

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID



# Guía de Configuración de Google Cloud para Máster NTIC de la UCM

Autores: Ismael Yuste y  
Pablo J. Villacorta

Actualizado Mayo 2021

## Índice

Google Cloud Dataproc	3
DATAPROC	3
FUNCIONES DE DATAPROC	4
Prueba gratuita de Google Cloud	7
Google Cloud Shell	11
CLOUD SHELL	11
Google Cloud Dataproc	12
Crear un Bucket de Google Cloud Storage	12
Crear un Cluster	15
Pasos	15
Operar con un Cluster (Acceso a Jupyter y SSH)	17
Borrar un Cluster	19

### DATAPROC

Método rápido, fácil y rentable de ejecutar Apache Spark y Apache Hadoop

#### Apache Hadoop y Apache Spark nativos de la nube

Dataproc es un servicio en la nube rápido, fácil de usar y totalmente gestionado para ejecutar clústeres de Apache Spark y Apache Hadoop de una manera más sencilla y rentable. Las operaciones que antes llevaban horas o días tardan apenas unos minutos o segundos. Además, solo se paga por los recursos que se utilizan (facturación por segundos). Dataproc también se integra con facilidad con otros servicios de Google Cloud Platform (GCP), de modo que tienes a tu disposición una plataforma potente y completa para procesar datos, analizarlos y realizar tareas de aprendizaje automático.

#### Procesamiento de datos rápido y escalable

Puedes crear rápidamente clústeres de Dataproc y cambiar su tamaño en cualquier momento (desde tres nodos a cientos de ellos). Así te despreocupas de que los flujos de procesamiento de tus datos sobrepasen los clústeres. Como cada acción de clúster tarda menos de 90 segundos de media, dispones de más tiempo para centrarte en la información valiosa y puedes supervisar la infraestructura más rápido.

#### Precios asequibles

Dataproc ha adoptado los principios de Google Cloud Platform, por lo que se beneficia de una estructura de precios de bajo coste muy fácil de entender basada en el uso real (medido por segundo). Además, los clústeres de Dataproc pueden incluir instancias interrumpibles con un coste menor, descuentos por uso confirmado y por uso continuado, lo que significa que dispones de clústeres muy potentes por un precio total incluso más bajo.

#### Ecosistema de código abierto

Con Dataproc, podrás utilizar las herramientas, las bibliotecas y la documentación de Spark y Hadoop. Además, como ofrece actualizaciones frecuentes de las versiones nativas de Spark, Hadoop, Pig y Hive, no tienes que aprender a utilizar herramientas ni API nuevas para empezar a usarlo. Además, puedes mover proyectos o flujos de procesamiento ETL sin necesidad de volver a desarrollarlos.

## **FUNCIONES DE DATAPROC**

Dataproced es un servicio Apache Spark y Apache Hadoop gestionado, rápido, fácil de usar y de bajo coste.

### **Gestión automática de clústeres**

Como el despliegue, el almacenamiento de registros y la supervisión son procesos gestionados, puedes centrarte en los datos en lugar de en los clústeres, que son estables, escalables y rápidos con Dataproced.

### **Clústeres de tamaño ajustable**

Crea y escala rápidamente clústeres con varios tipos de máquinas virtuales, tamaños de disco, número de nodos y opciones de red.

### **Autoescalado de clústeres**

El autoescalado de Dataproced es un mecanismo de automatización de la gestión de los recursos de clústeres que permite que se añadan y quiten automáticamente trabajadores del clúster (es decir, nodos).

### **Integración en la nube**

Está integrado en Cloud Storage, BigQuery, Bigtable, Stackdriver Logging, Stackdriver Monitoring y AI Hub, por lo que disfrutas de una plataforma de datos completa y sólida.

### **Gestionar versiones**

Gracias a la gestión de versiones de imágenes, puedes cambiar entre distintas versiones de Apache Spark, Apache Hadoop y otras herramientas.

### **Alta disponibilidad**

Ejecuta clústeres en el modo de alta disponibilidad con varios nodos maestros y configura tareas de reinicio en caso de fallo para que los clústeres y las tareas estén siempre disponibles.

