



---

## Cómo ejecutar Spark en Dataproc

# Temario

---

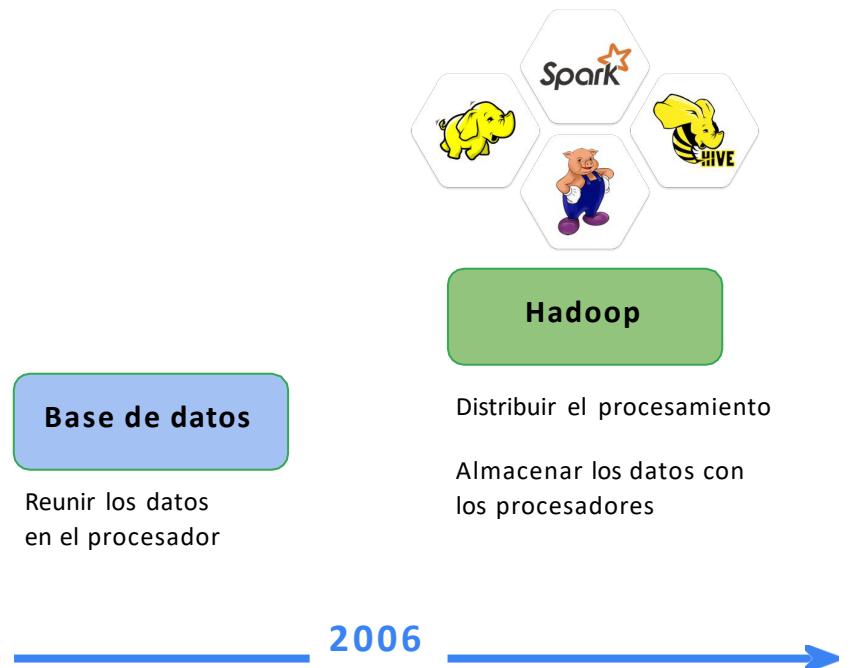
## Ecosistema de Hadoop

Cómo ejecutar Hadoop  
en Cloud Dataproc  
GCS en lugar del HDFS

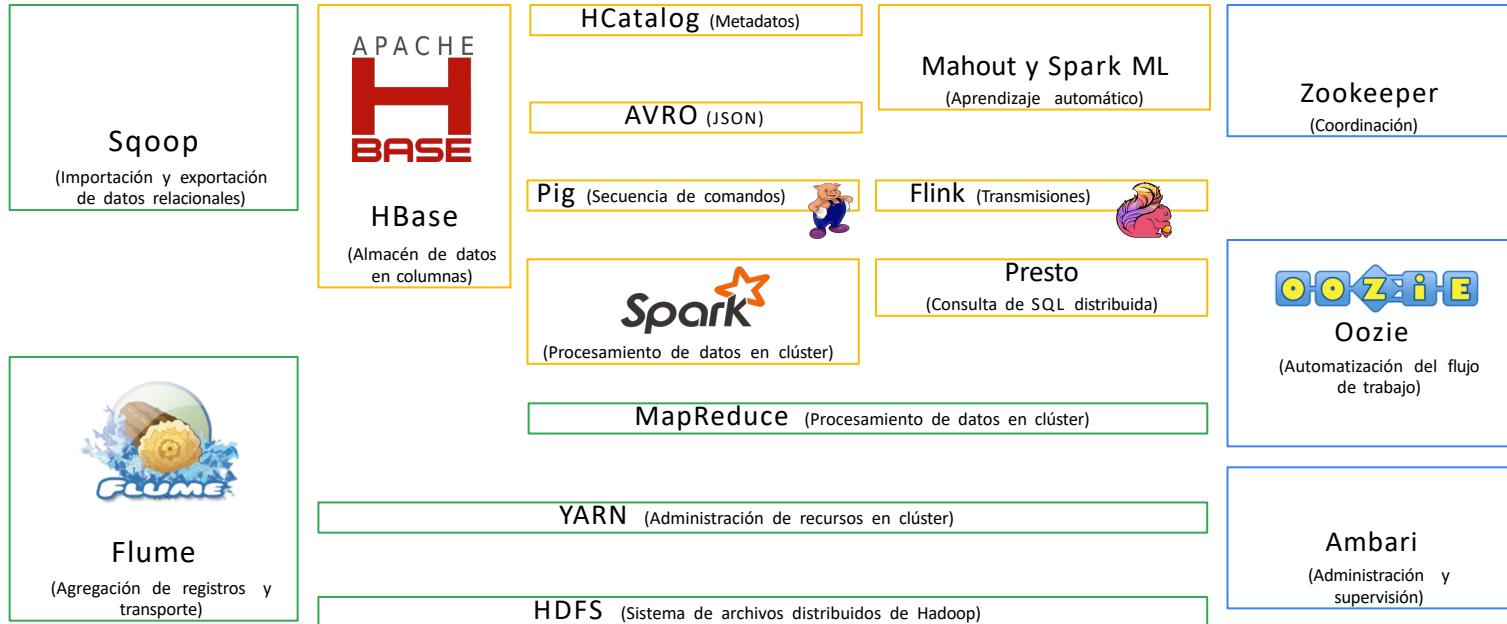
Cómo optimizar Dataproc

Lab

# El ecosistema de Hadoop se desarrolló debido a la necesidad de analizar grandes conjuntos de datos



# El ecosistema de Hadoop es muy popular para las cargas de trabajo de Big Data



# Los clústeres locales de Hadoop tienen varias limitaciones

-  No son flexibles.
-  Son difíciles de escalar con rapidez.
-  Tienen límites de capacidad.
-  No tienen separación entre el almacenamiento y los recursos de procesamiento.

# Cloud Dataproc simplifica las cargas de trabajo de Hadoop en GCP

-  Compatibilidad integrada para Hadoop  
Configuración y hardware administrados
-  Administración de versiones simplificada
-  Configuración de trabajo flexible

# Apache Spark es una forma popular, flexible y potente de procesar grandes conjuntos de datos



etcétera.



[spark.apache.org](http://spark.apache.org)

# Temario

---

Ecosistema de Hadoop

Cómo ejecutar Hadoop  
en Cloud Dataproc

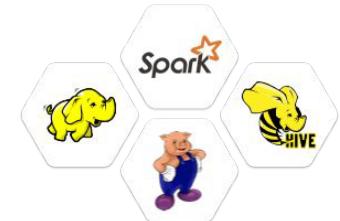
GCS en lugar del HDFS  
Cómo optimizar Dataproc

Lab

# Cloud Dataproc es un servicio administrado para ejecutar cargas de trabajo de procesamiento de datos de Hadoop y Spark



Google Cloud Platform



# Hay otras opciones de OSS disponibles en Cloud Dataproc

Spark (predeterminado)	Hive (predeterminado)	HDFS (predeterminado)
Pig (predeterminado)	Zeppelin	Zookeeper
Kafka	Hue	Tez
Presto	Anaconda	Proxy de Cloud SQL
Jupyter	Apache Flink	Cloud Datalab
IPython	Oozie	Sqoop
Mucho más...		

