

**CIENCIA DE DATOS:**  
**APRENDE LOS FUNDAMENTOS**  
**DE MANERA PRÁCTICA**



**LABORATORIO 01**  
**PREPARAR AMBIENTES**

**Juan Chipoco**  
mindquasar@gmail.com

# ÍNDICE

OBJETIVO.....	4
INSTALACION CLUSTER.....	5

## OBJETIVO



Realizar la instalacion de un Cluster Hadoop para realizar los laboratorios subsiguientes.

Dependiendo de los recursos de hardware del alumno la instalacion puede realizarse on-premise usando Cloudera o en la nube usando GCP (Google Cloud Platform). Las practicas subsiguientes se podran realizar en al ambiente que el alumno haya escogido.

### ***Requisitos para instalacion Cloudera en PC o laptop:***

La forma más fácil de instalar Hadoop es utilizar una de las máquinas virtuales que proporciona Cloudera en su pagina web. Las máquinas virtuales (*Cloudera QuickStart VMs*) traen todo el entorno ya configurado, ahorrando mucho tiempo. Están disponibles para VMWare, KVM y VirtualBox.

- Si su laptop tiene 12 GB RAM o 16GB RAM asignen 8 GB al cloudera quickstart para cada maquina virtual.
- Si su laptop tiene 8 GB RAM asignen 4 GB al cloudera quickstart para cada maquina virtual.

El problema fundamental de dichas máquinas virtuales es que requieren bastante memoria RAM (recomendado un mínimo de **4GB dedicados al guest** según Cloudera). Aunque pueden funcionar asignándoles menos memoria, el desempeño se reduce bastante y notarás más esperas.

La otra opción es, si dispones de un sistema operativo GNU/Linux, instalar Hadoop. Una forma fácil de realizar la instalación es utilizar *Cloudera manager installer*.

### ***Requisitos para instalacion en Google Cloud Platform:***

Tener un cuenta de gmail.

Activar la cuenta en GCP, se requiere una tarjeta de debito. Google solo validara la tarjeta sin cobrar nada y nos dara 300 dolares gratis para probar sus herramientas durante 3 meses. Aquí realizaremos la instalacion de Dataproc con 1 master y 2 workers.

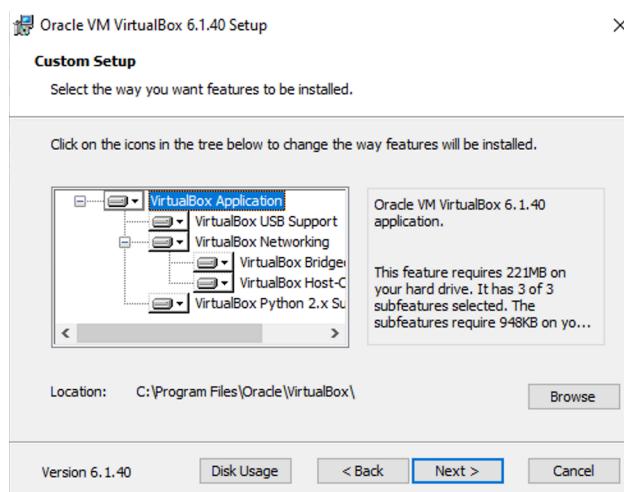
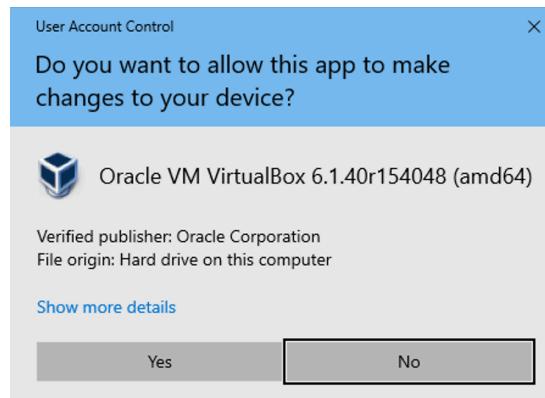
## INSTALACION CLUSTER