### **Seguridad empresarial**

Al crear un clúster de Cloud Dataproced, puedes habilitar el modo seguro de Hadoop a través de Kerberos añadiendo una configuración de seguridad. GCP y Dataproced ofrecen también otras prestaciones de seguridad que contribuyen a proteger tus datos. Algunas de las funciones de seguridad específicas de GCP más utilizadas con Dataproced son el encriptado en reposo predeterminado, OS Login, los Controles de Servicio de VPC y las claves de encriptado gestionadas por el cliente (CMEK)

### **Eliminación programada de clústeres**

Para evitar que se te cobre por clústeres inactivos, puedes usar la eliminación programada de Cloud Dataproc, que te permite deshacerte de clústeres cuando llevan un tiempo especificado inactivos, en un momento futuro o tras un periodo concreto.

### **Configuración manual o automática**

Dataproc configura automáticamente el hardware y el software, pero también te ofrece control manual.

### **Herramientas de desarrollo**

Dispones de varios métodos para gestionar los clústeres, como una interfaz web intuitiva, el SDK de Google Cloud, las API RESTful y el acceso SSH.

### **Acciones de inicialización**

Ejecuta acciones de inicialización para instalar o personalizar la configuración y las bibliotecas necesarias cuando crees clústeres.

### **Componentes opcionales**

Instala o configura componentes opcionales en el clúster. Estos componentes están integrados con los de Dataproc y ofrecen entornos plenamente configurados para Zeppelin, Druid, Presto y otros componentes de software libre relacionados con el ecosistema de Apache Hadoop y Apache Spark.

### **Imágenes personalizadas**

Los clústeres de Cloud Dataproc se pueden aprovisionar con una imagen personalizada que incluye tus paquetes de sistema operativo Linux preinstalados.

### **Máquinas virtuales flexibles**

Los clústeres pueden usar tipos de máquinas personalizadas y máquinas virtuales interrumpibles para que su tamaño se adapte a tus necesidades en todo momento.

### **Pasarela de componentes y acceso a cuadernos**

La pasarela de componentes de Dataproc te otorga acceso seguro en un clic a las interfaces web de componentes opcionales y predeterminadas de Cloud Dataproc que se ejecutan en el clúster.

### **Plantillas de flujo de trabajo**

Las plantillas de flujo de trabajo de Dataproc son un mecanismo útil para gestionar y ejecutar flujos de trabajo. Estas plantillas son configuraciones de flujos de trabajo reutilizables que definen un gráfico de tareas con información sobre dónde ejecutar dichas tareas.

Referencia: <https://cloud.google.com/dataproc>

Precios 0,010\$-0,064\$ por hora, por máquina.

## PRECIOS DE DATAPROC

Dataproc conlleva una pequeña tarifa incremental por CPU virtual en las instancias de Compute Engine que tu clúster utilice<sup>1</sup>.

Bélgica (europe-west1) ▾

Tipo de máquina	Precio
<b>Máquinas estándar</b> <i>1-64 CPU virtuales</i>	\$0.010 - \$0.640
<b>Máquinas de memoria elevada</b> <i>2-64 CPU virtuales</i>	\$0.020 - \$0.640
<b>Máquinas con un gran número de CPU</b> <i>2-64 CPU virtuales</i>	\$0.020 - \$0.640
<b>Máquinas personalizadas</b> <i>Basadas en el uso de vCPU y de memoria</i>	\$0.010/ vCPU hour

Si pagas en una moneda que no sea el dólar estadounidense, se aplicarán los precios que figuran para tu divisa en los [SKU de Cloud Platform](#).

<sup>1</sup> Dataproc conlleva una pequeña tarifa incremental por CPU virtual en las instancias de Compute Engine que tu clúster utilice mientras esté operativo. Otros recursos que use Dataproc, como la red de Compute Engine, BigQuery y Cloud Bigtable, se facturan a medida que se consumen. Consulta la [guía de precios](#) para obtener información más detallada.

