Dataproc: Qwik Start - Console

Overview

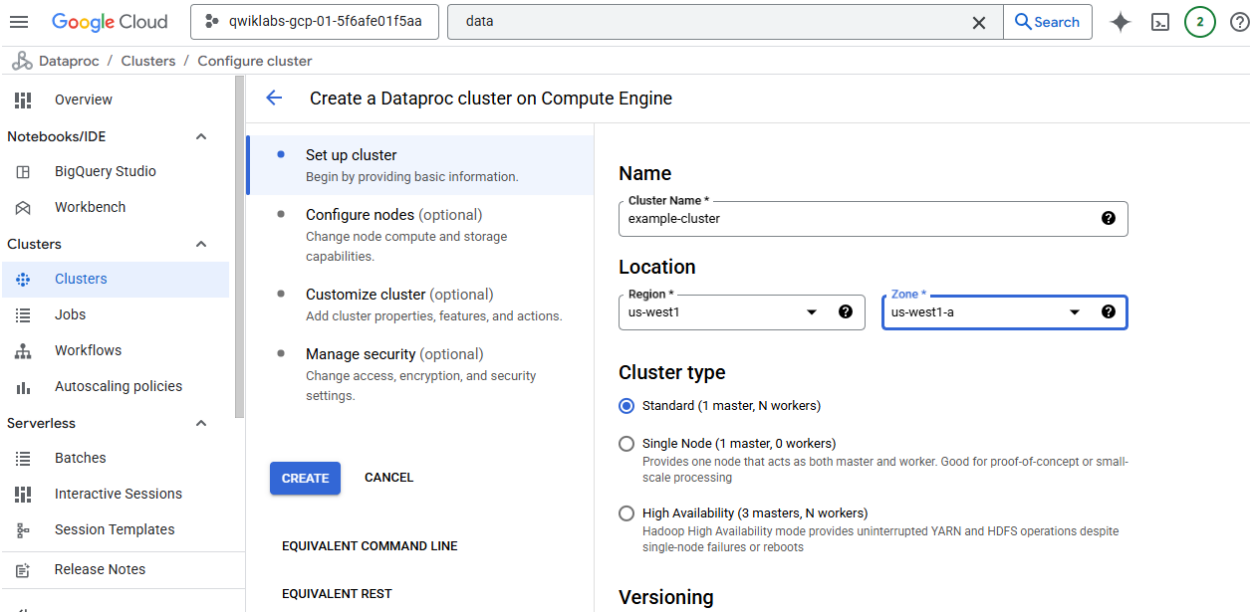
This lab demonstrates how to use **Google Cloud Dataproc**, a managed **Apache Spark** and **Hadoop** service, to create a cluster, run a Spark job, and modify the cluster size. The key objectives were:

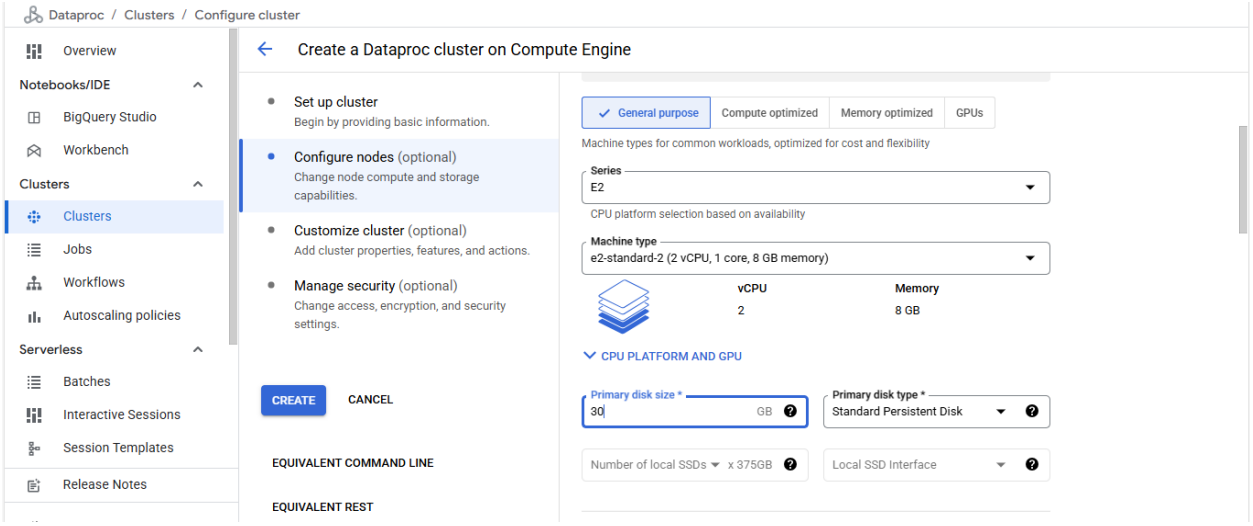
1. **Create a Dataproc cluster** in the Google Cloud Console.
2. **Run a simple Apache Spark job** to estimate the value of Pi.
3. **Modify the number of workers** in the cluster.

