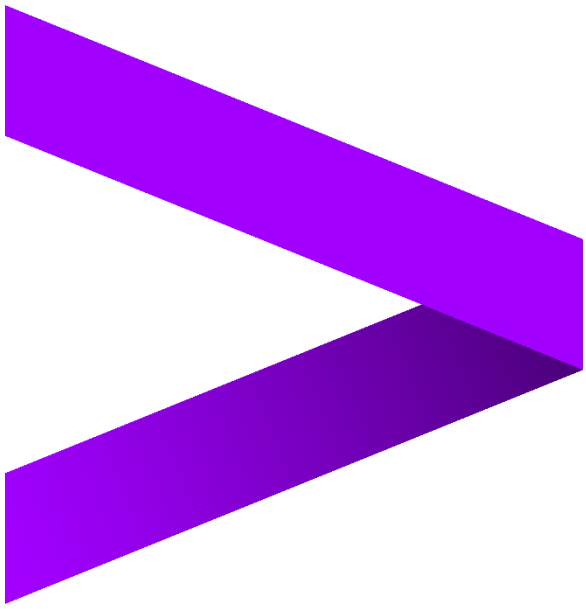
accenture

Learning and Knowledge Management

Dataproc

To run a simple Apache Spark Job



- **Definition:** Apache Spark is a powerful open-source unified analytics engine for large-scale data processing.

- **Purpose**: To create and execute Apache Spark Job with Dataproc

- Follow the link to complete the Lab

  - Click on the link: https://console.cloud.google.com
  - Once opened, provide your login credentials – Accenture id.
  - Follow the below steps to complete you hands-on.

  Steps:

  A. To create Dataproc Cluster

  1. From the navigation menu, select Dataproc and navigate to Clusters
  2. Click on Create Cluster option
  3. Click Create for Compute Engine
  4. Provide the configurations details like
     a. Name = example-cluster
     b. Choose your region and zone
     c. Machine series = E2
     d. Machine type = e2-standard-2
     e. Number of Worker Nodes = 2
     f. Primary Disk Size = 30 GB
     g. Internal IP Only = Deselect "Configure all instances to have only internal IP addresses"
  5. Click Create to create the Cluster

  B. To Submit a job

  1. Click Jobs on the left pane, this will help to switch to Dataproc jobs view, then click on Submit Job
  2. Provide the following details to the respective fields –
     a. Choose your region
     b. Cluster = example-cluster
     c. Job Type = Spark
     d. Main Class or Jar = org.apache.spark.examples.SparkPi

e. Jar files = file:///usr/lib/spark/examples/jars/spark-examples.jar
f. Arguments = 1000 (This will set the number of tasks)

3. Click on Submit

C. To view the Job Output

1. Click the JOB ID in the Jobs list
2. Scroll the output pane to view the output of the calculate value of Pi
3. This job has now calculated the rough value of Pi.

D. To update the cluster to modify the number of workers

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**.
7. The cluster is now modified accordingly