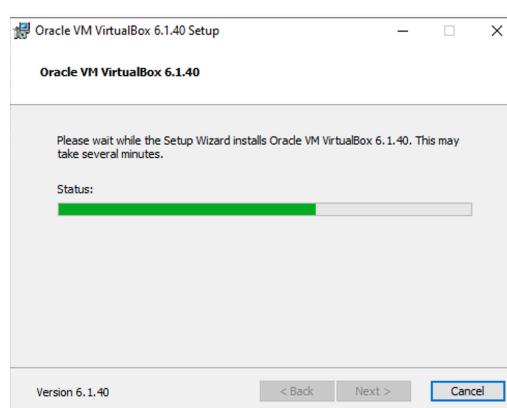
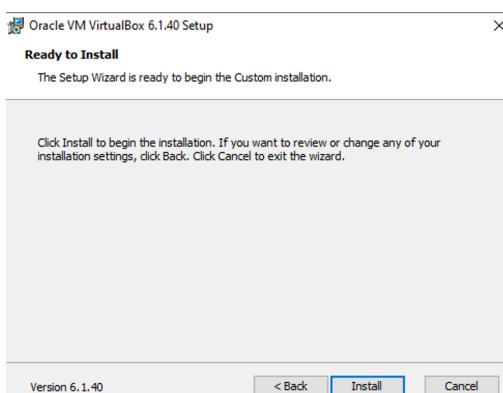
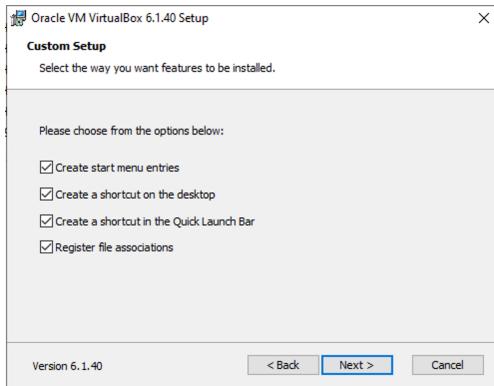
*Instalacion from scratch*

Primero bajaremos virtual box desde el site de Oracle:

<https://www.virtualbox.org/wiki/Downloads>

Seleccionamos Windows hosts:

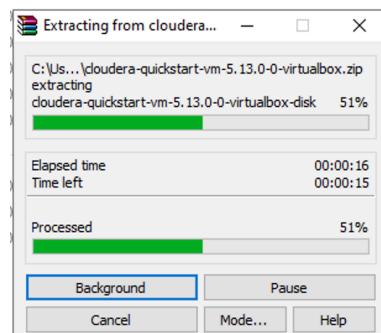






A continuacion bajemos la version quick start de cloudera:

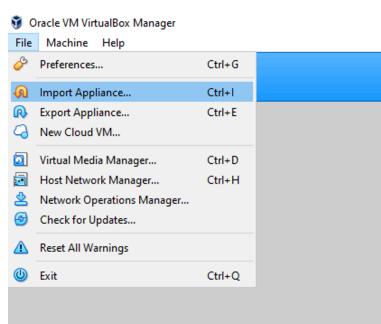
Lo descargamos y desempaqueamos con winrar:

