

# GOOGLE CLOUD PLATAFORM, HADOOP, SPARK

Proceso de creación de un clúster de Hadoop, y ejecución de un trabajo de Spark en el clúster creado

GCP, Hadoop y Spark

## Contenido

Google Cloud Plataform, Hadoop, Spark.....	2
Inicio de sesión para obtener 300 dolares de crédito.....	2
Creación de un cluster de Hadoop en GCP .....	5
requisito.....	5
Configuración de la red .....	5
Probar el clúster.....	10
Agregar un Spark job .....	12
Referencias.....	16

Cristian ARM

## Google Cloud Plataform, Hadoop, Spark

Inicio de sesión para obtener 300 dolares de crédito

Ingresar a la URL <https://console.developers.google.com/freetrial> donde le solicitara una cuenta de Google



Google

Acceder

Ir a Google Cloud Platform

Correo electrónico o teléfono

¿Olvidaste el correo electrónico?

¿Esta no es tu computadora? Usa una ventana privada para acceder: [Más información](#)

[Crear cuenta](#) [Siguiente](#)

Seleccionar País y Aceptar condiciones y continuar



Google Cloud Platform

Te damos la bienvenida, [redacted]

Crea y administra tus instancias, discos, redes y otros recursos de Google Cloud Platform desde un solo lugar.

País

Colombia

Condiciones del Servicio

☒ Acepto las [Condiciones del Servicio de Google Cloud Platform](#) y las de [las API y los servicios aplicables](#).

Actualizaciones por correo electrónico

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

[ACEPTAR Y CONTINUAR](#)

Diligencie el formulario para activar la prueba gratis

Prueba Google Cloud Platform de manera gratuita

## Paso 1 de 2



[Redacted Name]

[CAMBIAR DE CUENTA](#)

País

Colombia

Condiciones del Servicio

☒ [Leí y acepto las Condiciones del Servicio de la prueba gratuita de Google Cloud Platform.](#)

Debes seleccionar para continuar

[CONTINUAR](#)

### Accede a todos los productos de Cloud Platform

Obtén todo lo que necesitas para compilar y ejecutar tus apps, sitios web y servicios, incluidos Firebase y la API de Google Maps.

### Obtén \$300 en crédito gratis

Regístrate y obtén \$300 para gastar en Google Cloud Platform durante los próximos 12 meses.

### Sin cargos automáticos cuando finaliza la prueba gratuita

Te pedimos tu tarjeta de crédito para asegurarnos de que no eres un robot. No se te cobrará a menos que se actualice a una cuenta de pagos de forma manual.

Diligenciar el tipo de cuenta que desea usar y los datos personales

Prueba Google Cloud Platform de manera gratuita

## Paso 2 de 2

Información del cliente



Tipo de cuenta

Individual



Nombre y dirección

[Redacted Address]

Colombia

Tipo de pago



Pagos automáticos

Pagará este servicio solo después de acumular costos, mediante un cargo automático que se realizará si alcanza su límite de facturación o 30 días después de su último pago automático (lo que ocurra primero).

Forma de pago



Número de tarjeta



[Redacted Card Number]

MM

AA

CVC

Nombre del titular

[Redacted Name]

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

[INICIAR PRUEBA GRATUITA](#)

### Accede a todos los productos de Cloud Platform

Obtén todo lo que necesitas para compilar y ejecutar tus apps, sitios web y servicios, incluidos Firebase y la API de Google Maps.

### Obtén \$300 en crédito gratis

Regístrate y obtén \$300 para gastar en Google Cloud Platform durante los próximos 12 meses.

### Sin cargos automáticos cuando finaliza la prueba gratuita

Te pedimos tu tarjeta de crédito para asegurarnos de que no eres un robot. No se te cobrará a menos que se actualice a una cuenta de pagos de forma manual.

Se generará un cargo equivalente a 1 dólar por validación de la tarjeta de crédito

áreas comunes para ponerte en marcha rápidamente

Qué se aborda

LISTA D

Google Cloud Platform

**Te damos la bienvenida,** [REDACTED]

Gracias por registrarte. Tu prueba gratuita incluye un crédito de \$300 para gastar durante los próximos 12 meses. Si te quedas sin crédito, no te preocupes; no se te cobrará a menos que [actives la facturación automática](#).

