

Installation DataProc cluster

Overview

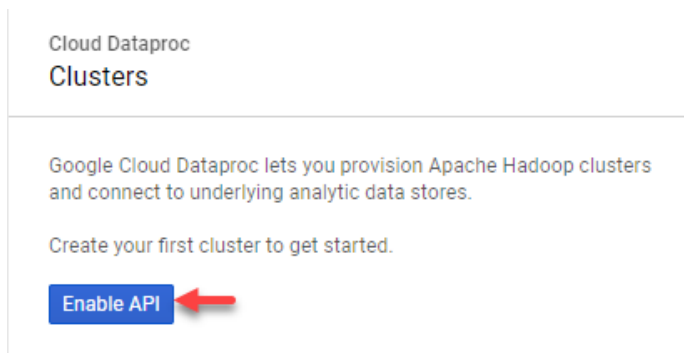
Structured data has a useful organization or schema. Unstructured data includes not only data that is without a schema, but also data that has some structure, but that structure is not useful for the intended analysis or query.

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

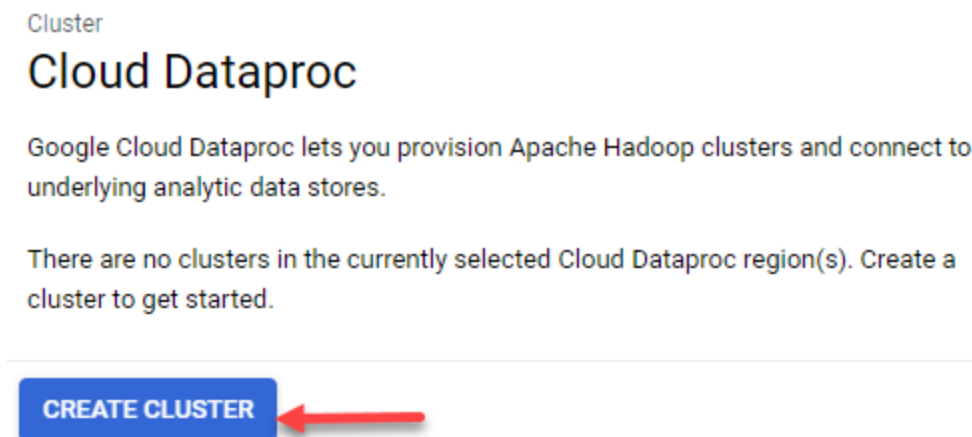
Task 1. Create a Dataproc Cluster

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

[←](#) Create a cluster

- **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.

CREATE CANCEL

Equivalent [REST](#) or [command line](#)

Name

Cluster Name
cluster-dataproc ?

Location

Region
us-central1 ?

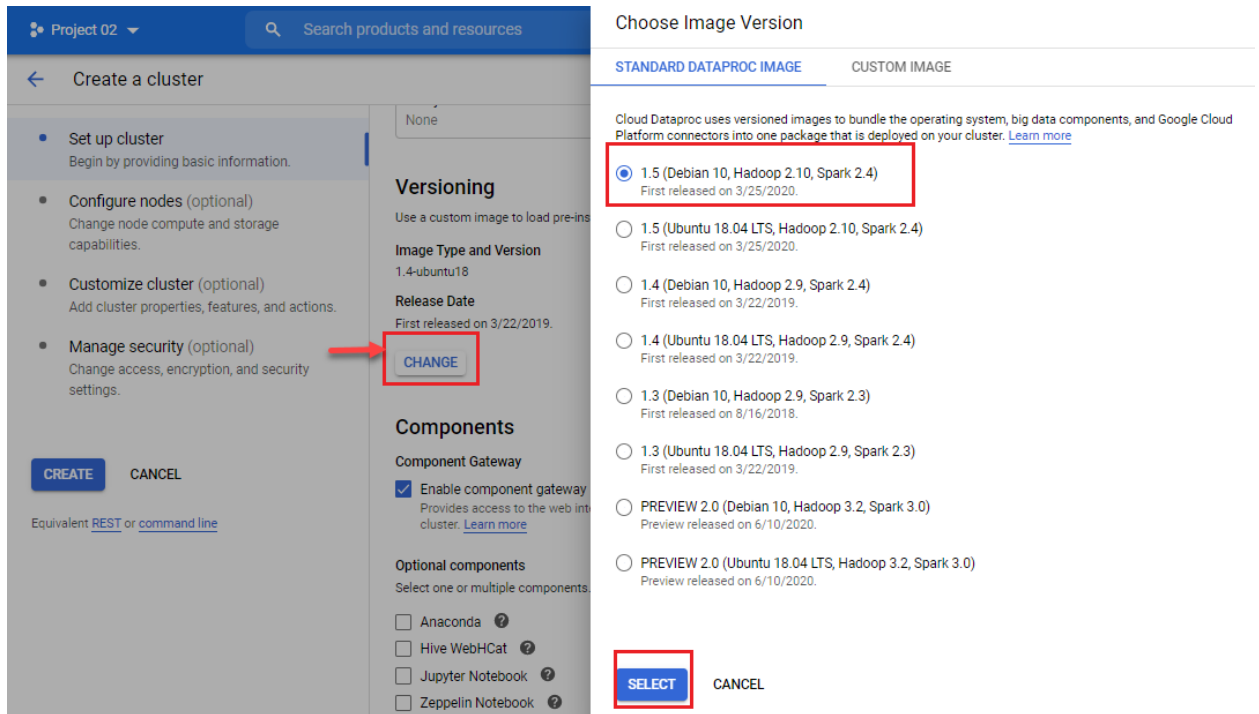
Zone
us-central1-b ?

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



4. Check Component gateway

Scroll down

Components

Component Gateway

- ☒ Enable component gateway
Provides access to the web interfaces of default and selected optional components on the cluster. [Learn more](#)

Optional components

Select one or multiple components. [Learn more](#)

- ☐ Anaconda ?
- ☐ Hive WebHCat ?
- ☐ Jupyter Notebook ?
- ☐ Zeppelin Notebook ?
- ☐ Druid ?
- ☐ Presto ?
- ☒ ZooKeeper ?
- ☐ Ranger ?
- ☒ HBase ?
- ☐ Flink ?
- ☐ Docker ?
- ☐ Solr ?

In Configure node

← Create a cluster

- 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.

CREATE CANCEL

Equivalent [REST](#) or [command line](#)

Master node

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

Machine family

GENERAL-PURPOSE COMPUTE-OPTIMIZED MEMORY-OPTIMIZED

Machine types for common workloads, optimized for cost and flexibility

Series
N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type
n1-standard-4 (4 vCPU, 15 GB memory)

vCPU 4 Memory 15 GB


✓ CPU PLATFORM AND GPU

Primary disk size (min 10GB)
200 GB

Primary disk type
Standard Persistent Disk

Number of local SSDs *
0 x 375GB

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.

✓ cluster-dataproc

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.

MONITORING	JOB	VM INSTANCES	CONFIGURATION	WEB INTERFACES
Filter instances				
Name ↑	Role			
✓ cluster-dataproc-m	Master			SSH

8. Click on SSH to open the terminal