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 Platform (GCP) 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

Qwiklabs setup

What you'll need

To complete this lab, you'll need:

- Access to a standard internet browser (Chrome browser recommended).
- Time. Note the lab's **Completion** time in Qwiklabs. This is an estimate of the time it should take to complete all steps. Plan your schedule so you have time to complete the lab. Once you start the lab, you will not be able to pause and return later (you begin at step 1 every time you start a lab).
- The lab's **Access** time is how long your lab resources will be available. If you finish your lab with access time still available, you will be able to explore the Google Cloud Platform or work on any section of the lab that was marked "if you have time". Once the Access time runs out, your lab will end and all resources will terminate.
- You **DO NOT** need a Google Cloud Platform account or project. An account, project and associated resources are provided to you as part of this lab.
- If you already have your own GCP account, make sure you do not use it for this lab.
- If your lab prompts you to log into the console, **use only the student account provided to you by the lab**. This prevents you from incurring charges for lab activities in your personal GCP account.

Start your lab

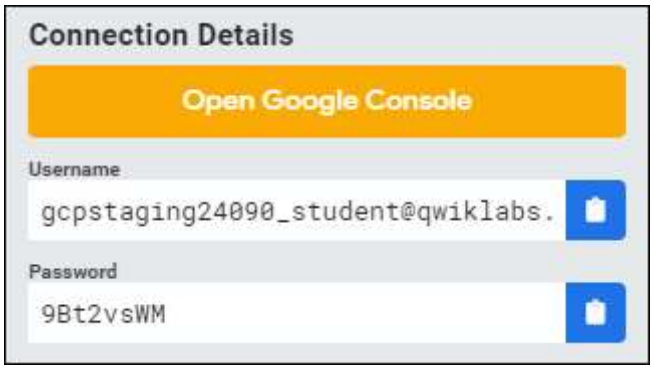
When you are ready, click **Start Lab**. You can track your lab's progress with the status bar at the top of your screen.

> **Important:** What is happening during this time? Your lab is spinning up GCP resources for you behind the scenes, including an account, a project, resources within the project, and permission for you to control the

resources needed to run the lab. This means that instead of spending time manually setting up a project and building resources from scratch as part of your lab, you can begin learning more quickly.

Find Your Lab's GCP Username and Password

To access the resources and console for this lab, locate the Connection Details panel in Qwiklabs. Here you will find the account ID and password for the account you will use to log in to the Google Cloud Platform:



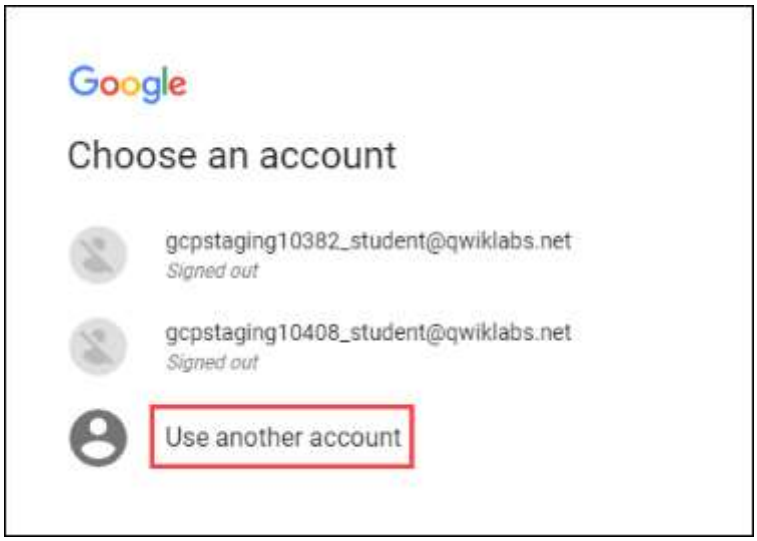
If your lab provides other resource identifiers or connection-related information, it will appear on this panel as well.

Google Cloud Platform Console

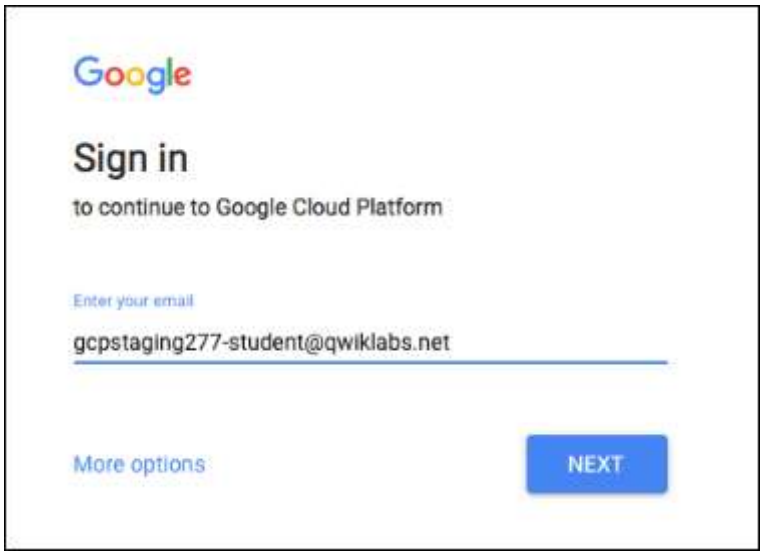
Log in to Google Cloud Console

Using the Qwiklabs browser tab/window or the separate browser you are using for the Qwiklabs session, copy the Username from the Connection Details panel and click the "Open Google Console" button.

You'll be asked to choose an account. Click **Use another account**.



