

Guia de Configuracion de Google Cloud para Máster NTIC de la UCM

Autor: Ismael Yuste

Actualizado Marzo 2020

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)	22
Borrar un Cluster	23
Crear un cluster en línea de comandos	24

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

Dataprocc 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 Dataprocc.

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 Dataprocc 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 Dataprocc, puedes habilitar el modo seguro de Hadoop a través de Kerberos añadiendo una configuración de seguridad. GCP y Dataprocc ofrecen también otras prestaciones de seguridad que contribuyen a proteger tus datos. Algunas de las funciones de seguridad específicas de GCP más utilizadas con

Dataproc 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 utilice¹.

Bélgica (europe-west1) ▾	
Tipo de máquina	Precio
Máquinas estándar 1-64 CPU virtuales	\$0.010 - \$0.640
Máquinas de memoria elevada 2-64 CPU virtuales	\$0.020 - \$0.640
Máquinas con un gran número de CPU 2-64 CPU virtuales	\$0.020 - \$0.640
Máquinas personalizadas Basadas en el uso de vCPU y de memoria	\$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](#).

¹ 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 12 meses y con un crédito de 300\$.

Después de ese periodo, podemos seleccionar activar la facturación para usar Google Cloud de pago para nuestros proyectos.

La url para inicial el alta de la prueba gratuita es cloud.google.com/free.

Necesitaremos una cuenta de gmail, y una tarjeta de crédito, en la que no se hará ningún cargo si no activamos la facturación.

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 maquina de administración preparada por Google

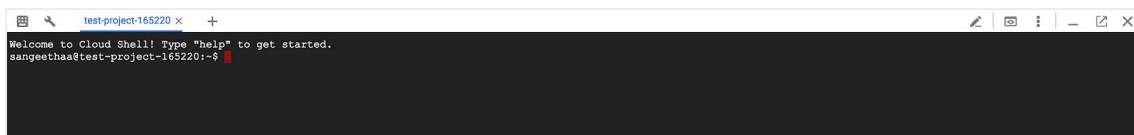
Google Cloud Shell te ofrece acceso a tus recursos en la nube mediante la línea de comandos directamente desde el navegador. Puedes administrar fácilmente tus proyectos y recursos sin tener que instalar en tu sistema el SDK de Google Cloud ni otras herramientas. Con Cloud Shell, la herramienta de línea 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.

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

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

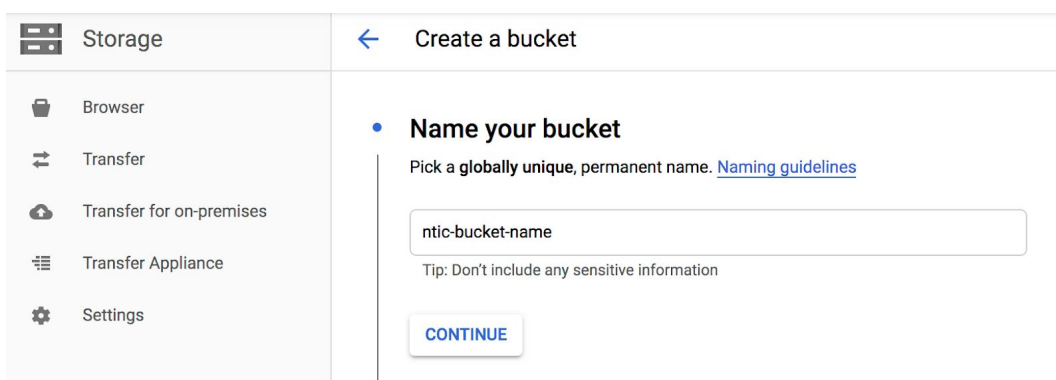
Google Cloud Dataproc

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, 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. Ir a Google Cloud Storage.
4. Seguir los pasos de creación de un Bucket asignando nombre, modo (region), localización (europe-west-1), tipo (standard) y resto de opciones por defecto.