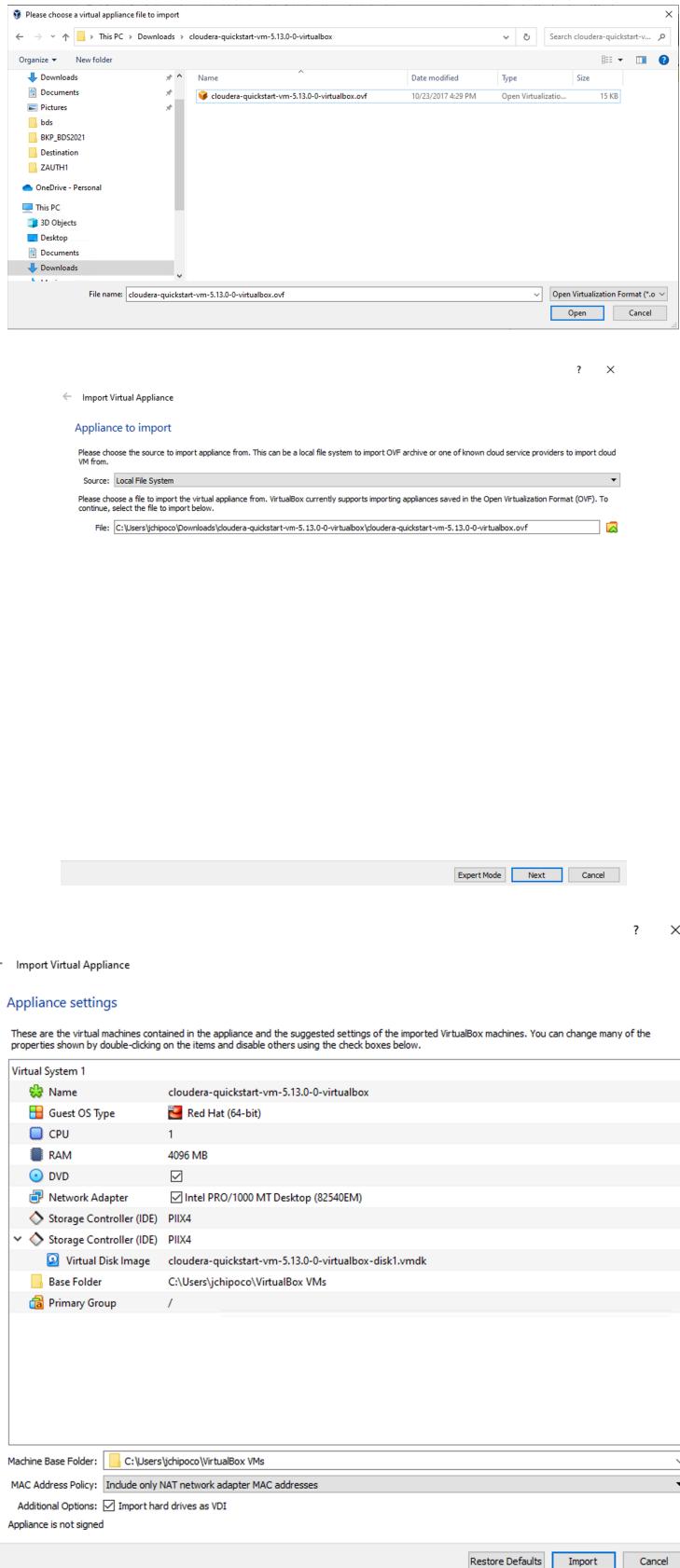


Una vez desempaqueado tendremos los siguientes archivos:

(C:) > Users > jchipoco > Downloads > cloudera-quickstart-vm-5.13.0-0-virtualbox	
Name	Date modified
cloudera-quickstart-vm-5.13.0-0-virtualbox.ovf	10/23/2017 4:29
cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vmdk	10/23/2017 4:34

Ahora regresamos a nuestra Oracle VM

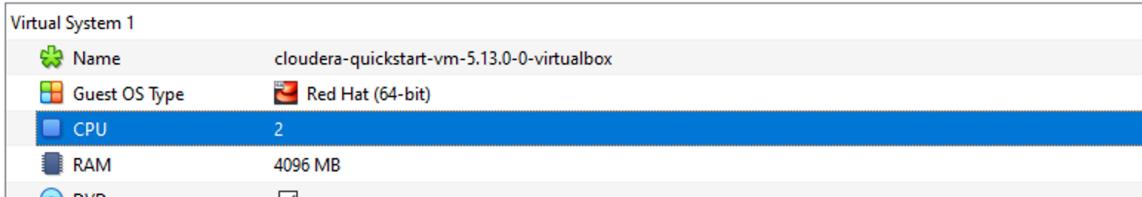




← Import Virtual Appliance

### Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.



Virtual System 1	
Name	cloudera-quickstart-vm-5.13.0-0-virtualbox
Guest OS Type	Red Hat (64-bit)
CPU	2
RAM	4096 MB
DVD	

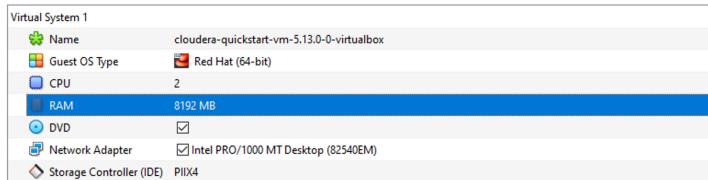
Si su laptop tiene 12 GB RAM o 16GB RAM asignen 8 GB al cloudera quickstart

Si su laptop tiene 8 GB RAM asignen 4 GB al cloudera quickstart

← Import Virtual Appliance

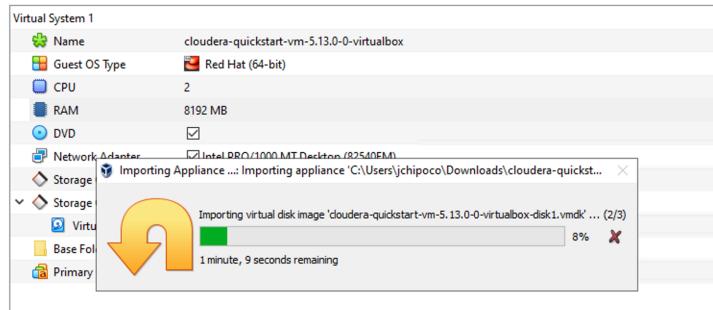
### Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.