Paste in the Username, and then the Password as prompted:

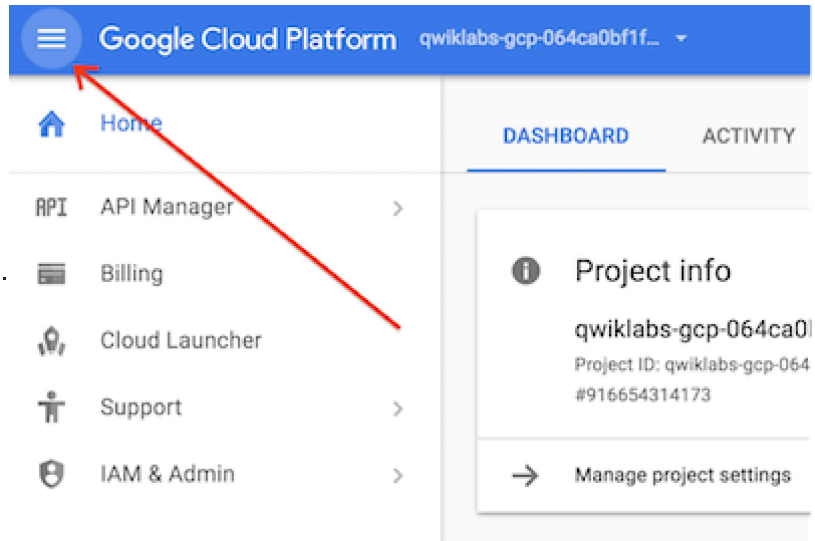


Accept the terms and conditions.

Since this is a temporary account, which you will only have access to for this one lab:

- Do not add recovery options
- Do not sign up for free trials

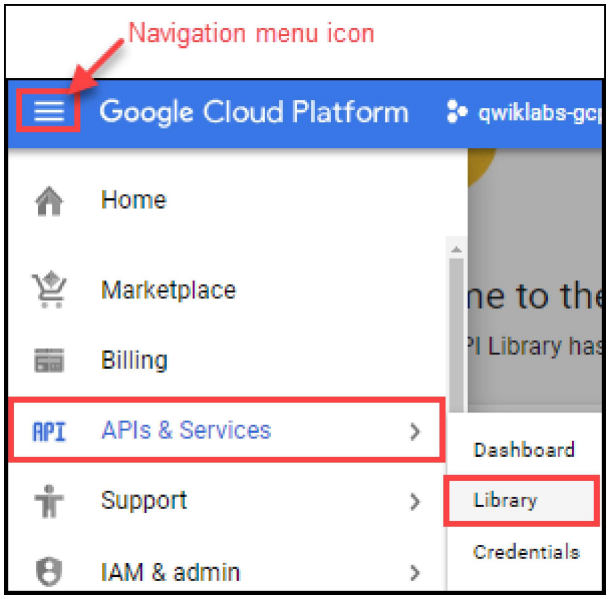
Note: You can view the menu with a list of GCP Products and Services by clicking the **Navigation menu** at the top-left next to “Google Cloud Platform”.



Confirm Cloud Dataproc API is enabled

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

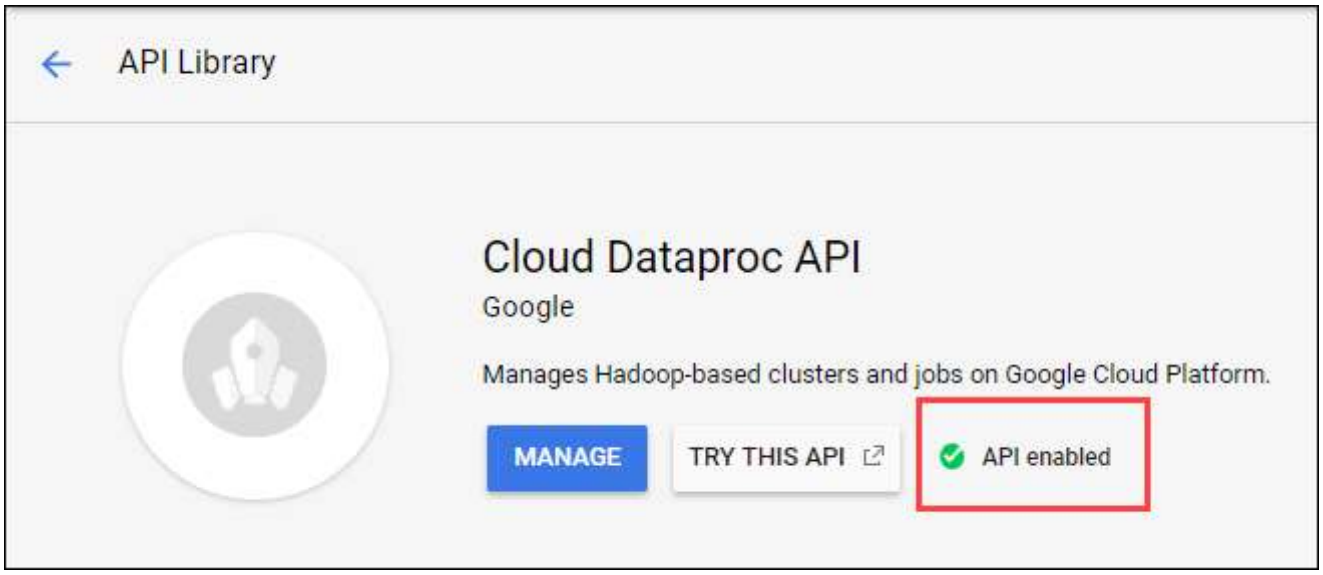
Click **Navigation menu** > **APIs & Services** > **Library**:



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:



Create a cluster

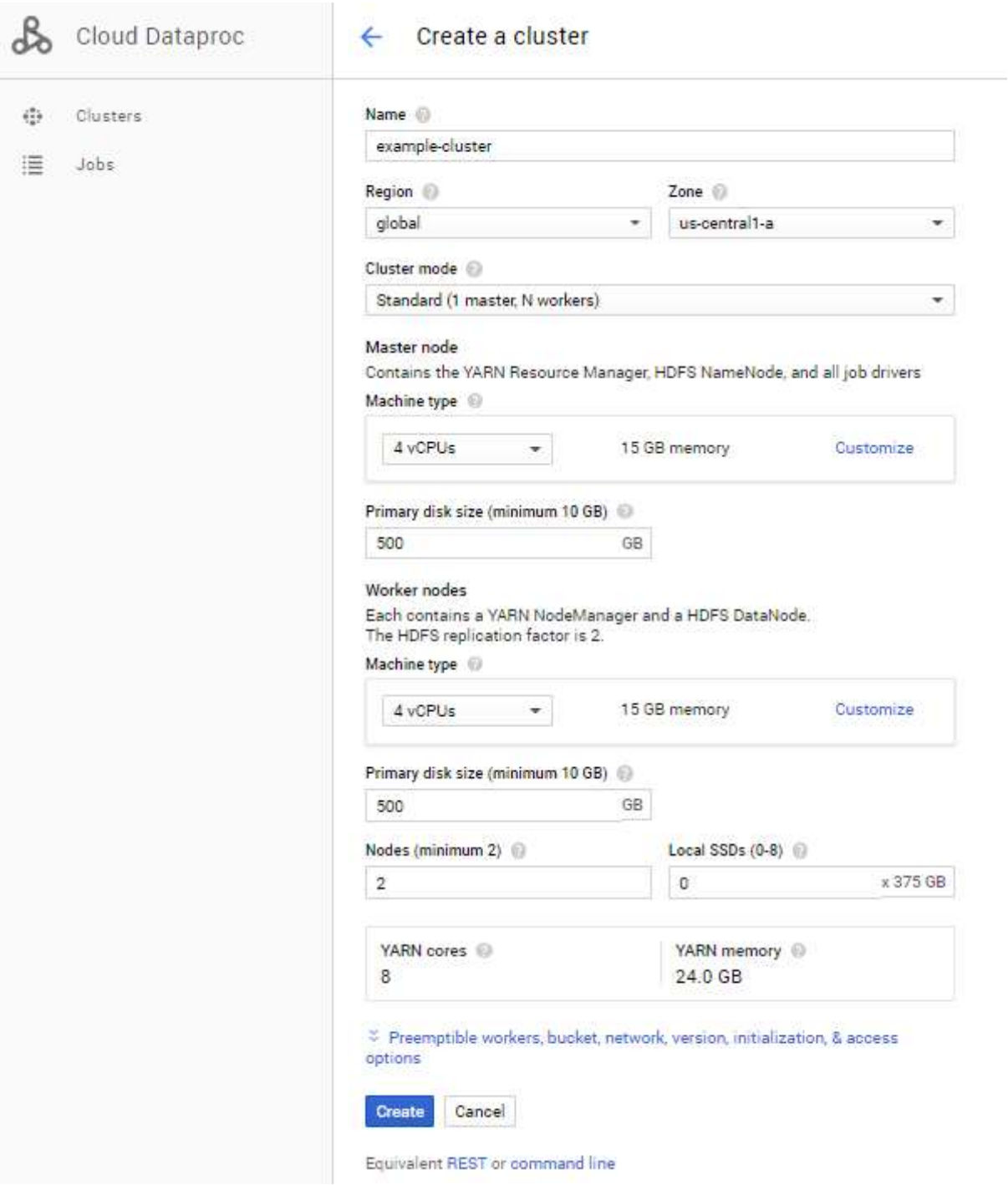
In the Cloud Platform Console, select **Navigation menu** > **Dataproc** > **Clusters**, then click **Create cluster**.

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

Field	Value
Name	example-cluster

Region	global
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 Google Cloud Storage) and metadata storage locations utilized by Cloud Dataproc within the user-specified region.



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. If you have completed the task successfully you will granted with an assessment score.