The screenshot shows the Google Cloud Storage 'Create a bucket' page. On the left is a sidebar with a 'Storage' header and a menu containing 'Browser', 'Transfer', 'Transfer for on-premises', 'Transfer Appliance', and 'Settings'. The main content area has a back arrow and the title 'Create a bucket'. Below this is a section titled 'Name your bucket' with a bullet point. It instructs the user to 'Pick a globally unique, permanent name' and provides 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 section is a blue '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 shows the Google Cloud Storage console. On the left is a sidebar with 'Storage' and a 'Browser' section containing 'Transfer', 'Transfer for on-premises', 'Transfer Appliance', and 'Settings'. The main area is titled 'Bucket details' for 'ntic-bucket-name'. It has tabs for 'Objects', 'Overview', 'Permissions', and 'Bucket Lock'. Below the tabs are buttons for 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete'. A search bar says 'Filter by prefix...'. Below that, it says 'Buckets / ntic-bucket-name'. At the bottom, 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.'

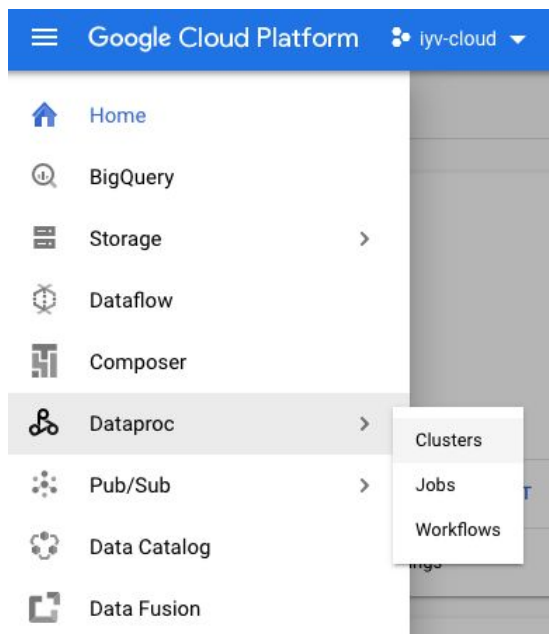
CREAR UN CLUSTER

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

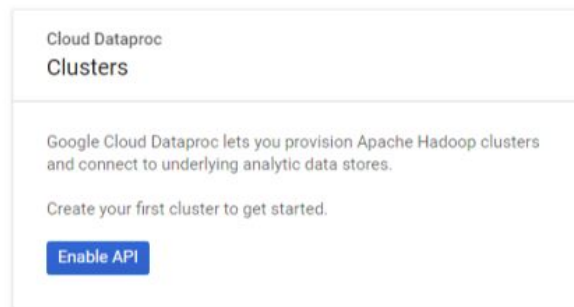
Asumimos que el alumno tiene una cuenta de gmail o Gsuite con acceso a Google Cloud, y está utilizando algún tipo de facturación, como la prueba gratuita de Google Cloud.

Pasos

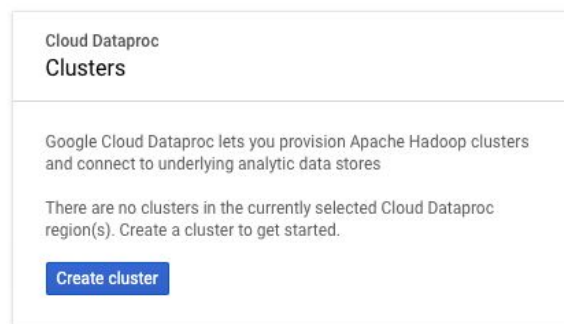
1. Abrir la consola de Google Cloud.
 - a. <https://console.cloud.google.com/>
2. Accedemos a Dataproc.
 - a. Buscamos en la barra de búsqueda (arriba centrado) o usamos el selector gráfico. Dataproc



