**Google Dataproc**

Table of Contents

[The Basics 1](#_Toc524534161)

[List all clusters on a project 1](#_Toc524534162)

[Create a cluster 2](#_Toc524534163)

[Submit a job 2](#_Toc524534164)

[Update a cluster 2](#_Toc524534165)

[Delete a cluster 2](#_Toc524534166)

[Working with PySpark 2](#_Toc524534167)

[Add python PySpark file to Google Storage 2](#_Toc524534168)

[Submit a job 2](#_Toc524534169)

[Workflow Template Basics 3](#_Toc524534170)

[Create a Workflow Template 3](#_Toc524534171)

[Add a Managed Cluster to the Template 3](#_Toc524534172)

[List all Workflow Templates 3](#_Toc524534173)

[Adding Jobs to the Template 3](#_Toc524534174)

[Instantiate the Workflow 4](#_Toc524534175)

[Monitoring and listing Templates 4](#_Toc524534176)

[Canceling a Workflow 4](#_Toc524534177)

[Delete Workflow Template 4](#_Toc524534178)

[Example Workflow with Initialization Step 4](#_Toc524534179)

[Create a Workflow Template 4](#_Toc524534180)

[Add Managed Cluster 4](#_Toc524534181)

[Add a PySpark Job 5](#_Toc524534182)

[Instantiate the Workflow 5](#_Toc524534183)

[References 6](#_Toc524534184)

# The Basics

## **List all clusters on a project**

|  |
| --- |
| gcloud dataproc clusters --project=single-portal-216120 list |

## **Create a cluster**

|  |
| --- |
| gcloud dataproc clusters create example-cluster2 \  --num-workers=2 \  --num-masters=1 \  --zone=us-central1-a \  --master-machine-type n1-standard-1 \  --worker-machine-type n1-standard-1 \  --project=single-portal-216120 |

## **Submit a job**

|  |
| --- |
| gcloud dataproc --project=single-portal-216120 \  jobs submit spark --cluster example-cluster2 \  --class org.apache.spark.examples.SparkPi \  --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000 |

## **Update a cluster**

|  |
| --- |
| gcloud dataproc --project=single-portal-216120 \  clusters update example-cluster2 --num-workers 3 |

## **Delete a cluster**

|  |
| --- |
| gcloud dataproc --project=single-portal-216120 \  clusters delete example-cluster |

# Working with PySpark

## **Add python PySpark file to Google Storage**

|  |
| --- |
| gsutil cp estimate\_pi.py gs://pyspark-dataproc/estimate\_pi.py |

## **Submit a job**

(Make sure your cluster is up and running)

|  |
| --- |
| gcloud dataproc --project=single-portal-216120 \  jobs submit pyspark --cluster estimate-pi \  gs://pyspark-dataproc/estimate\_pi.py |

**estimate\_pi.py**

|  |
| --- |
| import random  from pyspark import SparkContext  NUM\_SAMPLES = 1000  def inside(p):  x, y = random.random(), random.random()  return x\*x + y\*y < 1  sc = SparkContext()  count = sc.parallelize(c=range(0, NUM\_SAMPLES)).filter(inside).count()  print ("Pi is roughly %f" % (4.0 \* count / NUM\_SAMPLES)) |

# Workflow Template Basics

## **Create a Workflow Template**

|  |
| --- |
| gcloud beta dataproc workflow-templates --project=single-portal-216120 \  create wf-template |

## **Add a Managed Cluster to the Template**

|  |
| --- |
| gcloud beta dataproc workflow-templates set-managed-cluster wf-template \  --project=single-portal-216120 \  --num-workers 2 \  --num-masters=1 \  --zone=us-central1-a \  --master-machine-type n1-standard-1 \  --worker-machine-type n1-standard-1 \  --cluster-name estimate-pi |

## **List all Workflow Templates**

|  |
| --- |
| gcloud beta dataproc workflow-templates list --project=single-portal-216120 |

