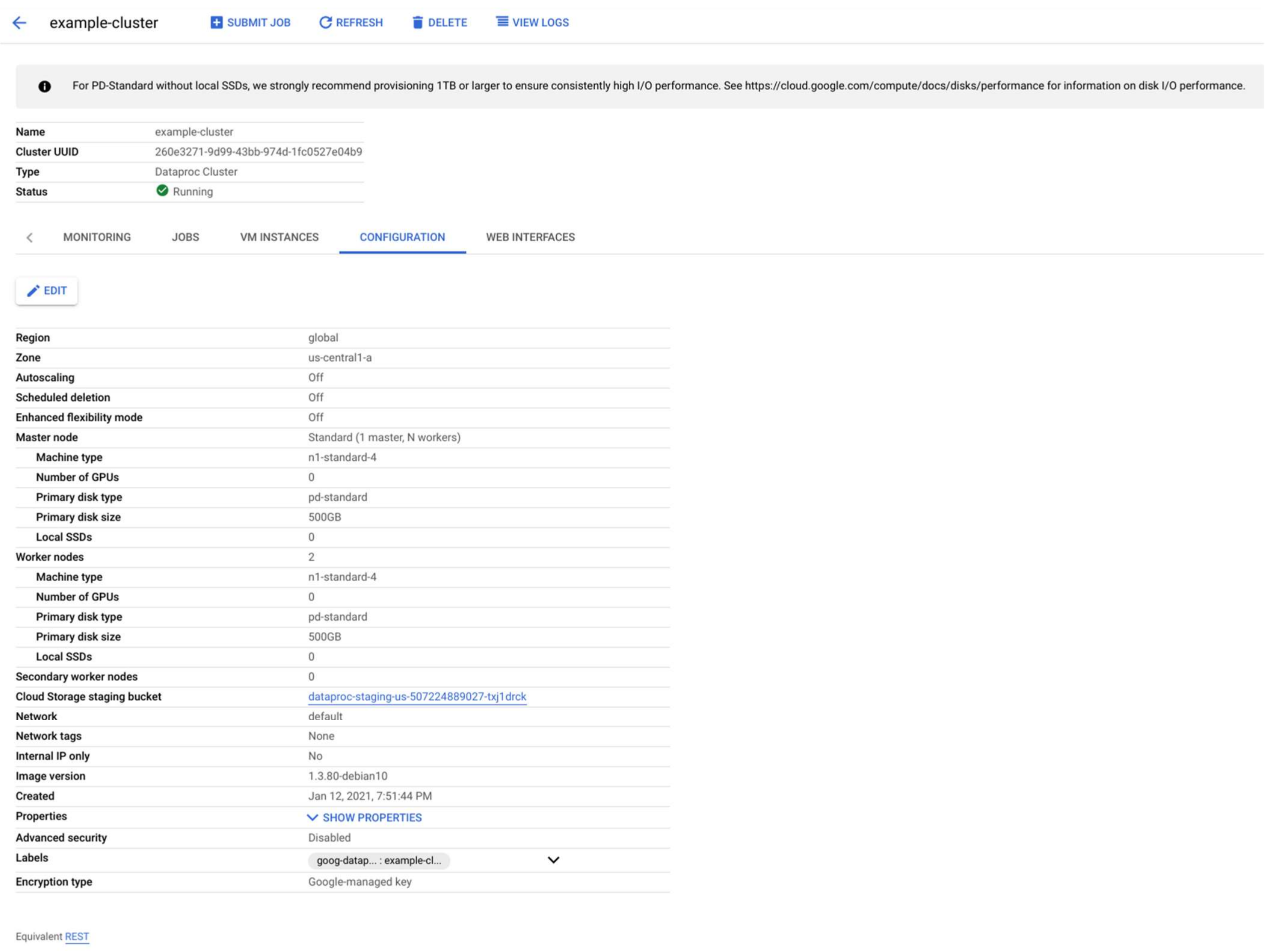
Job output is complete

Your job has successfully calculated a rough value for pi!

Update a cluster

To change the number of worker instances in your cluster:

1. Select **Clusters** in the left navigation pane to return to the Dataproc Clusters view.
2. Click **example-cluster** in the **Clusters** list. By default, the page displays an overview of your cluster's CPU usage.
3. Click **Configuration** to display your cluster's current settings.



example-cluster

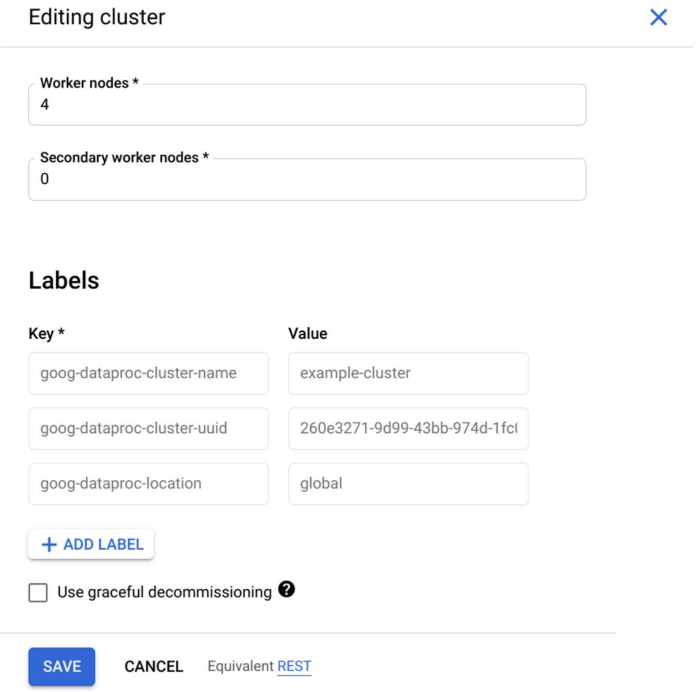
For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