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



4. Una vez habilitada el API, volvemos a Dataproc y creamos el cluster. Podemos hacerlo en cualquiera de los dos botones “create cluster”



5. Nos aparece una pantalla de creación del cluster como esta.

Dataproc	← Create a cluster
<ul style="list-style-type: none"> Clusters Jobs Workflows 	<p>Name ?</p> <input type="text" value="cluster-3d99"/> <p>Region ? Zone ?</p> <p>us-central1 ▼ us-central1-b ▼</p> <p>Cluster mode ?</p> <p>Standard (1 master, N workers) ▼</p>

6. Empezamos a tomar decisiones sobre la configuración del cluster.
 - a. Nombre del Cluster.
 - b. Región. Recomendamos Europa, especialmente europe-west1.
 - c. Cluster mode. Recomendamos Standard.

Name ?

Region ? **Zone** ?

europa-west1 europa-west1-d

Cluster mode ?

Standard (1 master, N workers)

d. Master: Tipo de máquina. Recomendamos n1-standard-1. Disco 500GB.

Master node

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

Machine configuration ?

Machine family

General-purpose

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-1 (1 vCPU, 3.75 GB memory)

	vCPU	Memory
	1	3.75 GB

⌵ CPU platform and GPU

Primary disk size (minimum 15 GB) ? **Primary disk type** ?

500 GB Standard persistent disk

e. Workers. Tipo de máquina. Recomendamos n1-standard-1. Disco 100GB. Número: 2.

Worker nodes

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

Machine configuration


Machine family


General-purpose

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-1 (1 vCPU, 3.75 GB memory) ▼


 vCPU
1


 Memory
3.75 GB

⌵ CPU platform and GPU

Primary disk size (minimum 15 GB) 
100 GB


Primary disk type 
Standard persistent disk ▼

Nodes (minimum 2) 
2

Local SSDs (0-8) 
0 x 375 GB

Esto genera un cluster con 1 nodo maestro y dos esclavos, y 2 cores de YARN y una memoria de 6 GB.

YARN cores 
2

YARN memory 
6 GB

f. Component gateway: activar.

Autoscaling policy ? (Optional)

☐ Enable autoscaling on the cluster.

This project does not currently have any applicable policy to enable autoscaling in this region. [Learn how to create autoscaling policy.](#)

Component gateway

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

g. Avanzado: (para ello hay que desplegar la opción)

⌵ [Advanced options](#)

Create

Cancel

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

Preemptible worker nodes ?

Each contains a YARN NodeManager. HDFS does not run on preemptible nodes. Machine type is copied from the Worker section.

Nodes ?

0

Enhanced flexibility mode ? (Optional)

☐ Enable enhanced flexibility mode.
Only supported by image version 1.4.

Network ?

default

Subnetwork ?

default (10.132.0.0/20)

Network tags ? (Optional)

Internal IP only

☐ Configure all instances to have only internal IP addresses. [Learn more](#)

- i. Cloud Storage staging bucket. (el creado anteriormente).
- ii. Imagen: 1.4 (Debian 9, Hadoop 2.9, Spark 2.4)
- iii. Optional components. Anaconda y Jupyter.

Cloud Storage staging bucket (Optional) ?

 ntic-bucket-name	Browse
--	--------

Image ?

Cloud Dataproc image version: 1.4 (Debian 9,
Hadoop 2.9, Spark 2.4)
First released on 3/22/2019.

Change

Optional components (Optional)

Install optional open source components on the cluster. [Learn more](#)

Selected components	ANACONDA
Selected components	JUPYTER
<p>Edit</p>	

Initialization actions (Optional) ?

+ Add initialization action

Project access ?

☐ Allow API access to all Google Cloud services in the same project. [Learn more](#)

Cluster properties (Optional) ?

