Installation DataProc cluster

Overview

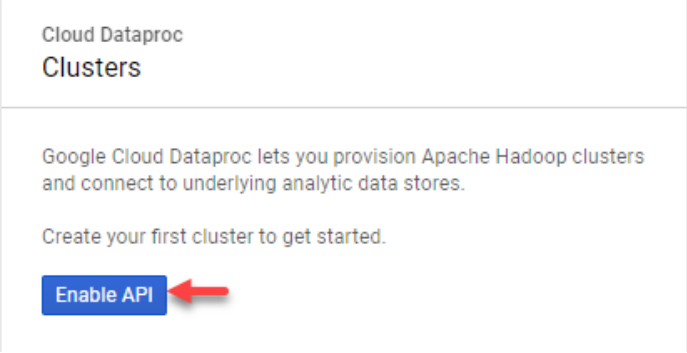
Structured data has a useful organization or schema. Unstructured data includes not only data that is without a schema, but also data that has some structure, but that structure is not useful for the intended analysis or query.

In this, you will learn about the infrastructure created by Dataproc and relate it to Hadoop operations.

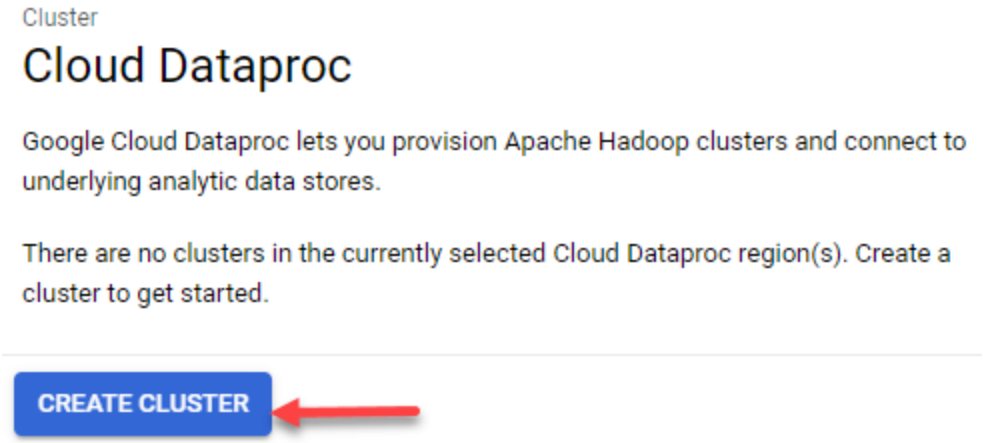
Task 1. Create a Dataproc Cluster

Create a Dataproc Hadoop Cluster customized to use the Google Cloud API