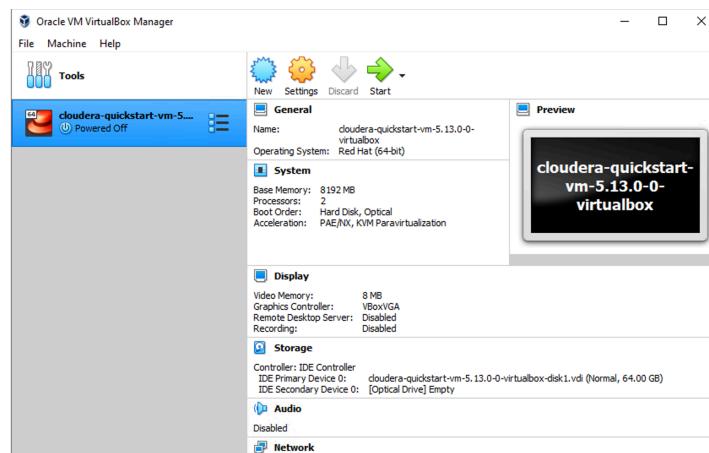
Virtual System 1	
Name	cloudera-quickstart-vm-5.13.0-0-virtualbox
Guest OS Type	Red Hat (64-bit)
CPU	2
RAM	8192 MB
DVD	<input checked="" type="checkbox"/>
Network Adapter	<input checked="" type="checkbox"/> Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)	PiIX4

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.



Virtual System 1	
Name	cloudera-quickstart-vm-5.13.0-0-virtualbox
Guest OS Type	Red Hat (64-bit)
CPU	2
RAM	8192 MB
DVD	<input checked="" type="checkbox"/>
Network Adapter	<input checked="" type="checkbox"/> Intel PRO/1000 MT Desktop (82540EM)
Storage	<ul style="list-style-type: none"> <li>Importing Appliance ... Importing appliance 'C:\Users\jchipoco\Downloads\cloudera-quickstart-vm-5.13.0-0-virtualbox' ... (2/3)</li> </ul>
Base Folder	Primary

Importing virtual disk image 'cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vmdk' ... (2/3)  
 8%   
 1 minute, 9 seconds remaining



Oracle VM VirtualBox Manager

File Machine Help

Tools New Settings Discard Start

cloudera-quickstart-vm-5... Powered Off

**General**

- Name: cloudera-quickstart-vm-5.13.0-0-virtualbox
- Operating System: Red Hat (64-bit)

**System**

- Base Memory: 8192 MB
- Processor: 2
- Boot Order: Hard Disk, Optical
- Acceleration: PAE/NV, KVM Paravirtualization

**Display**

- Video Memory: 8 MB
- Graphics Controller: VBoxVGA
- Remote Desktop Server: Disabled
- Recording: Disabled

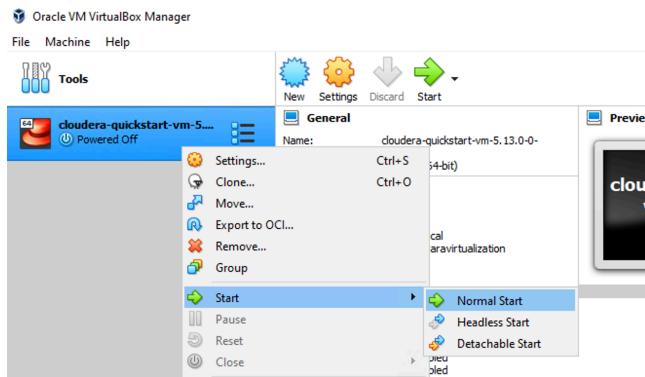
**Storage**

- Controller: IDE Controller
- IDE Primary Device 0: cloudera-quickstart-vm-5.13.0-0-virtualbox-disk1.vdi (Normal, 64.00 GB)
- IDE Secondary Device 0: [Optical Drive] Empty