ENTENDIDO

tos

s de C

Otras opciones de procesamiento populares

n, los créditos  
uctos  
proyecto  
ar los precios

Cristian ARM

## Creación de un cluster de Hadoop en GCP

### requisito

Obtener la ip publica con que navega en internet, en cualquier buscador pregunta por cual es mi IP

### Configuración de la red

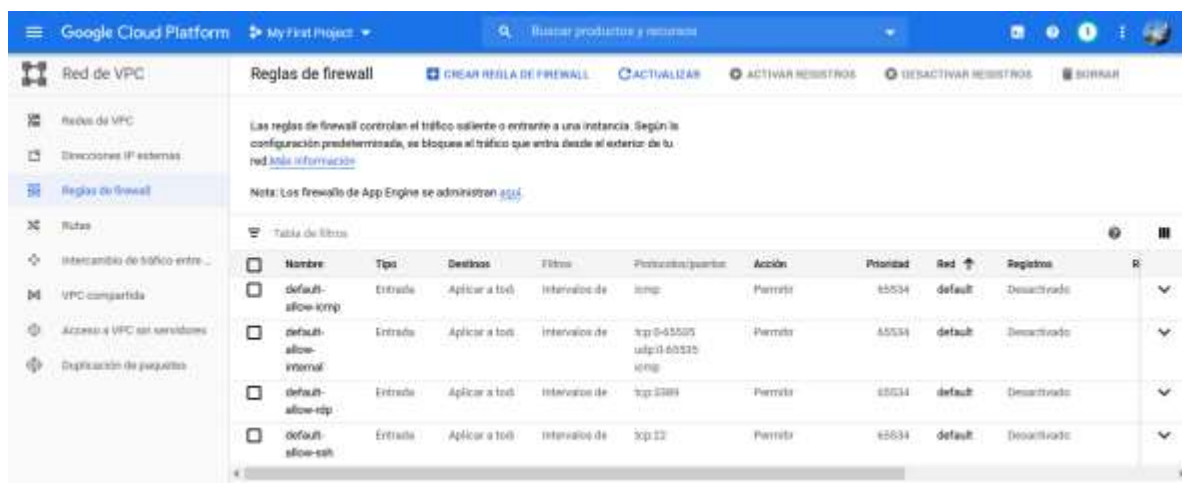
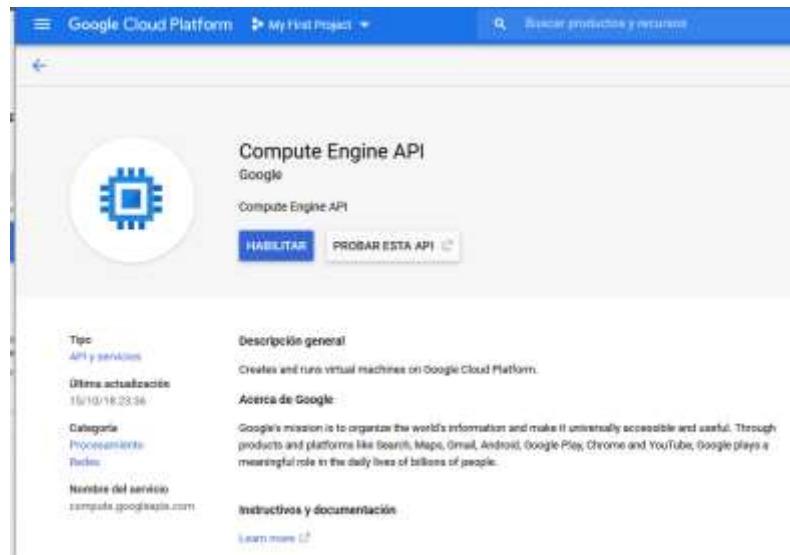
Ingresamos al dashboard en **console.cloud.google.com**



Damos clic en el menú hamburguesa, buscamos Networking, VPC network y luego Firewall Rules