1. In the Console, on the **Navigation menu** ( ) click **Dataproc** > **Clusters**.
   Note: Enable API

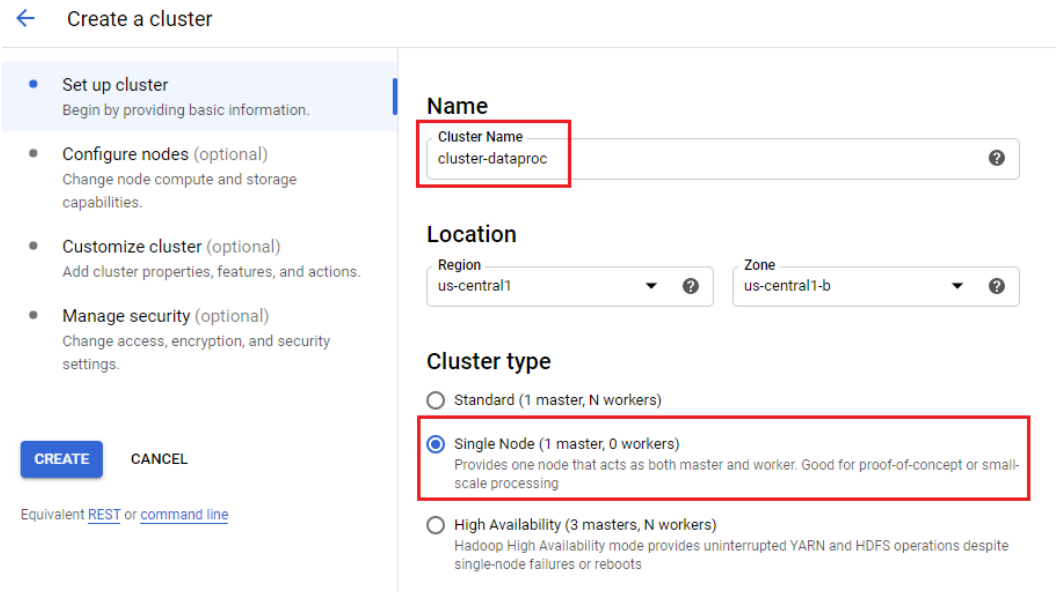


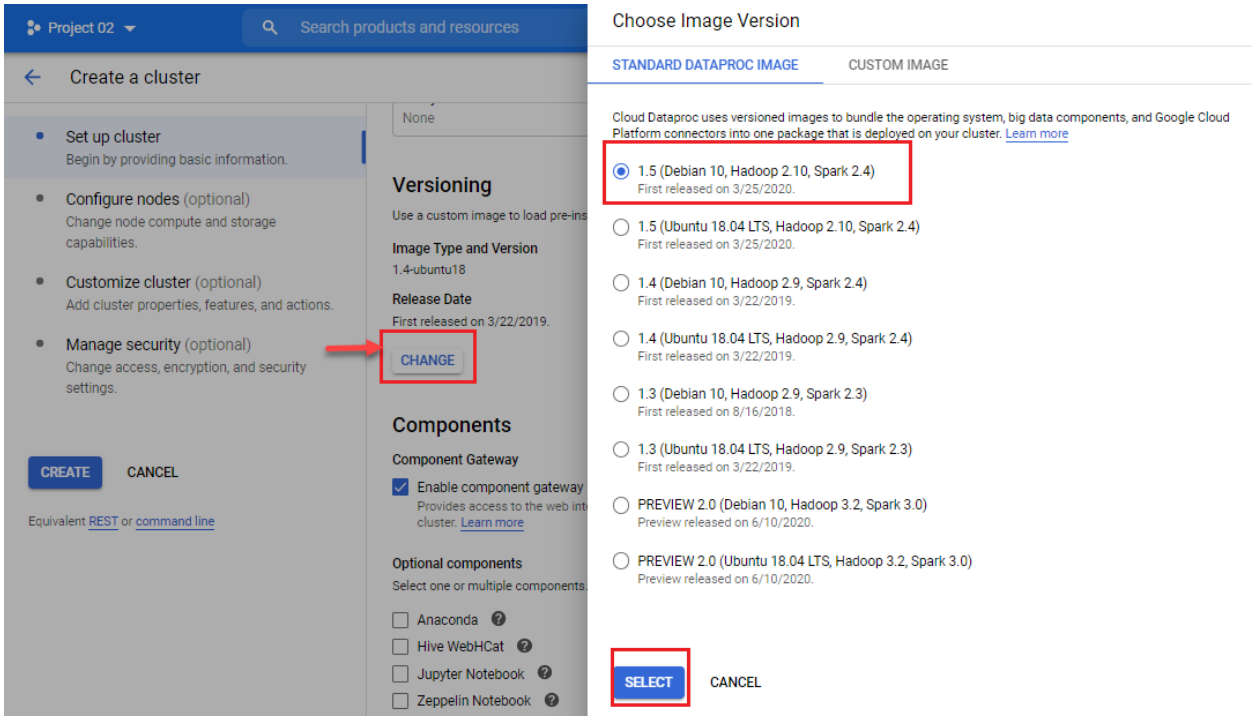
2. Click **Create Cluster**.



3. Specify the following, and leave the remaining settings as their defaults:

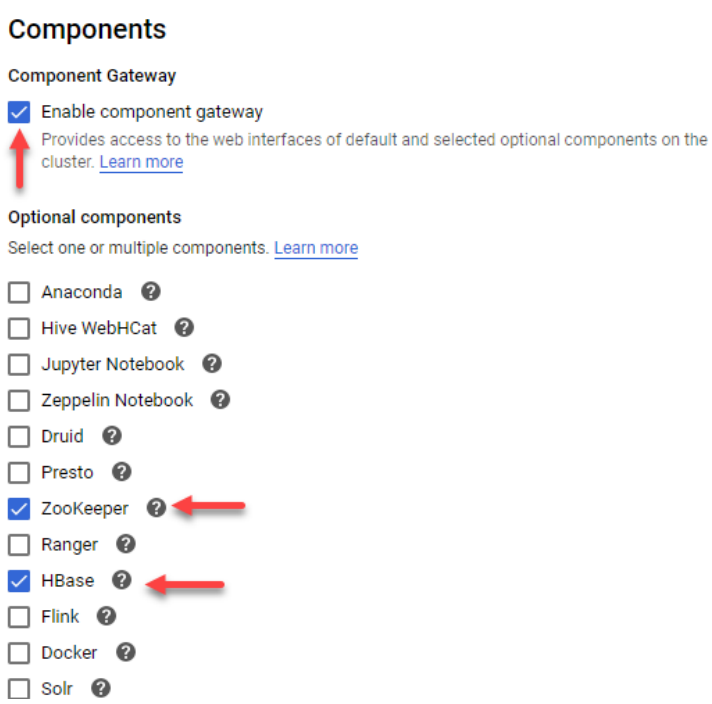
Property	Value (type value or select option as specified)
Name	`cluster-dataproc`
Region	**<your region>**
Zone	**<your zone>**
Cluster mode	**Standard (1 Master, 0 workers)**
(Master node) Machine type	`n1-standard-4`
(Master node) Primary disk size	**200 GB**