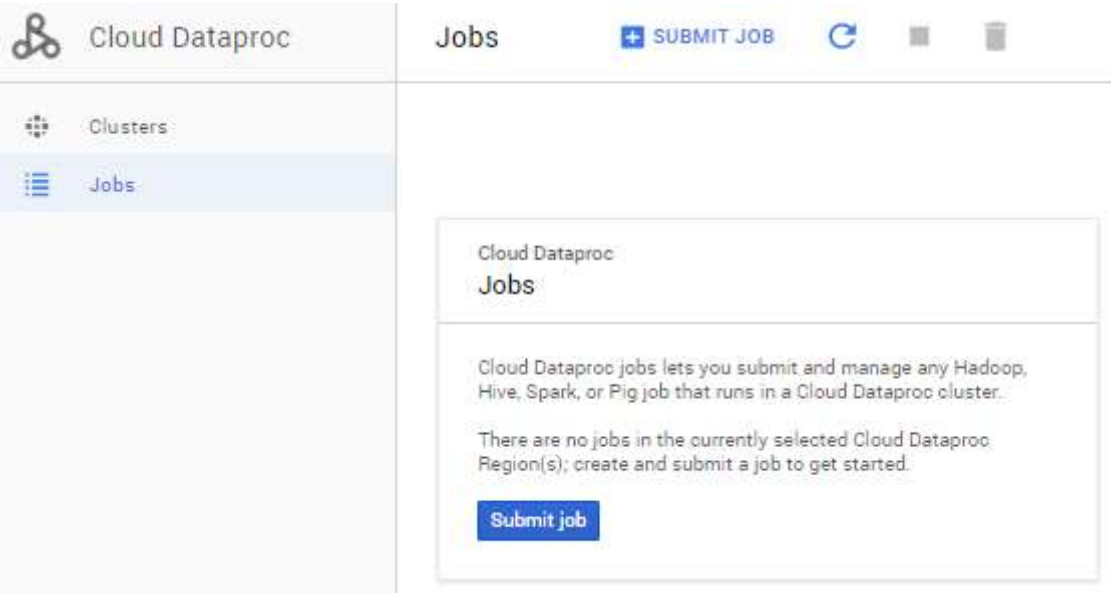
Create a Dataproc cluster

Check my progress

Submit a job

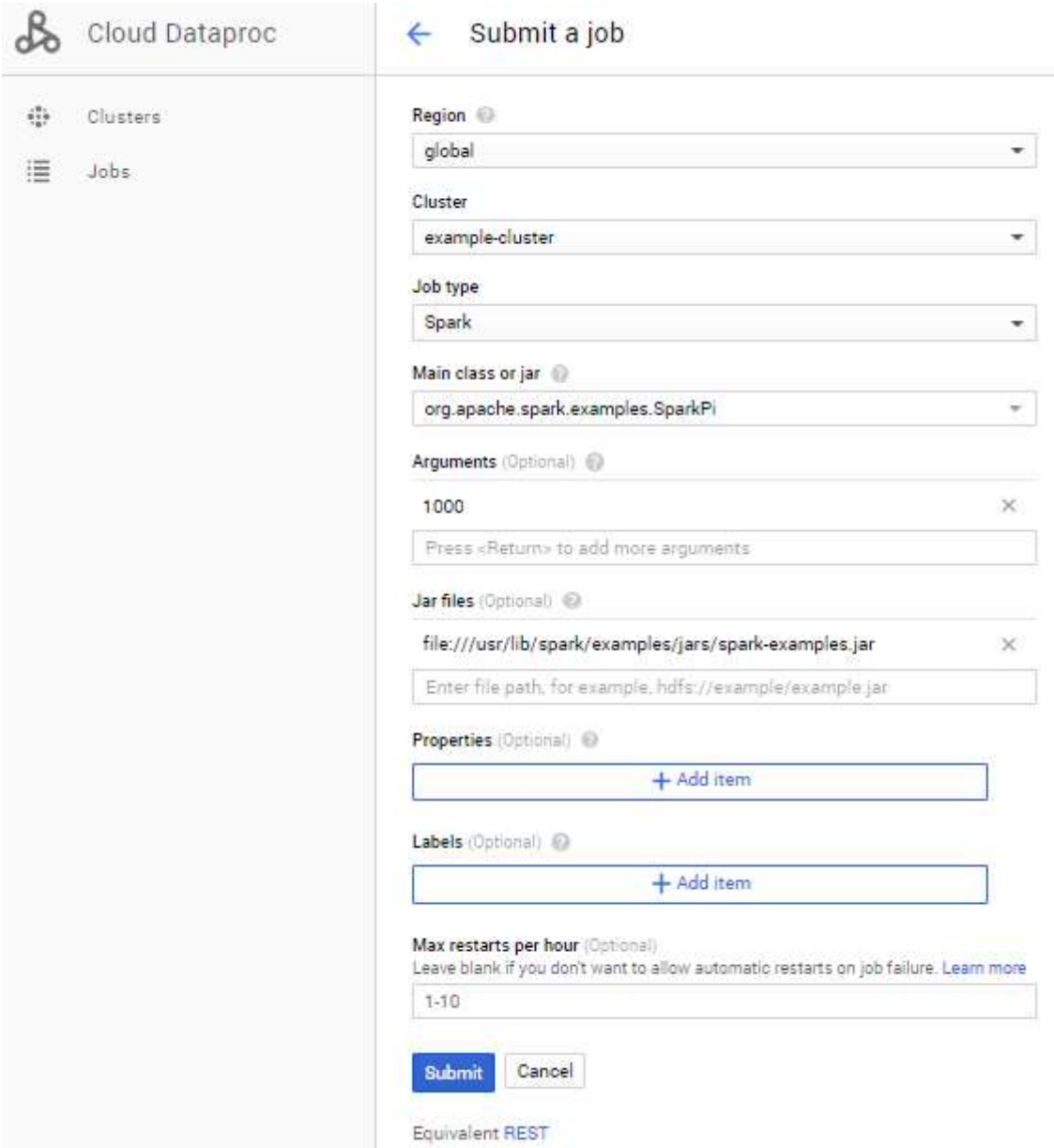
To run a sample Spark job:

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.

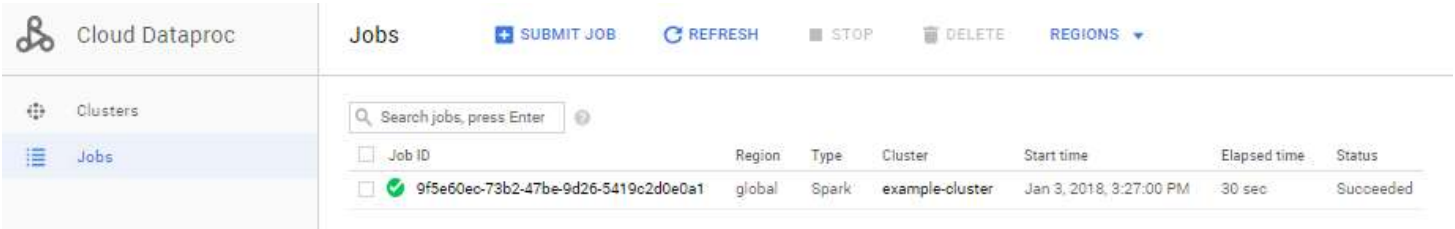
Field	Value
Cluster	example-cluster
Job type	Spark
Main class or jar	org.apache.spark.examples.SparkPi
Arguments	1000 (This sets the number of tasks.)
Jar file	file:///usr/lib/spark/examples/jars/spark-examples.jar



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.



