Dataproc

Dataproc

- Dataproc is a managed Apache Spark and Apache Hadoop service that lets you take advantage of open-source data tools for batch processing, querying, streaming, and machine learning.
- Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them.
- With less time and money spent on administration, you can focus on your jobs and your data.

Advantages of Dataproc

Low Cost

- Dataproc is priced at only 1 cent per virtual CPU in your cluster per hour, on top of the other Cloud Platform resources you use.
- Dataproc clusters can include preemptible instances that have lower compute prices, reducing your costs even further.
- Instead of rounding your usage up to the nearest hour, Dataproc charges you only for what you really use with second-by-second billing.

Advantages of Dataproc

Super Fast

- Dataproc clusters are quick to start, scale, and shutdown, with each of these operations taking 90 seconds or less, on average.
- Spend less time waiting for clusters and more hands-on time working with your data.

Advantages of Dataproc

Integrated

- Dataproc has built-in integration with other Google Cloud Platform services, such as BigQuery, Cloud Storage, Cloud Bigtable, Cloud Logging, and Cloud Monitoring
- It is more than just a Spark or Hadoop cluster - a complete data platform.
- Ex. Dataproc can be used to effortlessly ETL terabytes of raw log data directly into BigQuery for business reporting.

Advantages of Dataproc

Managed

- Use Spark and Hadoop clusters without needing an administrator or special software
- Easily interact with clusters and jobs through Google Cloud console, Cloud SDK, or Dataproc REST API
- Turn off clusters when not in use to avoid costs on idle resources
- Data remains safe as Dataproc integrates with Cloud Storage, BigQuery, and Cloud Bigtable

Advantages of Dataproc

Simple

- No need to learn new tools or APIs; existing projects move to Dataproc without redevelopment
- Spark, Hadoop, Pig, and Hive are frequently updated for faster productivity

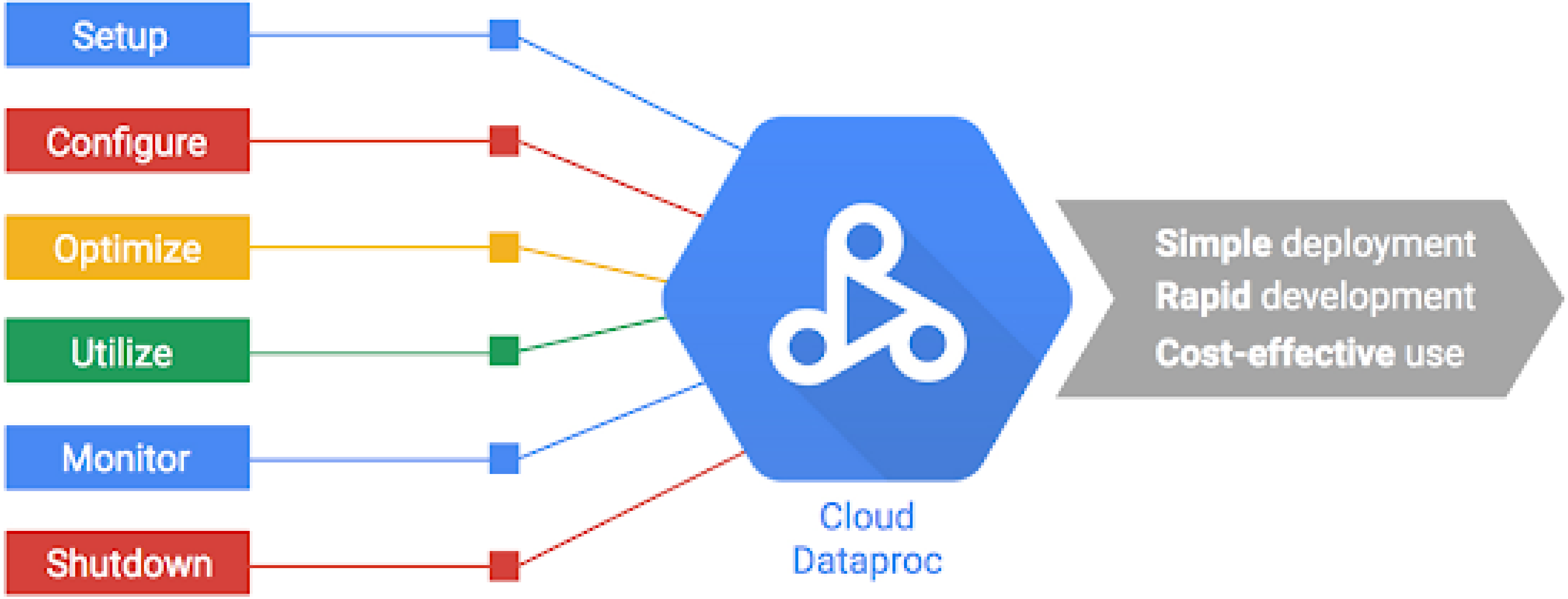
Dataproc Cluster Types

- Standard (1 master, N workers)
- Single Node (1 master, 0 workers)
- High Availability (3 masters, N workers)

Dataproc Cluster Types

- Jobs Supported
  - Hadoop
  - Spark
  - SparkR
  - PySpark
  - SparkSQL
  - Hive

Cluster Activities



Cluster Naming Rules

- A cluster name must start with a lowercase letter.
- Up to 51 characters allowed: lowercase letters, numbers, and hyphens.
- A name cannot end with a hyphen.

Cluster Naming Rules

- A cluster name must start with a lowercase letter.
- Up to 51 characters allowed: lowercase letters, numbers, and hyphens.
- A name cannot end with a hyphen.