Name	example-cluster
Cluster UUID	260e3271-9d99-43bb-974d-1fc0527e04b9
Type	Dataproc Cluster
Status	✓ Running

Region	global
Zone	us-central1-a
Autoscaling	Off
Scheduled deletion	Off
Enhanced flexibility mode	Off
Master node	Standard (1 master, N workers)
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Worker nodes	2
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Secondary worker nodes	0
Cloud Storage staging bucket	dataproc-staging-us-507224889027-txj1drck
Network	default
Network tags	None
Internal IP only	No
Image version	1.3.80-debian10
Created	Jan 12, 2021, 7:51:44 PM
Properties	SHOW PROPERTIES
Advanced security	Disabled
Labels	goog-datap... : example-cl...
Encryption type	Google-managed key

Equivalent REST

4. Click **Edit**. The number of worker nodes is now editable.
5. Enter **4** in the **Worker nodes** field.
6. Click **Save**.



Editing cluster

Worker nodes *
4

Secondary worker nodes *
0

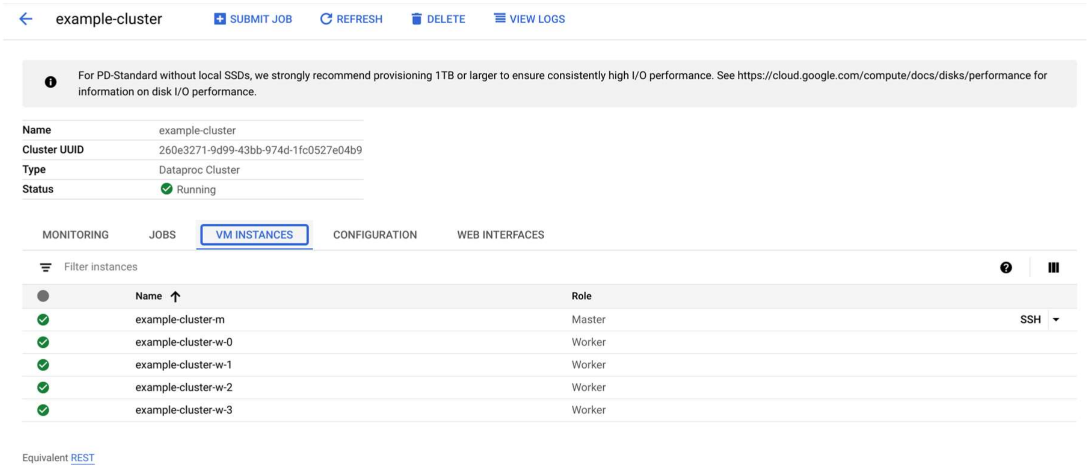
Labels

Key *	Value
goog-dataproc-cluster-name	example-cluster
goog-dataproc-cluster-uuid	260e3271-9d99-43bb-974d-1fc(
goog-dataproc-location	global

□ Use graceful decommissioning

Equivalent REST

Your cluster is now updated. Check out the number of VM instances in the cluster:



For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

Name	example-cluster
Cluster UUID	260e3271-9d99-43bb-974d-1fc0527e04b9
Type	Dataproc Cluster
Status	✓ Running

	Name ↑	Role
✓	example-cluster-m	Master
✓	example-cluster-w-0	Worker
✓	example-cluster-w-1	Worker
✓	example-cluster-w-2	Worker
✓	example-cluster-w-3	Worker

Test Completed Task

Click **Check my progress** to verify your performed task.