Test Completed Task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

Submit a job

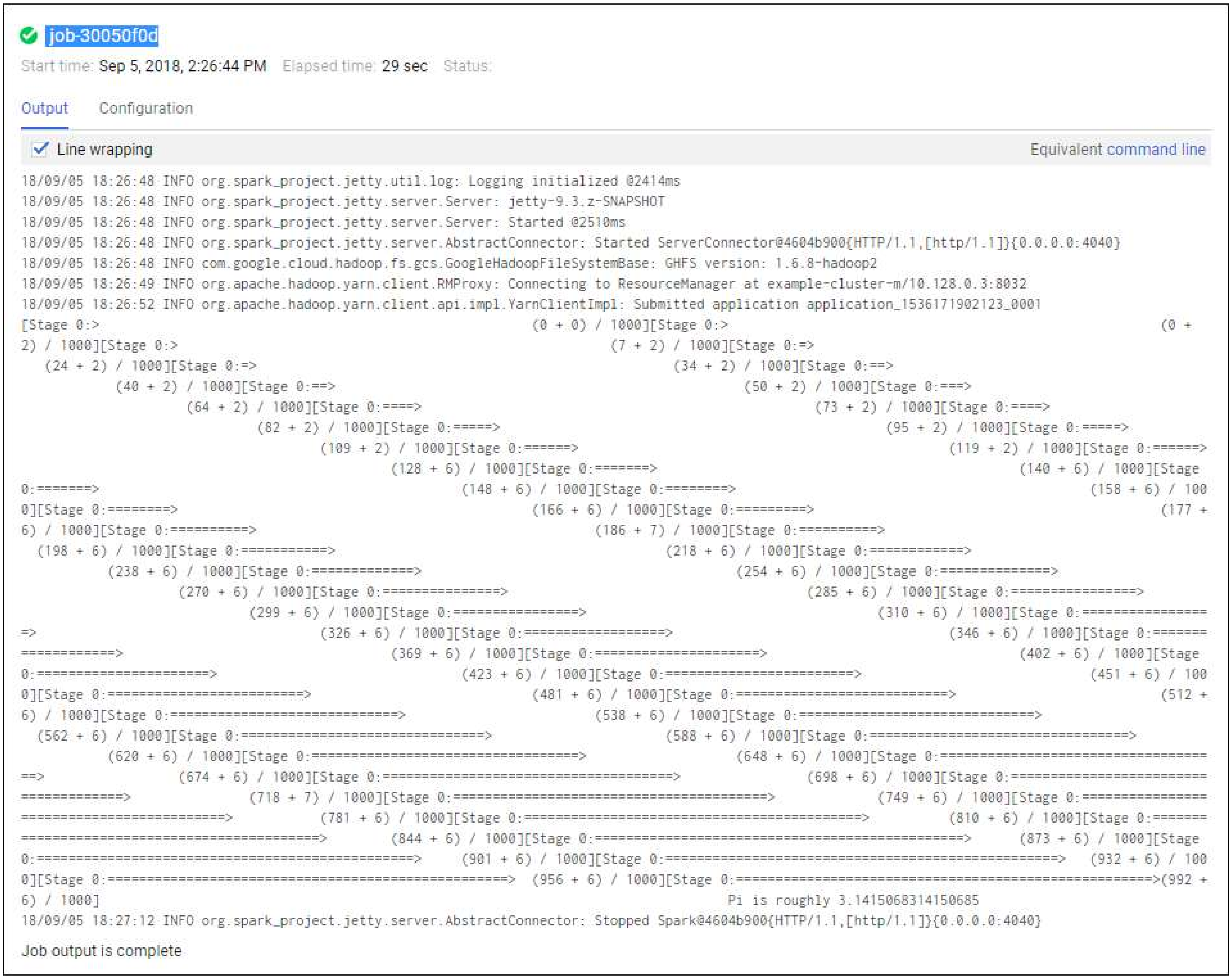
Check my progress

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-30050f0d

Start time: Sep 5, 2018, 2:26:44 PM Elapsed time: 29 sec Status:

```
18/09/05 18:26:48 INFO org.spark_project.jetty.util.log: Logging initialized @2414ms
18/09/05 18:26:48 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
18/09/05 18:26:48 INFO org.spark_project.jetty.server.Server: Started @2510ms
18/09/05 18:26:48 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@4604b900{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/09/05 18:26:48 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.8-hadoop2
18/09/05 18:26:49 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at example-cluster-m/10.128.0.3:8032
18/09/05 18:26:52 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1536171902123_0001
[Stage 0:>                                                       (0 + 0) / 1000][Stage 0:>                                                       (0 +
2) / 1000][Stage 0:>                                                       (7 + 2) / 1000][Stage 0:=>
   (24 + 2) / 1000][Stage 0:=>                                                     (34 + 2) / 1000][Stage 0:==>
            (40 + 2) / 1000][Stage 0:==>                                                    (50 + 2) / 1000][Stage 0:===>
                     (64 + 2) / 1000][Stage 0:====>                                                  (73 + 2) / 1000][Stage 0:====>