Cluster Region

- A Compute Engine region must be specified for the cluster (e.g., us-east1, europe-west1).
- Regional isolation ensures that VM instances, metadata, and Cloud Storage remain within the selected region.
- Region details can be viewed with the gcloud compute regions list command.
- Within a region, resources are further divided into zones (e.g., us-east1-b).
- Zones provide redundancy and fault isolation for cluster components.
- Cluster nodes are typically distributed across zones for high availability.

Cluster Connectivity

- A Dataproc cluster requires full internal IP networking cross-connectivity between master and worker VMs.
- The default VPC network provides this connectivity automatically.
- Networking can be customized for advanced configurations such as private clusters.

Master and Worker Nodes

- Master node manages cluster metadata, scheduling, and monitoring.
- Worker nodes run the actual distributed data processing tasks.
- Additional secondary workers can be added for preemptible, cost-saving compute capacity.

Machine Types and Scaling

- Each node in a cluster is assigned a Compute Engine machine type (e.g., c4-standard-4).
- Machine types define CPU, memory, and storage capacity.
- Clusters can scale by adding or removing workers dynamically.

Storage Integration

- Dataproc integrates with Cloud Storage for storing input, output, and temporary data.
- Cluster data persists in Cloud Storage, even if the cluster is deleted.
- HDFS is supported but typically used only for temporary storage.

Security and IAM

- Cluster access and management are secured with Identity and Access Management (IAM).
- Encryption of data at rest and in transit is enabled by default.
- Service accounts control permissions for clusters accessing other Google Cloud services.

Cost and Lifecycle Management

- Clusters can be turned off when not needed, saving costs on idle resources.
- Preemptible VMs lower compute costs.
- Auto-deletion policies can be configured to shut down clusters after jobs are completed.

Preemptible VMs

- **Short-Lived Instances:** Virtual machines that can run for a maximum of 24 hours.
- **Cost Savings:** Cheaper than standard VMs.
- **Cluster Use:** Often added as secondary worker nodes in Dataproc clusters to handle extra processing.
- **Preemption:** Can be stopped by Google Cloud at any time if resources are needed elsewhere.
- Best For Batch jobs, big data analytics, and machine learning workloads that can tolerate interruptions.

HDFS Integration in Dataproc

- Dataproc integrates with Apache Hadoop and uses the Hadoop Distributed File System (HDFS) for on-cluster storage.
- Additionally, Dataproc installs an HDFS-compatible Cloud Storage connector, enabling Blob storage (GCS) to be used in parallel with HDFS.
- Data can be uploaded to or downloaded from clusters using either HDFS or Cloud Storage.
- VM boot disks (hosting HDFS data) are tied to cluster lifetime: they are deleted when the Dataproc cluster is deleted.
- To preserve data beyond the cluster lifespan, it is recommended to store files in Cloud Storage.