**Audio**

- Disabled

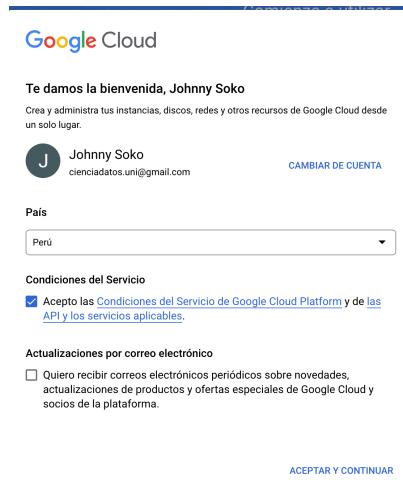
**Network**



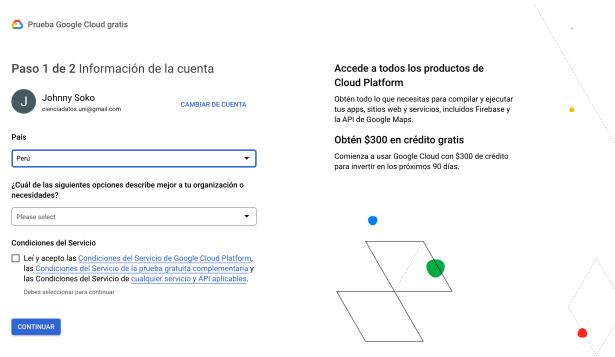
## Instalacion en GCP

Una vez creada una cuenta de gmail vamos a la siguiente ruta:

<https://console.cloud.google.com/>



Presionamos Activar en la parte superior derecha para activar la cuenta



### Paso 1 de 2 Información de la cuenta

Johnny Soko  
cieniciadatos.uni@gmail.com [CAMBIAR DE CUENTA](#)

País  
Perú

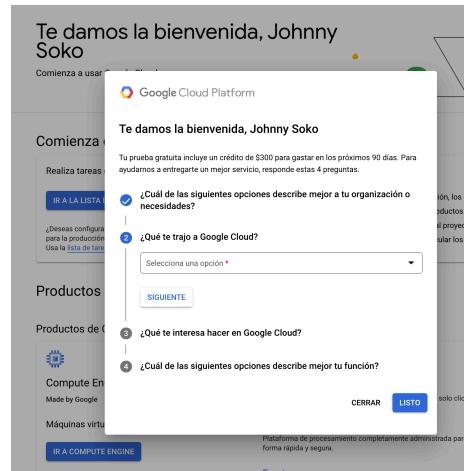
¿Cuál de las siguientes opciones describe mejor a tu organización o necesidades?  
Please select  
Otra actividad

Condiciones del Servicio  
 [Leí y acepto las Condiciones del Servicio de Google Cloud Platform, las Condiciones del Servicio de la prueba gratuita complementaria y las Condiciones del Servicio de cualquier servicio y API aplicables.](#)  
Debes seleccionar para continuar

CONTINUAR

Provincia/región  
Municipalidad Metropolitana de Lima

[INICIAR PRUEBA GRATUITA](#)



Google Cloud Platform

### Te damos la bienvenida, Johnny Soko

Tu prueba gratuita incluye un crédito de \$300 para gastar en los próximos 90 días. Para ayudarnos a entregarte un mejor servicio, responde estas 4 preguntas.

1. ¿Cuál de las siguientes opciones describe mejor a tu organización o necesidades?

2. ¿Qué te trajo a Google Cloud?

3. ¿Qué te interesa hacer en Google Cloud?

Sitios web Apps para dispositivos móviles  
 Copia de seguridad del almacenamiento Análisis de datos  
 Inteligencia artificial/aprendizaje automático Desarrollo de juegos  
 Creación de contenidores Administración de datos  
 Máquinas virtuales (VM) Google Maps  
 Otras API (p. ej., Text-to-Speech, Speech-to-Text, Vision)

4. ¿Cuál de las siguientes opciones describe mejor tu función?

object=natural-name-371403

Buscar dataproc

PRODUCTOS Y PÁGINAS

- Dataproc
- Clústeres Dataproc
- Federación Dataproc
- Flujos de trabajo Dataproc

Al seleccionar Dataproc, nos saldrá el siguiente mensaje:

La API de Dataproc no está habilitada en este proyecto. Redireccionando al panel de administración de la API de Dataproc.

## Habilitamos la API de Dataproc:

[Detalles del producto](#)



**Cloud Dataproc API**  
[Google Enterprise API](#)

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

**HABILITAR** **PROBAR ESTA API** 

Haz clic para habilitar esta API.

**DESCRIPCIÓN GENERAL** **PRECIOS** **DOCUMENTACIÓN**

Ahora ya podemos crear nuestro Cluster:

Clúster

### Cloud Dataproc

Google Cloud Dataproc te permite aprovisionar clústeres de Apache Hadoop y conectarte a almacenes de datos de análisis subyacentes.

No hay clústeres en las regiones de Cloud Dataproc seleccionadas actualmente. Crea un clúster para comenzar.