# Utilice acciones de inicialización para agregar otro software al clúster en el inicio

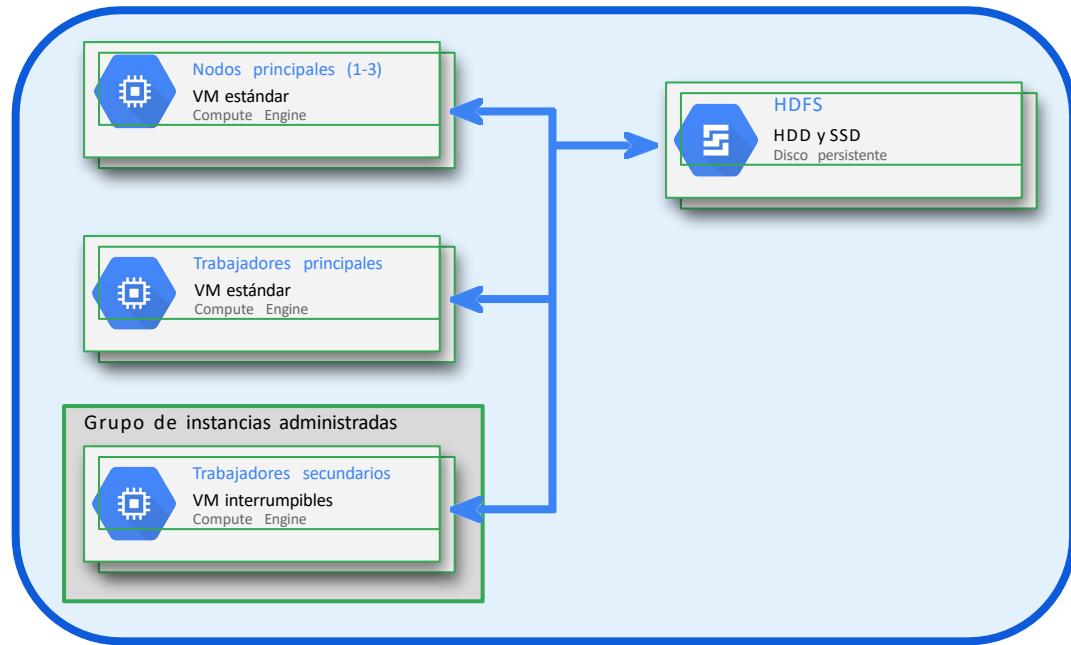
Utilice **acciones de inicialización** para instalar componentes adicionales en el clúster.



```
gcloud dataproc clusters create <CLUSTER_NAME> \  
    --initialization-actions gs://$MY_BUCKET/hbase/  
        hbase.sh \  
    --num-masters 3 --num-workers 2
```

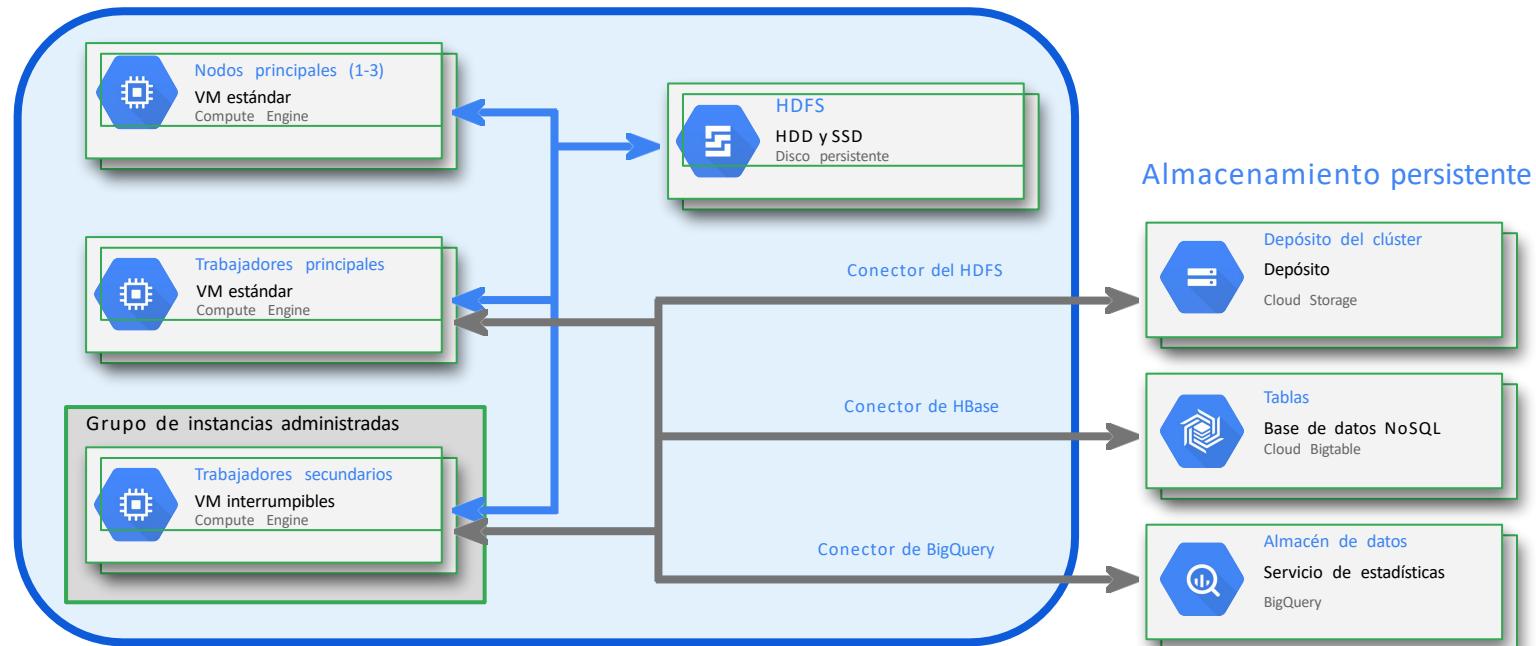
# Un clúster de Dataproc tiene nodos principales, trabajadores y el HDFS

Clúster de Dataproc



# El clúster de Dataproc puede leer o escribir en productos de almacenamiento de GCP

Clúster de Dataproc



# Cómo usar Cloud Dataproc

Ajustes

Configuración

Optimización

Uso

Supervisión

Crear un clúster

Console  
comando gcloud/archivo YAML  
Plantilla de Deployment Manager  
API de REST de SDK de Cloud