## Prueba gratuita de Google Cloud

Google Cloud ofrece una prueba gratuita por **3 meses** y un crédito gratuito de **300 dólares**. En la práctica, no vamos a agotar este saldo. Apenas llegaremos a gastar unos 20 o 30 dólares. Transcurrido ese período o consumidos los 300 dólares (lo que ocurra antes), sólo podremos seguir usando Google Cloud si actualizamos (explícitamente, en el botón habilitado para ello) nuestra cuenta de prueba a una cuenta de pago

La URL para inicial el alta de la prueba gratuita es [cloud.google.com/free](https://cloud.google.com/free).

Necesitaremos una cuenta de Gmail, y una tarjeta de crédito. No se hará ningún cargo hasta que, una vez agotado el crédito gratuito de 300 dólares, hagamos click expresamente en



The image is a screenshot of the Google Cloud Platform (GCP) website's free tier page. At the top, the heading "Nivel gratuito de Google Cloud Platform" is displayed in a large, dark font. Below the heading, a subtext reads "Aprende y crea en GCP gratis." A prominent blue button labeled "Comenzar gratis" is centered below the subtext. The page features two main components of the free tier, separated by a plus sign. The first component, "12 meses", is accompanied by a yellow cloud icon and states: "Crédito gratuito de \$300 para comenzar a usar cualquier producto de GCP." The second component, "Siempre gratuito", is accompanied by a green dollar sign icon and states: "Límites de uso gratuito en productos seleccionados para clientes que cumplan con los requisitos, durante y después de la prueba gratuita. La oferta está sujeta a cambios."

El alta en la prueba gratuita, consta de dos pasos.

## Prueba Google Cloud Platform de manera gratuita

### Paso 1 de 2

#### País

España

#### Condiciones del Servicio

- ☒ Acepto las [Condiciones del Servicio de Google Cloud Platform](#) y las de [las API y los servicios aplicables](#). También leí y acepto las [Condiciones del Servicio de la prueba gratuita de Google Cloud Platform](#).

Debes seleccionar para continuar

#### Actualizaciones por correo electrónico

- ☐ Quiero recibir correos electrónicos periódicos sobre novedades, actualizaciones de productos y ofertas especiales de Google Cloud y Google Cloud Partners.

CONTINUAR



## Paso 2 de 2

### Información del cliente



Tipo de cuenta ⓘ



Individual ▼



Nombre y dirección ⓘ

Nombre

User 1

Línea 1 de la dirección

Calle Gran Via

Línea 2 de la dirección

Código postal

28020 ⓘ

Ciudad

Madrid

Provincia/región

Madrid ▼

Número telefónico (Opcional)

## Tipo de pago



### Pagos automáticos mensuales

Pagará este servicio todos los meses en la fecha de vencimiento del pago, mediante un cargo automático.

## Forma de pago ⓘ



### Detalles de la tarjeta



La dirección de la tarjeta de crédito o débito es la misma que figura arriba.

**INICIAR PRUEBA GRATUITA**

## Google Cloud Shell

Cloud Shell es una línea de comandos, en una máquina virtual en la nube, que nos permite lanzar comandos bash en la nube, sin necesidad de instalar nada en nuestro ordenador. Podemos usarla, por ejemplo, para lanzar el cluster de Dataproc que describimos en la siguiente s

### CLOUD SHELL

Administra tus aplicaciones e infraestructura desde la línea de comandos en cualquier navegador

*Tu máquina de administración preparada por Google*

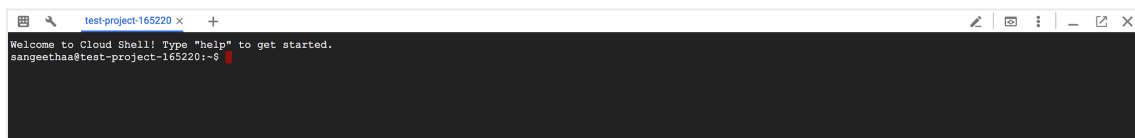
Google Cloud Shell te ofrece acceso a tus recursos en la nube mediante la línea de comandos directamente desde el navegador. Puedes administrar fácilmente tus proyectos y recursos sin tener que instalar en tu sistema el SDK de Google Cloud ni otras herramientas. Con Cloud Shell, la herramienta de línea de comandos gcloud del SDK de Google Cloud y otras utilidades esenciales están siempre disponibles, actualizadas y completamente autenticadas para cuando las necesites.