Use cluster properties to add or modify configuration files when creating a cluster.
[Learn more](#)

+ Add cluster property

Metadata (Optional)

Add additional metadata for instances that run in your cluster. [Learn more](#)

+ Add metadata

Advanced Security (Optional)

Enabling Kerberos and Hadoop Secure Mode will provide user authentication, isolation, and encryption inside a Cloud Dataproc cluster. [Learn more](#)

☐ Enable Kerberos and Hadoop Secure Mode.

Labels (Optional) ?

+ Add label

- iv. Schedule deletion: nos permite definir cuando eliminamos el cluster por si nos olvidamos de apagarlo. 6h.

Scheduled deletion (Optional)

[Learn how to use scheduled deletion](#)

- ☒ Delete on a fixed time schedule
☐ Delete cluster at a specified future time
☒ Delete after elapsed time since creation

6 Hours

- ☐ Delete after a cluster idle time period without submitted jobs

The cluster will be deleted 6 hours after creation

Encryption

Data is encrypted automatically. Select an encryption key management solution.

- ☒ Google-managed key
No configuration required
☐ Customer-managed key
Manage via Google Cloud Key Management Service

7. Una vez elegidas todas las opciones, hacemos clic en el boton de Crear para crear el cluster.

[^](#) Less

Create

Cancel

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

8. El cluster se creará en unos 90-120 segundos, y una vez activo, aparecerá como veis en la imagen.

Dataproc

Clusters

Jobs

Workflows

Clusters

CREATE CLUSTER

REFRESH

DELETE

REGIONS

SHOW INFO PANEL

Search clusters, press Enter

<input type="checkbox"/>	Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
<input type="checkbox"/>	ntic-cluster-name	europa-west1	europa-west1-d	2	On	ntic-bucket-name	Feb 25, 2020, 7:09:27 PM	Running

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	europa-west1	europa-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.

Monitoring Jobs VM Instances Configuration Web Interfaces

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	VM Instances	Configuration	Web Interfaces
✓	ntic-cluster-name-m	Master	SSH	
✓	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

CREATE CLUSTER

REFRESH

DELETE

REGIONS

Search clusters, press Enter

<input checked="" type="checkbox"/> Name ^	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <u>ntic-cluster-name</u>	europe-west1	europe-west1-d	2	On	<u>ntic-bucket-name</u>

Confirm deletion

Deleting cluster ntic-cluster-name will delete this cluster and all of its data. You cannot undo this later.

CANCEL OK

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

CREAR UN CLUSTER EN LÍNEA DE COMANDOS

Para automatizar la creación de cluster para cada día, podemos usar la línea de comandos (lanzar comandos bash en Cloud Shell), haciendo esta tarea más ligera, dado que sólo tenemos que cambiar las opciones si nos interesa añadir por ejemplo mas memoria, o más máquinas a nuestro cluster.

Aqui teneis un ejemplo del cluster creado en este tutorial, para que lo uséis.

Las variables en MAYÚSCULAS, las tenéis que sustituir por los valores de vuestro proyecto.

```
gcloud beta dataproc clusters create CLUSTER_NAME
--enable-component-gateway --bucket ntic-bucket-name --region
europe-west1 --subnet default --zone europe-west1-d
--master-machine-type n1-standard-1 --master-boot-disk-size 500
--num-workers 2 --worker-machine-type n1-standard-1
--worker-boot-disk-size 100 --image-version 1.4-debian9
--optional-components ANACONDA,JUPYTER --max-age 21600s
--project PROJECT_NAME
```


ANEXO: CREACIÓN DEL CLUSTER EN LA INTERFAZ ACTUALIZADA

Recientemente Google ha hecho ligeros cambios en la interfaz gráfica con la que se crea el cluster. Incluimos nuevas capturas para evitar confusión.

PASO 1 - SET UP CLUSTER

Configurar las opciones marcadas en rojo. En el caso de *Location*, la Zona concreta se selecciona automáticamente al cambiar la región a *europa-west-1*. Podemos dejar la Zona que se haya marcado por defecto.

Además en la sección **Versioning** hay que cambiar la imagen para que sea versión 1.4-debian10 (aunque 1.4 Ubuntu también debe funcionar).

