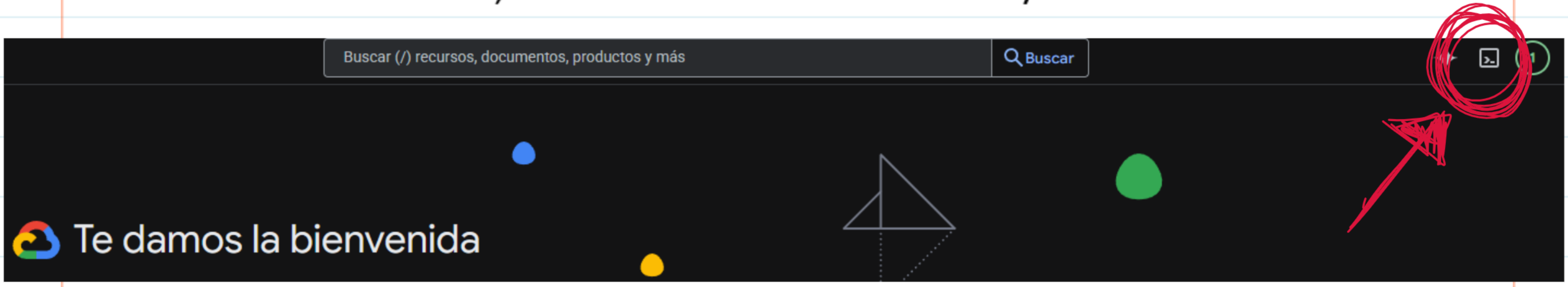
Cómo crear un clúster desde código, estable y que prenda y apague bonito.

Ve a tu consola en GCP, busca el ícono de "Cloud Shell y dale click"



Después, se va a abrir una consolita dentro del navegador, autoriza el acceso y escribe el siguiente código (que en el repo está en creacion_cluster.txt)

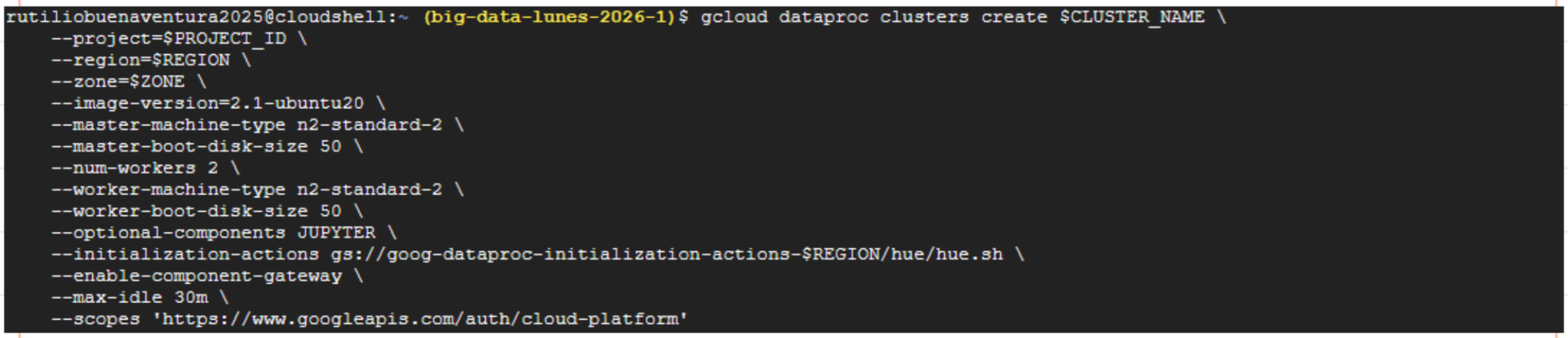
1

```
export PROJECT_ID=$(gcloud config get-value project)
export REGION=us-central1
export ZONE=us-central1-a
export CLUSTER_NAME=hive-learning-cluster
```

2

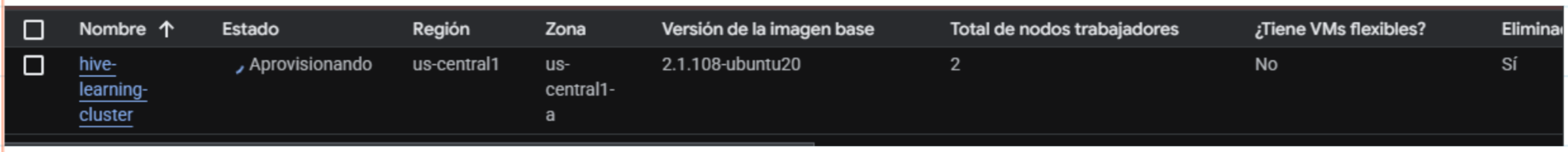
```
gcloud dataproc clusters create $CLUSTER_NAME \
   --project=$PROJECT_ID \
   --region=$REGION \
   --zone=$ZONE \
   --image-version=2.1-ubuntu20 \
   --master-machine-type n2-standard-2 \
   --master-boot-disk-size 100 \
   --num-workers 2 \
   --worker-machine-type n2-standard-2 \
   --worker-boot-disk-size 100 \
   --optional-components JUPYTER,ZEPPELIN \
   --enable-component-gateway \
   --scopes 'https://www.googleapis.com/auth/cloud-platform'
```