Inicia Cloud Shell

Haz clic en el botón Activar Cloud Shell en la parte superior de la ventana de la consola.



Se abrirá una sesión de Cloud Shell en un marco nuevo en la parte inferior de la consola, que mostrará una ventana emergente con una línea de comandos. Es posible que esta sesión tarde unos segundos en inicializarse.



Tu sesión de Cloud Shell está lista para usarse.

Guía de Inicio rápido: <https://cloud.google.com/shell/docs/quickstart>

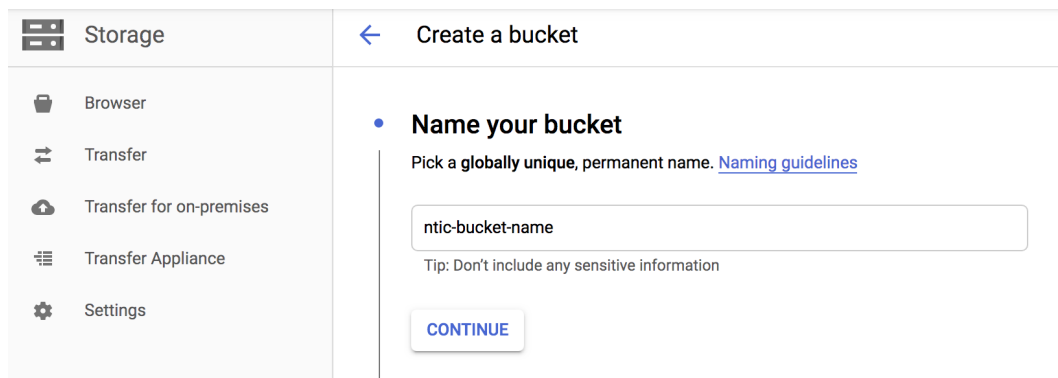
Referencia: <https://cloud.google.com/shell>

### CREAR UN BUCKET DE GOOGLE CLOUD STORAGE

Para persistir los datos y el código que ejecutemos en nuestro Dataproc, vamos a utilizar un Bucket de Google Cloud Storage. Es básicamente, un almacenamiento persistente de tipo HDFS (aunque la tecnología **no** es HDFS), en la nube de Google, que nos va a permitir interactuar con el Cluster de Dataproc, y que aunque este se cree y se destruya, no va a desaparecer, sino que será un volumen persistente asociado a nuestro cluster efímero.

Para crearlo, seguiremos estos pasos.

1. Abrir la consola de Google Cloud.
  - a. <https://console.cloud.google.com/>
2. Activar el API.
3. En el menu de la izquierda, ir a Storage → Browser.
4. Seguir los pasos de creación de un Bucket asignando nombre, modo (**region**), localización (**europa-west-1**), tipo (**standard**) y resto de opciones por defecto.



The screenshot shows the Google Cloud Storage console interface for creating a new bucket. On the left is a sidebar menu with options: Storage, Browser, Transfer, Transfer for on-premises, Transfer Appliance, and Settings. The main area is titled 'Create a bucket' and contains a section 'Name your bucket'. It prompts the user to 'Pick a globally unique, permanent name' with a link to 'Naming guidelines'. A text input field contains the placeholder 'ntic-bucket-name'. Below the field is a tip: 'Tip: Don't include any sensitive information'. At the bottom of the form is a 'CONTINUE' button.

- **Choose where to store your data**

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [Learn more](#)

**Location type**

- ☒ **Region**  
Lowest latency within a single region
- ☐ **Multi-region**  
Highest availability across largest area
- ☐ **Dual-region**  
High availability and low latency across 2 regions


**Location**

europa-west1 (Belgium) ▼

CONTINUE

- **Choose a default storage class for your data**

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. [Learn more](#)

- ☒ **Standard**   
Best for short-term storage and frequently accessed data
- ☐ **Nearline**  
Best for backups and data accessed less than once a month
- ☐ **Coldline**  
Best for disaster recovery and data accessed less than once a quarter
- ☐ **Archive**  
Best for long-term digital preservation of data accessed less than once a year

CONTINUE

- **Choose how to control access to objects**

**Access control**

- ☒ **Fine-grained**  
Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)
- ☐ **Uniform**  
Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

