



**BIG DATA**  
**ACADEMY**

# LABORATORIO DATABRICKS ON AZURE

FORMADOR: ALONSO MELGAREJO  
[alonsoraulmgs@gmail.com](mailto:alonsoraulmgs@gmail.com)

**LABORATORIO DATABRICKS ON AZURE****ENLACE A PORTAL AZURE**<https://portal.azure.com/>**1. CREAR GRUPO DE RECURSOS**

1.1. Seleccionamos la opción de servicios



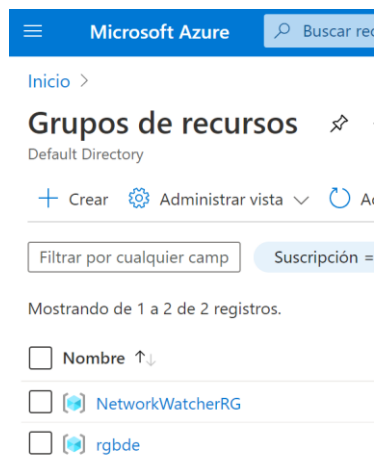
Servicios de



1.2. Seleccionamos la opción "Grupos de recursos"



1.3. Seleccionamos la opción "Crear"



## 1.4. Configuramos:

<b>Suscripción</b>	Nombre de tu suscripción
<b>Grupo de recursos</b>	gr-databricks-bda
<b>Región</b>	(US) Centro-Sur de EE. UU.

Damos clic en “Revisar y crear”

Datos básicos   Etiquetas   Revisar y crear

Grupo de recursos - Contenedor que incluye los recursos relacionados para un grupo de recursos puede contener todos los recursos de la solución o solamente administrar en grupo. Debe decidir cómo quiere asignar los recursos a los grupos que resulte más pertinente para su organización. [Más información](#)

**Detalles del proyecto**

Suscripción \* ⓘ

Grupo de recursos \* ⓘ

**Detalles del recurso**

Región \* ⓘ

**Revisar y crear**   < Anterior   Siguiente: Etiquetas >

## 1.5. Damos clic en “Crear”

**Crear un grupo de recursos** ...

✓ Validación superada.

Datos básicos   Etiquetas   Revisar y crear

**Datos básicos**

Suscripción

Grupo de recursos

Región

**Etiquetas**

**Crear**   < Anterior   Siguiente >   [Descargar una](#)

## 2. HABILITAR SERVICIO DE DATABRICKS

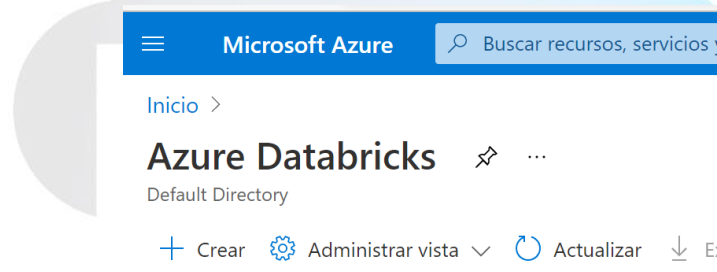
2.1. En el buscador de servicios de Azure, buscamos el servicio:

Databricks

Hacemos clic sobre él



2.2. Seleccionamos la opción “Crear”



2.3. Configuramos

<b>Suscripción</b>	Nombre de tu suscripción
<b>Grupo de recursos</b>	gr-databricks-bda
<b>Nombre del área de trabajo</b>	ws-databricks-bda
<b>Región</b>	Centro-Sur de EE. UU.
<b>Plan de tarifa</b>	<p><b>SI AÚN NO HEMOS USADO NUESTRA CUENTA GRATUITA:</b> Evaluación (Premium: DBU gratis durante 14 días)</p> <p><b>SI YA HEMOS USADO NUESTRA CUENTA GRATUITA:</b> Estándar (Apache Spark, seguro con Azure AD)</p>

Damos clic en “Revisar y crear”