En el caso que no lo tenga habilitado solicitar autorización



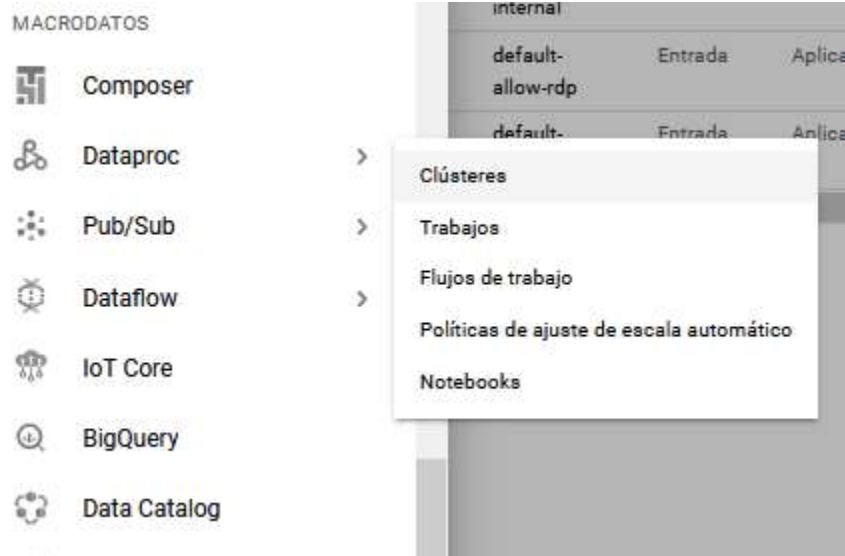
Creamos una nueva regla, tener en cuenta:

- Colocar un nombre
- En la parte de destinos, seleccionar "All instances in the network"
- En la parte de filtros de fuentes, colocar la ip que obtuvo al inicio
- Para los protocolos, colocar 8088,50070,8080 en puertos tcp específicos
- Click en crear

## Configuración del Dataproc

Ir al menú hamburguesa Biga Data, Dataproc y seleccionar cluster





Crear un nuevo cluster con las siguientes condiciones

- Asignar un nombre
- La localización deje la sugerida
- Para el nodo principal,
  - para la maquina seleccione Serie N1, n1-standart-4
  - Para almacenamiento 15 GB, persistencia estándar
- Para el nodo principal,
  - para la maquina seleccione Serie N1, n1-standard-4
  - Para almacenamiento 15 GB, persistencia estándar
  - Cantidad de nodos mínimo 2

**Nodo principal**  
Contiene el administrador de recursos YARN, HDFS NameNode y todos los controladores de trabajos

**Configuración de la máquina**

**Familia de máquinas**

Uso general

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

**Series**

N1

Con la tecnología de la plataforma de CPU Intel Skylake o uno de sus predecesores

**Tipo de máquina**

n1-standard-4 (4 CPU virtuales, 15 GB de memoria)



CPU virtual	Memoria
4	15 GB

Plataforma de CPU y GPU

Tamaño del disco principal (mínimo 15 GB) 15 GB

Tipo de disco principal Disco persistente estándar

**Nodos trabajadores**

Cada uno contiene un YARN NodeManager y un HDFS DataNode.  
El factor de replicación HDFS es 2.

**Configuración de la máquina****Familia de máquinas****Uso general**

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

**Series**

N1

Con la tecnología de la plataforma de CPU Intel Skylake o uno de sus predecesores

**Tipo de máquina**

n1-standard-4 (4 CPU virtuales, 15 GB de memoria)



CPU virtual  
4

Memoria  
15 GB

**Plataforma de CPU y GPU****Tamaño del disco principal (mínimo 15 GB)**

15

GB

**Tipo de disco principal**

Disco persistente estándar

**Nodos (mínimo 2)**

2

**SSD locales (0 a 8)**

0

x 375 GB

**Núcleos YARN**

8

**Memoria YARN**

24 GB

Dar click en crear, despues de unos segundos, aparecera en verde indicando que finalizo

Google Cloud Platform My First Project Buscar productos y recursos

Dataproc Clústeres + CREAR CLÚSTER ACTUALIZAR BORRAR REGIONES OCULTAR PANEL DE INFORMACIÓN

Busca clústeres y presiona Intro

<input type="checkbox"/>	Nombre ↑	Región	Zona	Total de nodos trabajadores	Eliminación
<input type="checkbox"/>	cluster-a7ee	us-central1	us-central1-b	2	Desactivado

No se seleccionaron clústeres.

PERMISOS ETIQUETAS

Selección por lo menos un recurso.