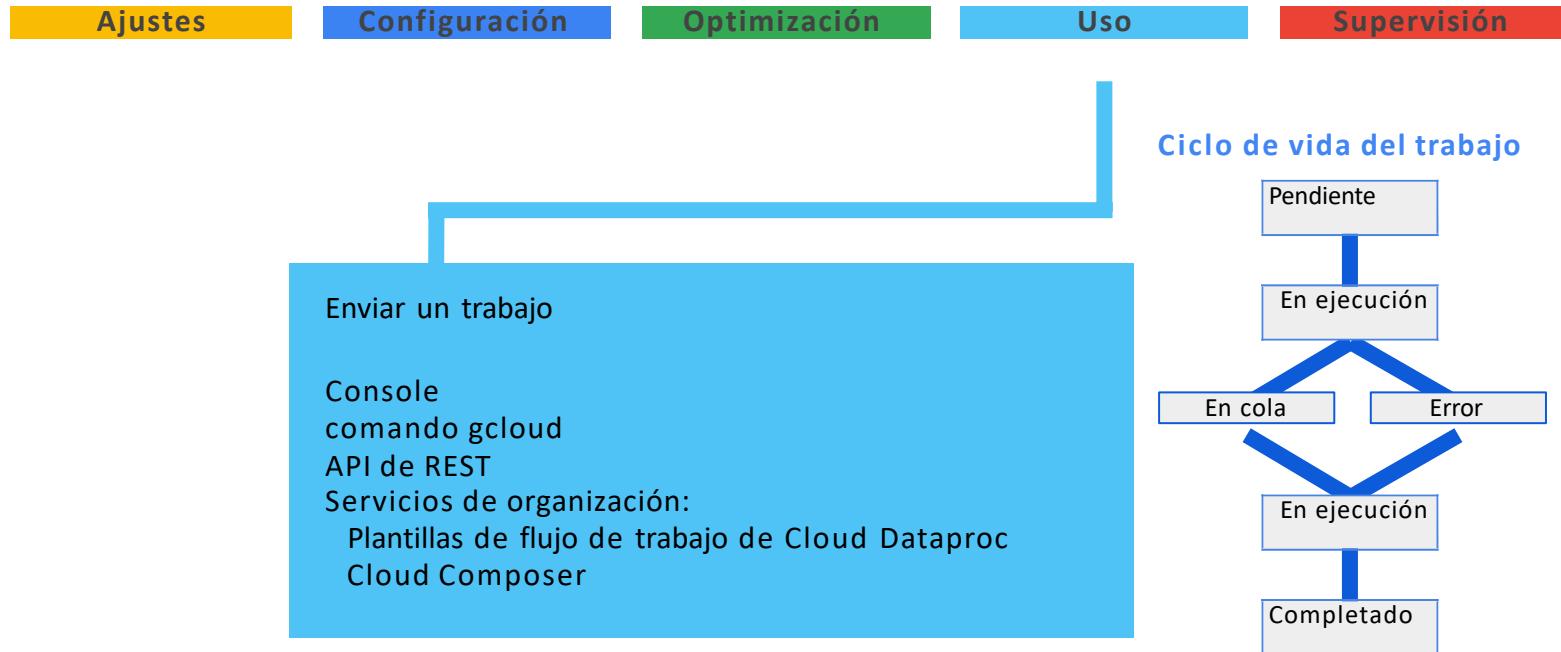
# Configuración

Ajustes	Configuración	Optimización	Uso	Supervisión
<b>Opciones de clúster</b>	Región y zona Extremo regional o global Versión de Dataproc (la versión predeterminada es la más reciente)	Nodo único (1:0) Estándar (1:n) Alta disponibilidad (3:n)		
	Componentes opcionales	Propiedades del clúster Etiquetas del usuario		
<b>Opciones del nodo principal</b>	Núcleos de CPU virtual Disco principal y tipo de disco	SSD locales		
<b>Nodos trabajadores</b>	Cantidad mínima de nodos trabajadores	<b>Opciones de VM</b> Acciones de inicialización Metadatos de VM		
<b>Nodos interrumpibles</b>	Cantidad máxima (la cantidad predeterminada es 0)			

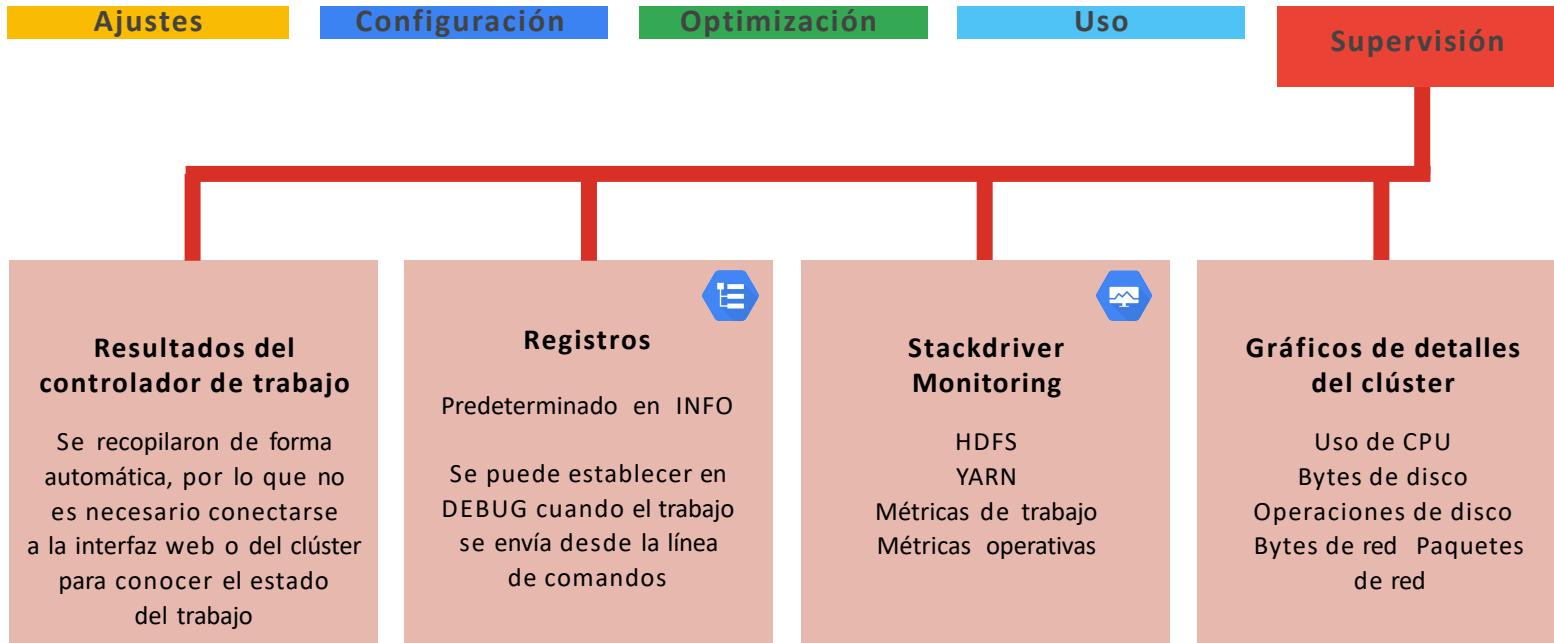
# Optimización

Ajustes	Configuración	Optimización	Uso	Supervisión
VM interrumpibles		Costo más bajo		
Tipos personalizados de máquina		Asignación eficiente de recursos para cargas de trabajo consistentes		
Plataforma de CPU mínima		Distribución coherente de la carga de trabajo: rendimiento de la CPU virtual		
Imágenes personalizadas		Tiempo más rápido para alcanzar un estado operativo		
Disco de arranque persistente SSD		Tiempo de arranque más rápido		
GPU conectadas		Procesamiento más rápido para algunas cargas de trabajo		
Versión de Dataproc		Especificar para evitar cambios o establecer de forma predeterminada la más reciente		

# Uso: Envío de trabajos



# Supervisión a través de Console y Stackdriver



# Temario

---

Ecosistema de Hadoop

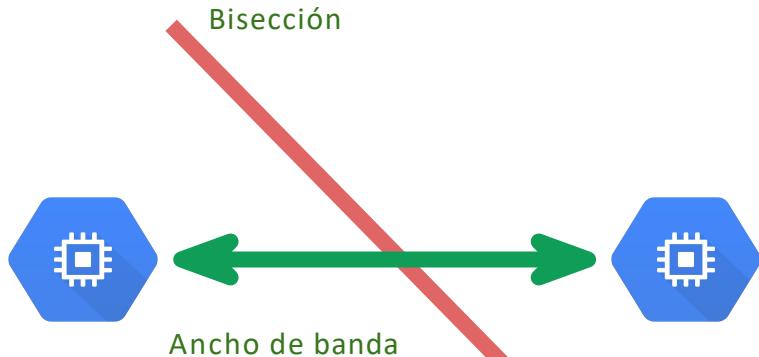
Cómo ejecutar Hadoop  
en Cloud Dataproc