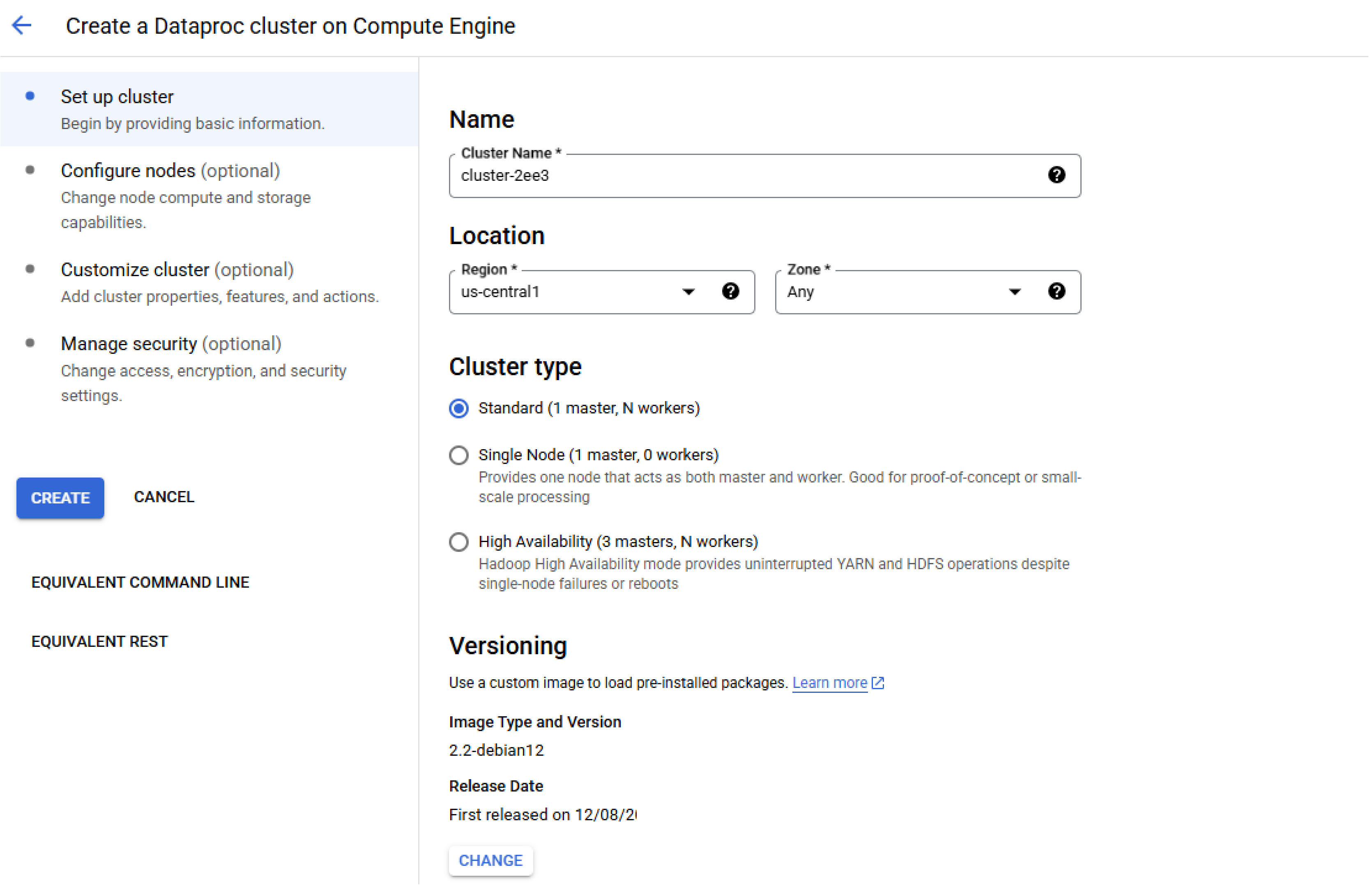
Scaling a Dataproc Cluster

- Scaling a Dataproc cluster involves increasing or decreasing the number of worker nodes—both primary and secondary.
- This change can be made even while jobs are running.
- Horizontal scaling adjusts the number of worker nodes.
- Vertical scaling (changing machine types, CPUs, memory) is not supported once the cluster is created.
- Scaling applies to primary worker nodes or secondary (e.g., preemptible) workers, or both.
- New nodes must match the machine type of existing workers.

Autoscaling for Smart Scaling

- Rather than manual updates, Autoscaling can automatically adjust worker counts.
- Autoscaling uses preconfigured policies to adjust resource levels based on workload.

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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 COMMAND LINE

EQUIVALENT REST

Name

Cluster Name *: cluster-2ee3

Location

Region *: us-central1

Zone *: Any

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

Versioning

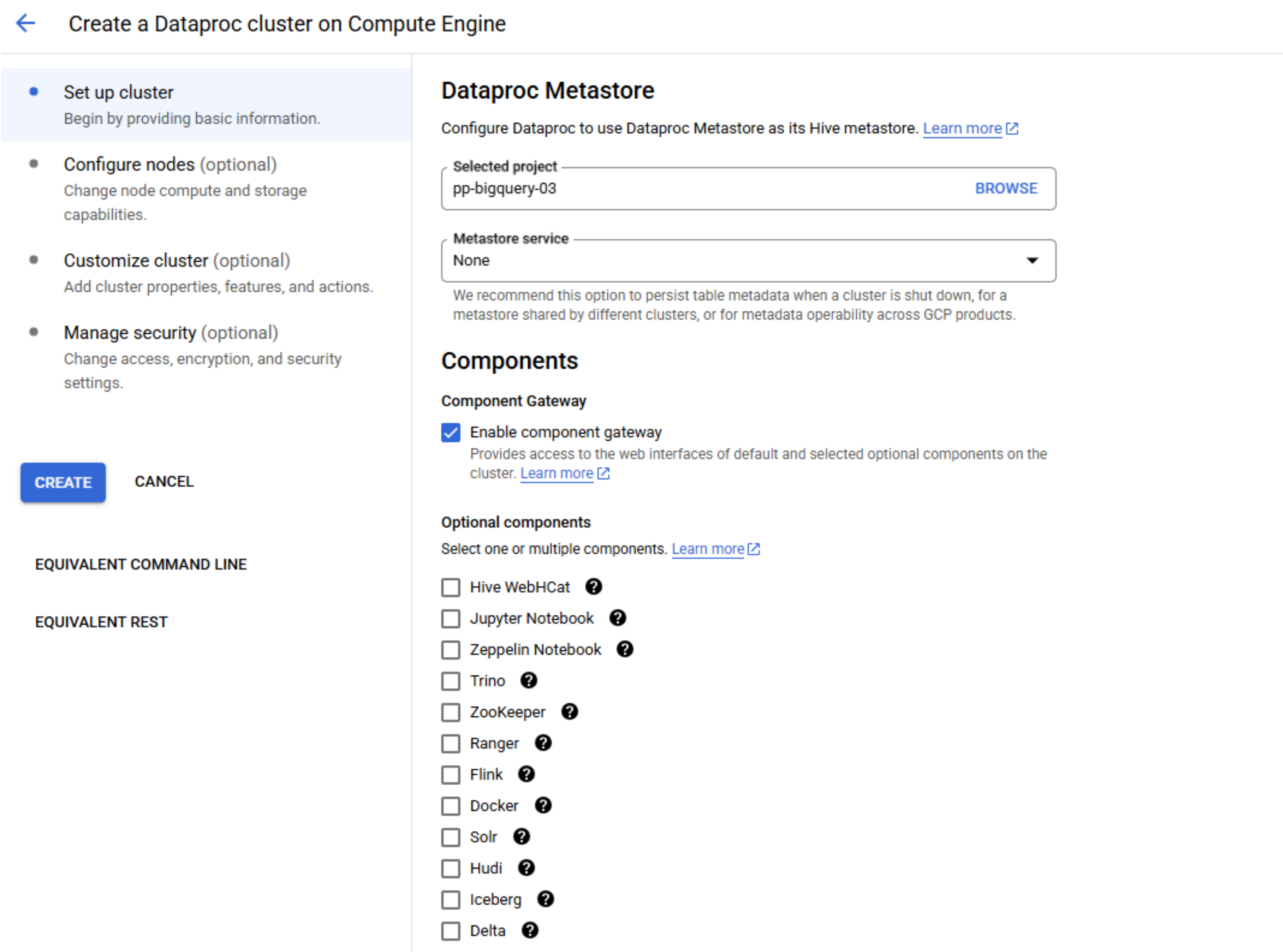
Use a custom image to load pre-installed packages. Learn more

Image Type and Version
2.2-debian12

Release Date
First released on 12/08/2

CHANGE

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

Dataproc Metastore

Configure Dataproc to use Dataproc Metastore as its Hive metastore. Learn more

Selected project: pp-bigquery-03

Metastore service: None

We recommend this option to persist table metadata when a cluster is shut down, for a metastore shared by different clusters, or for metadata operability across GCP products.