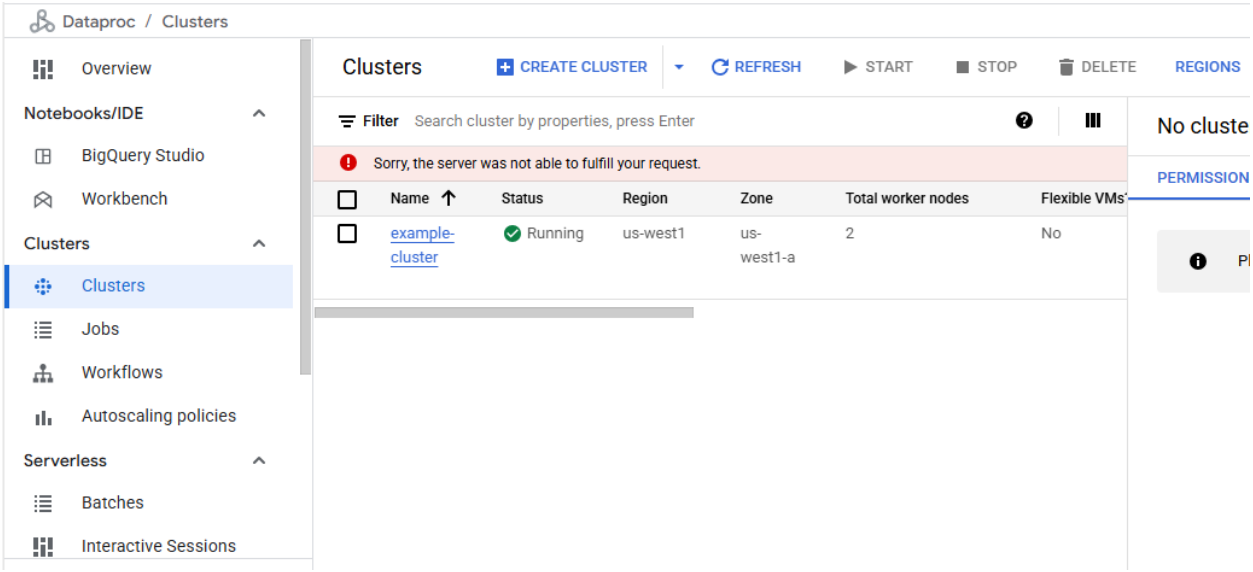
Task 1: Create a Cluster

Steps:

1. Navigated to **Dataproc > Clusters** in the Google Cloud Console.
2. Clicked **Create Cluster** with the following configuration:
    a. **Name:** example-cluster
    b. **Region & Zone:** Default selected
    c. **Machine Series:** E2
    d. **Machine Type (Master & Workers):** e2-standard-2
    e. **Primary Disk Size:** 30 GB (for both Master and Workers)
    f. **Worker Nodes:** 2
    g. **Internal IP Only:** Disabled (to allow external access)
3. Clicked **Create** and waited for the cluster status to change from **Provisioning** to **Running**.