GCS en lugar del HDFS

Cómo optimizar Dataproc

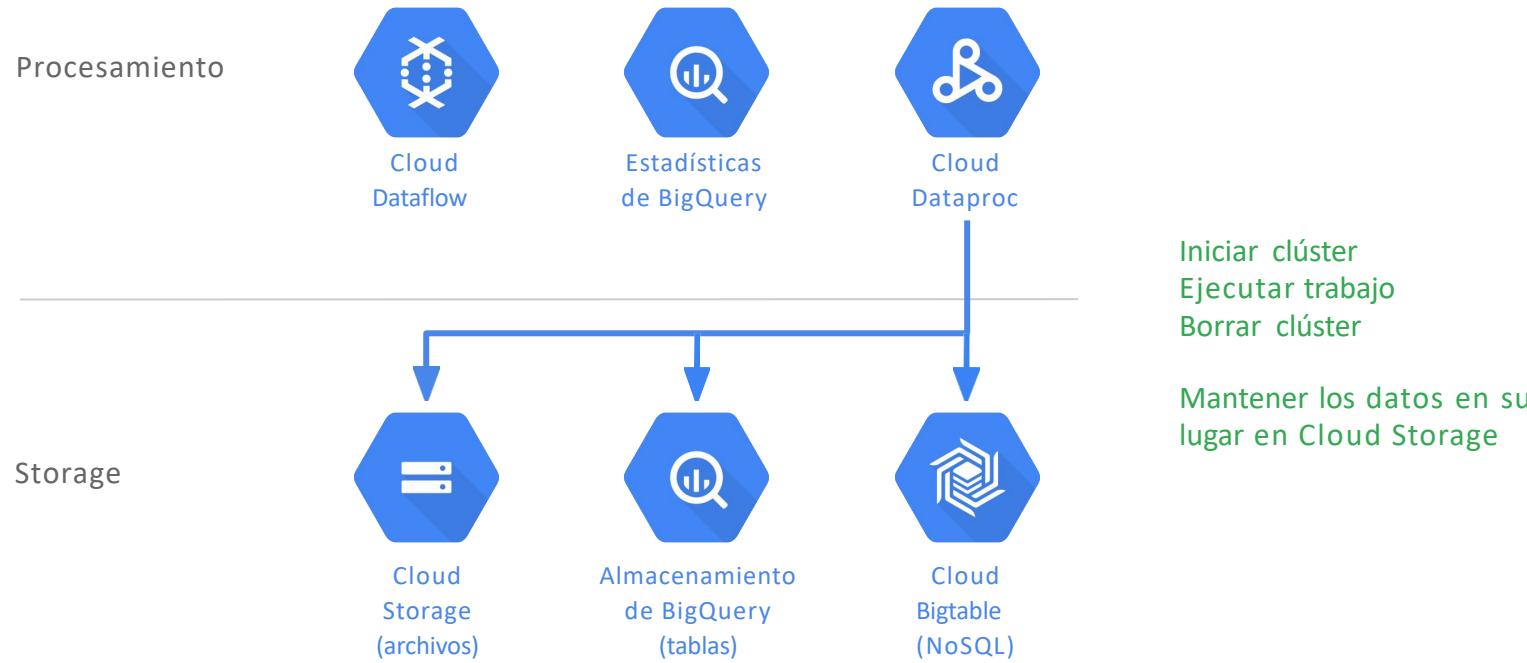
Lab

# El ancho de banda en Petabit cambia las reglas del juego para el Big Data

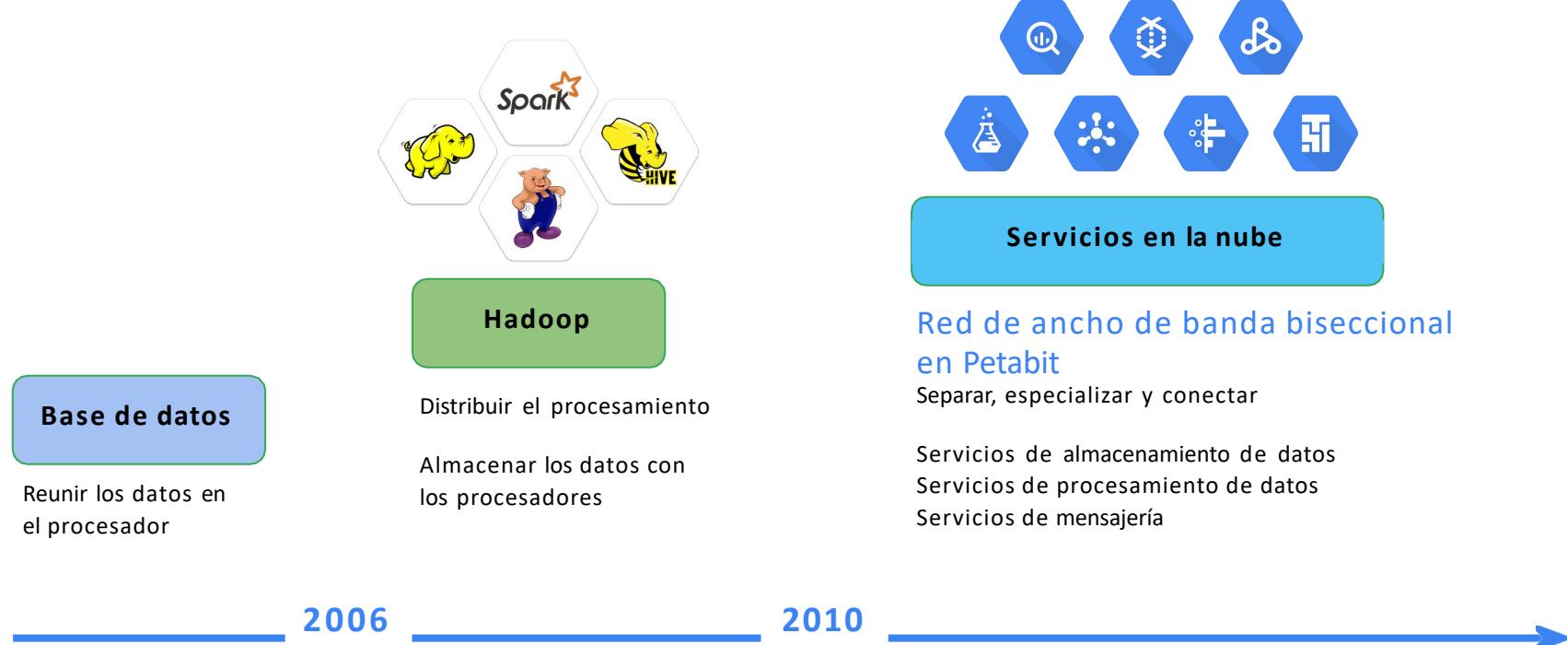


Procesar los datos donde están sin copiarlos

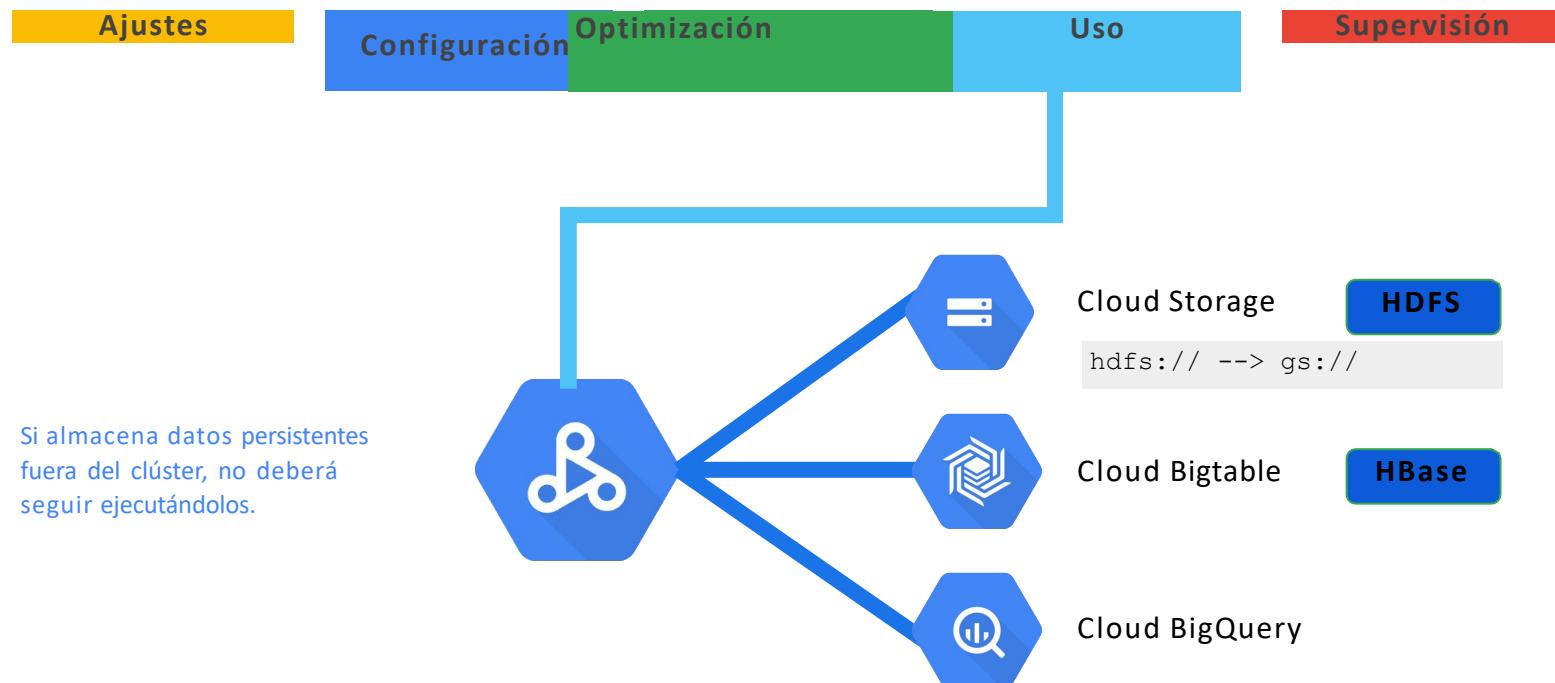
# En GCP, Jupyter y Colossus hacen posible la separación del procesamiento y el almacenamiento



# La separación del procesamiento y el almacenamiento permite mejores opciones



# El almacenamiento fuera del clúster es la puerta de entrada a la eficiencia



# Use Cloud Storage en lugar del HDFS con Cloud Dataproc

Ajustes

Configuración

Optimización

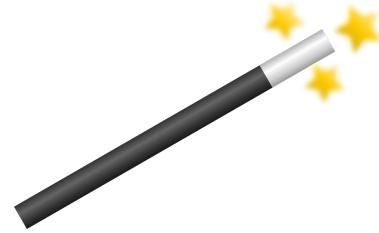
Uso

Supervisión



Cloud Storage es un servicio distribuido.