Por último, en la sección **Components** hay que marcar **Enable component gateway**, y en la parte inferior (Optional components) marcar las casillas **Anaconda** y **Jupyter Notebook**.

The screenshot shows the 'Create a cluster' interface in Google Cloud Platform. The left sidebar has a 'Set up cluster' tab highlighted with a red box. The main form has several sections with red boxes highlighting specific options:

- Name:** Cluster Name (uomcluster)
- Location:** Region (europa-west-1) and Zone (europa-west-1-b) are highlighted with a red box.
- Cluster type:** Standard (1 master, N workers) is selected.
- Autoscaling:** Policy (None)
- Versioning:** Image Type and Version (1.4-debian10) is highlighted with a red box. Below it, the Release Date (First released on 22/03/2019) and a CHANGE button are also highlighted with a red box.
- Components:** Component Gateway (Enable component gateway) is checked and highlighted with a red box.
- Optional components:** Anaconda and Jupyter Notebook are checked and highlighted with a red box.

PASO 2 – CONFIGURE NODES

Sólo es necesario configurar los nodos para **n1-standard-1**, que es una configuración más austera y más barata que la que viene seleccionada por defecto. El resto de opciones se pueden dejar sin tocar.

← Create a cluster

- Set up cluster
Create Cluster
- Configure nodes (optional)**
Nodes
- Customise cluster (optional)
Customise
- Manage security (optional)
Manage network security

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-OPTIMISED

MEMORY-OPTIMISED

Machine types for common workloads, optimised for cost and flexibility


Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-1 (1 vCPU, 3.75 GB memory)



vCPU

1

Memory

3.75 GB

✓ CPU PLATFORM AND GPU

Primary disk size (min 10 GB)

500

GB

Primary disk type

Standard Persistent Disk

Number of local SSDs *

0

x 375GB

Worker nodes

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

Machine family

GENERAL-PURPOSE

COMPUTE-OPTIMISED

MEMORY-OPTIMISED

Machine types for common workloads, optimised for cost and flexibility


Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-1 (1 vCPU, 3.75 GB memory)




vCPU

1

Memory

3.75 GB

26

 **ntic**
master
revolucionamos la comunicación

Dr. Pablo J. Villacorta Iglesias

PASO 3 – CUSTOMIZE CLUSTER

Aquí hay que marcar la **Scheduled deletion** (borrado planificado) con la casilla **Delete on a fixed time schedule** para que Dataproc desmonte automáticamente el cluster pasado un cierto período de tiempo. Es aconsejable fijar ese tiempo en, al menos, 6 horas para que esté levantado durante toda la sesión de clase, aunque en la imagen siguiente hemos fijado 2 h.

FUNDAMENTAL: Además, en la parte inferior, hay que fijar como **File storage** el bucket (espacio de almacenamiento permanente) de Google Cloud Storage que habíamos creado al principio, antes de empezar a crear el cluster. Pinchamos en **Browse** y marcamos nuestro bucket.

← Create a cluster

- Set up cluster
Create Cluster
- Configure nodes (optional)
Nodes
- **Customise cluster (optional)**
Customise
- Manage security (optional)
Manage network security

CREATE CANCEL

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

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.

+ ADD LABELS

Cluster properties

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

+ ADD PROPERTIES

Initialisation actions

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

+ ADD INITIALISATION ACTION

Custom cluster metadata

Add custom metadata to cluster instances. [Learn more](#)

+ ADD METADATA

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 cluster at a specified future time

☒ Delete after elapsed time since creation

Timeout * 2 Hours ▾

☐ Delete after a cluster idle time period without submitted jobs

The cluster will be deleted 2 hours after creation

File storage

File storage bucket ucmbucket **BROWSE**

Optional cloud storage bucket to be used for storing cluster job dependencies, job driver output and cluster configuration files.

PASO 4 - MANAGE SECURITY: en esta sección no hay que cambiar ninguna opción.