**CREATE CLUSTER**

**Crea un clúster de Dataproc**

Selecciona el servicio de infraestructura que deseas usar.

**Clúster en Compute Engine** **CREAR**  
 Crea el clúster en Compute Engine.

**Clúster en GKE** **CREAR**  
 Crea el clúster en Google Kubernetes Engine (GKE).

**CANCELAR**

**Control de versiones**  
 Usa una imagen personalizada para cargar paquetes preinstalados. [Más información](#)

**Tipo de imagen y versión**  
 2.0-debian10

**Fecha de lanzamiento**  
 Primera actualización el 22/1/2021

**CAMBiar**

**Componentes**  
 Puerta de enlace del componente

**Habilitar puerta de enlace de componentes**  
 Proporciona acceso a las interfaces web de componentes predeterminados y opcionales seleccionados en el clúster. [Más información](#)

**Componentes opcionales**  
 Selecciona uno o varios componentes. [Más información](#)

- Anaconda 
- Hive WebHCat 
- Jupyter Notebook 
- Zeppelin Notebook 
- Druid 
- Presto 
- ZooKeeper 
- Ranger 
- HBase 
- Flink 
- Docker 
- Solr 

Crea un clúster de Dataproc en Compute Engine

flexibilidad

- Configura el clúster
 

Comienza por proporcionar información básica.
- Configura los nodos (opcional)
 

Cambia las capacidades de procesamiento y almacenamiento del nodo.
- Personaliza el clúster (opcional)
 

Agrega propiedades, funciones y acciones del clúster.
- Administrar seguridad (opcional)
 

Cambia la configuración de acceso, encriptación y seguridad.

**CREAR** CANCELAR

LÍNEA DE COMANDOS EQUIVALENTE

Nodos trabajadores

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

Familia de máquinas

USO GENERAL OPTIMIZADA PARA PROCESAMIENTO CON OPTIMIZACIÓN DE MEMORIA GPU

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

Serie N1

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

TIPO DE MAQUINA: n1-standard-2 (2 CPU virtuales, 7.5 GB de memoria)

vCPU 2 Memory 7.5 GB

PLATAFORMA DE CPU Y GPU

Number of worker nodes \* 4

Tamaño del disco principal \* 50 Primary disk type Standard Persistent Disk

Cantidad de SSD Io... 0 Interface de SSD local SCS

Para continuar debemos crear un Bucket en nuestro Cloud Storage dado que lo que estamos haciendo es crear un cluster para computing no para Storage.

Google Cloud EduConsultores storage

Cloud Storage Buckets CREAR ACTUALIZAR

Buckets Supervisión NUEVO Configuración

Obtén una vista previa del nuevo panel de supervisión de Cloud Storage

TRY NOW

Consulta las recomendaciones de seguridad

Aplica recomendaciones de seguridad a tus buckets para protegerlos mejor. En la columna de estadísticas de seguridad tienen permisos excesivos.

VER EN TABLA MÁS INFORMACIÓN

Filtro Filtrar depósitos

Nombre	Fecha de creación	Tipo de ubicación	Ubicación	Default storage class	Última modificación
No hay filas para mostrar					

Cloud Storage Crear un bucket

Buckets Supervisión NUEVO Configuración

Asigna un nombre a tu bucket

Seleccióna un nombre permanente globalmente único. [Lineamientos para asignar nombre](#)

Ej: 'example', 'example\_bucket-1', or 'example.com'

Supervisión: No incluye información sensible

ETIQUETAS (OPCIONAL)

CONTINUAR

Información útil

Precios de ubicación

Las tarifas de almacenamiento varían según la clase de almacenamiento de los datos y la ubicación de los buckets. [Detalles de precios](#)

Configuración actual: Multi-region / Standard

Elemento	Costo
us (varias regiones en Estados Unidos)	\$0.026 por GB al mes
Con replicación predeterminada	\$0.020 por GB escrito

ESTIMAR COSTO MENSUAL

Elige dónde almacenar tus datos

Ubicación: us (varias regiones en Estados Unidos)

Tipo de ubicación: Multi-region

Elige una clase de almacenamiento para tus datos

Clase de almacenamiento predeterminada: Standard

Elige cómo controlar el acceso a los objetos

Prevención del acceso público: Activada

Control de acceso: Uniforme

Elige cómo proteger los datos de objeto

Herramientas de protección: Ninguno

Encriptación de datos: Google-managed key

**CREAR** CANCELAR

[Crear un bucket](#)