Elimina los cuellos de botella tradicionales y los puntos únicos de fallas.

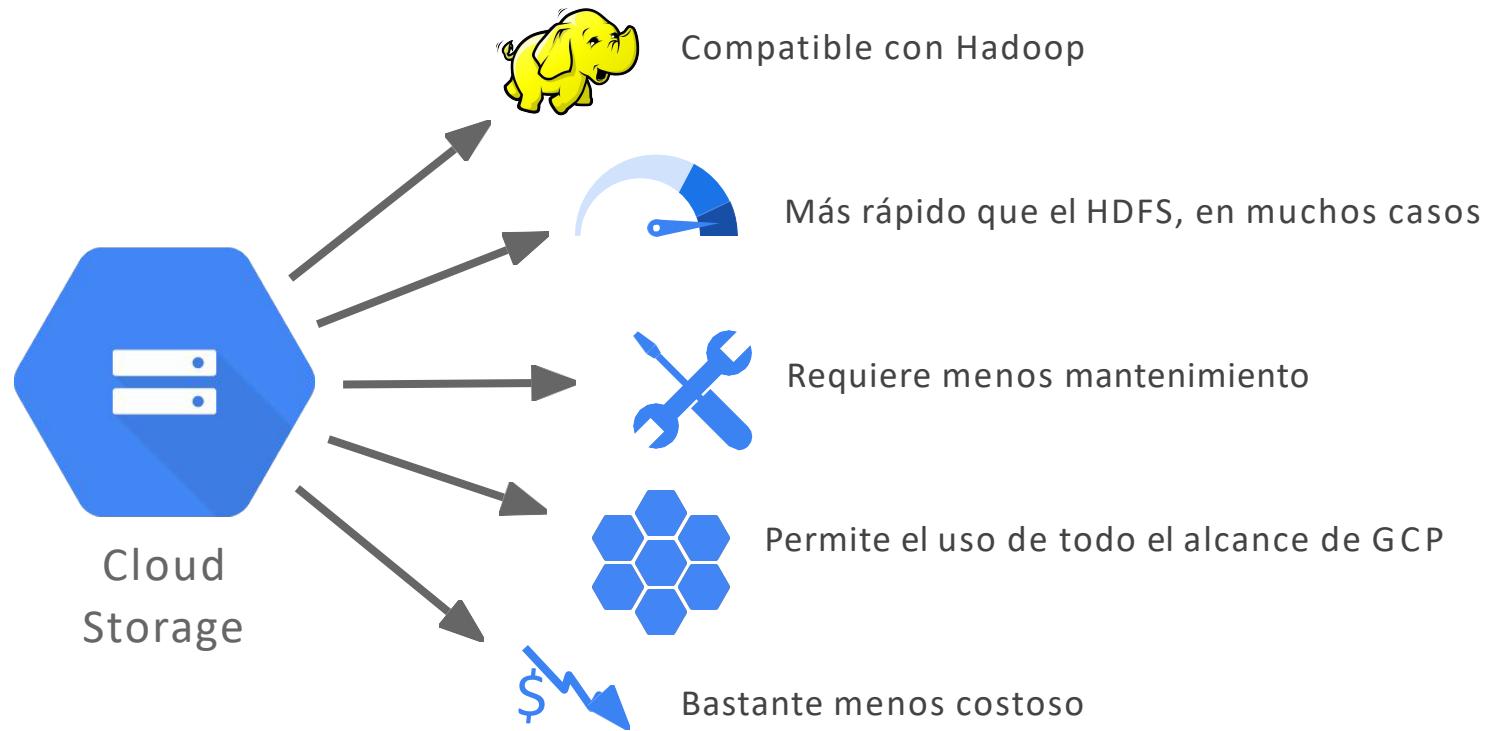


Los directorios se simulan, por lo que cambiar el nombre de un directorio implica cambiar el nombre de todos los objetos.\*



Los objetos no admiten "agregar".

# Utilice Cloud Storage como el almacén de datos persistentes



# Cloud Storage es un reemplazo directo del HDFS



Interfaces de Hadoop FileSystem: compatibles con "HCFS" (Sistema de archivos compatible con Hadoop)  
File[Input|Output]Format, SparkContext.textFile, etc., simplemente funciona



Cloud Dataflow: Anulación de duplicación, orden y ventana

# Prácticas recomendadas de rendimiento

## Cómo optimizar las operaciones masivas/paralelas



Evite lecturas pequeñas; use bloques grandes donde sea posible.



Evite iterar de forma secuencial sobre muchos directorios anidados en un trabajo único.



# Temario

---

Ecosistema de Hadoop

Cómo ejecutar Hadoop  
en Cloud Dataproc

GCS en lugar del HDFS

Cómo optimizar Dataproc

Lab

# El HDFS local es necesario a veces

- El HDFS es una buena opción en los siguientes casos:
- Sus trabajos requieren muchas operaciones de metadatos.
- Modifica los datos del HDFS con frecuencia o cambia el nombre de los directorios.
- Usa mucho la operación de agregar en los archivos del HDFS.
- Tiene cargas de trabajo que implican E/S pesadas.
- Tiene cargas de trabajo de E/S que son especialmente sensibles a la latencia.

# GCP proporciona diferentes opciones de almacenamiento para distintos trabajos



Cloud  
Storage

- Almacén de datos primario para GCP
- Datos no estructurados



Cloud  
Bigtable

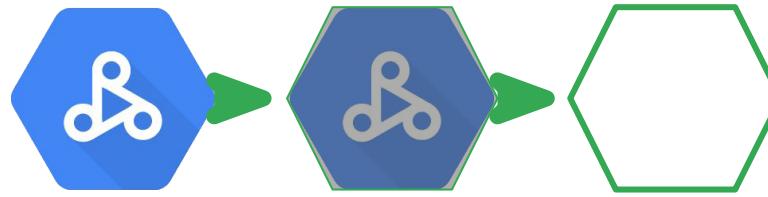
- Cantidad grande de datos dispersos
- Compatible con HBase
- Latencia baja
- Escalabilidad alta



BigQuery

- Almacenamiento de datos
- La API de Storage lo hace más rápido que antes
- Podría enviar consultas a BigQuery, mediante la refactorización el trabajo

# Eliminación programada del clúster



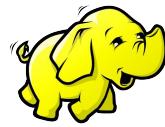
Inactivo

Marca de tiempo

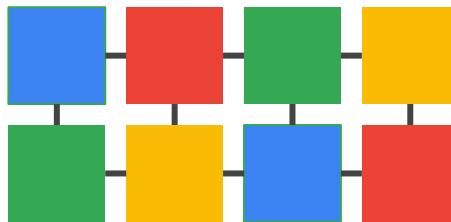
Duración

Mínimo de 10 minutos  
Máximo de 14 días  
Nivel de detalle de 1 segundo

# Con los clústeres efímeros, solo paga por lo que usa



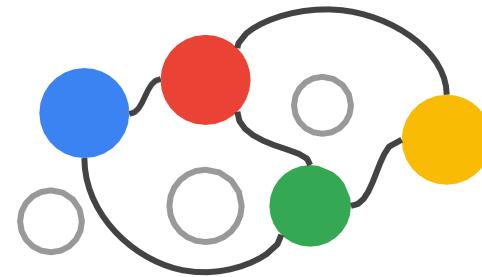
Clústeres persistentes



Los recursos están activos en todo momento.  
Usted paga de forma constante por todos  
los clústeres disponibles.

