

## Laboratorio Data Processing with Google Dataproc

# Laboratorio Data Processing with Google Dataproc

### **Laboratorio**

Cree un archivo TXT o CSV

Subalo a un cloud storage

Cree un cluster de Dataproc

Use el archivo y aplica un conjunto de transformaciones y acciones

En Dataproc cree el cluster con nombre: **cluster-8c48**

Google Cloud Dataproc Clústeres

Trabajos alojados en clústeres

- Clústeres
- Trabajos
- Flujos de trabajo

Filtro: Busca clústeres y presiona Intro

| Nombre                       | Estado       | Región   | Zona       | Total de nodos trabajadores | Eliminación programada | Bucket de etapa de pruebas de Cloud Storage                     | Fecha de creación |
|------------------------------|--------------|----------|------------|-----------------------------|------------------------|---|-------------------|
| <a href="#">cluster-8c48</a> | En ejecución | us-west4 | us-west4-a | 2                           | Desactivado            | <a href="#">dataproc-staging-us-west4-916192672684-stvwblxe</a> | 13 jun 2023, 1    |

Google Cloud Dataproc Detalles del clúster

Trabajos alojados en clústeres

- Clústeres
- Trabajos
- Flujos de trabajo
- Políticas de ajuste de escala...

Sin servidores

- Lotes

Servicios de Metastore

- Metastore
- Federación

Servicios públicos

- Intercambio de componentes
- Workbench

SUPERVISIÓN TRABAJOS INSTANCIAS DE VM CONFIGURACIÓN INTERFACES WEB

Túnel SSH

Crea un túnel SSH para conectarte a una interfaz web

Puerta de enlace de los componentes

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

- [YARN ResourceManager](#)
- [MapReduce Job History](#)
- [Spark History Server](#)
- [HDFS NameNode](#)
- [YARN Application Timeline](#)
- [HiveServer2 \(cluster-8c48-m\)](#)
- [Tez](#)
- [Jupyter](#)
- [JupyterLab](#)

En Cloud Storage cree el bucket con nombre: **bucket\_matapiam01**

Recibidos (14.901) - matapiam01

bucket\_matapiam01 - Detalles de

cluster-8c48 - Interfaces - Detalle

ikgle7ktivfnlm7gdbmc7qy4qy-dc

console.cloud.google.com/storage/browser/bucket\_matapiam01;tab=objects?forceOnBucketsSortingFiltering=true&hl=es&project=datapath-380101&...

Google Cloud

Datapath

bucket

Buscar

ACTUALIZAR

APRENDIZAJE

Buckets

Supervisión

Configuración

bucket\_matapiam01

Ubicación

Clase de almacenamiento

Acceso público

Protección

us (varias regiones en Estados Unidos)

Standard

No público

Ninguno

OBJETOS

CONFIGURACIÓN

PERMISOS

PROTECCIÓN

CICLO DE VIDA

OBSERVABILIDAD

INFORMES DE INVENTARIO

Depósitos

bucket\_matapiam01

SUBIR ARCHIVOS

SUBIR CARPETA

CREAR CARPETA

TRANSFERIR LOS DATOS

ADMINISTRAR CONSERVACIONES

DESCARGAR

BORRAR

Filtrar solo por prefijo de nombre

Filtro

Filtrar objetos y carpetas

Mostrar datos borrados

| Nombre  | Tamaño | Tipo       | Fecha de creación    | Clase de almacenamiento | Última modificación  | Acceso público | Historial de version |
|---------|--------|------------|----------------------|-------------------------|----------------------|----------------|----------------------|
| Lab.txt | 164 B  | text/plain | 13 jun 2023 22:17:02 | Standard                | 13 jun 2023 22:17:02 | No público     | -                    |

Google Cloud console showing details for the object **Lab.txt** in the bucket **bucket\_matapiam01**.