Cristian ARM

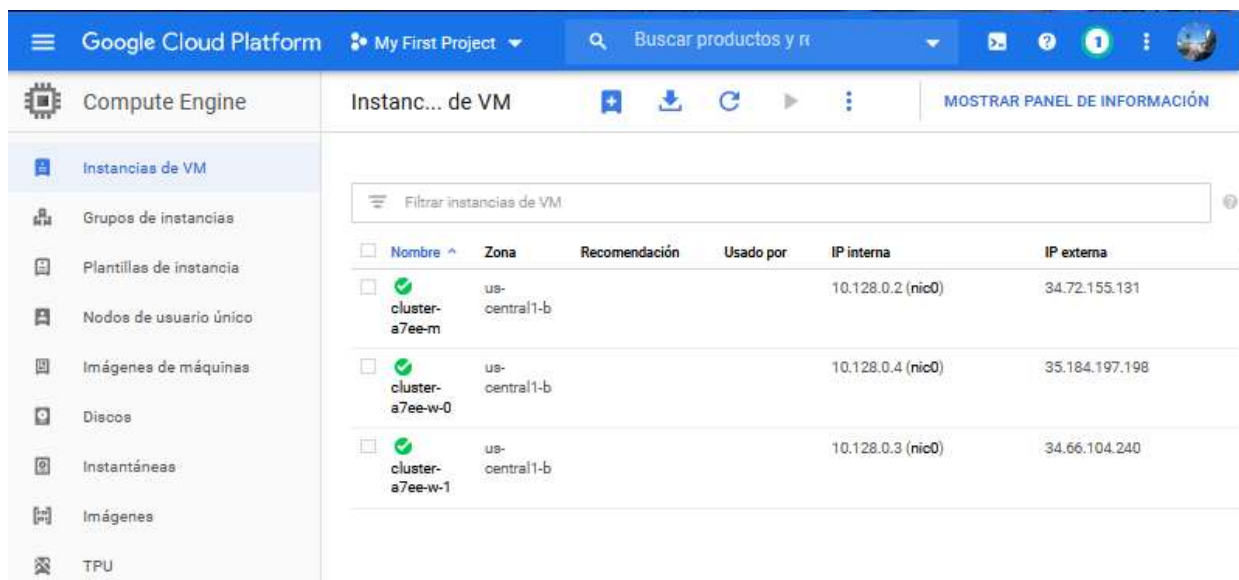
## Probar el clúster

En el menú hamburguesa, seleccione procesamiento, Compute Engine

### PROCESAMIENTO

-  App Engine
-  Compute Engine
-  Kubernetes Engine
-  Cloud Functions

Mire que la ip externa de su nodo maestro (esta en el listado de su cluster que termina en -m)



The screenshot shows the Google Cloud Platform interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'My First Project', a search bar, and notification icons. The left sidebar shows the 'Compute Engine' menu with options like 'Instancias de VM', 'Grupos de instancias', 'Plantillas de instancia', 'Nodos de usuario único', 'Imágenes de máquinas', 'Discos', 'Instantáneas', 'Imágenes', and 'TPU'. The main content area is titled 'Instancias de VM' and displays a table of VM instances. The table has columns for 'Nombre', 'Zona', 'Recomendación', 'Usado por', 'IP interna', and 'IP externa'. Three instances are listed, all with a green checkmark icon, indicating they are healthy. The first instance is 'cluster-a7ee-m', the second is 'cluster-a7ee-w-0', and the third is 'cluster-a7ee-w-1'. The first instance is the master node, as indicated by the '-m' suffix.

Nombre	Zona	Recomendación	Usado por	IP interna	IP externa
cluster-a7ee-m	us-central1-b			10.128.0.2 (nic0)	34.72.155.131
cluster-a7ee-w-0	us-central1-b			10.128.0.4 (nic0)	35.184.197.198
cluster-a7ee-w-1	us-central1-b			10.128.0.3 (nic0)	34.66.104.240