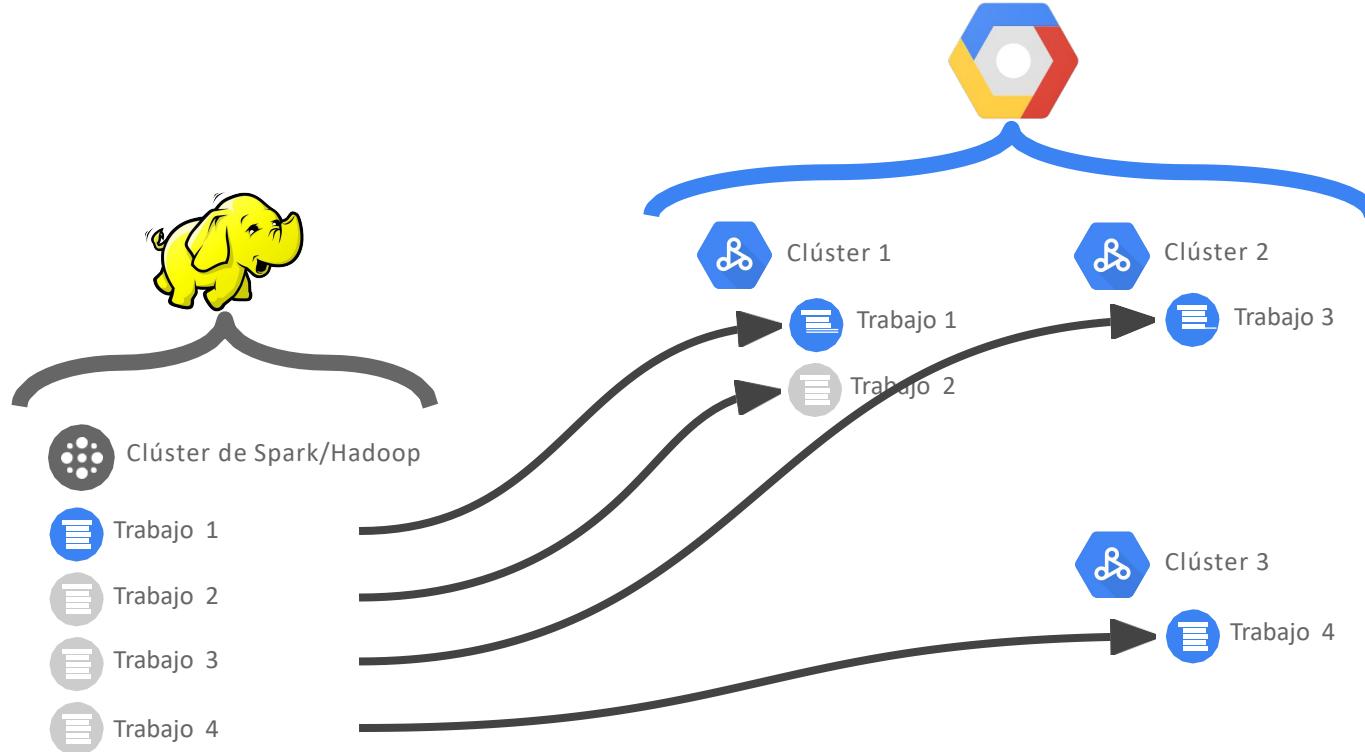


Clústeres efímeros

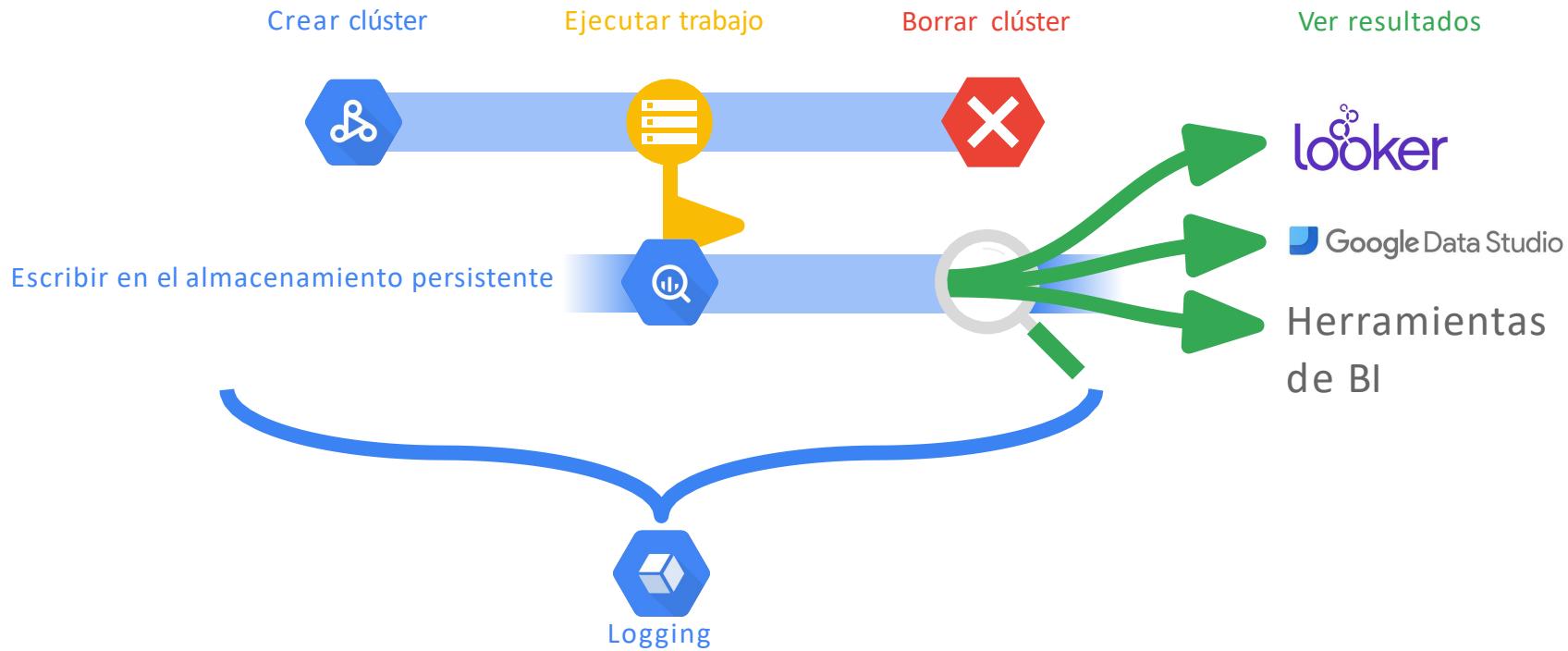


Los recursos necesarios solo están activos  
cuando se usan. Solo paga por lo que usa.