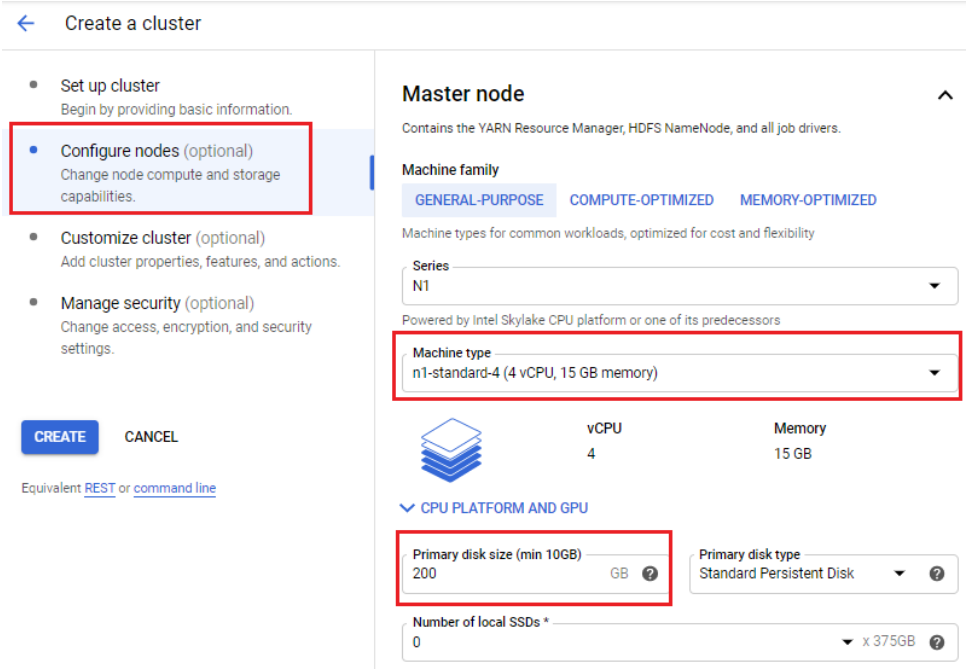


4. Check Component gateway

Scroll down

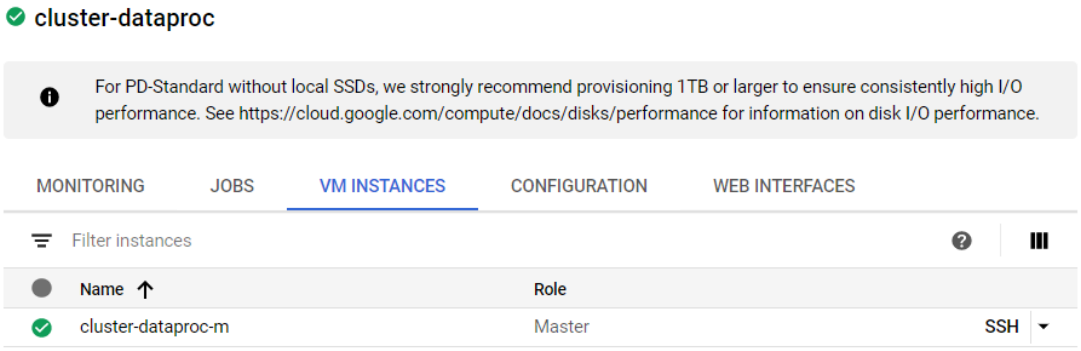


In Configure node



5. Click **Create**.
6. The cluster will take several minutes to become operational. In the Console, on the **Navigation menu** ( ) click **Dataproc** > **Clusters**.
7. Click on your cluster, cluster-dataproc. Then click on the VM Instances tab.

   The instances will become operational before the hadoop software has completed initialization. When a checkmark in a green circle appears next to the name of the cluster, it is operational.



8. Click on SSH to open the terminal