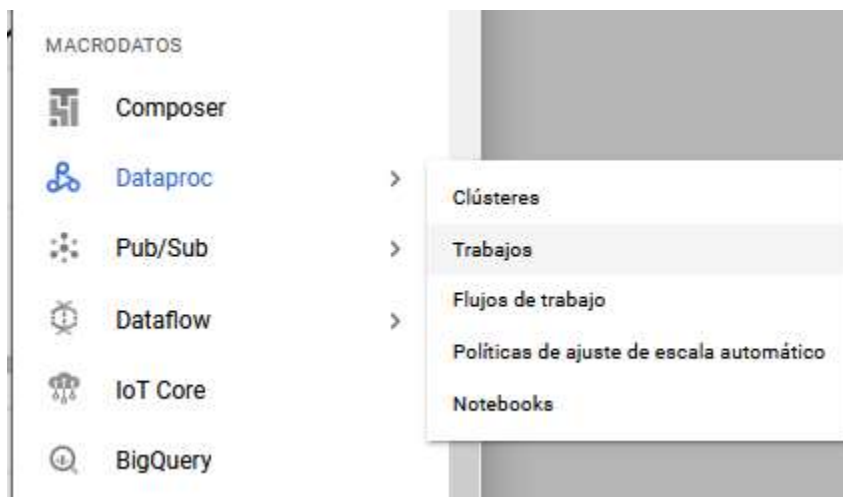
Desde su navegador digite la dirección externa con el puerto de la aplicación



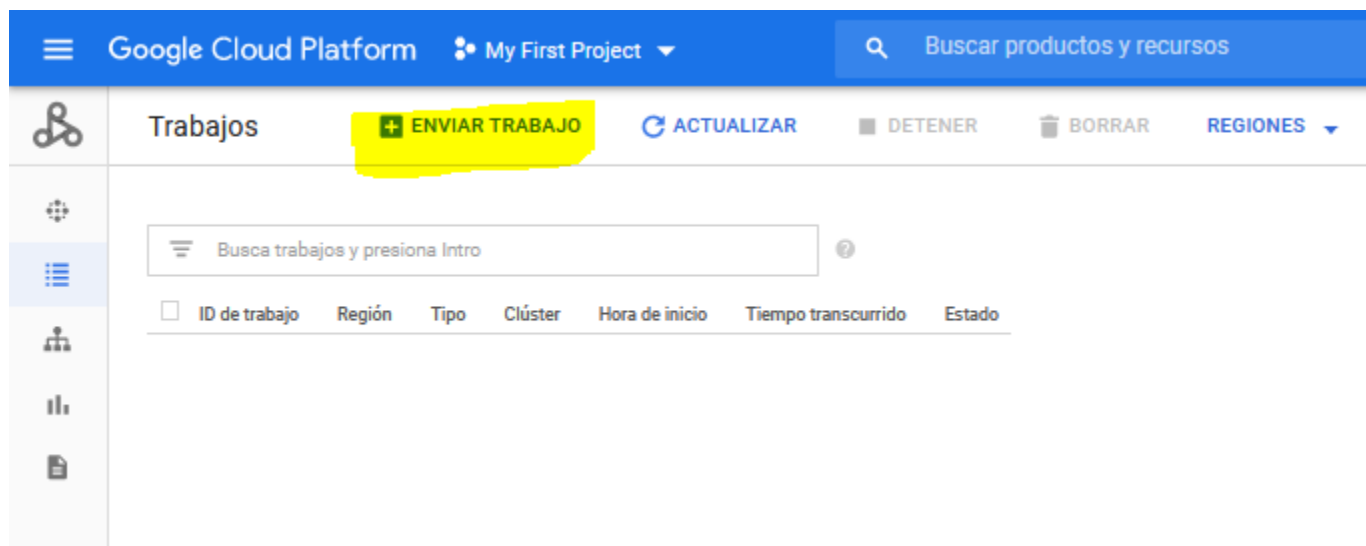
Cristian ARM

## Agregar un Spark job

Ingresar a Macrodatos, Dataproc y Trabajos



Seleccionar enviar trabajo



Diligencias los siguientes datos

- Id del tranabjo lo puede cambiar o dejar le asignado
- La región debe se la misma de donde creo el cluster
- En Cluster, seleccione el nombre del cluster
- En tipo de trabajo seleccione Spark
- En archivo jar o clase principal, digite `org.apache.spark.examples.SparkPi`
- En argumentos digite 1000
- En archivo jar digite <file:///usr/lib/spark/examples/jars/spark-examples.jar>

Google Cloud Platform My First Project 🔍 Buscar productos y recursos

← Enviar un trabajo

ID de trabajo  
job-ea46f0a1

Región  
us-central1

Clúster  
cluster-a7ee

Tipo de trabajo  
Spark

Archivo jar o clase principal  
org.apache.spark.examples.SparkPi

Argumentos (Opcional)  
1000  
Presiona <Intro> para agregar más argumentos

Archivos jar (Opcional)  
file:///usr/lib/spark/examples/jars/spark-examples.jar  
Ingresa la ruta del archivo; por ejemplo, hdfs://ejemplo/ejemplo.jar