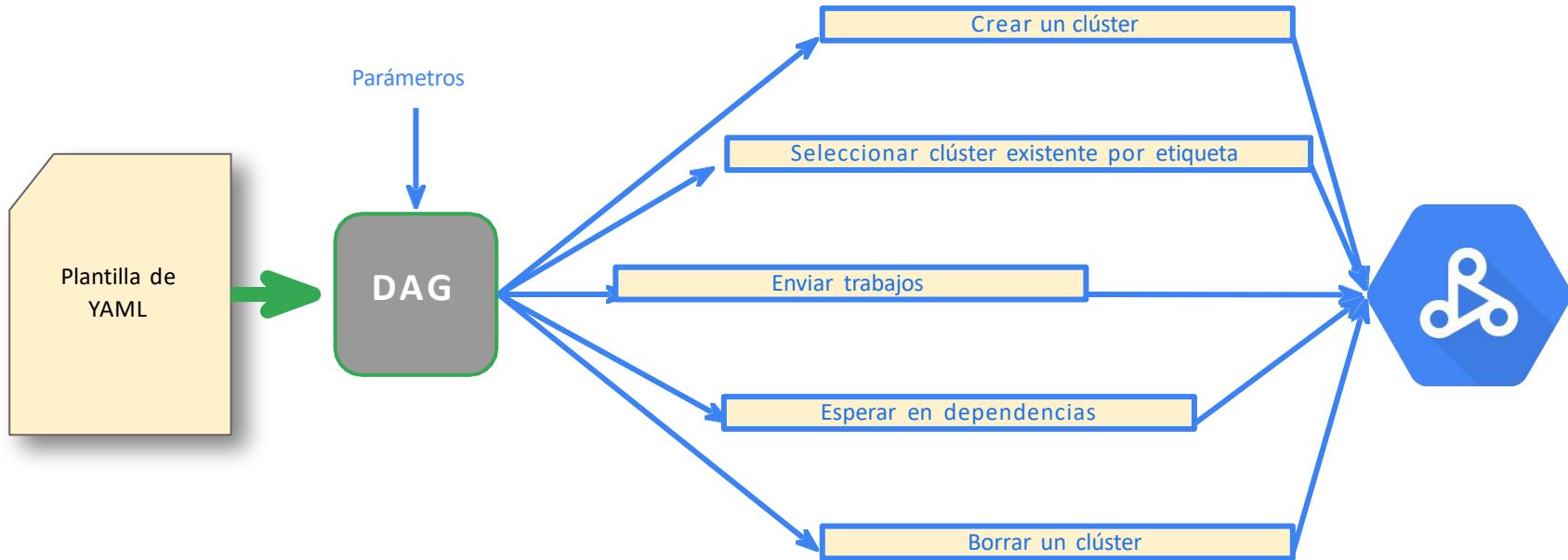
# División de los clústeres y los trabajos



# Cómo usar clústeres efímeros para el ciclo de vida de un trabajo



# Plantilla de flujo de trabajo de Cloud Dataproc



# Plantillas de flujo de trabajo de Cloud Dataproc

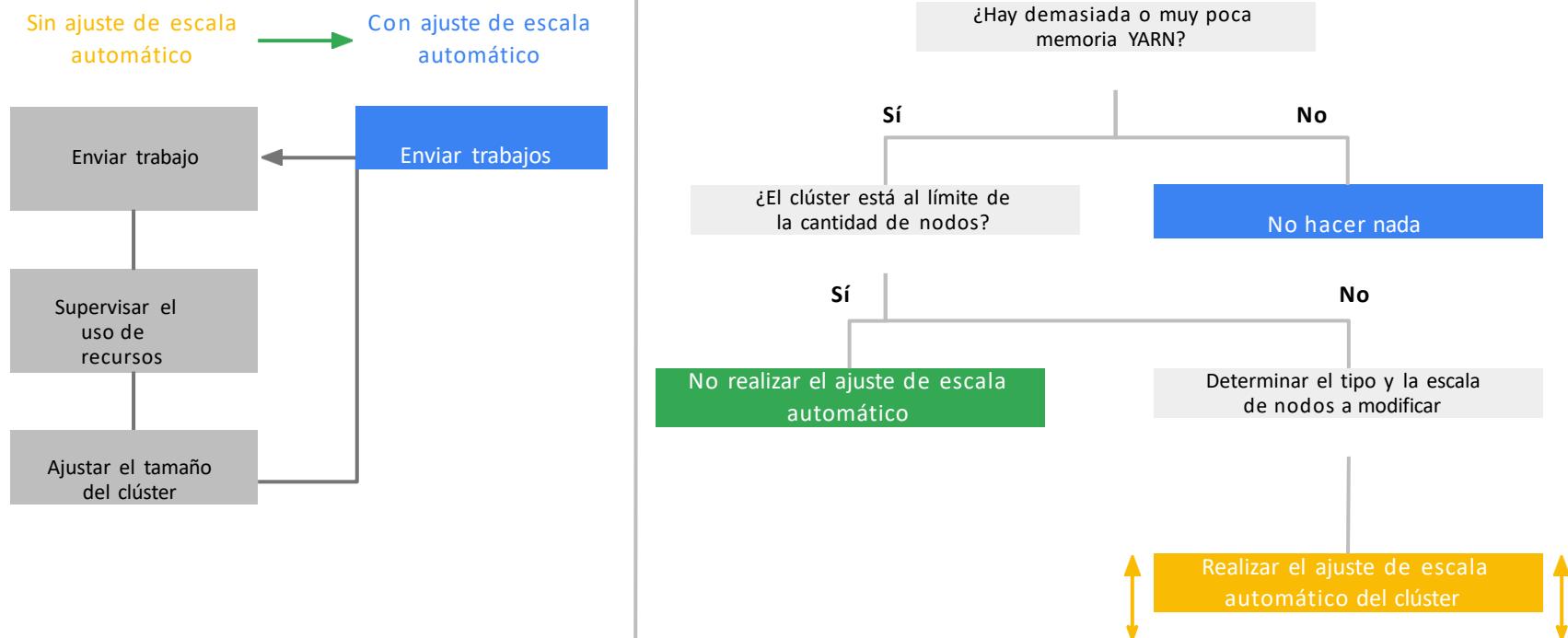
```
# the things we need pip-installed on the cluster
STARTUP_SCRIPT=gs://${BUCKET}/sparktobq/
startup_script.sh
echo "pip install --upgrade --quiet google-compute-engine google-cloud-storage matplotlib" >
/tmp/startup_script.sh
gsutil cp /tmp/startup_script.sh $STARTUP_SCRIPT

# create new cluster for job
gcloud dataproc workflow-templates set-managed-cluster $TEMPLATE \
--master-machine-type $MACHINE_TYPE \
--worker-machine-type $MACHINE_TYPE \
--initialization-actions $STARTUP_SCRIPT \
--num-workers 2 \
--image-version 1.4 \
--cluster-name $CLUSTER

# steps in job
gcloud dataproc workflow-templates add-
job \ pyspark gs://${BUCKET}/
spark_analysis.py \
--step-id create-report \
--workflow-template $TEMPLATE \
-- --bucket=${BUCKET}

# submit workflow template
gcloud beta dataproc workflow-templates instantiate $TEMPLATE
```

# Flujo de trabajo del ajuste de escala automático de Cloud Dataproc



# El ajuste de escala automático de Cloud Dataproc proporciona capacidad flexible

Clúster con muchos trabajos o un solo trabajo grande

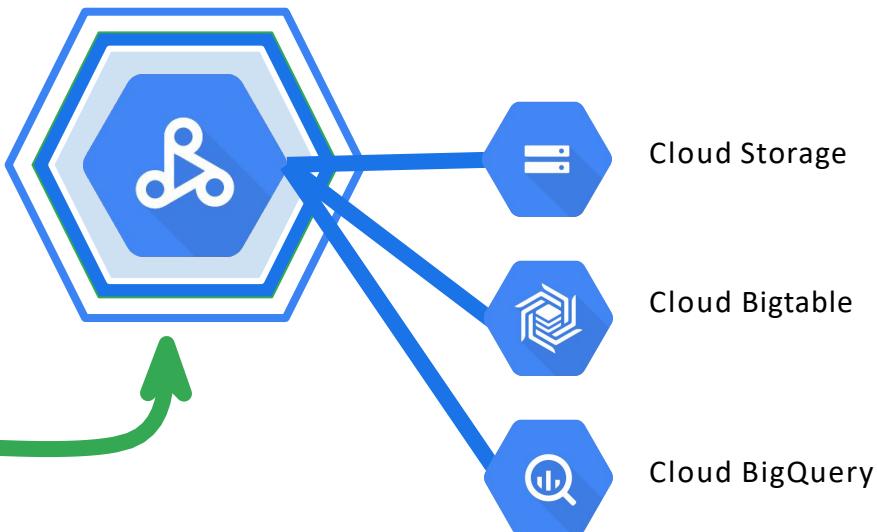
No para clústeres inactivos o escala a cero

