

# Automate Data Capture at Scale with Document AI Challenge Lab

GSP103



- Dataproc offers fully managed Apache Spark and Hadoop clusters that deploy in seconds and scale dynamically for cost efficiency.
- This lab demonstrates creating Dataproc clusters directly through the Google Cloud Console interface.
- You'll execute a simple Apache Spark job within the provisioned cluster environment.
- Hands-on experience includes monitoring job execution and cluster performance metrics.
- Finally, resize the cluster by adjusting worker node count to match workload demands.

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

Dashboard Catalog Paths Collections

End Lab 00:43:50 Time limit

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.

Open Google Cloud console

Username: student-01-dc6f0136a6db Password: QNhbk1VlyLvq Project ID: quicklabs-gcp-03-32450d4a

Student Resources: Dataproc: Qwik Start - Qwiklabs

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:

1. Click **Navigation menu > APIs & Services > Library**:
2. Type **Cloud Dataproc** in the **Search for APIs & Services** dialog. The console will display the Cloud Dataproc API in the search results.
3. Click on **Cloud Dataproc API** to display the status of the API. If the API is not already enabled, click the **Enable** button.

Once the API is enabled, proceed with the lab instructions.

Permission to Service Account

To assign storage permission to the service account, which is required for creating a cluster:

1. Go to **Navigation menu > IAM & Admin > IAM**.

Lab instructions and tasks 0/100

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Next >

< Previous

3 18°C Mostly sunny

Search

ENG IN

08:30 PM 27-12-2025

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

End Lab 00:43:38 Time limit

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.

Open Google Cloud console

Username: student-01-dc6f0136a6db (copy)

Password: QNhbk1VlyLvq (copy)

Project ID: quicklabs-gcp-03-32450d4a (copy)

Student Resources: Dataproc: Qwik Start - Qwiklabs

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

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

Once the API is enabled, proceed with the lab instructions.

## Permission to Service Account

To assign storage permission to the service account, which is required for creating a cluster:

1. Go to **Navigation menu > IAM & Admin > IAM**.
2. Click the pencil icon on the `compute@developer.gserviceaccount.com` service account.
3. Click on the **+ ADD ANOTHER ROLE** button. Select role **Storage Admin**.

Once you've selected the **Storage Admin** role, click on **Save**.

0/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next



API API Library – APIs & Services - q

https://console.cloud.google.com/apis/library?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05

API APIs & Services / API Library

← API Library

## Welcome to the API Library

The API Library has documentation, links, and a smart search experience.

Cloud Dataproc

- cloud dataproc api
- cloud dataproc control api

Maps

View all (33)

Filter Type to filter

Visibility

- Public (504)
- Private (3)

Category

- Analytics (11)

Maps SDK for Android Google Maps for your native Android app.

Maps SDK for iOS Google Maps for your native iOS app.

Maps JavaScript API Google Maps for your website

Places API Google Enterprise API Get detailed information about 100 million places

18°C 3 Mostly sunny

Search

ENG IN

08:31 PM 27-12-2025

API APIs & Services – qwiklabs-gcp-03-32450d4add05 +

https://console.cloud.google.com/apis/library/browse?project=qwiklabs-gcp-03-32450d4add05&q=cloud%20dataproc%20api

Free AI Paraphrasing... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05

API APIs & Services / API Library / Browse

API Library API Library

cloud dataproc api

3 results

Visibility Filter Type to filter

Public (3)

Category Filter Type to filter

Big data (1)

Healthcare & Life Sciences (1)

Google Enterprise APIs (2)

Other (1)

Cloud Dataproc Control API Google Manages internal resources for Google Cloud Dataproc.

Cloud Dataproc API Google Enterprise API ⓘ Manages Hadoop-based clusters and jobs on Google Cloud Platform.

Cloud Healthcare API Google Enterprise API ⓘ The Cloud Healthcare API bridges the gap between care systems and applications built on Google Cloud. By supporting standards-driven data formats and protocols of existing healthcare technologies, the Cloud Healthcare API connects your data to advanced Google Cloud capabilities, including data processing with Cloud Dataproc, scalable analytics with BigQuery, and machine learning wit...

https://console.cloud.google.com/apis/library/dataproc.googleapis.com?project=qwiklabs-g...

