

Dataproc: Qwik Start - Command Line

Table of Contents

1. Introduction
2. Task 1: Creating a Dataproc Cluster
3. Task 2: Submitting a Spark Job
4. Task 3: Scaling the Cluster
5. Key Learnings
6. Conclusion

1. Introduction

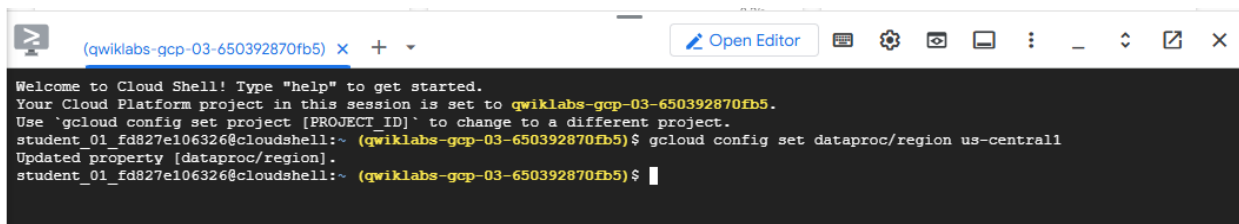
This lab demonstrates how to use **Google Cloud Dataproc** to:

- Create and manage **Apache Spark/Hadoop clusters**
- Submit and monitor **Spark jobs**
- Dynamically **scale worker nodes**

Task 1. Create a cluster

1. In Cloud Shell, run the following command to set the Region:

`gcloud config set dataproc/region Region`



```
(qwiklabs-gcp-03-650392870fb5) x + v Open Editor
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to qwiklabs-gcp-03-650392870fb5.
Use 'gcloud config set project [PROJECT_ID]' to change to a different project.
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5) $ gcloud config set dataproc/region us-central1
Updated property [dataproc/region].
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5) $
```

2. Dataproc creates staging and temp buckets that are shared among clusters in the same region. Since we're not specifying an account for Dataproc to use, it will use the Compute Engine default service account, which doesn't have storage bucket permissions by default. Let's add those.

- First, run the following commands to grab the PROJECT_ID and PROJECT_NUMBER:

PROJECT_ID=\$(gcloud config get-value project) && gcloud config set project \$PROJECT_ID

PROJECT_NUMBER=\$(gcloud projects describe \$PROJECT_ID --format='value(projectNumber)')

```
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5) $ PROJECT_ID=$(gcloud config get-value project) && \
gcloud config set project $PROJECT_ID

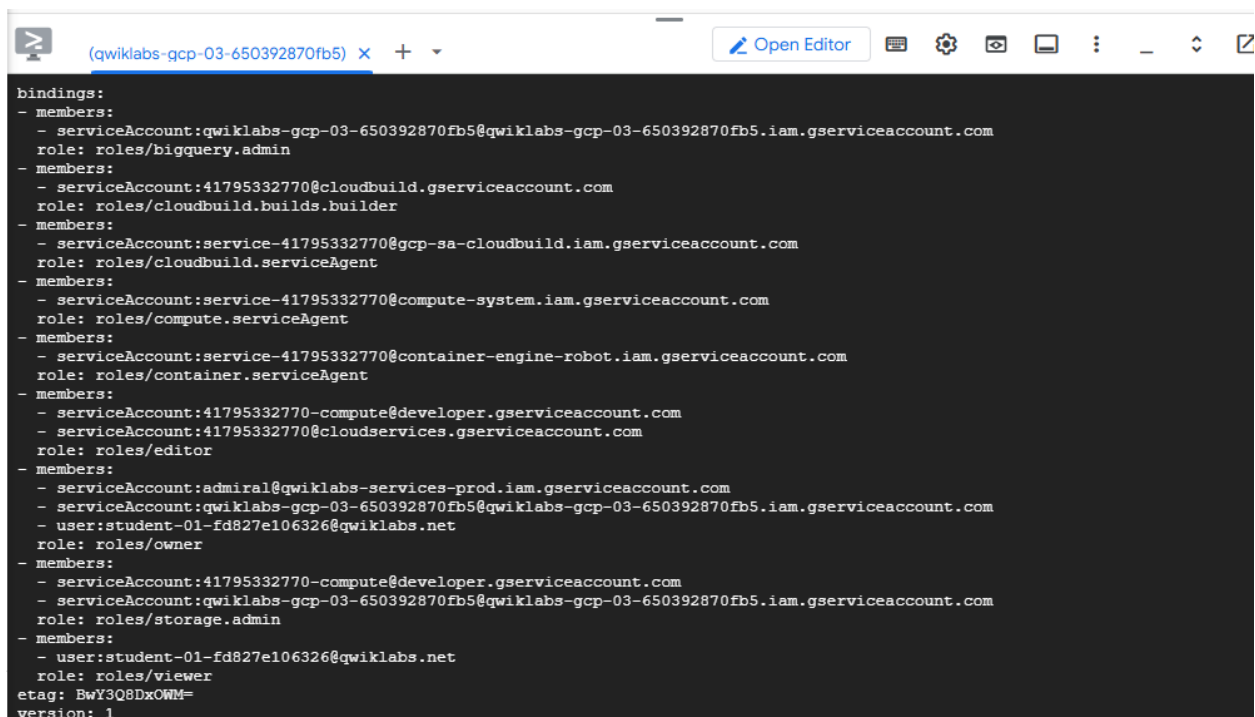
PROJECT_NUMBER=$(gcloud projects describe $PROJECT_ID --format='value(projectNumber)')
Your active configuration is: [cloudshell-18641]
Updated property [core/project].
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5) $
```