CONTINUE

CREATE

CANCEL

5. Esto nos genera un bucket que podemos utilizar para nuestros cluster.

The screenshot displays the Google Cloud Storage console interface. On the left is a sidebar with navigation options: Storage, Browser, Transfer, Transfer for on-premises, Transfer Appliance, and Settings. The main content area is titled 'Bucket details' for the bucket 'ntic-bucket-name'. It includes tabs for Objects, Overview, Permissions, and Bucket Lock. Action buttons for 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete' are present. A search bar labeled 'Filter by prefix...' is also visible. Below these elements, a message states: 'There are no live objects in this bucket. If you have object versioning enabled, this bucket may contain noncurrent versions of objects, which aren't visible in the console. You can list noncurrent objects by using the gsutil command line or the APIs.'

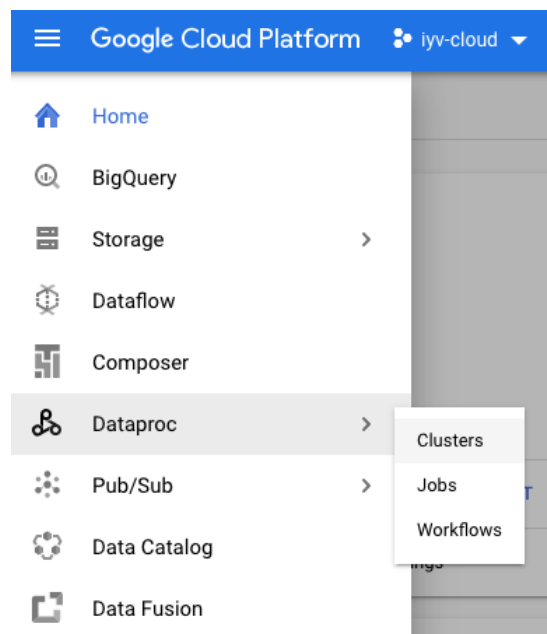
## CREAR UN CLUSTER

Vamos a mostrar los pasos a seguir para crear un cluster de Dataproc con la configuración necesaria para el Master.

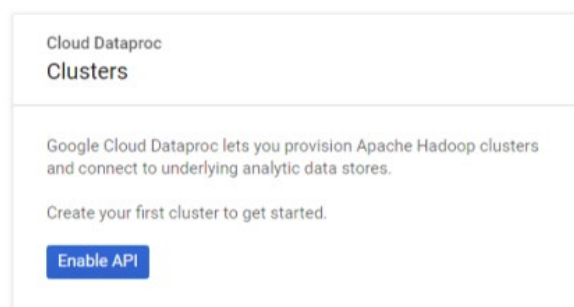
Asumimos que el alumno tiene una cuenta de gmail o Gsuite con acceso a Google Cloud, y está utilizando algún tipo de facturación, como la prueba gratuita de Google Cloud.

### Pasos

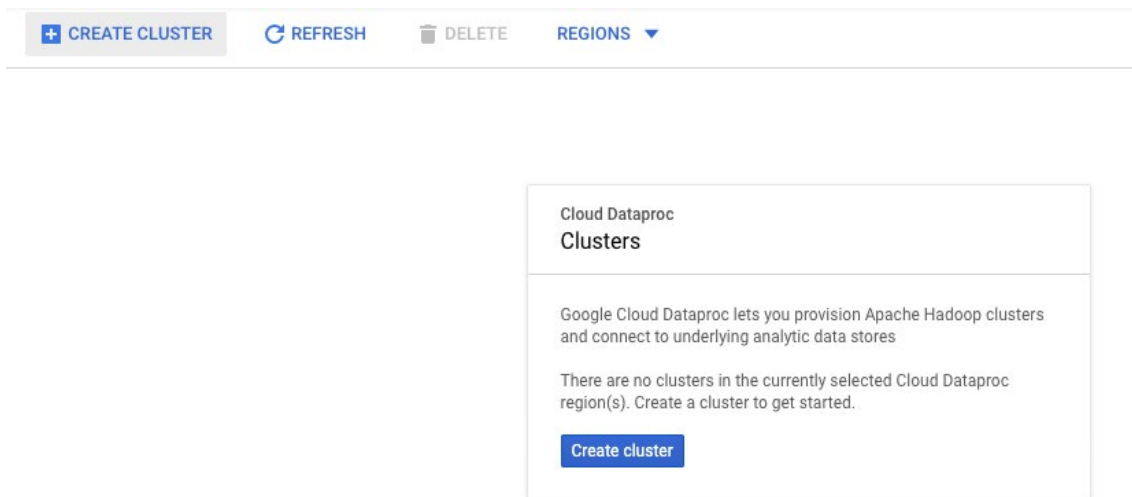
1. Abrir la consola de Google Cloud.
  - a. <https://console.cloud.google.com/>
2. Accedemos a Dataproc.
  - a. Buscamos en la barra de búsqueda (arriba centrado) o usamos el selector gráfico. Dataproc