Components

Component Gateway

☑ Enable component gateway

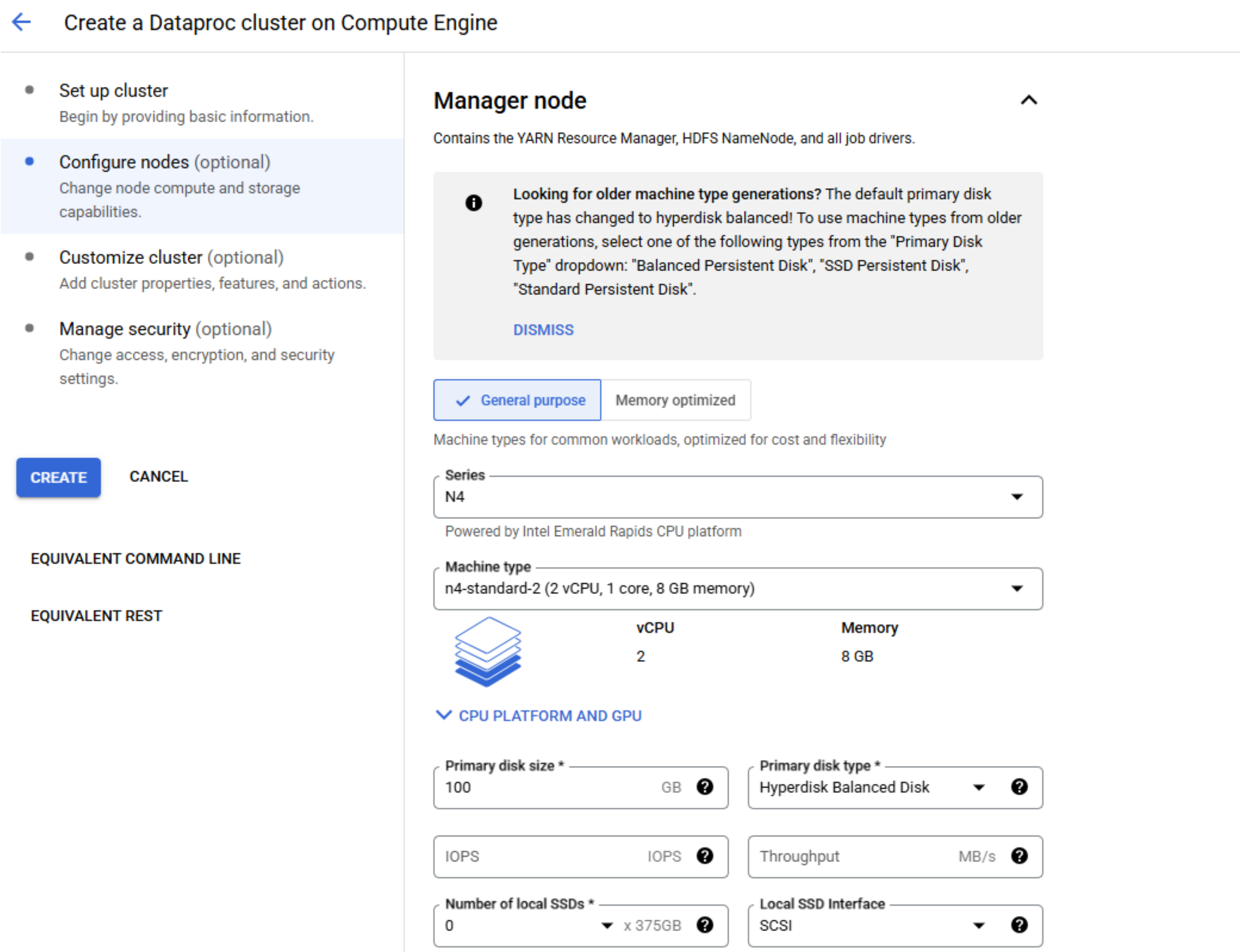
Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

Optional components

Select one or multiple components. Learn more

- □ Hive WebHCat
- □ Jupyter Notebook
- □ Zeppelin Notebook
- □ Trino
- □ ZooKeeper
- □ Ranger
- □ Flink
- □ Docker
- □ Solr
- □ Hudi
- □ Iceberg
- □ Delta

Create a Dataproc Cluster



← Create a Dataproc cluster on Compute Engine

- **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 COMMAND LINE

EQUIVALENT REST

Manager node

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

Looking for older machine type generations? The default primary disk type has changed to hyperdisk balanced! To use machine types from older generations, select one of the following types from the "Primary Disk Type" dropdown: "Balanced Persistent Disk", "SSD Persistent Disk", "Standard Persistent Disk".

DISMISS

✓ General purpose | Memory optimized

Machine types for common workloads, optimized for cost and flexibility

Series: N4

Powered by Intel Emerald Rapids CPU platform

Machine type: n4-standard-2 (2 vCPU, 1 core, 8 GB memory)