**Asigna un nombre a tu bucket**  
 Nombre: cienciadatos-bucket-2022

**Elige dónde almacenar tus datos**  
 Esta opción permanente define la ubicación geográfica de tus datos y afecta el costo, el rendimiento y la disponibilidad. [Más información](#)

**Tipo de ubicación**

- Multi-region  
Máxima disponibilidad en el área más amplia
- Dual-region  
Alta disponibilidad y baja latencia en 2 regiones
- Región  
Latencia mínima dentro de una sola región

us-east1 (Carolina del Sur)

[CONTINUAR](#)

**Elige una clase de almacenamiento para tus datos**  
 Clase de almacenamiento predeterminada: Standard

**Elige cómo controlar el acceso a los objetos**  
 Prevención del acceso público: Desactivada  
 Control de acceso: Uniforme

**Elige cómo proteger los datos de objeto**  
 Herramientas de protección: Ninguna  
 Encriptación de datos: Google-managed key

[CREAR](#) [CANCELAR](#)

[Cloud Storage](#)

[Buckets](#) [Supervisión](#) [Nuevo](#) [Configuración](#)

[Detalles del bucket](#)

cisciendatos-bucket-2022

Ubicación	Clase de almacenamiento	Acceso público	Protección
us-east1 (Carolina del Sur)	Standard	No público	Ninguna

[OBJETOS](#) [CONFIGURACIÓN](#) [PERMISOS](#) [PROTECCIÓN](#) [CICLO DE VIDA](#) [OBSERVABILIDAD](#) [NUEVO](#)

Depósitos > cisciendatos-bucket-2022

[SUBIR ARCHIVOS](#) [SUBIR CARPETA](#) [CREAR CARPETA](#) [TRANSFERRIR LOS DATOS](#) [ADMINISTRAR CONSERVACIONES](#) [DESCARGAR](#) [BORRAR](#)

Filtrar solo por prefijo de nombre:  [Filtrar](#) [Filtrar objetos y carpetas](#)

[Mostrar datos borrados](#)

Nombre  Tamaño  Tipo  Fecha de creación  Clase de almacenamiento  Última modificación  Acceso público  Historial de versiones  Descripción  Fecha de vencimiento del

No hay files para mostrar

[SUBIR ARCHIVOS](#) [CREAR TRABAJO DE TRANSFERENCIA](#)

[Filtro](#) Filtrar depósitos

Nombre	Fecha de creación	Tipo de ubicación	Ubicación	Default storage class	Última modificación	Acceso público
cienciadatos-bucket-2022	11 dic 2022 23:31:25	Region	us-east1	Standard	11 dic 2022 23:31:25	No público

Retornamos a Dataproc

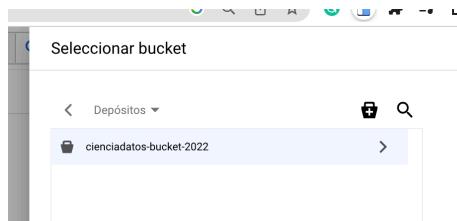
[Borrar luego de un tiempo de inactividad del clúster sin trabajos enviados](#)

### Bucket de etapa de pruebas de Cloud Storage

 Bucket de etapa de pruebas de almacenamiento [EXPLORAR](#)

El bucket de etapa de pruebas de Cloud Storage que se usará para almacenar las dependencias de trabajos clústeres, los resultados del controlador de trabajos y los archivos de configuración de clústeres.

Seleccionamos nuestro bucket (deposito)



### Bucket de etapa de pruebas de Cloud Storage

Bucket de etapa de pruebas de almacenamiento

	cienciadatos-bucket-2022	<b>EXPLORAR</b>
--	--------------------------	-----------------

El bucket de etapa de pruebas de Cloud Storage que se usará para almacenar las dependencias de trabajos clústeres, los resultados del controlador de trabajos y los archivos de configuración de clústeres.

Finalmente presionamos Crear:

[Crea un clúster de Dataproc en Compute Engine](#)

• Configura el clúster

• Configura los nodos (opcional)

• Personaliza el clúster (opcional)

• Administrar seguridad (opcional)

**Acceso al proyecto**

Habilita el alcance de la plataforma de nube para este clúster [Learn more](#)

Clave de encriptación administrada por Google No se requiere configuración

Clave de encriptación administrada por el cliente (CMK) Administrar a través de Google Cloud Key Management Service

**Autenticación personal del clúster**

Habilita la autenticación del clúster personal de Dataproc para permitir que las cargas de trabajo interactivas del clúster se ejecuten de forma segura con tu identidad de usuario final. [Más información](#)