## **Adding Jobs to the Template**

|  |
| --- |
| gcloud beta dataproc workflow-templates add-job pyspark \  --step-id foo \  --project single-portal-216120 \  --workflow-template wf-template gs://pyspark-dataproc/estimate\_pi.py |

## **Instantiate the Workflow**

|  |
| --- |
| gcloud beta dataproc workflow-templates instantiate wf-template \  --project single-portal-216120 |

## **Monitoring and listing Templates**

|  |
| --- |
| gcloud dataproc operations list \  --project single-portal-216120 \  --filter "labels.goog-dataproc-operation-type=WORKFLOW AND status.state=RUNNING" |

## **Canceling a Workflow**

(Couldn’t get this working... maybe because it was such a quick job)

Get the workflow id from above, *Monitoring and listing Templates*

|  |
| --- |
| gcloud beta dataproc operations cancel \  --project single-portal-216120 \  8c9a4831-6419-4f9b-ab01-b78a93a9bf5e |

## **Delete Workflow Template**

|  |
| --- |
| gcloud beta dataproc workflow-templates delete wf-template \  --project single-portal-216120 |

# Example Workflow with Initialization Step

## **Create a Workflow Template**

|  |
| --- |
| gcloud beta dataproc workflow-templates --project=single-portal-216120 \  create wf-template-init |

## **Add Managed Cluster**

|  |
| --- |
| gcloud beta dataproc workflow-templates set-managed-cluster wf-template-init \  --project=single-portal-216120 \  --num-workers 2 \  --num-masters=1 \  --zone=us-central1-a \  --master-machine-type n1-standard-1 \  --worker-machine-type n1-standard-1 \  --initialization-actions gs://pyspark-dataproc/init.sh \  --cluster-name estimate-pi |

**init.sh**

|  |
| --- |
| #!/usr/bin/env bash  gsutil -m cp -r gs://dataproc-initialization-actions/conda/bootstrap-conda.sh .  gsutil -m cp -r gs://dataproc-initialization-actions/conda/install-conda-env.sh .  chmod 755 ./\*conda\*.sh  # Install Miniconda / conda  ./bootstrap-conda.sh  # Update conda root environment with specific packages in pip and conda  CONDA\_PACKAGES='pandas scikit-learn'  PIP\_PACKAGES='tqdm'  CONDA\_PACKAGES=$CONDA\_PACKAGES PIP\_PACKAGES=$PIP\_PACKAGES ./install-conda-env.sh |

## **Add a PySpark Job**

|  |
| --- |
| gcloud beta dataproc workflow-templates add-job pyspark \  --step-id bar \  --project single-portal-216120 \  --workflow-template wf-template-init gs://pyspark-dataproc/estimate\_pi\_tqdm.py |

**estimate\_pi\_tqdm.py**

|  |
| --- |
| import random  from tqdm import tqdm  from pyspark import SparkContext  NUM\_SAMPLES = 1000  def inside(p):  for n in tqdm(range(1000)):  x, y = random.random(), random.random()  return x\*x + y\*y < 1  sc = SparkContext()  count = sc.parallelize(c=range(0, NUM\_SAMPLES)).filter(inside).count()  print ("Pi is roughly %f" % (4.0 \* count / NUM\_SAMPLES)) |

## **Instantiate the Workflow**

|  |
| --- |
| gcloud beta dataproc workflow-templates instantiate wf-template-init \  --project single-portal-216120 |

# References

**Quick start using the console**

<https://cloud.google.com/dataproc/docs/quickstarts/quickstart-console>

**Quick start using gcloud command line**

<https://cloud.google.com/dataproc/docs/quickstarts/quickstart-gcloud>

**PySpark cheat sheet**

<https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_Cheat_Sheet_Python.pdf>

**Workflow Template Overview**

<https://cloud.google.com/dataproc/docs/concepts/workflows/overview>

**Initialization Actions**

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>

**Miniconda on Dataproc**

<https://github.com/GoogleCloudPlatform/dataproc-initialization-actions/tree/master/conda>