aquí puedes usar 3



```
rutiliobuenaventura2025@cloudshell:~ (big-data-lunes-2026-1)$ gcloud dataproc clusters create $CLUSTER_NAME \
    --project=$PROJECT_ID \
    --region=$REGION \
    --zone=$ZONE \
    --image-version=2.1-ubuntu20 \
    --master-machine-type n2-standard-2 \
    --master-boot-disk-size 50 \
    --num-workers 2 \
    --worker-machine-type n2-standard-2 \
    --worker-boot-disk-size 50 \
    --optional-components JUPYTER \
    --initialization-actions gs://goog-dataproc-initialization-actions-$REGION/hue/hue.sh \
    --enable-component-gateway \
    --max-idle 30m \
    --scopes 'https://www.googleapis.com/auth/cloud-platform'
```

Verás en tu pestaña de "Dataproc" que el cluster se está aprovisionando



□	Nombre ↑	Estado	Región	Zona	Versión de la imagen base	Total de nodos trabajadores	¿Tiene VMs flexibles?	Elimina
□	hive-learning-cluster	Aprovisionando	us-central1	us-central1-a	2.1.108-ubuntu20	2	No	Sí

Qué está sucediendo? Está creando un cluster con 1 Namenode y 2 Workers, con acceso a Jupyter

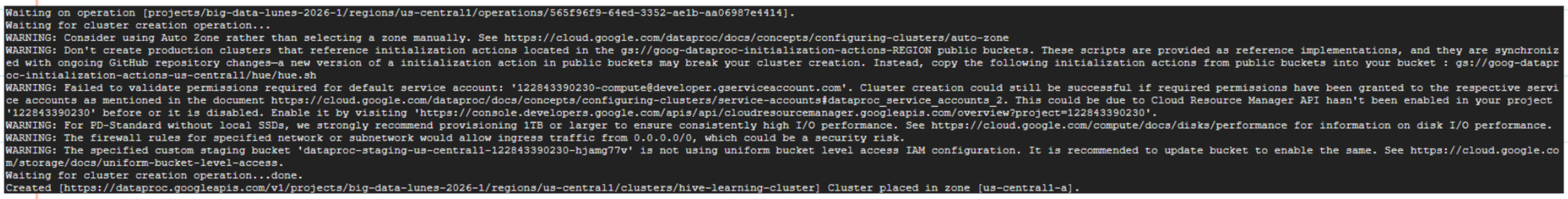
Cuando el cluster esté en ejecución, listo! ahora... si se te olvida borrarlo, o cerrarlo, y no usas HIVE para tus tablas, pero te gusta PySpark.

puedes agregar estas líneas al final del código, antes de "scopes"

```
--expiration-time 3h \
--max-idle 1h \
```

Esto significa que si no lo usas en 1 hora, o su vida llegó a las 3 horas, se borrará automáticamente.

Te van a salir unos warnings, no te preocupes, dejalos ser. Si sale un error, llama a tu profe de confianza... preferentemente yo =/.



```
Waiting on operation [projects/big-data-lunes-2026-1/regions/us-central1/operations/565f96f9-64ed-3352-ae1b-aa06987e4414].
Waiting for cluster creation operation...
WARNING: Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Don't create production clusters that reference initialization actions located in the gs://goog-dataproc-initialization-actions-REGION public buckets. These scripts are provided as reference implementations, and they are synchroniz
ed with ongoing GitHub repository changes—a new version of a initialization action in public buckets may break your cluster creation. Instead, copy the following initialization actions from public buckets into your bucket : gs://goog-datapr
oc-initialization-actions-us-central1/hue/hue.sh
WARNING: Failed to validate permissions required for default service account: '122843390230-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permissions have been granted to the respective servi
ce accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project
'122843390230' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=122843390230'.
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.
WARNING: The firewall rules for specified network or subnetwork would allow ingress traffic from 0.0.0.0/0, which could be a security risk.
WARNING: The specified custom staging bucket 'dataproc-staging-us-central1-122843390230-hjamg77v' is not using uniform bucket level access IAM configuration. It is recommended to update bucket to enable the same. See https://cloud.google.co
m/storage/docs/uniform-bucket-level-access.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/big-data-lunes-2026-1/regions/us-central1/clusters/hive-learning-cluster] Cluster placed in zone [us-central1-a].
```