Habilitar

**Multiusuario seguro**

Habilita el modo seguro de multiusuario para permitir que cuentas de servicios de Dataproc puedan controlar un clúster con varios usuarios. Asigna uno o más roles de la lista de autorizaciones a los permisos adecuados para actuar en nombre de todos los usuarios. [Más información](#)

Habilitar

**Modo seguro de Kerberos y Hadoop**

Habilita el modo seguro de Kerberos y Hadoop para proporcionar autenticación de usuario, cifrado y encriptación en un clúster de Dataproc. [Más información](#)

Habilitar

**CREAR** **CANCELAR**

Clústeres							<b>+ CREATE CLUSTER</b>	<b>ACTUALIZAR</b>	<b>INICIAR</b>	<b>DETENER</b>	<b>BORRAR</b>	<b>REGIONES</b>	<b>+ 5 ALERTAS RECOMENDADAS</b>
<b>Filtro</b> Busca clústeres y presiona Intro													
<input type="checkbox"/>	Nombre	↑	Estado	Región	Zona	Total de nodos trabajadores	Eliminación programada	Bucket de etapa de pruebas de Cloud Storage					
<input type="checkbox"/>	cluster-37ba		Aprovisionamiento	us-central1	us-central1-1	4	Desactivado	cienciadatos-bucket-2022					

Clústeres							<b>+ CREATE CLUSTER</b>	<b>ACTUALIZAR</b>	<b>INICIAR</b>	<b>DETENER</b>	<b>BORRAR</b>	<b>REGIONES</b>	<b>+ 5 ALERTAS RECOMENDADAS</b>
<b>Filtro</b> Busca clústeres y presiona Intro													
<input type="checkbox"/>	Nombre	↑	Estado	Región	Zona	Total de nodos trabajadores	Eliminación programada	Bucket de etapa de pruebas de Cloud Storage					
<input type="checkbox"/>	cluster-37ba		En ejecución	us-central1	us-central1-1	4	Desactivado	cienciadatos-bucket-2022					

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone>

**SUPERVISIÓN** **TRABAJOS** **INSTANCIAS DE VM** **CONFIGURACIÓN** **INTERFACES WEB**

**Nombre:** cluster-37ba  
**UUID del clúster:** 434c918d-98d3-4974-a980-374b9ffb71aa  
**Tipo:** Clúster de Dataproc  
**Estado:** En ejecución

**GUARDAR COMO PANTALLA PERSONALIZADA** **RESTABLECER EL ZOOM**

**Memoria YARN**: 250GB (Allocated: 0, Available: 249GB, Reserved: 0)

**Memoria YARN pendiente**: 1GB (Pending: 0)

**Administradores de nodos YARN**: 4 (Active: 4, Decommissioned: 0, Lost: 0, New: 0, Rebooted: 0, Shutdown: 0, Unhealthy: 0)

**Capacidad HDFS**: 150GB

**Uso de CPU**: 50% (UTC-6: 22:30, 22:40, 23:00, 23:50, 12:00, 00:10)

**Bytes de red**: 400KB/s (300KB/s)

**Nombre:** cluster-37ba  
**UUID del clúster:** 434c918d-98d3-4974-a980-374b9ffb71aa  
**Tipo:** Clúster de Dataproc  
**Estado:** En ejecución

**SUPERVISIÓN** **TRABAJOS** **INSTANCIAS DE VM** **CONFIGURACIÓN** **INTERFACES WEB**

**Filtro:** Filtrar instancias

Nombre	Rol
cluster-37ba-m	Instancia principal
cluster-37ba-w-0	Trabajador
cluster-37ba-w-1	Trabajador
cluster-37ba-w-2	Trabajador
cluster-37ba-w-3	Trabajador

**REST EQUIVALENTE**

**Nombre:** cluster-37ba  
**UUID del clúster:** https://ssh.cloud.google.com/v2/ssh/projects/educaconsultores/zones/us-central1-c/instances/cluster-37ba-m?authuser=0&hl...  
**Tipo:** https://ssh.cloud.google.com/v2/ssh/projects/educaconsultores/zones/us-central1-c/instances/cluster-37ba-m?authuser=0&hl...  
**Estado:** SSH en el navegador

**SUPERVISIÓN**

**Filtro:** Filtrar instancias

Linux cluster-37ba-m 5.10.0-0.deb10.16-amd64 #1 SMP Debian 5.10.127-2-bpo10+1 (2022-07-28) 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.  
 last login: Mon Dec 12 05:15:26 2022 from 35.235.244.32  
 jchipoco@cluster-37ba-m: ~

**REST EQUIVALENTE**