vCPU	Memory
2	8 GB

CPU PLATFORM AND GPU

Primary disk size *: 100 GB

Primary disk type *: Hyperdisk Balanced Disk

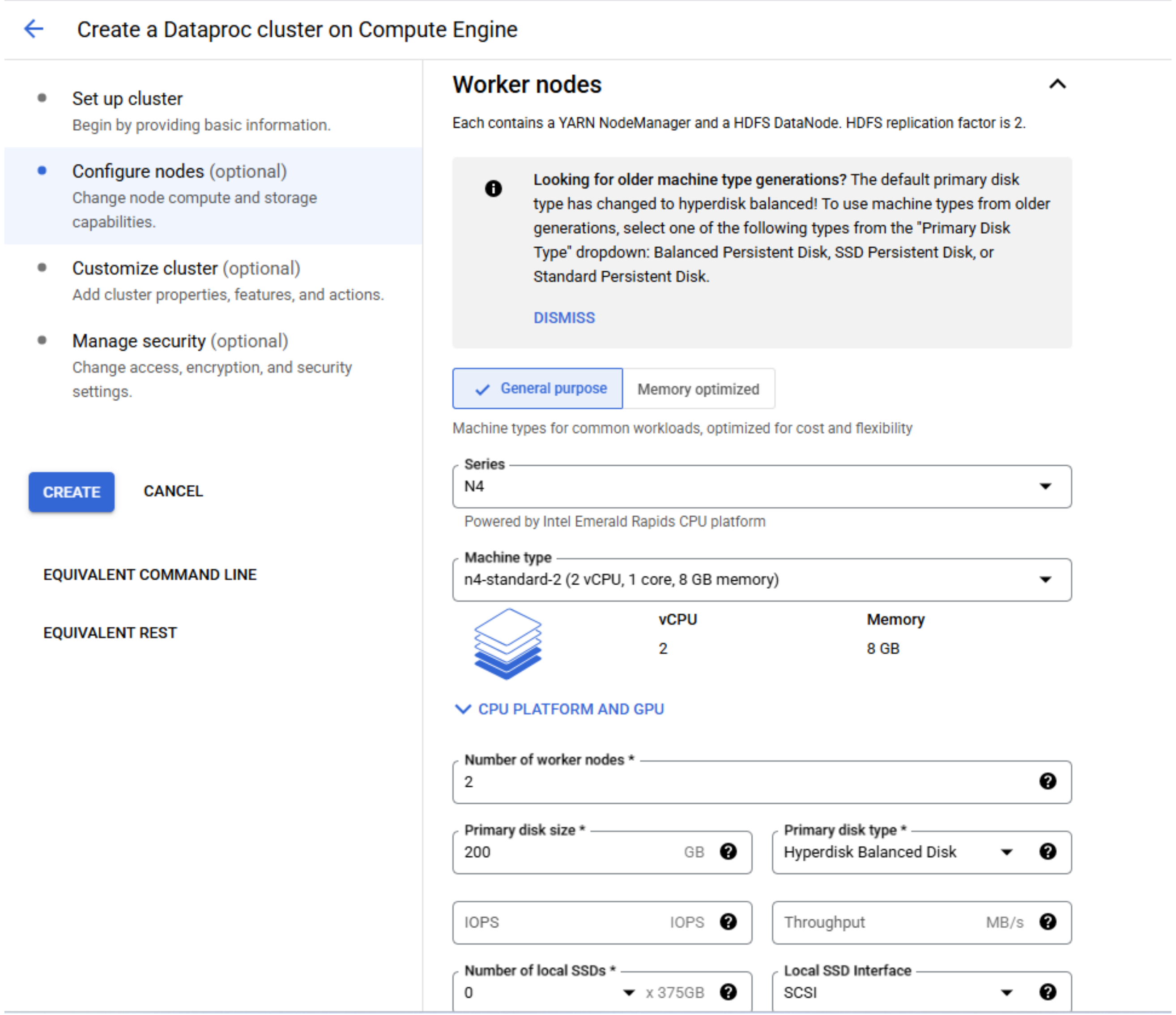
IOPS: IOPS

Throughput: MB/s

Number of local SSDs *: 0 x 375GB

Local SSD Interface: SCSI

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

Looking for older machine type generations? The default primary disk type has changed to hyperdisk balanced! To use machine types from older generations, select one of the following types from the "Primary Disk Type" dropdown: Balanced Persistent Disk, SSD Persistent Disk, or Standard Persistent Disk.

Machine types for common workloads, optimized for cost and flexibility

Series: N4

Powered by Intel Emerald Rapids CPU platform

Machine type: n4-standard-2 (2 vCPU, 1 core, 8 GB memory)