- Now run the following command to give the Storage Admin role to the Compute Engine default service account:

gcloud projects add-iam-policy-binding \$PROJECT_ID \

--member=serviceAccount:\$PROJECT_NUMBER-
compute@developer.gserviceaccount.com \

--role=roles/storage.admin



```
(qwiklabs-gcp-03-650392870fb5) x + - Open Editor
bindings:
- members:
  - serviceAccount:qwiklabs-gcp-03-650392870fb5@qwiklabs-gcp-03-650392870fb5.iam.gserviceaccount.com
  role: roles/bigquery.admin
- members:
  - serviceAccount:41795332770@cloudbuild.gserviceaccount.com
  role: roles/cloudbuild.builds.builder
- members:
  - serviceAccount:service-41795332770@gcp-sa-cloudbuild.iam.gserviceaccount.com
  role: roles/cloudbuild.serviceAgent
- members:
  - serviceAccount:service-41795332770@compute-system.iam.gserviceaccount.com
  role: roles/compute.serviceAgent
- members:
  - serviceAccount:service-41795332770@container-engine-robot.iam.gserviceaccount.com
  role: roles/container.serviceAgent
- members:
  - serviceAccount:41795332770-compute@developer.gserviceaccount.com
  - serviceAccount:41795332770@cloudservices.gserviceaccount.com
  role: roles/editor
- members:
  - serviceAccount:admiral@qwiklabs-services-prod.iam.gserviceaccount.com
  - serviceAccount:qwiklabs-gcp-03-650392870fb5@qwiklabs-gcp-03-650392870fb5.iam.gserviceaccount.com
  - user:student-01-fd827e106326@qwiklabs.net
  role: roles/owner
- members:
  - serviceAccount:41795332770-compute@developer.gserviceaccount.com
  - serviceAccount:qwiklabs-gcp-03-650392870fb5@qwiklabs-gcp-03-650392870fb5.iam.gserviceaccount.com
  role: roles/storage.admin
- members:
  - user:student-01-fd827e106326@qwiklabs.net
  role: roles/viewer
etag: BwY3Q8DxOWM=
version: 1
```

3. Enable Private Google Access on your subnet by running the following command:

gcloud compute networks subnets update default --region=REGION --enable-private-ip-google-access

```
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5)$ gcloud compute networks subnets update default --region=us-central1 --enable-private-ip-google-access
Updated [https://www.googleapis.com/compute/v1/projects/qwiklabs-gcp-03-650392870fb5/regions/us-central1/subnetworks/default].
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5)$
```

4. Run the following command to create a cluster called example-cluster with e2-standard-4 VMs and default Cloud Dataproc settings:

gcloud dataproc clusters create example-cluster --worker-boot-disk-size 500 --worker-machine-type=e2-standard-4 --master-machine-type=e2-standard-4

```
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5)$ gcloud dataproc clusters create example-cluster --worker-boot-disk-size 500 --worker-machine-type=e2-standard-4 --master-machine-type=e2-standard-4
Waiting on operation [projects/qwiklabs-gcp-03-650392870fb5/regions/us-central1/operations/fb6becc1-d9bd-3b96-bdc4-8ab4c9ad4533].
Waiting for cluster creation operation...
WARNING: Failed to validate permissions required for default service account: '41795332770-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permissions have been granted to the respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to specifying a cross-project VM service account in the cluster creation request and not configuring it properly. To configure a cross-project VM service account follow the instructions at https://cloud.google.com/iam/docs/impersonating-service-accounts#attaching-different-project as not configuring properly could result in cluster creation failures during later stages.
WARNING: 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.
WARNING: The firewall rules for specified network or subnetwork would allow ingress traffic from 0.0.0.0/0, which could be a security risk.
WARNING: No image specified. Using the default image version. It is recommended to select a specific image version in production, as the default image version may change at any time.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/qwiklabs-gcp-03-650392870fb5/regions/us-central1/clusters/example-cluster] Cluster placed in zone [us-central1-c].
```