3. La primera vez que accedemos a Dataproc, nos solicita activar el **API de Dataproc**. Esto es necesario para la creación automática del cluster.



- Una vez habilitada el API, volvemos a Dataproc para crear el cluster. Vemos que existe una interfaz gráfica para hacerlo paso a paso en cualquiera de los dos botones "Create cluster", pero ciertas opciones de personalización que vamos a utilizar no están disponibles desde esa interfaz, por lo tanto NO utilizaremos ninguno de esos dos botones.



Lo que haremos será abrir la línea de comandos de Google Cloud (*Google Cloud Shell*) haciendo click en el botón de la parte superior:

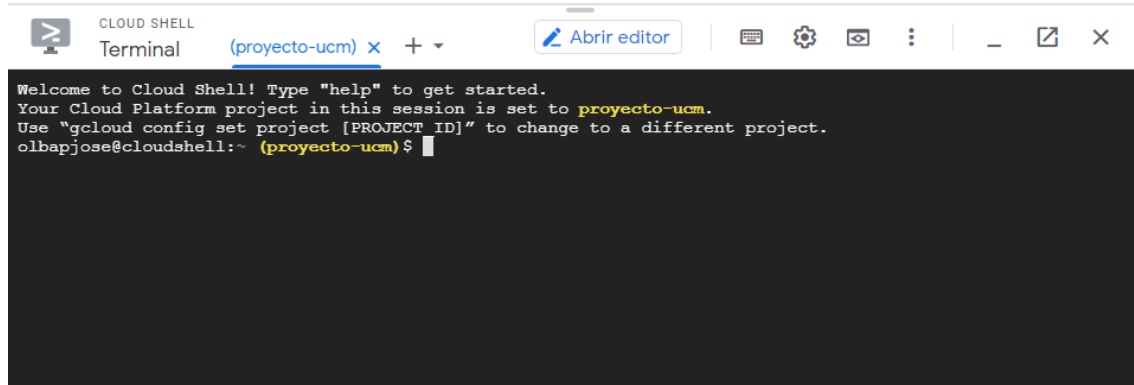


- La forma más rápida de crear un cluster, y que utilizaremos en el curso, es copiar y pegar el siguiente comando en la Google Cloud Shell, cambiando lo necesario (marcado en **rojo**) para adaptarlo al identificador de nuestro proyecto (aparece en amarillo entre paréntesis al abrir la consola de Google Cloud Shell) y al nombre de nuestro bucket de Google Cloud Storage. El nombre del cluster puede elegirse libremente. Está configurado para que el cluster se desmantele pasadas unas 8 horas (30000 segundos).

```
gcloud beta dataproc clusters create nombrecluster --enable-component-gateway --bucket nombrebucket --region europe-west1 --zone europe-west1-c --master-machine-type n1-standard-2 --master-boot-disk-size 500 --num-workers 2 --worker-machine-type n1-standard-2 --worker-boot-disk-size 500 --image-version 1.4-debian10 --properties ^#^spark:spark.jars.repositories=https://repos.spark-packages.org/#spark:spark.jars.packages=graphframes:graphframes:0.7.0-spark2.4-s_2.11,org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.0 --optional-components ANACONDA,JUPYTER,ZOOKEEPER --max-age 30000s --initialization-actions 'gs://goog-dataproc-initialization-actions-europe-west1/kafka/kafka.sh' --project identificadorproyecto
```