                              (82 + 2) / 1000][Stage 0:=====>                                                 (95 + 2) / 1000][Stage 0:=====>
                                      (109 + 2) / 1000][Stage 0:======>                                               (119 + 2) / 1000][Stage 0:======>
                                               (128 + 6) / 1000][Stage 0:=======>                                              (140 + 6) / 1000][Stage
0:=======>                                              (148 + 6) / 1000][Stage 0:========>                                             (158 + 6) / 100
0][Stage 0:========>                                             (166 + 6) / 1000][Stage 0:=========>                                            (177 +
6) / 1000][Stage 0:==========>                                           (186 + 7) / 1000][Stage 0:==========>
  (198 + 6) / 1000][Stage 0:===========>                                          (218 + 6) / 1000][Stage 0:============>
           (238 + 6) / 1000][Stage 0:=============>                                        (254 + 6) / 1000][Stage 0:==============>
                    (270 + 6) / 1000][Stage 0:===============>                                      (285 + 6) / 1000][Stage 0:================>
                             (299 + 6) / 1000][Stage 0:================>                                     (310 + 6) / 1000][Stage 0:================
=>                                    (326 + 6) / 1000][Stage 0:==================>                                   (346 + 6) / 1000][Stage 0:=======
============>                                  (369 + 6) / 1000][Stage 0:=====================>                                (402 + 6) / 1000][Stage
