Dataproc introduction - Command Line

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 gcloud on the Google Cloud 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.

Create a cluster

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

```
gcloud config set dataproc/region us-central1
```

Run the following command to create a cluster called `example-cluster` with default Cloud Dataproc settings:

```
gcloud dataproc clusters create example-cluster --worker-boot-disk-size 500 --worker-machine-type n1-standard-2
```

If asked to confirm a zone for you cluster. Enter **Y**. Your cluster will build for a couple of minutes.

```
cloud_user_p_f4a944a5@cloudshell:~ (playground-s-11-07b820a4)$ gcloud dataproc clusters create example-cluster --worker-boot-disk-size 500 --worker-machine-type n1-standard-2
Waiting on operation [projects/playground-s-11-07b820a4/regions/us-central1/operations/6e1badd2-ca73-382e-9234-452009d4ddfd].
Waiting for cluster creation operation...
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.
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.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/playground-s-11-07b820a4/regions/us-central1/clusters/example-cluster] Cluster placed in zone [us-central1-a].
```

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
```

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`

The job's running and final output is displayed in the terminal window:

```
22/02/03 19:58:27 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7923f5b3{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [f8d7ed2cd3d648b0883fc78060807c0e] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-853758339733-wyhbzspz/google-cloud-dataproc-metainfo/d9938084-adc9-4f13-ac3b-1cfb82d0456b/jobs/f8d7
ed2cd3d648b0883fc78060807c0e/
driverOutputResourceUri: gs://dataproc-staging-us-central1-853758339733-wyhbzspz/google-cloud-dataproc-metainfo/d9938084-adc9-4f13-ac3b-1cfb82d0456b/jobs/f8
d7ed2cd3d648b0883fc78060807c0e/driveroutput
jobUuid: 403c73de-ef1a-3963-bb38-603ac8e3fc4c
placement:
  clusterName: example-cluster
  clusterUuid: d9938084-adc9-4f13-ac3b-1cfb82d0456b
reference:
  jobId: f8d7ed2cd3d648b0883fc78060807c0e
  projectId: playground-s-11-07b820a4
sparkJob:
  args:
  - '1000'
  jarFileUris:
  - file:///usr/lib/spark/examples/jars/spark-examples.jar
  mainClass: org.apache.spark.examples.SparkPi
status:
  state: DONE
  stateStartTime: '2022-02-03T19:58:30.638082Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-02-03T19:57:45.350237Z'
- state: SETUP_DONE
  stateStartTime: '2022-02-03T19:57:45.377869Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-02-03T19:57:45.770656Z'
yarnApplications:
- name: Spark Pi
  progress: 1.0
  state: FINISHED
  trackingUrl: http://example-cluster-m:8088/proxy/application_1643918216672_0001/
```

Update a cluster

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

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

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

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

Now you can create a Dataproc cluster and adjust the number of workers from the `gcloud` command line on the Google Cloud.