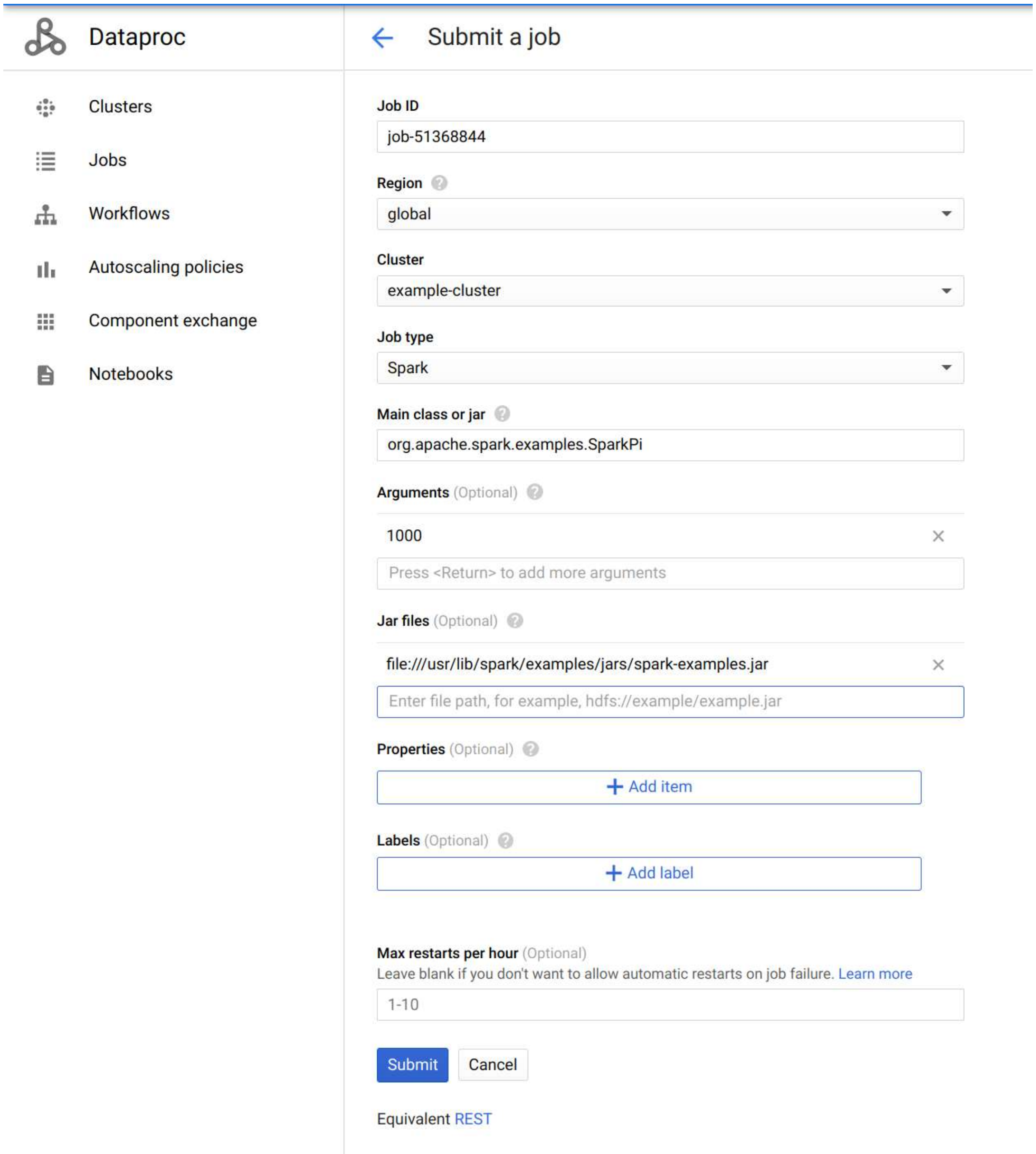
Update a cluster

Check my progress

To rerun the job with the updated cluster, you would click **Jobs** in the left pane, then click **SUBMIT JOB**.

Set the same fields you set in the **Submit a job** section:

Field	Value
Cluster	example-cluster
Job type	Spark
Main class or jar	org.apache.spark.examples.SparkPi
Arguments	1000 (This sets the number of tasks.)
Jar file	file:///usr/lib/spark/examples/jars/spark-examples.jar



Job ID
job-51368844

Region
global

Cluster
example-cluster

Job type
Spark

Main class or jar
org.apache.spark.examples.SparkPi

Arguments (Optional)
1000
Press <Return> to add more arguments

Jar files (Optional)
file:///usr/lib/spark/examples/jars/spark-examples.jar
Enter file path, for example, hdfs://example/example.jar

Properties (Optional)
+ Add item

Labels (Optional)
+ Add label

Max restarts per hour (Optional)
Leave blank if you don't want to allow automatic restarts on job failure. Learn more
1-10

Click **Submit**.

Test your Understanding

Below are multiple-choice questions to reinforce your understanding of this lab's concepts. Answer them to the best of your abilities.

Which type of Dataproc job is submitted in the lab?
Spark

Dataproc helps users process, transform and understand vast quantities of data.
True

Congratulations!

Now you know how to use the Cloud Console to create and update a Dataproc cluster and then submit a job in that cluster.

Finish Your Quest

Continue your Qwiklabs Baseline: Data, ML, AI or Data Engineering Quest. A Quest is a series of related labs that form a learning path. Completing this Quest earns you the badge above, to recognize your achievement. You can make your badge (or badges) public and link to them in your online resume or social media account. Enroll in a Quest and get immediate completion credit if you've taken this lab. See other available Qwiklabs Quests.

Next Steps / Learn More

This lab is also part of a series of labs called Qwik Starts. These labs are designed to give you a little taste of the many features available with Google Cloud. Search for "Qwik Starts" in the lab catalog to find the next lab you'd like to take!

Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.
Manual Last Updated March 03, 2021
Lab Last Tested March 03, 2021