Google Cloud Marketplace Terms of Service ⓘ

18°C 3 Mostly sunny

Search

ENG IN

08:32 PM 27-12-2025



## Cloud Dataproc API

[Google Enterprise API](#)

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

[Manage](#)[Try this API](#)

API Enabled

[Overview](#)[Documentation](#)[Related Products](#)

### Overview

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

### Additional details

Type: [SaaS & APIs](#)

Last product update: 7/22/22

Category: [Google Enterprise APIs](#)

Service name: dataproc.googleapis.com

### Tutorials and documentation

IAM – IAM & Admin – qwiklabs-gcp-03-32450d4add05

https://console.cloud.google.com/iam-admin/iam?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud Project: qwiklabs-gcp-03-32450d4add05

Search (/) for resources, docs, products, and more

Search

IAM & Admin / IAM

IAM

Allow Deny Recommendations history

Permissions for project "qwiklabs-gcp-03-32450d4add05"

These permissions affect this project and all of its resources. [Learn more](#)

Include Google-provided role grants

View by principals View by roles

Grant access Remove access

Filter: compute@developer.gserviceaccount.com

Enter property name or value

Type	Principal ↑	Name	Role	Security insights
	1011811515964-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	<a href="#">Edit principal</a>

18°C 18°C Mostly sunny

Search

ENG IN 08:33 PM 27-12-2025

IAM – IAM & Admin – qwiklabs-gcp-03-32450d4add05 +

https://console.cloud.google.com/iam-admin/iam?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud Project: qwiklabs-gcp-03-32450d4add05 Search (/) for resources

IAM & Admin / IAM

**IAM**

Allow Deny Recommendations history

### Permissions for project "qwiklabs-gcp-03-32450d4add05"

These permissions affect this project and all of its resources. [Learn more](#)

[View by principals](#) [View by roles](#)

[Grant access](#) [Remove access](#)

Filter: compute@developer.gserviceaccount.com

Type	Principal
<input type="checkbox"/>	compute@developer.gserviceaccount.com
<input type="checkbox"/>	1011811515964-compute@developer.gserviceaccount.com

Edit access to "qwiklabs-gcp-03-32450d4add05"

Principal: 1011811515964-compute@developer.gserviceaccount.com Project: qwiklabs-gcp-03-32450d4add05

### Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role: Editor IAM condition (optional)

+ Add IAM condition

View, create, update, and delete most Google Cloud resources. See the list of included permissions.

+ Add another role

Help me choose roles

Save Test changes Cancel

18°C Mostly sunny

Search

08:33 PM 27-12-2025 ENG IN

IAM – IAM & Admin – qwiklabs-gcp-03-32450d4add05

https://console.cloud.google.com/iam-admin/iam?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud Project: qwiklabs-gcp-03-32450d4add05 Search (/) for resources

IAM & Admin / IAM

**IAM**

Allow Deny Recommendations history

### Permissions for project "qwiklabs-gcp-03-32450d4add05"

These permissions affect this project and all of its resources. [Learn more](#)

View by principals View by roles

Grant access Remove access

Filter: compute@developer.gserviceaccount.com

Type Principal ↑

1011811515964-compute@developer.gserviceaccount.com

Principal: 1011811515964-compute@developer.gserviceaccount.com

Project: qwiklabs-gcp-03-32450d4add05

### Edit access to "qwiklabs-gcp-03-32450d4add05"

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role: Editor IAM condition (optional) [+ Add IAM condition](#)

View, create, update, and delete most Google Cloud resources. See the list of included permissions.

Role: Storage Admin IAM condition (optional) [+ Add IAM condition](#)

Grants full control of buckets and objects.

+ Add another role

Help me choose roles

Save Test changes Cancel

Summary of changes

Roles removed n/a

Role added Storage Admin

**Test changes**

18°C 3 Mostly sunny ENG IN 08:34 PM 27-12-2025

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

Dashboard Catalog Paths Collections

End Lab 00:38:14 Time limit

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.

Open Google Cloud console

Username: student-01-dc6f0136a6db Password: QNhbk1VlyLvq Project ID: quicklabs-gcp-03-32450d4a

Student Resources: Dataproc: Qwik Start - Qwiklabs

Task 1. Create a cluster

1. In the Cloud Platform Console, select Navigation menu > View all products > Dataproc > Clusters, then click Create cluster.