Google Cloud
qwiklabs-gcp-01-5f6afe01f5aa
data
Search
Dataproc / Clusters / Configure cluster
Overview
Notebooks/IDE
BigQuery Studio
Workbench
Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive Sessions
Session Templates
Release Notes
Create a Dataproc cluster on Compute Engine
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.
CREATE
CANCEL
EQUIVALENT COMMAND LINE
EQUIVALENT REST
Name
Cluster Name *
example-cluster
Location
Region *
us-west1
Zone *
us-west1-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
Versioning

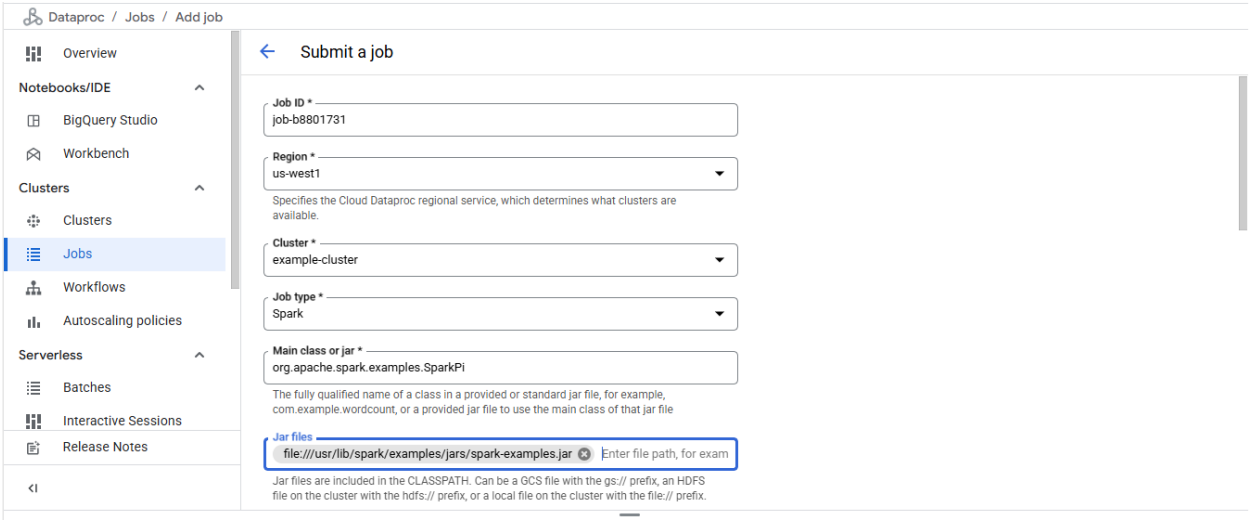

Dataproc / Clusters / Configure cluster
Overview
Notebooks/IDE
BigQuery Studio
Workbench
Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Interactive Sessions
Session Templates
Release Notes
Create a Dataproc cluster on Compute Engine
Set up cluster
Begin by providing basic information.
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.
CREATE
CANCEL
EQUIVALENT COMMAND LINE
EQUIVALENT REST
General purpose
Compute optimized
Memory optimized
GPUs
Machine types for common workloads, optimized for cost and flexibility
Series
E2
CPU platform selection based on availability
Machine type
e2-standard-2 (2 vCPU, 1 core, 8 GB memory)
vCPU
2
Memory
8 GB
CPU PLATFORM AND GPU
Primary disk size *
30
GB
Primary disk type *
Standard Persistent Disk
Number of local SSDs
x 375GB
Local SSD Interface