vCPU	Memory
2	8 GB

Number of worker nodes *: 2

Primary disk size *: 200 GB

Primary disk type *: Hyperdisk Balanced Disk

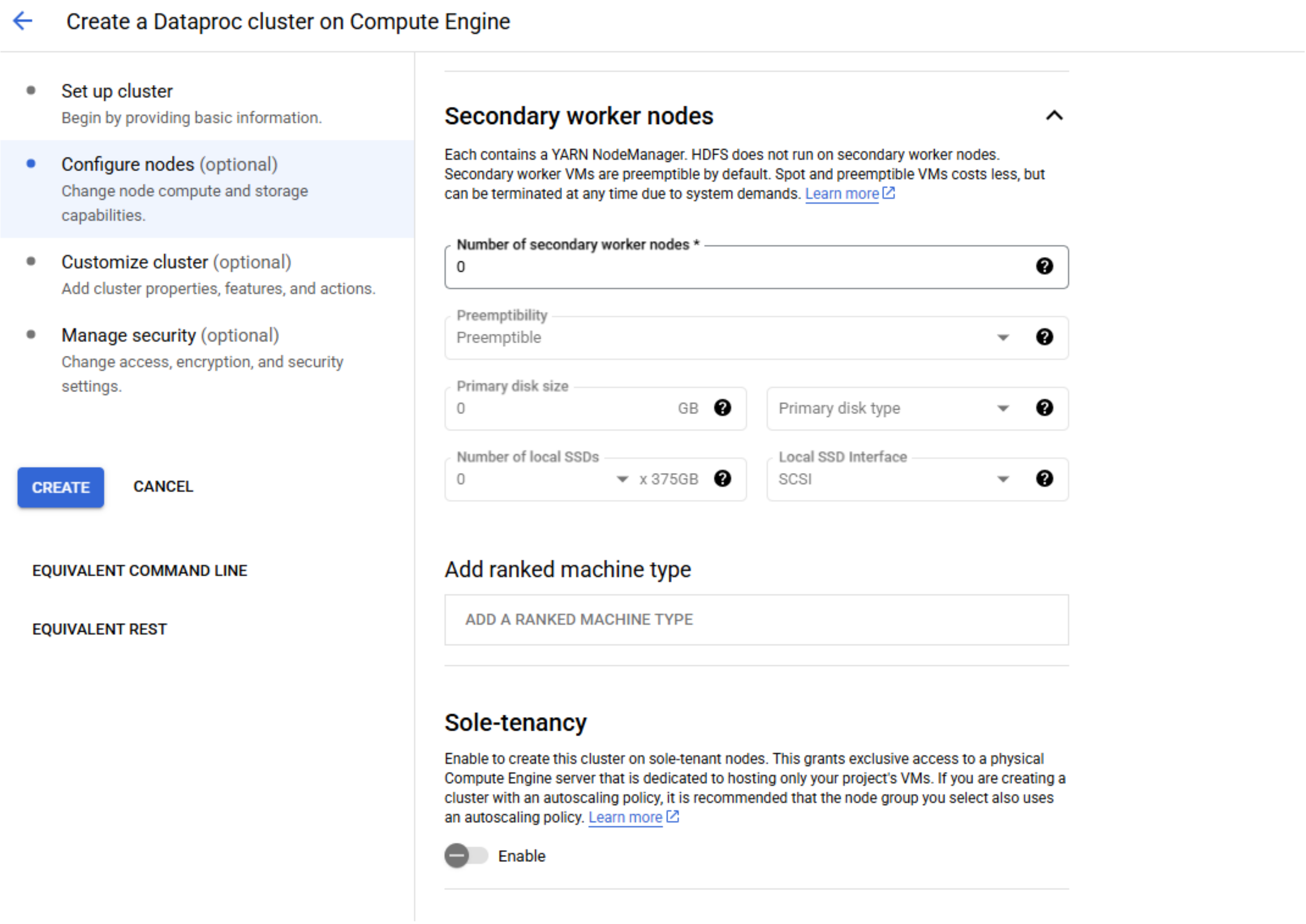
IOPS: IOPS

Throughput: MB/s

Number of local SSDs *: 0 x 375GB

Local SSD Interface: SCSI

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

Secondary worker nodes

Each contains a YARN NodeManager. HDFS does not run on secondary worker nodes. Secondary worker VMs are preemptible by default. Spot and preemptible VMs costs less, but can be terminated at any time due to system demands. Learn more