2. Click Create for Cluster on Compute Engine.

3. Set the following fields for your cluster and accept the default values for all other fields:

Note: In the Configure nodes section ensure both the Master node and Worker nodes are set to the correct Machine Series and Machine Type. If the E2 series is not displayed, verify that you have selected "Standard Persistent Disk" as the Primary Disk type option.

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

0/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

Primary DISK type option.

Field	Value
Name	example-cluster
Region	us-central1
Zone	us-central1-c
Primary disk type (Manager Node)	Standard Persistent Disk
Machine Series (Manager Node)	E2
Machine Type (Manager Node)	e2-standard-2
Primary disk size (Manager Nodes)	30 GB
Number of Worker Nodes	2
Primary disk type (Worker Node)	Standard Persistent Disk
Machine Series (Worker Nodes)	E2
Machine Type (Worker Nodes)	e2-standard-2

End Lab 00:38:02 Time limit

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.](#)

Open Google Cloud console

Username: student-01-dc6f0136a6db6e33 Password: QNhbk1VlyLvq Project ID: qwiklabs-gcp-03-32450d4a

Student Resources: Dataproc: Qwik Start - Qwiklabs

0/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Prepared Data for ML APIs on Google Cloud > Dataproc: Qwik Start - Console

Contents

Dataproc: Qwik Start - Console

End Lab 00:37:53 Time limit

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.](#)

Open Google Cloud console

Username: student-01-dc6f0136a6db (copy)

Password: QNhbk1VlyLvq (copy)

Project ID: quicklabs-gcp-03-32450d4a (copy)

Node) Standard Persistent Disk

Machine Series (Manager Node) E2

Machine Type (Manager Node) e2-standard-2

Primary disk size (Manager Nodes) 30 GB

Number of Worker Nodes 2

Primary disk type (Worker Node) Standard Persistent Disk

Machine Series (Worker Nodes) E2

Machine Type (Worker Nodes) e2-standard-2

Primary disk size (Worker Nodes) 30 GB

Internal IP only Deselect "Configure all instances to have only internal IP addresses"

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-central1 or europe-west1, to isolate resources (including

0/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next

Dataproc – qwiklabs-gcp-03-324 x +

https://console.cloud.google.com/dataproc/overview?referrer=search&project=qwiklabs-gcp-03-32450d4add05 | 1

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc X Search

Dataproc / Overview

Overview

## Dataproc

Dataproc is a fully managed and highly scalable service for running Apache Spark, Apache Hadoop, Apache Flink, Presto, and 30+ open source tools and frameworks. Use Dataproc for data lake modernization, ETL, and secure data science, at scale, integrated with Google Cloud, at a fraction of the cost. [Learn more about Dataproc](#)

## Serverless for Apache Spark

Google Serverless for Apache Spark (formerly Dataproc Serverless) enables you to run Spark workloads without provisioning and managing your own clusters. Only submit your code, and (optionally) Spark parameters. The service will run the workload on a managed Google Cloud infrastructure, autoscaling as needed. Use Serverless for ETL, interactive data science at scale, without infrastructure management, while paying only for the job's runtime. [Learn more about serverless Spark](#)

+ Grant access



18°C Mostly sunny

Search

ENG IN

08:37 PM 27-12-2025

Clusters – Dataproc – qwiklabs-gcp-03-32450d4add05 +

https://console.cloud.google.com/dataproc/clusters?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc X Search

Dataproc / Clusters

Clusters Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Clusters Create cluster Refresh Start Stop Delete + 5 recommended alerts Regions Learn

Cloud Dataproc

Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

Create cluster

18°C 3 Mostly sunny Search ENG IN 08:37 PM 27-12-2025

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore 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.

Name Cluster Name \* example-cluster

Location Region \* us-central1 Zone \* us-central1-c

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

Create Cancel

Equivalent command line

Equivalent REST

3 18°C Mostly sunny

Search

ENG IN

08:39 PM 27-12-2025

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore 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.

Manager node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers.

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

Create Cancel

Equivalent command line

Equivalent REST

18°C 3 Mostly sunny

Search

ENG IN 08:42 PM 27-12-2025

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore 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.

Primary disk size \* 30 GB Primary disk type \* Standard Persistent Disk

Number of local SSDs x 375GB Local SSD Interface

### Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

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)

Image	vCPU	Memory
	2	8 GB

18°C Mostly sunny 08:42 PM 27-12-2025 ENG IN

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore 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.