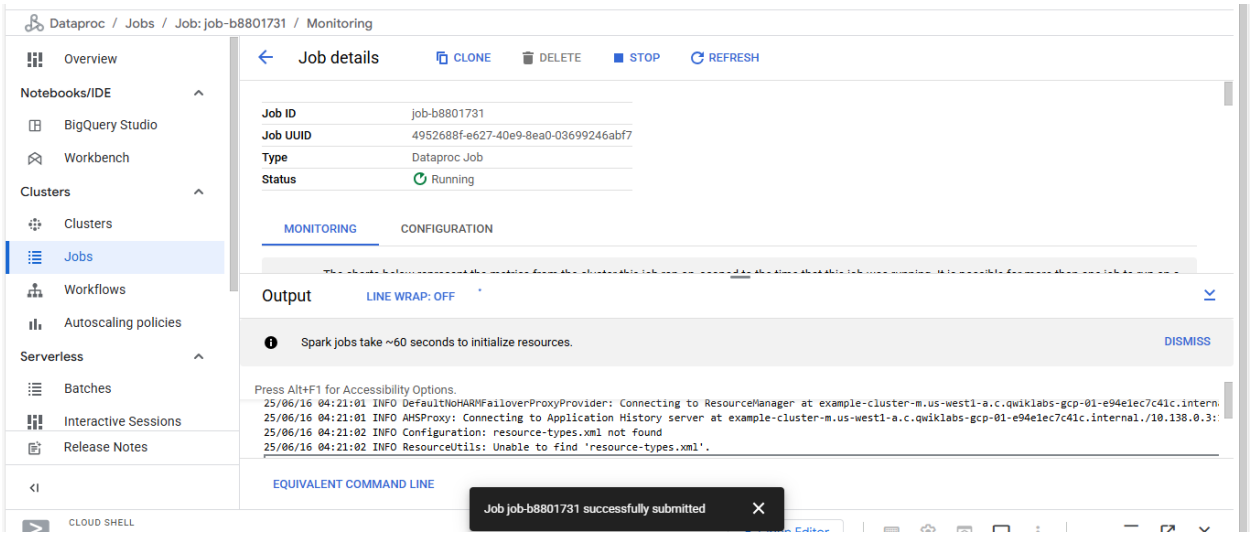


Task 2: Submit a Job

Steps:

1. Navigated to **Dataproc > Jobs**.
2. Clicked **Submit Job** with the following settings:
    a. **Cluster:** example-cluster
    b. **Job Type:** Spark
    c. **Main Class:** org.apache.spark.examples.SparkPi
    d. **Jar File:** file:///usr/lib/spark/examples/jars/spark-examples.jar
    e. **Arguments:** 1000 (number of tasks for Pi estimation)
3. Clicked **Submit**.

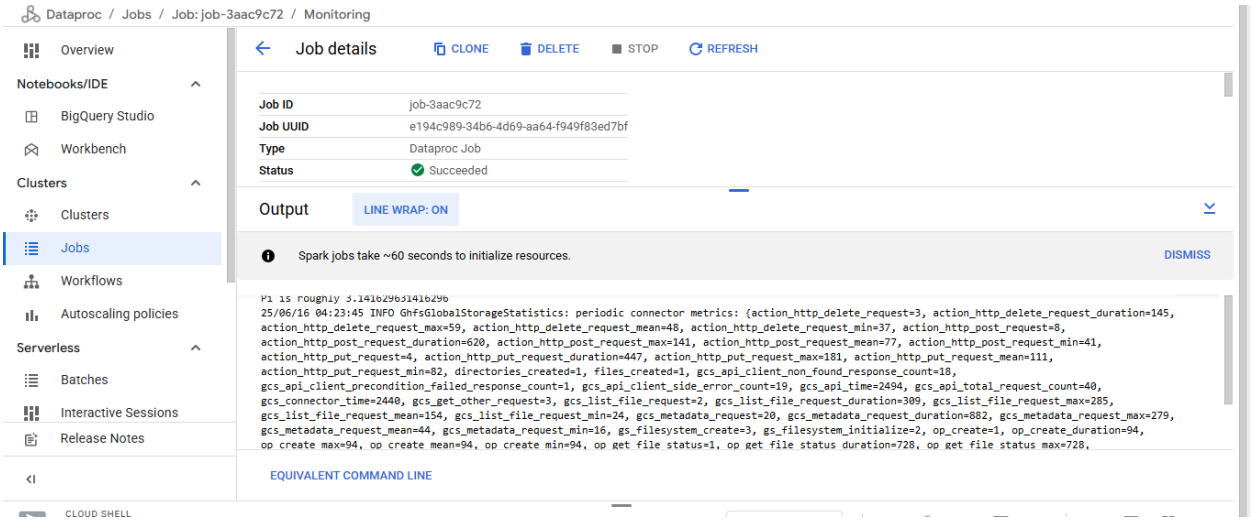




Task 3. View the job output

To see your completed job's output:

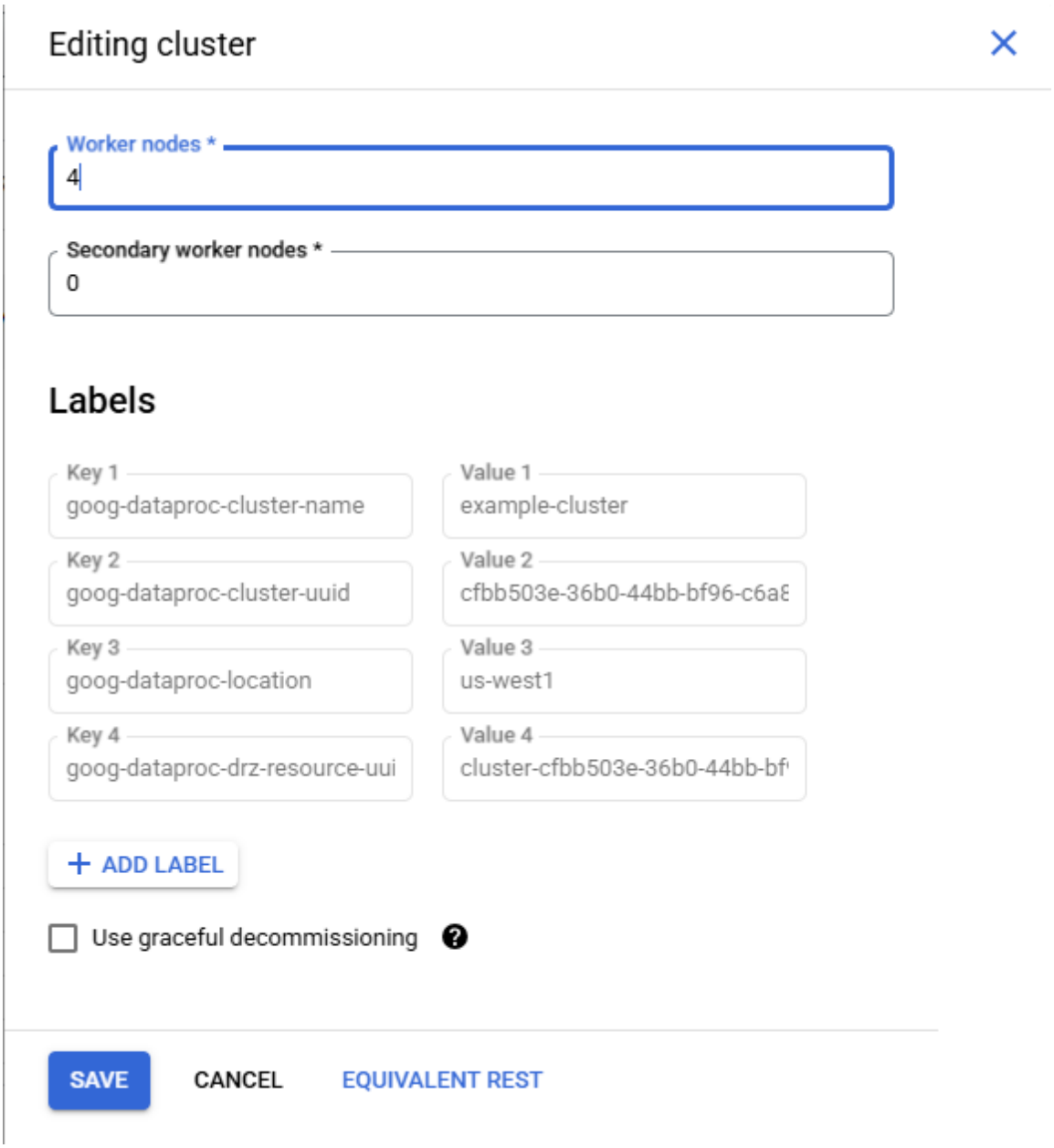
1. Click the job ID in the **Jobs** list.
2. Select **LINE WRAP** to ON or scroll all the way to the right to see the calculated value of Pi. Your output, with **LINE WRAP** ON, should look something like this:



Task 4. Update a cluster to modify the number of workers

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



1. To rerun the job with the updated cluster, you would click **Jobs** in the left pane, then click **SUBMIT JOB**.
2. Set the same fields you set in the **Submit a job** section:

Field	Value
Region	____

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

3. Click **Submit**.

Conclusion

- Successfully **created a Dataproc cluster** and ran a **Spark job** to estimate Pi.
- **Modified the cluster size** from 2 to 4 workers, demonstrating **scalability**.
- The lab reinforced key concepts of **managed Spark/Hadoop clusters** in Google Cloud.