El ajuste de escala automática se basa en las métricas YARN de Hadoop

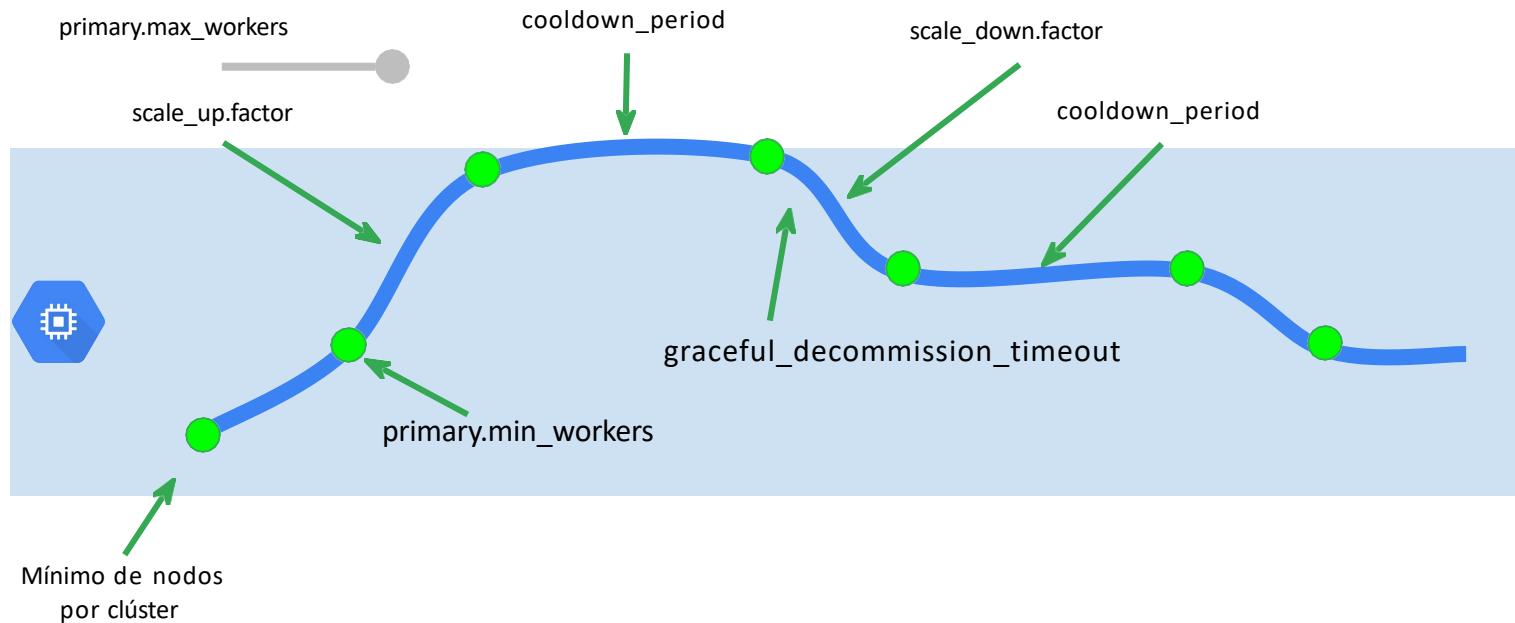
Datos externos

No para el HDFS en clúster

No para la transmisión estructurada de Spark



# Cómo funciona el ajuste de escala automático de Cloud Dataproc

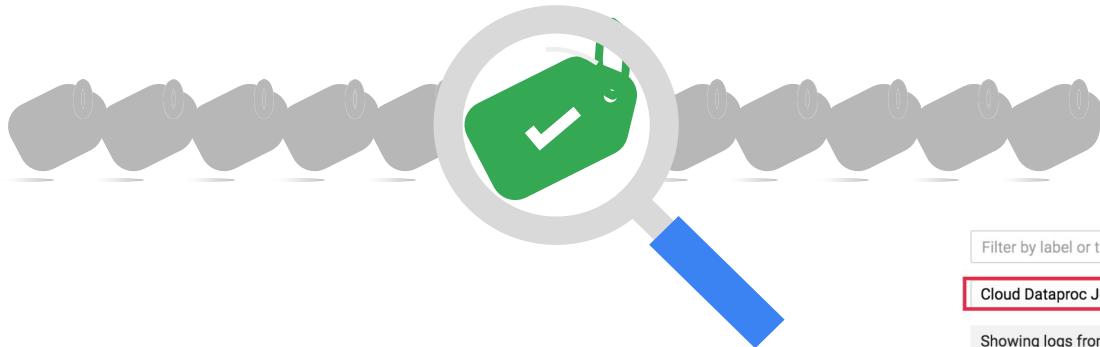


# Cómo usar Stackdriver Logging y la supervisión del rendimiento

The screenshot shows the Google Cloud Stackdriver Logging interface. At the top, there are buttons for 'CREATE METRIC' and 'CREATE EXPORT'. Below that is a search bar labeled 'Filter by label or text search' and a dropdown menu set to 'Cloud Dataproc Cluster'. To the right of the search bar are filters for 'All logs', 'Any log level', and a time range from 'Last 24 hours' to 'Jump to now'. The main area displays log entries for a 'Cloud Dataproc Cluster'. A message at the top of the log list states: 'No older entries found matching current filter in the last 24 hours.' The log entries show a sequence of events related to a job submission:

```
2017-09-20T11:42:20.000Z dataproc SubmitJob global:564f57524dc84d369f4f914f39f5d666 dimosthenis@google.com created new context for 564f57524dc84d369f4f914f39f5d666
2017-09-20T11:42:20.000Z dataproc StartNewJobId '564f57524dc84d369f4f914f39f5d666'
2017-09-20T11:42:20.000Z dataproc StartNewDriver 'spark-submit, --conf, spark.yarn.tags=dataproc_hash_5d4159bf'
2017-09-20T11:42:20.000Z dataproc PingDriverOutputTo gs://dataproc-107cf59b-2f38-47e6-98fa-bde4f549c397-e666 completed with exit code 0
2017-09-20T11:42:20.000Z dataproc PingDriverOutputTo gs://dataproc-107cf59b-2f38-47e6-98fa-bde4f549c397-e666 completed with exit code 0
2017-09-20T11:42:20.000Z dataproc LogEvent 24dc84d369f4f914f39f5d666
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-2ae5
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-7bad
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-7eab
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-93a3
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-e3e3
2017-09-20T11:42:20.000Z dataproc LogEvent cluster-fb15
2017-09-20T11:42:20.000Z dataproc LogEvent All cluster_uuid
2017-09-20T11:42:20.000Z dataproc LogEvent 79451d69-a30f-43a4-ad75-c78b6e0016ae
```

# Cómo crear etiquetas en clústeres y trabajos para encontrar registros con mayor rapidez



Filter by label or text search

Cloud Dataproc Job dataproc.job.driver

Showing logs from all time (PDT)

2019-04-03 09:34:16.478 PDT Pi is roughly 3.1417569

{  
  insertId: "le8i240nizp188ay9"  
  labels: {...} dataproc.job.driver  
  logName: "projects/google.com:hadoop-cloud-dev/log"  
  receiveTimestamp: "2019-04-03T16:34:19.778423350Z"  
  resource: {...}  
  textPayload: "Pi is roughly 3.1417569514175696"  
  timestamp: "2019-04-03T16:34:16.478380936Z"  
}

2019-04-03 09:34:16.000 PDT Stopped Spark@19569ebd{...}

2019-04-03 09:33:45.000 PDT Submitted application a...

# Cómo establecer el nivel de registro

Puede establecer el nivel de registro del controlador con el siguiente comando gcloud:

```
gcloud dataproc jobs submit hadoop --driver-log-levels
```

Establezca el nivel de registro para el resto de la aplicación desde el contexto de Spark.

Por ejemplo:

```
spark.sparkContext.setLogLevel("DEBUG")
```

# Supervise sus trabajos

## Add Chart

GCE VM Instance - CPU usage for cluster-fb15

**METRIC**    **VIEW OPTIONS**

Find resource type and metric ②

Resource type: GCE VM Instance x

Metric: CPU usage x

Filter ②

goog-dataproc-cluster-name = cluster-fb15 x

Filter

Group By ②

Filter

Aggregation ②

none

▼ SHOW MORE OPTIONS

+ ADD METRIC