Number of local SSDs x 375GB Local SSD Interface

### Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

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	Memory
	2	8 GB

▼ CPU platform and GPU

Number of worker nodes \* 2

Equivalent command line

Equivalent REST

Create Cancel

18°C 18°C Mostly sunny

Search

ENG IN 08:42 PM 27-12-2025

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

← Create a Dataproc cluster on Compute Engine

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

Number of worker nodes \* 2

Primary disk size \* 30 GB Primary disk type \* Standard Persistent Disk

Number of local SSDs x 375GB Local SSD Interface

Secondary worker nodes

Create Cancel

Equivalent command line

Equivalent REST

18°C 3 Mostly sunny

Search

ENG IN 08:42 PM 27-12-2025

Create a Dataproc cluster on Compute Engine

https://console.cloud.google.com/dataproc/clustersAdd?computeInfra=gce&project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Dataproc Search

Dataproc / Clusters / Configure cluster

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore 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.

**Internal IP only**  
 Configure all instances to have only internal IP addresses. [Learn more](#)

**Labels**  
A list of key:value pairs to attach to the cluster for tracking.  
[+ Add labels](#)

**Cluster properties**  
Use cluster properties to add or modify configuration files when creating a cluster.  
[+ Add properties](#)

**Initialization actions**  
Use initialization actions to customize settings, install applications, or make other modifications to your cluster. Select scripts or executables that Cloud Dataproc will run when provisioning your cluster.  
[+ Add initialization action](#)

Create Cancel

Equivalent command line  
Equivalent REST

18°C Mostly sunny

Search

ENG IN 08:42 PM 27-12-2025

 DataProc / Clusters

Clusters		<a href="#">+ Create cluster</a>	<a href="#">Refresh</a>	<a href="#">Start</a>	<a href="#">Stop</a>	<a href="#">Delete</a>	Regions	<a href="#">+ 5 recom.</a>
<input type="text"/> Filter Search cluster by properties, press Enter <a href="#">?</a> <a href="#">☰</a>								
<input type="checkbox"/>	Name	Status	Region	Zone	Base image version	Total worker nodes	File	...
<input type="checkbox"/>	<a href="#">example-cluster</a>	Provisioning	us-central1	us-central1-c	2.2.71-debian12	2	N	<a href="#">...</a>

No clusters selected

## Permissions Labels

**i** Please select at least one resource.

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

End Lab 00:08:27 Time limit

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.

Open Google Cloud console

Username: student-01-dc6f0136a6db (copy)

Password: QNhbk1VlyLvq (copy)

Project ID: quicklabs-gcp-03-32450d4a (copy)

Student Resources: Dataproc: Qwik Start - Qwiklabs

Task 2. Submit a job

To run a sample Spark job:

1. Click **Jobs** in the left pane to switch to Dataproc's jobs view, then click **Submit job**.
2. Set the following fields to update Job. Accept the default values for all other fields:

Field	Value
Region	us-central1
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.)

50/100 Lab instructions and task

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next



Submit a job – Dataproc – qwikla... X +

https://console.cloud.google.com/dataproc/jobs/jobsSubmit?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (/) for resources, docs, products, and more Search

Dataproc / Jobs / Add job

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Submit a job

Job ID \* job-59e361f7

Region \* us-central1

Specifies the Cloud Dataproc regional service, which determines what clusters are available.

Cluster \* example-cluster

Job type \* Spark

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

The fully qualified name of a class in a provided or standard jar file, for example, com.example.wordcount, or a provided jar file to use the main class of that jar file

Jar files file:///usr/lib/spark/examples/jars/spark-examples.jar Enter file path

Jar files are included in the CLASSPATH. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Files

Files are included in the working directory of each executor. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

18°C Mostly sunny

Search

ENG IN 09:07 PM 27-12-2025

Submit a job – Dataproc – qwikl... X +

https://console.cloud.google.com/dataproc/jobs/jobsSubmit?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (/) for resources, docs, products, and more Search

Dataproc / Jobs / Add job

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Submit a job

org.apache.spark.examples.SparkPi

The fully qualified name of a class in a provided or standard jar file, for example, com.example.wordcount, or a provided jar file to use the main class of that jar file

Jar files file:///usr/lib/spark/examples/jars/spark-examples.jar Enter file path

Jar files are included in the CLASSPATH. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Files

Files are included in the working directory of each executor. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Archive files

Archive files are extracted in the Spark working directory. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix. Supported file types: .jar, .tar, .tar.gz, .tgz, .zip.