0:======================>                               (423 + 6) / 1000][Stage 0:========================>                             (451 + 6) / 100
0][Stage 0:=========================>                            (481 + 6) / 1000][Stage 0:===========================>                          (512 +
6) / 1000][Stage 0:=============================>                        (538 + 6) / 1000][Stage 0:==============================>
  (562 + 6) / 1000][Stage 0:===============================>                      (588 + 6) / 1000][Stage 0:=================================>
           (620 + 6) / 1000][Stage 0:==================================>                   (648 + 6) / 1000][Stage 0:==================================
==>                 (674 + 6) / 1000][Stage 0:=====================================>                (698 + 6) / 1000][Stage 0:=========================
============>                (718 + 7) / 1000][Stage 0:========================================>             (749 + 6) / 1000][Stage 0:================
==========================>           (781 + 6) / 1000][Stage 0:===========================================>          (810 + 6) / 1000][Stage 0:=======
======================================>        (844 + 6) / 1000][Stage 0:===============================================>      (873 + 6) / 1000][Stage
0:================================================>     (901 + 6) / 1000][Stage 0:==================================================>   (932 + 6) / 100
0][Stage 0:===================================================>  (956 + 6) / 1000][Stage 0:=====================================================>(992 +
6) / 1000]                                                                                Pi is roughly 3.1415068314150685
18/09/05 18:27:12 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@4604b900{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!

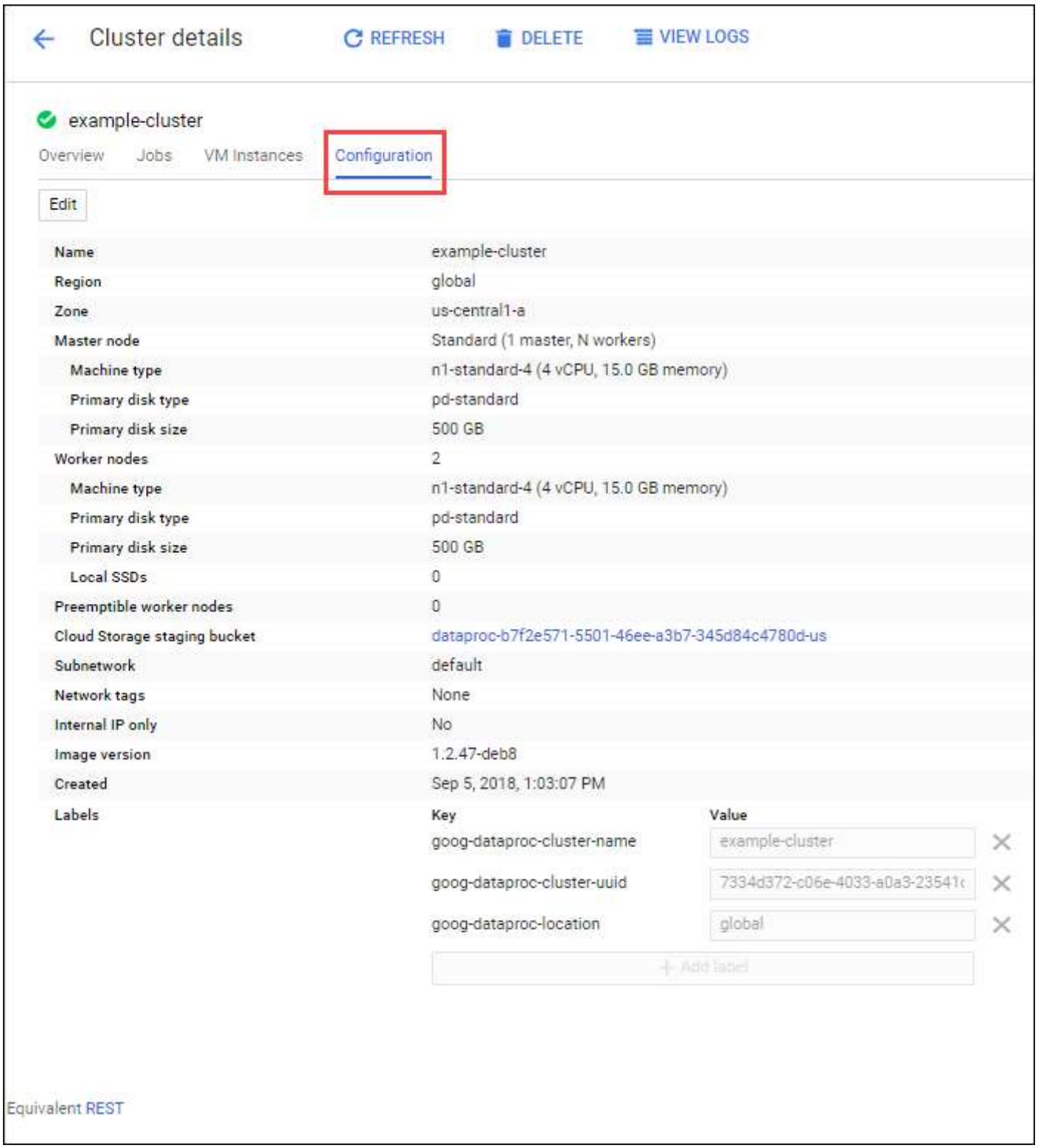
Update a cluster

To change the number of worker instances in your cluster:

. Select **Clusters** in the left navigation pane to return to the Dataproc Clusters view.

. Click **example-cluster** in the **Clusters** list. By default, the page displays an overview of your cluster's CPU usage.

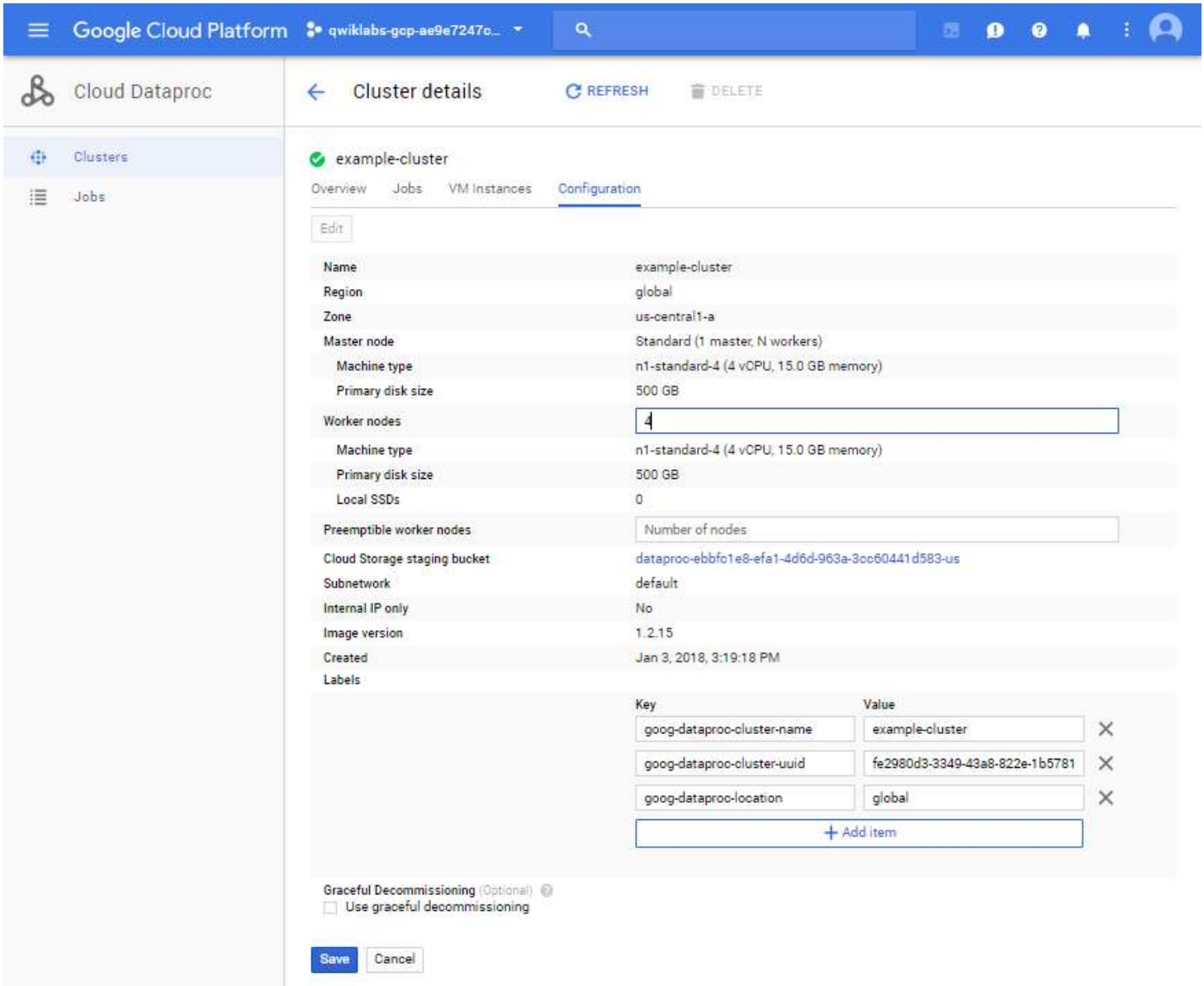
. Click **Configuration** to display your cluster's current settings.



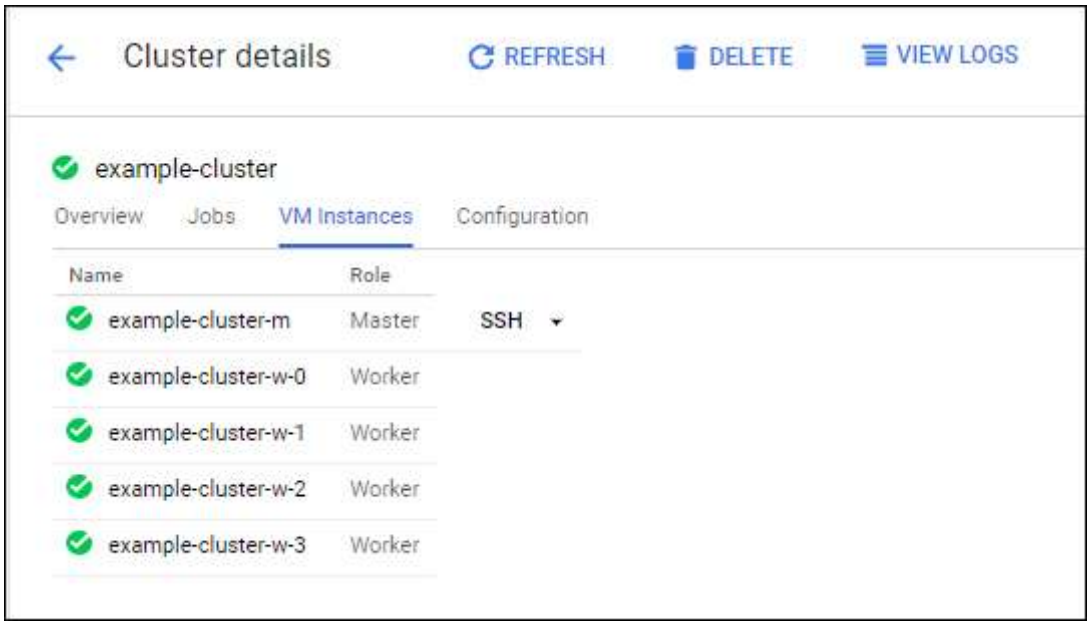
. Click **Edit**. The number of worker nodes is now editable.

. Enter **4** in the **Worker nodes** field.

. Click **Save.**



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



Test Completed Task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

Update a cluster

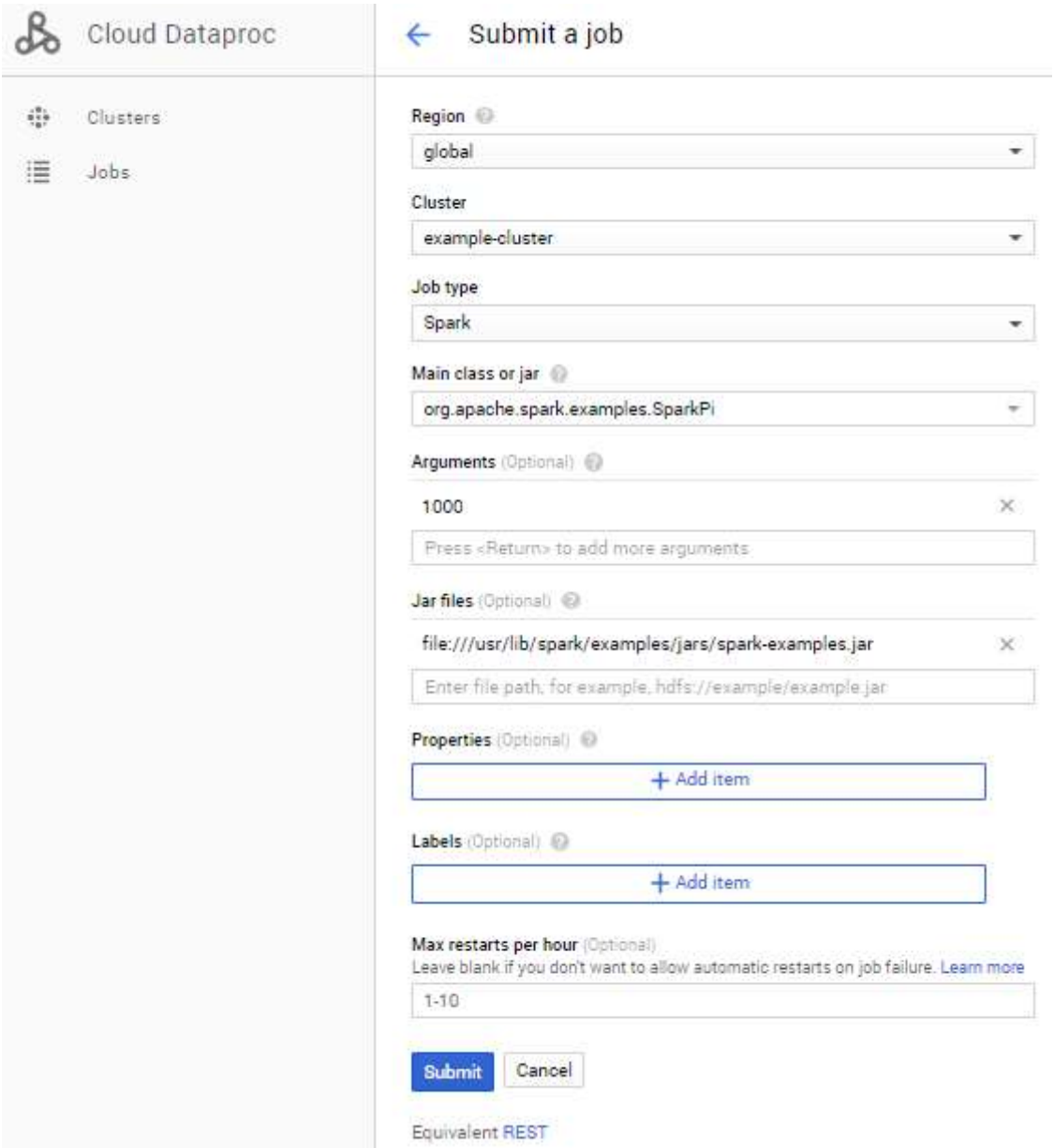
Check my progress

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



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?

- SparkSql
- PySpark
- Spark
- Hadoop
- Pig

Submit

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

True False

Congratulations!

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

Finish Your Quest

Continue your Quest with Baseline: Data, ML, AI orData Engineering. 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!