**Objeto publicado**

Depósitos > bucket\_matapiam01 > Lab.txt

OBJETO PUBLICADO HISTORIAL DE VERSIONES

DESCARGAR EDITAR METADATOS EDITAR ACCESO BORRAR

| Descripción general      |   |
|--------------------------|---|
| Tipo                     | text/plain  |
| Tamaño                   | 164 B   |
| Fecha y hora de creación | 13 jun 2023 22:17:02  |
| Última modificación      | 13 jun 2023 22:17:02  |
| Clase de almacenamiento  | Standard  |
| Tiempo personalizado     | —   |
| URL pública              | No aplicable  |
| URL autenticada          | <a href="https://storage.cloud.google.com/bucket_matapiam01/Lab.txt">https://storage.cloud.google.com/bucket_matapiam01/Lab.txt</a> |
| URI de gsutil            | gs://bucket_matapiam01/Lab.txt  |
| Permisos                 |   |
| Acceso público           | No público  |
| Protección               |   |
| Historial de versiones   | —   |
| Política de retención    | Ninguno   |
| Estado de conservación   | Ninguno   |
| Tipo de encriptación     | Google-managed key  |

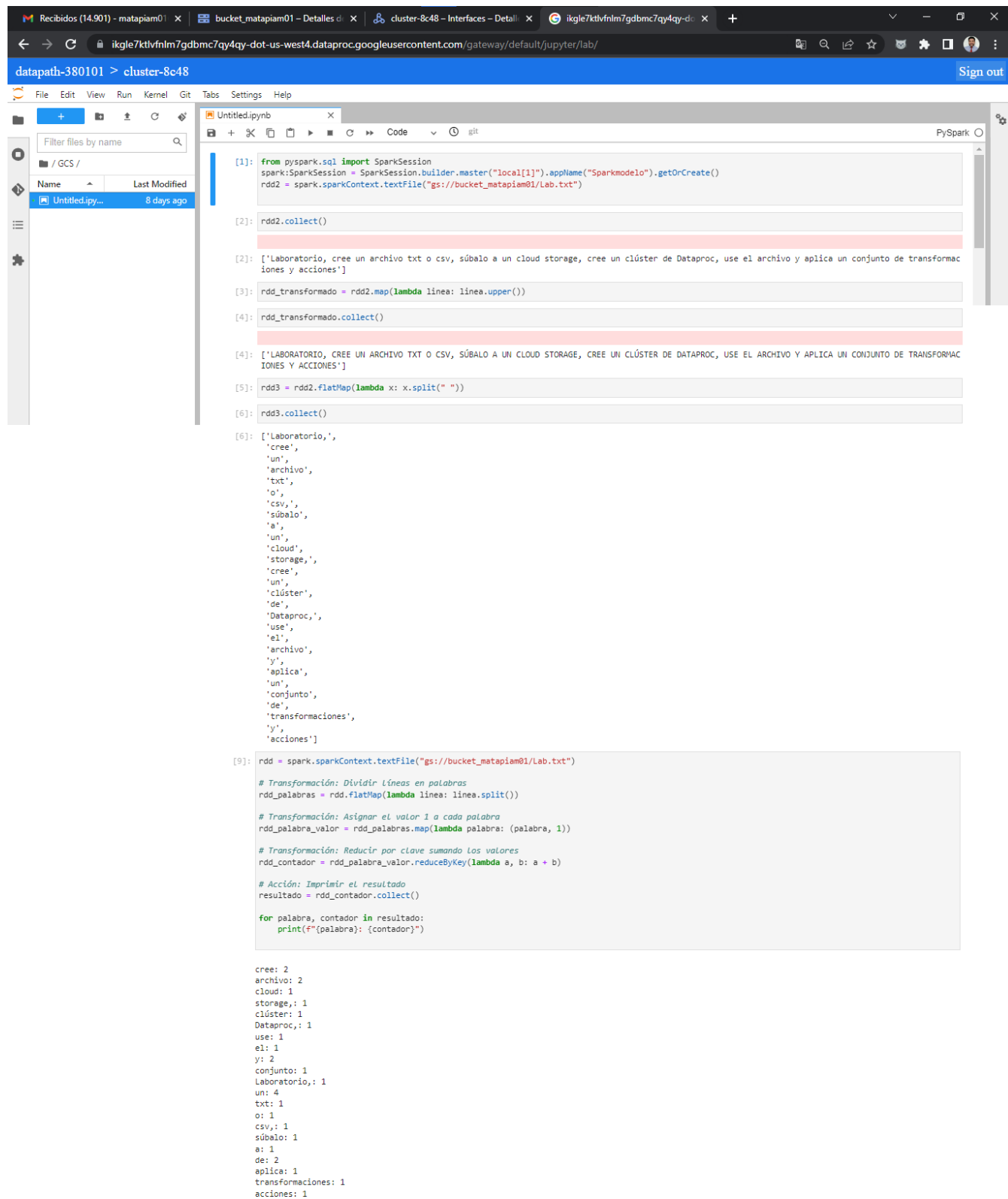
Contenido del archivo: **Lab.txt**

Download link for the file **Lab.txt**:

[ff07709e7b636f8be771f297ec69276ad9e1986ab27aeef1331c54-apidata.googleusercontent.com/download/storage/v1/b/bucket\\_matapiam01/o/Lab.txt?jk=...](https://ff07709e7b636f8be771f297ec69276ad9e1986ab27aeef1331c54-apidata.googleusercontent.com/download/storage/v1/b/bucket_matapiam01/o/Lab.txt?jk=...)

Laboratorio, cree un archivo txt o csv, sábelo a un cloud storage, cree un clÃster de Dataproc, use el archivo y aplica un conjunto de transformaciones y acciones

# Contenido del JupyterLab



The screenshot displays a JupyterLab environment within a web browser. The browser's address bar shows the URL: `ikgle7ktivnlm7gdbmc7qy4qy-dot-us-west4.dataproc.googleusercontent.com/gateway/default/jupyter/lab/`. The JupyterLab interface includes a left sidebar with a file explorer showing a directory named `/ GCS /` containing a file `Untitled.ipynb` modified 8 days ago. The main workspace shows the `Untitled.ipynb` notebook with the following code:

```
[1]: from pyspark.sql import SparkSession
spark:SparkSession = SparkSession.builder.master("local[1]").appName("Sparkmodelo").getOrCreate()
rdd2 = spark.sparkContext.textFile("gs://bucket_matapiam01/Lab.txt")

[2]: rdd2.collect()

[2]: ['Laboratorio, cree un archivo txt o csv, súbalo a un cloud storage, cree un clúster de Dataproc, use el archivo y aplica un conjunto de transformaciones y acciones']

[3]: rdd_transformado = rdd2.map(lambda linea: linea.upper())

[4]: rdd_transformado.collect()

[4]: ['LABORATORIO, CREE UN ARCHIVO TXT O CSV, SÚBALO A UN CLOUD STORAGE, CREE UN CLÚSTER DE DATAPROC, USE EL ARCHIVO Y APLICA UN CONJUNTO DE TRANSFORMACIONES Y ACCIONES']

[5]: rdd3 = rdd2.flatMap(lambda x: x.split(" "))

[6]: rdd3.collect()

[6]: ['Laboratorio,',
'cree',
'un',
'archivo',
'txt',
'o',
'csv,',
'súbalo',
'a',
'un',
'cloud',
'storage,',
'cree',
'un',
'clúster',
'de',
'Dataproc,',
'use',
'el',
'archivo',
'y',
'aplica',
'un',
'conjunto',
'de',
'transformaciones',
'y',
'acciones']

[9]: rdd = spark.sparkContext.textFile("gs://bucket_matapiam01/Lab.txt")

# Transformación: Dividir líneas en palabras
rdd_palabras = rdd.flatMap(lambda linea: linea.split())

# Transformación: Asignar el valor 1 a cada palabra
rdd_palabra_valor = rdd_palabras.map(lambda palabra: (palabra, 1))

# Transformación: Reducir por clave sumando los valores
rdd_contador = rdd_palabra_valor.reduceByKey(lambda a, b: a + b)

# Acción: Imprimir el resultado
resultado = rdd_contador.collect()

for palabra, contador in resultado:
    print(f"{palabra}: {contador}")
```

The output of the notebook shows the following word counts:

```
cree: 2
archivo: 2
cloud: 1
storage,: 1
clúster: 1
Dataproc,: 1
use: 1
el: 1
y: 2
conjunto: 1
Laboratorio,: 1
un: 4
txt: 1
o: 1
csv,: 1
súbalo: 1
a: 1
de: 2
aplica: 1
transformaciones: 1
acciones: 1
```