Arguments 1000 Press <Return> to add more arguments

Additional arguments to pass to the main class. Press Return after each argument.

Max restarts per hour

Leave blank if you don't want to allow automatic restarts on job failure. [Learn more](#)

Spark performance enhancements

3 18°C Mostly sunny

Search

ENG IN 09:07 PM 27-12-2025

job-59e361f7 - Summary - Job d X +

https://console.cloud.google.com/dataproc/jobs/job-59e361f7/summary?job=job-59e361f7&region=us-central1&project=qwiklabs-gcp-03-3245... | 1

Free AI Paraphrasing... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (/) for resources, docs, products, and more Search

Dataproc / Jobs / Job: job-59e361f7 / Summary

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

← Job details Clone Delete Stop Refresh

Job ID: job-59e361f7  
Job UUID: 7737f41a-5396-4a17-bb14-268d41194954  
Type: Dataproc Job  
Status: Running

Summary Monitoring Configuration

↳ Insights by Gemini Preview

Output Line wrap: Off

Spark jobs take ~60 seconds to initialize resources. Dismiss

Press Alt+F1 for Accessibility Options.

Equivalent command line

3 18°C Mostly sunny

Search

ENG IN 09:08 PM 27-12-2025

Google Skills Partner

Contents Dataproc: Qwik Start - Console

End Lab 00:04:44 Time limit

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.](#)

[Open Google Cloud console](#)

Username: student-01-dc6f0136a6db Password: QNhbk1VlyLvq Project ID: qwiklabs-gcp-03-32450d4a

Student Resources: [Dataproc: Qwik Start - Qwiklabs](#)

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:

Your job has successfully calculated a rough value for pi!

80/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

Previous Next

job-59e361f7 - Summary - Job d X +

https://console.cloud.google.com/dataproc/jobs/job-59e361f7/summary?job=job-59e361f7&region=us-central1&project=qwiklabs-gcp-03-3245... | 1

Free AI Paraphrasing... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (/) for resources, docs, products, and more Search

Dataproc / Jobs / Job: job-59e361f7 / Summary

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

← Job details Clone Delete Stop Refresh

Job ID: job-59e361f7  
Job UUID: 7737f41a-5396-4a17-bb14-268d41194954  
Type: Dataproc Job  
Status: Succeeded

Summary Monitoring Configuration

Output Line wrap: Off

Spark jobs take ~60 seconds to initialize resources. Dismiss

25/12/27 15:38:15 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-us-central1-1011811515964-1zvlofdz/o?delimiter=/&fields=item] 25/12/27 15:38:15 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state. 25/12/27 15:38:16 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-centr 25/12/27 15:38:29 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {action\_http\_delete\_request=1, action\_http\_delete\_request\_duration=82, action\_http\_delete\_request\_max=82, ac [CONTEXT ratelimit\_period="5 MINUTES" ] pi is roughly 3.141914991491498 25/12/27 15:38:52 INFO GhfsGlobalStorageStatistics: Detected potential high latency for operation op\_hflush. latencyMs=638; previousMaxLatencyMs=413; operationCount=15; context=gs://datapr 25/12/27 15:38:52 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action\_http\_patch\_request=0, files\_created=1, gcs\_api\_server\_timeout\_count=0, op\_get\_list\_status\_result\_si

Output is complete

Equivalent command line

3 18°C Mostly sunny

Search

ENG IN 09:09 PM 27-12-2025

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Contents Dataproc: Qwik Start - Console

Dashboard Catalog Paths Collections

End Lab 00:04:30 Time limit

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.

Open Google Cloud console

Username: student-01-dc6f0136a6db Password: QNhbk1VlyLvq Project ID: quicklabs-gcp-03-32450d4a

Student Resources: Dataproc: Qwik Start - Qwiklabs

80/100 Lab instructions and tasks

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

## 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**.

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

Test completed task

Previous Next

example-cluster – Configuration +

https://console.cloud.google.com/dataproc/clusters/example-cluster/configuration?region=us-central1&project=qwiklabs-gcp-03-32450d4add05 | 1

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search ( / ) for resources, docs, products, and more Search

Dataproc / Clusters / Cluster: example-cluster / Cluster configuration

Overview ← Cluster details + Submit Job Refresh Start Stop Delete View Logs

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone> More

Name example-cluster  
Cluster UUID c45bf331-1d5a-4358-b259-31775baf1854  
Type Dataproc Cluster  
Status Running