Number of secondary worker nodes *: 0

Preemptibility: Preemptible

Primary disk size: 0 GB

Primary disk type

Number of local SSDs: 0 x 375GB

Local SSD Interface: SCSI

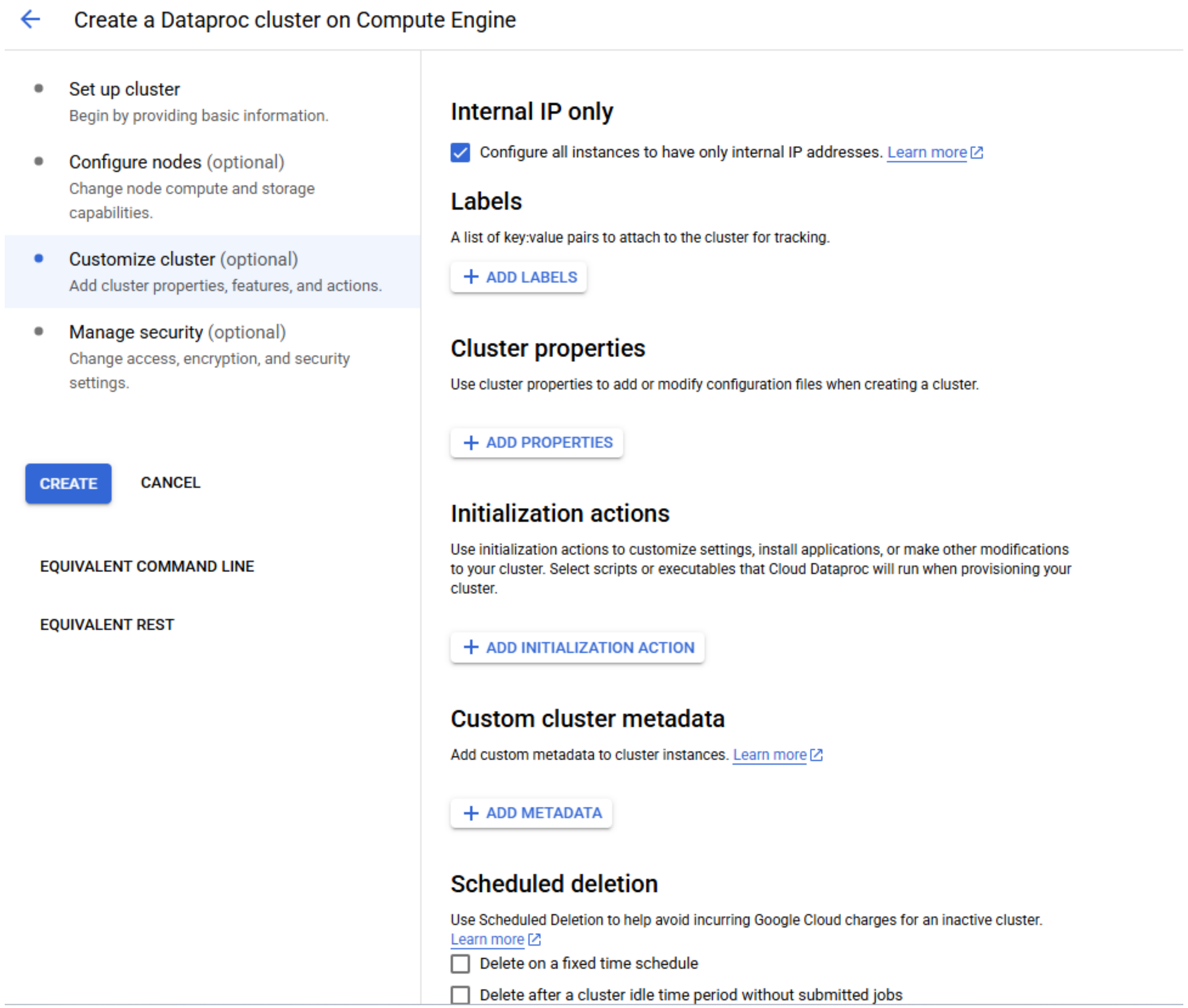
Add ranked machine type

Sole-tenancy

Enable to create this cluster on sole-tenant nodes. This grants exclusive access to a physical Compute Engine server that is dedicated to hosting only your project's VMs. If you are creating a cluster with an autoscaling policy, it is recommended that the node group you select also uses an autoscaling policy. Learn more

○ Enable

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

Internal IP only

☑ Configure all instances to have only internal IP addresses. Learn more

Labels

A list of key:value pairs to attach to the cluster for tracking.

Cluster properties

Use cluster properties to add or modify configuration files when creating a cluster.

Initialization actions

Use initialization actions to customize settings, install applications, or make other modifications to your cluster. Select scripts or executables that Cloud Dataproc will run when provisioning your cluster.

Custom cluster metadata

Add custom metadata to cluster instances. Learn more

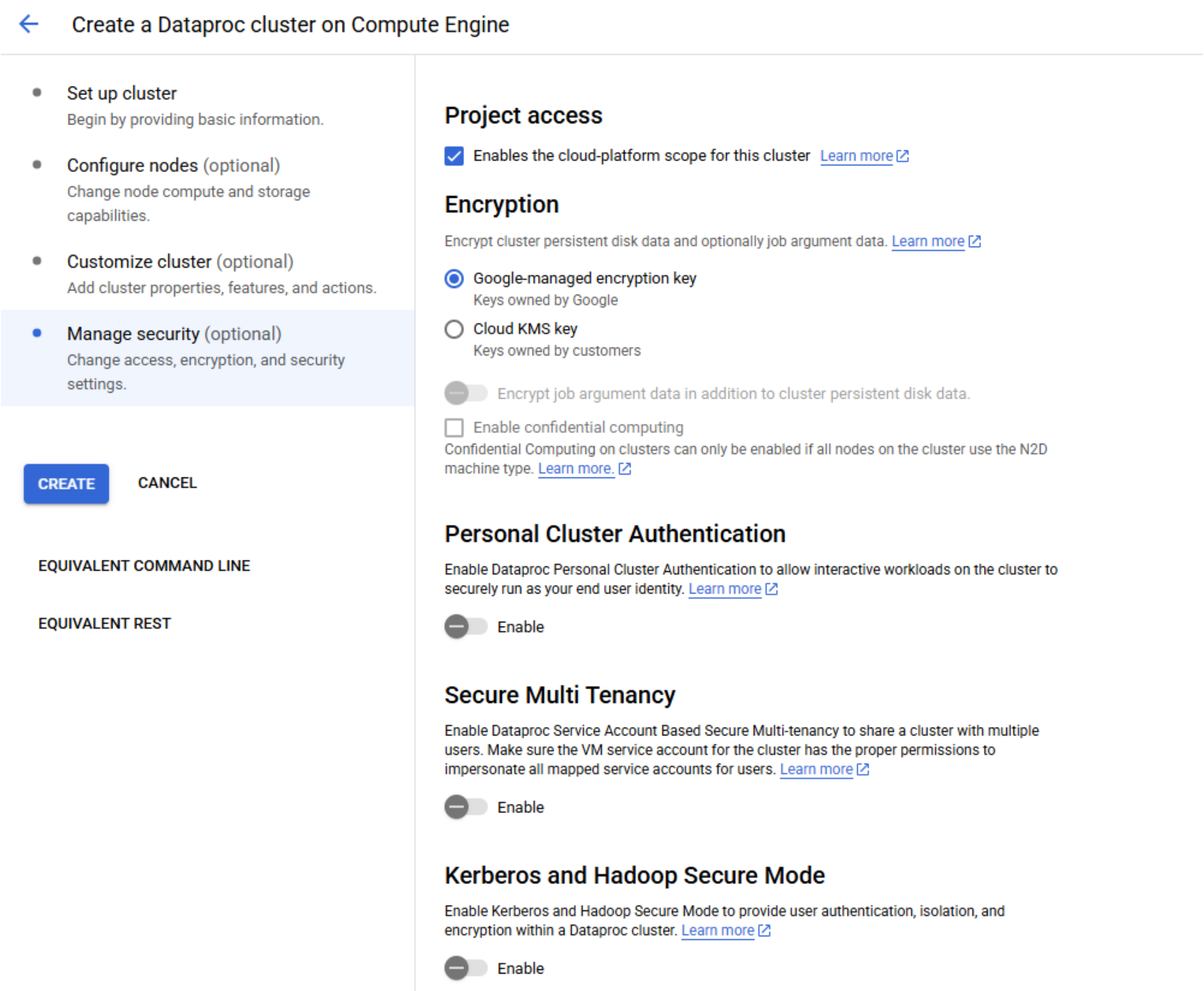
Scheduled deletion

Use Scheduled Deletion to help avoid incurring Google Cloud charges for an inactive cluster. Learn more

□ Delete on a fixed time schedule

□ Delete after a cluster idle time period without submitted jobs

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

Project access

☑ Enables the cloud-platform scope for this cluster Learn more

Encryption

Encrypt cluster persistent disk data and optionally job argument data. Learn more

● Google-managed encryption key
Keys owned by Google

○ Cloud KMS key
Keys owned by customers

Encrypt job argument data in addition to cluster persistent disk data.