Propiedades (Opcional)  
+ Agregar elemento

Etiquetas (Opcional)  
+ Agregar etiqueta

Cantidad máxima de reinicios por hora (Opcional)  
Deja el campo en blanco si no deseas que el sistema se reinicie automáticamente tras un error en el trabajo. [Más información](#)  
Entre 1 y 10

Enviar Cancelar

REST equivalente

Dar click, en las lista de tareas aparecen las tareas creadas, cuando termine de lo indicara el estado

Google Cloud Platform My First Project 🔍 Buscar productos y recursos




Trabajos + ENVIAR TRABAJO ACTUALIZAR DETENER BORRAR REGIONES






Busca trabajos y presiona Intro


<input type="checkbox"/>	ID de trabajo	Región	Tipo	Clúster	Hora de inicio	Tiempo transcurrido	Estado
<input checked="" type="checkbox"/>	job-ea46f0a1	us-central1	Spark	cluster-a7ee	29 may. 2020 12:17:57	33 s	Correcto

Cristian ARM

De click en el trabajo creado para ver la salida, si quiere ajustar el texto seleccione la casilla ajuste de línea, se puede apreciar el valor de pi

 Detalles del trabajo ACTUALIZAR CLONAR



 job-ea46f0a1

Hora de inicio: 29 may. 2020 12:17:57

Tiempo transcurrido: 33 s

Estado:

Resultados

Configuración

☒ Ajuste de línea

Línea de comandos equivalente

20/05/29 17:18:02 INFO org.spark\_project.jetty.util.log: Logging initialized @3129ms

20/05/29 17:18:02 INFO org.spark\_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown

20/05/29 17:18:02 INFO org.spark\_project.jetty.server.Server: Started @83262ms

20/05/29 17:18:02 INFO org.spark\_project.jetty.server.AbstractConnector: Started ServerConnector@771db12c[HTTP/1.1,[http/1.1]]{0.0.0.0:4040}

20/05/29 17:18:02 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pool

is in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.

20/05/29 17:18:03 INFO org.apache.hadoop.yarn.client.RMFProxy: Connecting to ResourceManager at cluster-a7ee-m/10.128.0.2:8032

20/05/29 17:18:04 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at cluster-a7ee-m/10.128.0.2:10200

20/05/29 17:18:06 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application\_1590769029914\_0001

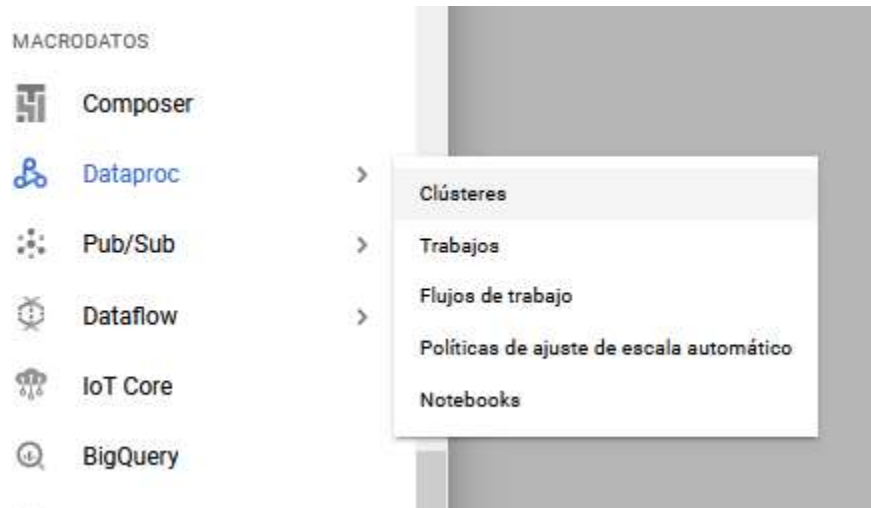
Pi is roughly 3.14140199141402

20/05/29 17:18:26 INFO org.spark\_project.jetty.server.AbstractConnector: Stopped Spark@771db12c[HTTP/1.1,[http/1.1]]{0.0.0.0:4040}