Summary Monitoring Jobs VM Instances Configuration Web Interfaces

Edit

Region us-central1  
Zone us-central1-c  
Image version 2.2.71-debian12  
Autoscaling Off  
Performance Enhancements  
Advanced optimizations Off  
Advanced execution layer Off  
Google Cloud Storage caching Off  
Dataproc Metastore None

https://console.cloud.google.com/dataproc/clusters/example-cluster/configuration?region=us-central1&project=qwiklabs-gcp-03-32450d4add05

3 18°C Mostly sunny Search Calendar Video Call File Microsoft Edge Mozilla Firefox Python ENG IN 09:10 PM 27-12-2025

example-cluster – Configuration +

https://console.cloud.google.com/dataproc/clusters/example-cluster/configuration?region=us-central1&project=qwiklabs-gcp-03-32450d4add05 | 1

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (/) for resources, docs, products, and more

Dataproc / Clusters / Cluster: example-cluster / Cluster configuration

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Cluster details ← Cluster details + Submit Job Refresh Start Stop Delete

Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dat

Name	example-cluster
Cluster UUID	c45bf331-1d5a-4358-b259-31775baf1854
Type	Dataproc Cluster
Status	Running

Summary Monitoring Jobs VM Instances Configuration Web Interfaces

Edit

Region	us-central1
Zone	us-central1-c
Image version	2.2.71-debian12
Autoscaling	Off

Performance Enhancements

Advanced optimizations	Off
Advanced execution layer	Off
Google Cloud Storage caching	Off

Dataproc Metastore

Scheduled deletion	None
--------------------	------

Editing cluster

Worker nodes \* 4

Secondary worker nodes \* 0

Labels

Key 1	Value 1
goog-dataproc-cluster-name	example-cluster
Key 2	Value 2
goog-dataproc-cluster-uuid	c45bf331-1d5a-4358-b259-31775baf1854
Key 3	Value 3
goog-dataproc-location	us-central1
Key 4	Value 4
goog-dataproc-drz-resource	cluster-c45bf331-1d5a-4358-b259-31775baf1854

+ Add label

Use graceful decommissioning ?

Save Cancel Equivalent REST

18°C 3 Mostly sunny

Search

ENG IN 09:10 PM 27-12-2025

 DataProc / Clusters

Clusters		+ Create cluster	Refresh	Start	Stop	Delete	Regions	+ 5 recent
<input type="text"/> Filter Search cluster by properties, press Enter <span>?</span> <span>☰</span>								
<span>!</span> Sorry, the server was not able to fulfill your request.								
<input type="checkbox"/>	Name <span>↑</span>	Status	Region	Zone	Base image version	Total worker nodes	Flexi	
<input type="checkbox"/>	<a href="#">example-cluster</a>	<span>Updating</span>	us-central1	us-central1-c	2.2.71-debian12	4	No	

No clusters selected

## Permissions Labels

**i** Please select at least one resource.

Prepare Data for ML APIs on Google Cloud

https://partner.skills.google/course\_templates/631/labs/594531

Free AI Paraphrasin... Transcript of Case St... GCP-LAB Hydra datawarehouse Products

Google Skills Partner

Assessment Completed!

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	us-central1
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**.

Check complete. Points earned: 20. Message: Assessment Completed!

Dashboard Catalog Paths Collections Student Resources

Open Google Cloud console

Username: student-01-dc6f0136a6db Password: QNhbk1VlyLvq Project ID: quicklabs-gcp-03-32450d4a

Next >

100/100

GSP103 Overview Setup and requirements Task 1. Create a cluster Task 2. Submit a job Task 3. View the job output Task 4. Update a cluster to modify the number of workers Task 5. Test your understanding Congratulations!

example-cluster – Configuration +

https://console.cloud.google.com/dataproc/jobs/jobsSubmit?project=qwiklabs-gcp-03-32450d4add05

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search (S) Search

Dataproc / Jobs / Add job

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Submit a job

Job ID \* job-22d7b989

Region \* us-central1

Specifies the Cloud Dataproc regional service, which determines what clusters are available.