□ Enable confidential computing
Confidential Computing on clusters can only be enabled if all nodes on the cluster use the N2D machine type. Learn more.

Personal Cluster Authentication

Enable Dataproc Personal Cluster Authentication to allow interactive workloads on the cluster to securely run as your end user identity. Learn more

Enable

Secure Multi Tenancy

Enable Dataproc Service Account Based Secure Multi-tenancy to share a cluster with multiple users. Make sure the VM service account for the cluster has the proper permissions to impersonate all mapped service accounts for users. Learn more

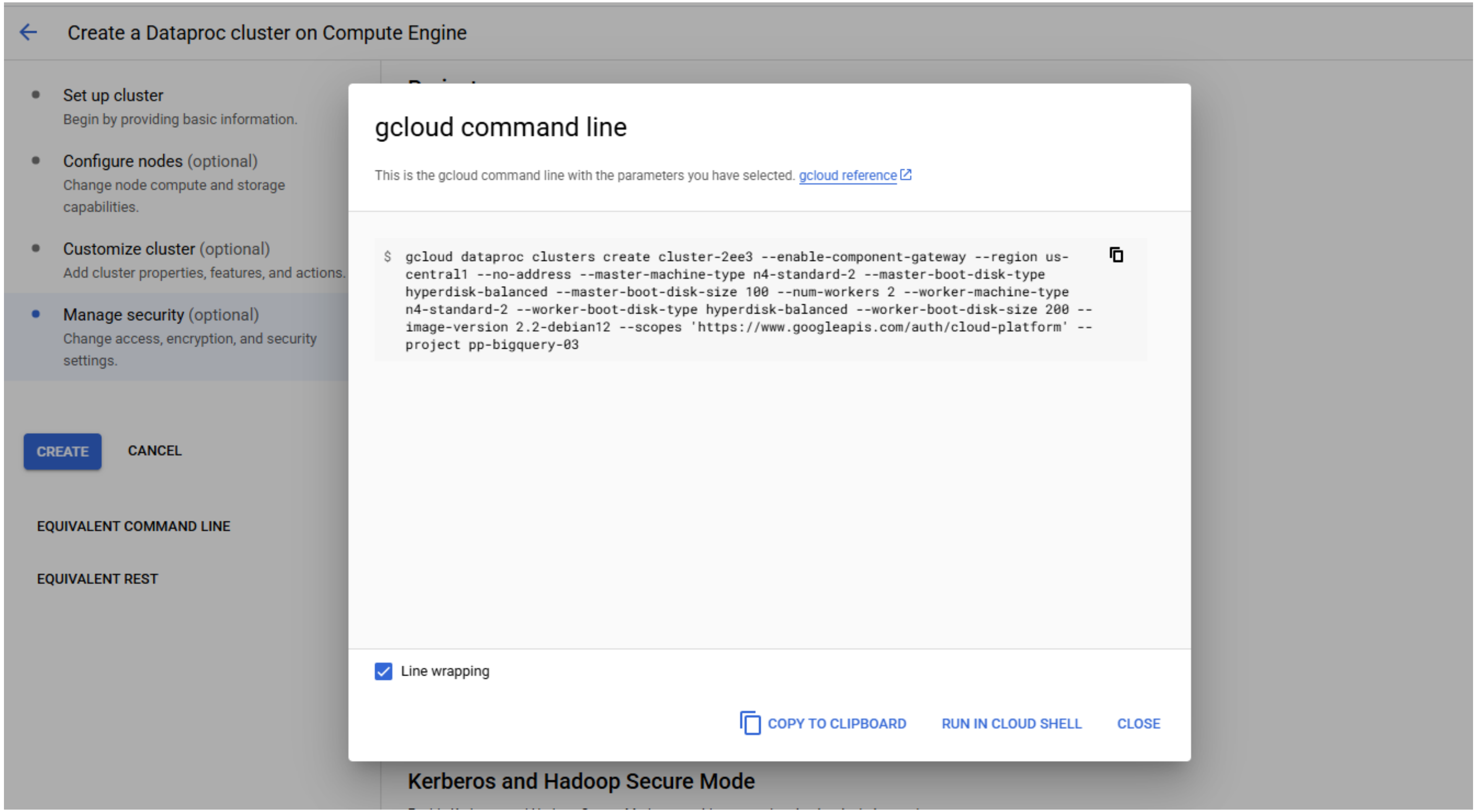
Enable

Kerberos and Hadoop Secure Mode

Enable Kerberos and Hadoop Secure Mode to provide user authentication, isolation, and encryption within a Dataproc cluster. Learn more

Enable

Create a Dataproc Cluster



Create a Dataproc cluster on Compute Engine

- 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.

gcloud command line

This is the gcloud command line with the parameters you have selected. gcloud reference

```
$ gcloud dataproc clusters create cluster-2ee3 --enable-component-gateway --region us-
  central1 --no-address --master-machine-type n4-standard-2 --master-boot-disk-type
  hyperdisk-balanced --master-boot-disk-size 100 --num-workers 2 --worker-machine-type
  n4-standard-2 --worker-boot-disk-type hyperdisk-balanced --worker-boot-disk-size 200 --
  image-version 2.2-debian12 --scopes 'https://www.googleapis.com/auth/cloud-platform' --
  project pp-bigquery-03
```

☑ Line wrapping

Kerberos and Hadoop Secure Mode