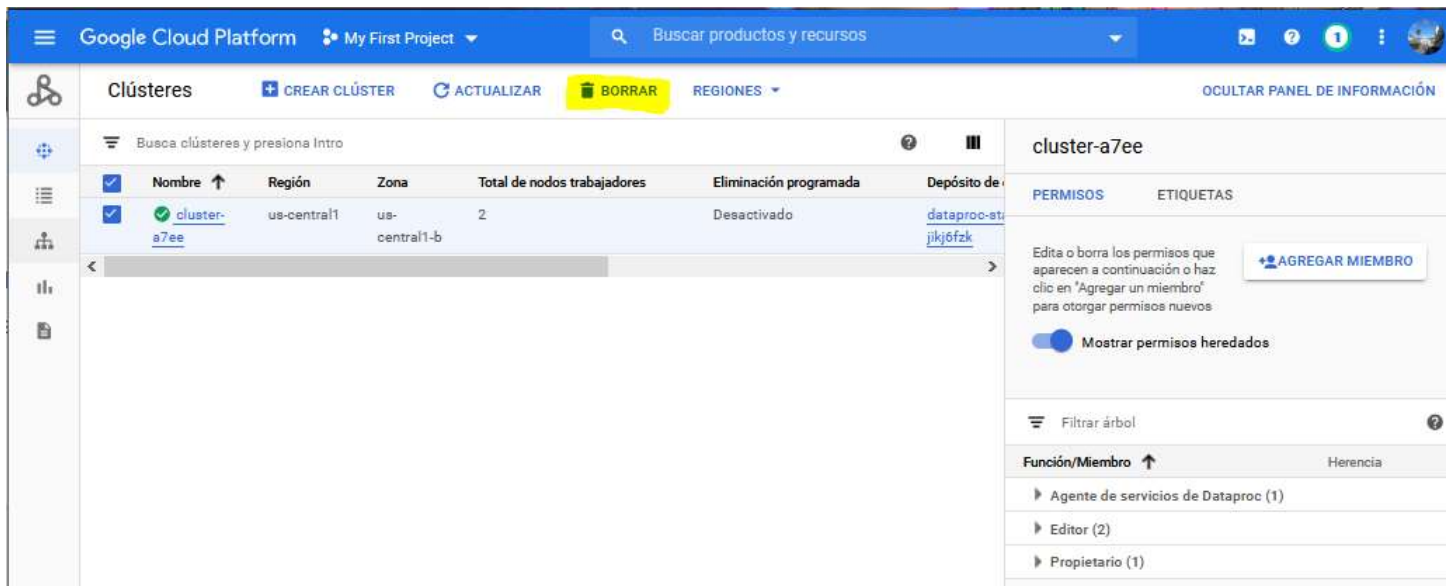
La salida del trabajo está completa

Cristian ARM  
Apagado de clusters

Air Macrodatos, DataProc, Cluster



Seleccionar el proyecto y dar click en borrar



## Confirmar eliminación

Si borras el clúster cluster-a7ee, se borrará este clúster y todos sus datos. No se puede deshacer esta acción.

CANCELAR CONFIRMAR



Cristian ARM

## Referencias

<https://medium.com/@rmache/big-data-with-spark-in-google-colab-7c046e24b3>

<https://medium.com/google-cloud/launch-a-hadoop-cluster-in-90-seconds-or-less-in-google-cloud-dataproc-b3acc1c02598>

<https://codelabs.developers.google.com/codelabs/cloud-dataproc-starter/index.html?index=..%2F..index#4>

<https://colab.research.google.com/drive/1EcotODzgSnLozSH3hDuBfZro6gJXY8lo>

<https://hackernoon.com/why-dataproc-googles-managed-hadoop-and-spark-offering-is-a-game-changer-9foed183fda3>

[https://colab.research.google.com/github/asifahmed90/pyspark-ML-in-Colab/blob/master/PySpark\\_Regression\\_Analysis.ipynb#scrollTo=sq8U3BtmhtRx](https://colab.research.google.com/github/asifahmed90/pyspark-ML-in-Colab/blob/master/PySpark_Regression_Analysis.ipynb#scrollTo=sq8U3BtmhtRx)

<https://gist.github.com/yahwang/d4086d8coca806a9d056d7efd709e2e6>