Cluster \* example-cluster

Job type \* Spark

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

The fully qualified name of a class in a provided or standard jar file, for example, com.example.wordcount, or a provided jar file to use the main class of that jar file

Jar files file:///usr/lib/spark/examples/jars/spark-examples.jar Enter file path

Jar files are included in the CLASSPATH. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Files

Files are included in the working directory of each executor. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

18°C Mostly sunny

Search

ENG IN 09:13 PM 27-12-2025

job-22d7b989 - Summary - Job +

https://console.cloud.google.com/dataproc/jobs/job-22d7b989/summary?job=job-22d7b989&region=us-central1&project=qwiklabs-gcp-03-32...

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search

Dataproc / Jobs / Job: job-22d7b989 / Summary

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Job details Clone Delete Stop Refresh

Job ID: job-22d7b989  
Job UUID: 1840cb07-731b-4a12-82fd-2569c056bcf4  
Type: Dataproc Job  
Status: Running

Summary Monitoring Configuration

↳ Insights by Gemini Preview

Output Line wrap: Off

Spark jobs take ~60 seconds to initialize resources. Dismiss

Press Alt+F1 for Accessibility Options.

```
25/12/27 15:44:19 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
25/12/27 15:44:20 INFO DataprocSparkPlugin: Registered 188 driver metrics
25/12/27 15:44:21 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at example-cluster-m.us-central1-c.c.qwiklabs-gcp-03-32450d4add05.internal./10.128.0.8:8032
25/12/27 15:44:21 INFO AHSProxy: Connecting to Application History server at example-cluster-m.us-central1-c.c.qwiklabs-gcp-03-32450d4add05.internal./10.128.0.8:10200
25/12/27 15:44:23 INFO Configuration: resource-types.xml not found
25/12/27 15:44:23 INFO ResourceUtils: Unable to find 'resource-types.xml'.
25/12/27 15:44:24 INFO YarnClientImpl: Submitted application application_1766849062921_0002
25/12/27 15:44:25 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at example-cluster-m.us-central1-c.c.qwiklabs-gcp-03-32450d4add05.internal./10.128.0.8:8032
```

Equivalent command line Job job-22d7b989 successfully submitted X

3 18°C Mostly sunny

Search

ENG IN 09:14 PM 27-12-2025

job-22d7b989 - Summary - Job +

https://console.cloud.google.com/dataproc/jobs/job-22d7b989/summary?job=job-22d7b989&region=us-central1&project=qwiklabs-gcp-03-32...

Free AI Paraphrasing... Transcript of Case Study GCP-LAB Hydra datawarehouse Products

Google Cloud qwiklabs-gcp-03-32450d4add05 Search s

Dataproc / Jobs / Job: job-22d7b989 / Summary

Overview Notebooks/IDE BigQuery Studio Workbench Clusters Clusters Jobs Workflows Autoscaling policies Serverless Batches Interactive Sessions Session Templates Metastore Services Metastore Release Notes

Job details Clone Delete Stop Refresh

Job ID: job-22d7b989  
Job UUID: 1840cb07-731b-4a12-82fd-2569c056bcf4  
Type: Dataproc Job  
Status: Succeeded

Output Line wrap: Off

Spark jobs take ~60 seconds to initialize resources.

25/12/27 15:44:23 INFO ResourceUtils: Unable to find 'resource-types.xml'.  
25/12/27 15:44:24 INFO YarnClientImpl: Submitted application application\_1766849062921\_0002  
25/12/27 15:44:25 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at example-cluster-m.us-central1-c.qwiklabs-gcp-03-32450d4add05.internal./10.128.0.8:8030  
25/12/27 15:44:28 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.  
25/12/27 15:44:29 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-centr  
25/12/27 15:44:47 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {action\_http\_delete\_request=2, action\_http\_delete\_request\_duration=71, action\_http\_delete\_request\_max=36, ac  
[CONTEXT ratelimit\_period="5 MINUTES" ]  
pi is roughly 3.141558871415589  
25/12/27 15:44:59 INFO GhfsGlobalStorageStatistics: Detected potential high latency for operation op\_hflush. latencyMs=672; previousMaxLatencyMs=594; operationCount=15; context=gs://datap  
25/12/27 15:44:59 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action\_http\_patch\_request=0, files\_created=1, gcs\_api\_server\_timeout\_count=0, op\_get\_list\_status\_result\_si

Output is complete

Equivalent command line

3 18°C Mostly sunny

Search

ENG IN

09:15 PM 27-12-2025