Debemos pegar el comando anterior en la Google Cloud Shell, que es la línea de comandos de Google para poder manejar (por ejemplo, crear, destruir, configurar, etc.) cualquier servicio de Google Cloud utilizando comandos específicos de Google. **Cuidado con los símbolos ^^ al copiar y pegar el comando anterior.** Cuando Google lo pregunte, pulsamos en "Autorizar". En la imagen siguiente, el identificador del proyecto sería proyecto-ucm, que se muestra en amarillo.



```
CLOUD SHELL
Terminal (proyecto-ucm) x + Abrir editor

Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to proyecto-ucm.
Use "gcloud config set project [PROJECT ID]" to change to a different project.
olbapjose@cloudshell:~ (proyecto-ucm) $
```

## OPERAR CON UN CLUSTER (ACCESO A JUPYTER Y SSH)

Para poder trabajar con el Cluster, sólo tenéis que hacer clic en el nombre del cluster una vez levantado.

<input type="checkbox"/> Name ^	Region	Zone
<input checked="" type="checkbox"/> ntic-cluster-name	europe-west1	europe-west1-d

Accederás a las opciones de monitorización, Trabajos (Jobs), VM (máquinas virtuales del cluster), Configuración y Interfaces Web.

Para acceder a Jupyter o JupyterLab, solo tienes que ir a las Web Interfaces y hacer clic en la opción deseada.

### SSH tunnel

Create an SSH tunnel to connect to a web interface

### Component gateway

[YARN ResourceManager](#) ↗

[HDFS NameNode](#) ↗

[MapReduce Job History](#) ↗

[YARN Application Timeline](#) ↗

[Spark History Server](#) ↗

[Tez](#) ↗

[Jupyter](#) ↗

[JupyterLab](#) ↗

Equivalent [REST](#)

Para acceder por SSH a la máquina Master, sólo tienes que ir a la pestaña VM Instances y hacer clic en abrir en una ventana del navegador, para poder acceder a una consola como la de la imagen.

Monitoring Jobs <u>VM Instances</u> Configuration Web Interfaces		
Name	Role	
✓ ntic-cluster-name-m	Master	SSH <input type="button" value="v"/>
✓ ntic-cluster-name-w-0	Worker	
✓ ntic-cluster-name-w-1	Worker	

Equivalent [REST](#)

```
ssh.cloud.google.com/projects/iyv-cloud/zones/europe-west1-d/instances/ntic-cluster-name-m?
connected, host fingerprint: ssh-rsa 0 0E:F0:CD:64:74:6D:5A:88:D6:5C:1C:14:1E:BA
3E:41:A3:3D:3E:5B:0C:3A:33:D6:A5:34:66:D1:3C:86:33:4F
linux ntic-cluster-name-m 4.19.0-0.bpo.6-amd64 #1 SMP Debian 4.19.67-2+deb10u2~b
po9+1 (2019-11-12) x86_64

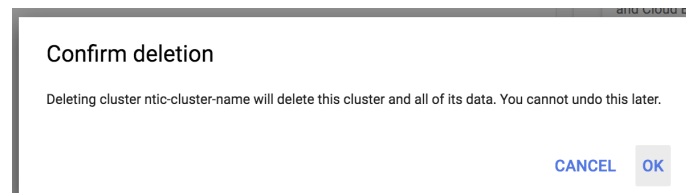
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
smael@ntic-cluster-name-m:~$
```

## BORRAR UN CLUSTER

Para borrar un cluster, sólo tenemos que seleccionarlo en la UI, y hacer clic en DELETE. Nos pedirá confirmar el borrado, y en unos minutos el mismo desaparecerá.

Clusters <span>+ CREATE CLUSTER</span> <span>REFRESH</span> <span>DELETE</span> <span>REGIONS ▼</span>						
<input type="text" value="Search clusters, press Enter"/>						
<input checked="" type="checkbox"/> Name ^	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	
<input checked="" type="checkbox"/> <span>ntic-cluster-name</span>	europe-west1	europe-west1-d	2	On	<u>ntic-bucket-name</u>	



Si hemos marcado en las opciones que nos borre el cluster a las 6h, esta acción será automática.