5. If asked to confirm a zone for your cluster. Enter **Y**.

Task 2. Submit a job

- Run this command to submit a sample Spark job that calculates a rough value for pi:

gcloud dataproc jobs submit spark --cluster example-cluster \

--class org.apache.spark.examples.SparkPi \

--jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000

```
student 01 fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5)$ gcloud dataproc jobs submit spark --cluster example-cluster \
--class org.apache.spark.examples.SparkPi \
--jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000
Job [78bd8f974427407bbf447cd2856a4b17] submitted.
Waiting for job output...
25/06/11 03:57:40 INFO SparkEnv: Registering MapOutputTracker
25/06/11 03:57:40 INFO SparkEnv: Registering BlockManagerMaster
25/06/11 03:57:40 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/06/11 03:57:40 INFO SparkEnv: Registering OutputCommitCoordinator
25/06/11 03:57:41 INFO MetricsConfig: Loaded properties from hadoop-metrics2.properties
25/06/11 03:57:41 INFO MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
25/06/11 03:57:41 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
25/06/11 03:57:41 INFO DataprocSparkPlugin: Registered 188 driver metrics
25/06/11 03:57:42 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at example-cluster-m.us-central1-c.c.qwikl
abs-gcp-03-650392870fb5.internal./10.128.0.2:8032
25/06/11 03:57:42 INFO AHSPProxy: Connecting to Application History server at example-cluster-m.us-central1-c.c.qwiklabs-gcp-03-6503
92870fb5.internal./10.128.0.2:10200
25/06/11 03:57:43 INFO Configuration: resource-types.xml not found
25/06/11 03:57:43 INFO ResourceUtils: Unable to find 'resource-types.xml'.
25/06/11 03:57:45 INFO YarnClientImpl: Submitted application application_1749614172631_0001
25/06/11 03:57:46 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at example-cluster-m.us-central1-c.c.qwikl
abs-gcp-03-650392870fb5.internal./10.128.0.2:8030
25/06/11 03:57:48 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists
with desired state.
25/06/11 03:57:49 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *
not* yet see flushed data for gs://dataproc-temp-us-central1-41795332770-xelwakkh/e8ac81b8-358a-4283-bf15-f6faca72e365/spark-job-hi
story/application_1749614172631_0001.inprogress [CONTEXT ratelimit_period="1 MINUTES" ]
```

The command specifies:

- That you want to run a **spark** job on the example-cluster cluster
- The class containing the main method for the job's pi-calculating application
- The location of the jar file containing your job's code
- The parameters you want to pass to the job—in this case, the number of tasks, which is 1000

Task 3. Update a cluster

1. To change the number of workers in the cluster to four, run the following command:

```
gcloud dataproc clusters update example-cluster --num-workers 4
```

```
(qwiklabs-gcp-03-650392870fb5) x + ▾ Open Editor
driverControlFilesUri: gs://dataproc-staging-us-central1-41795332770-tqxntljn/google-cloud-dataproc-metainfo/e8ac81b8-358a-4283-bf15-f6faca72e365/jobs/78bd8f974427407bbf447cd2856a4b17/
driverOutputResourceUri: gs://dataproc-staging-us-central1-41795332770-tqxntljn/google-cloud-dataproc-metainfo/e8ac81b8-358a-4283-bf15-f6faca72e365/jobs/78bd8f974427407bbf447cd2856a4b17/driveroutput
jobUuid: 5f75f7f3-274f-3d5a-bf79-0c4dbda8de6d
placement:
  clusterName: example-cluster
  clusterUuid: e8ac81b8-358a-4283-bf15-f6faca72e365
reference:
  jobId: 78bd8f974427407bbf447cd2856a4b17
  projectId: qwiklabs-gcp-03-650392870fb5
sparkJob:
  args:
    - '1000'
  jarFileUri:
    - file:///usr/lib/spark/examples/jars/spark-examples.jar
  mainClass: org.apache.spark.examples.SparkPi
status:
  state: DONE
  stateStartTime: '2025-06-11T03:58:14.910704Z'
statusHistory:
- state: PENDING
  stateStartTime: '2025-06-11T03:57:33.148535Z'
- state: SETUP_DONE
  stateStartTime: '2025-06-11T03:57:33.221298Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2025-06-11T03:57:33.616284Z'
yarnApplications:
- name: Spark Pi
  progress: 1.0
  state: FINISHED
trackingUrl: http://example-cluster-m.us-central1-c.c.qwiklabs-gcp-03-650392870fb5.internal.:8088/proxy/application_17496141726310001/
```

2. You can use the same command to decrease the number of worker nodes:

gcloud dataproc clusters update example-cluster --num-workers 2

```
student_01_fd827e106326@cloudshell:~ (qwiklabs-gcp-03-650392870fb5) $ gcloud dataproc clusters update example-cluster --num-workers 2
Waiting on operation [projects/qwiklabs-gcp-03-650392870fb5/regions/us-central1/operations/d54160b4-88a6-3ee6-958a-2e02563d87c9].
Waiting for cluster update operation...working..
```

5. Key Learnings

- ✓ **Serverless Spark:** Dataproc automates cluster management (no manual Hadoop setup).
- ✓ **Dynamic Scaling:** Workers can be added/removed in seconds.
- ✓ **Job Submission:** Spark jobs run via simple gcloud commands.
- ✓ **Cost Efficiency:** Pay only for resources used during job execution.

6. Conclusion

This lab successfully demonstrated:

- **Cluster creation** with custom machine types.
- **Spark job execution** (π estimation).
- **On-demand scaling** of worker nodes.

Future Work:

- Process larger datasets (e.g., log files).
- Integrate with **BigQuery** for analytics.