Revisar y crear

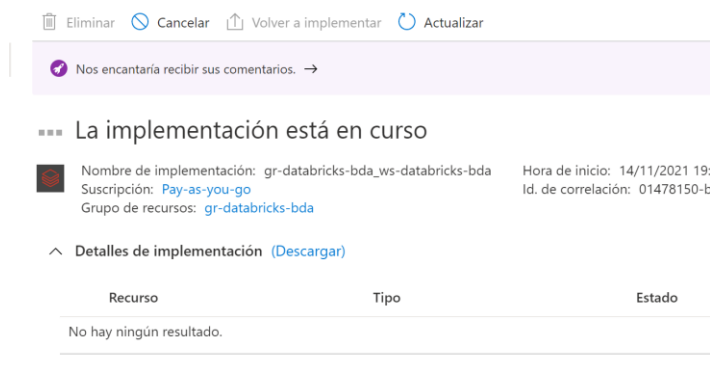
< A

## 2.4. Damos clic en “Crear”



La implementación del servicio comenzará a desplegarse **[TIEMPO: 5 MINUTOS]**

**Mientras el servicio se va desplegando, vamos a crear el storage en donde subiremos los datos a procesar**

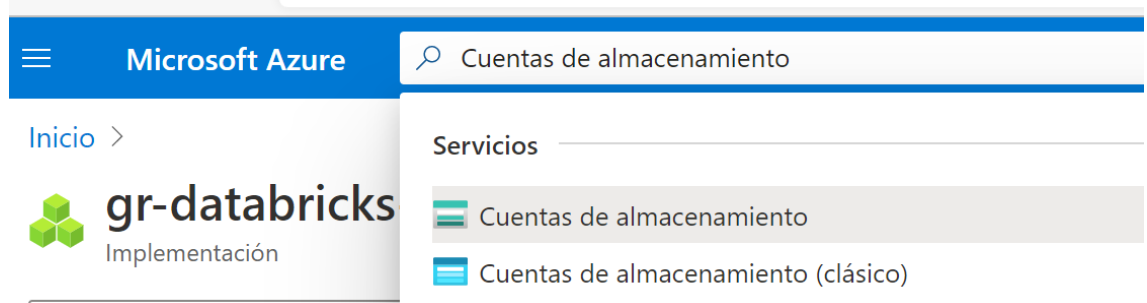


## 3. CREAR STORAGE

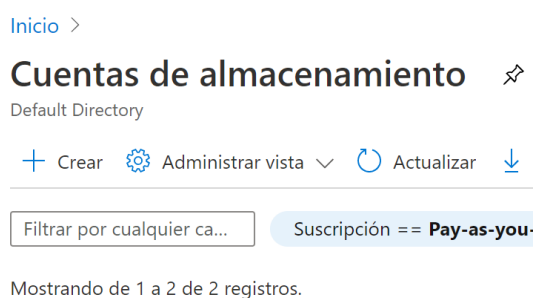
### 3.1. Buscamos el servicio

Cuentas de almacenamiento

Damos clic sobre él



### 3.2. Seleccionamos la opción “Crear”



### 3.3. Configuramos

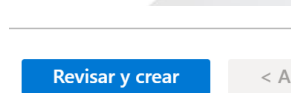
#### PESTAÑA “Datos básicos”

Suscripción	Nombre de tu suscripción
Grupo de recursos	gr-databricks-bda
Nombre de la cuenta de almacenamiento <b>IMPORTANTE: DEBE SER UN NOMBRE ÚNICO A NIVEL MUNDIAL</b>	azurestoragebdaarmg
Región	Centro-Sur de EE. UU.

#### PESTAÑA “Opciones avanzadas”

Habilitar el espacio de nombres jerárquico	Activamos la casilla
--	----------------------

Damos clic en “Revisar y crear”



### 3.4. Damos clic en “Crear”



La implementación del servicio comenzará a desplegarse **[TIEMPO: 1 MINUTO]**

## 4. SUBIR LOS ARCHIVOS A PROCESAR EN EL STORAGE

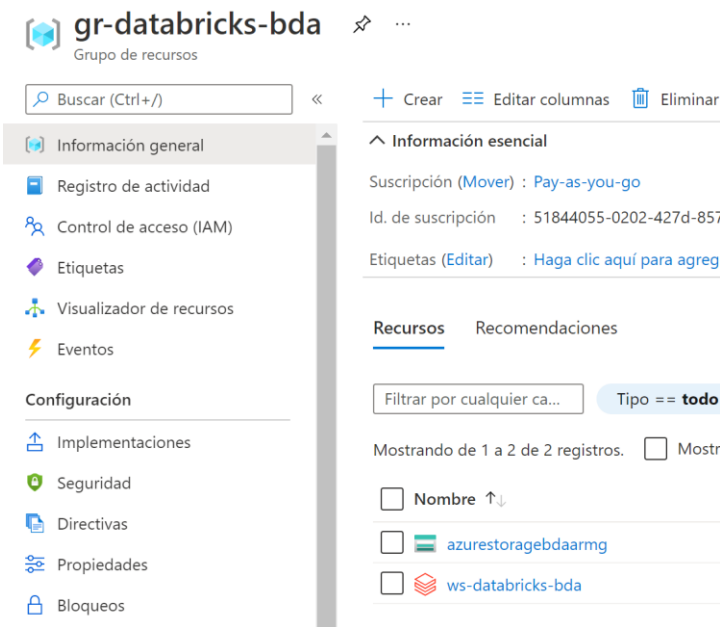
### 4.1. Buscamos el servicio

Grupos de recursos

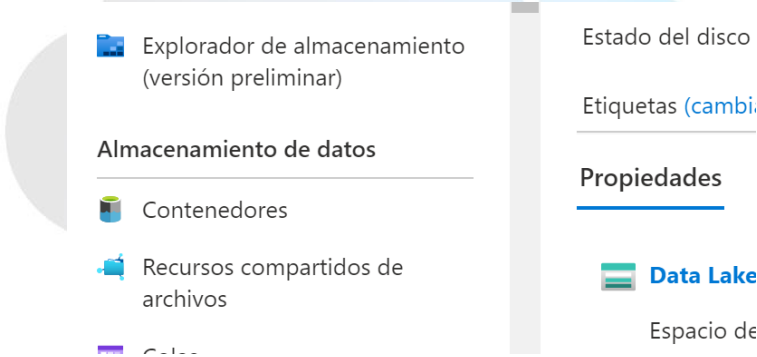
### 4.2. Seleccionamos el recurso

gr-databricks-bda

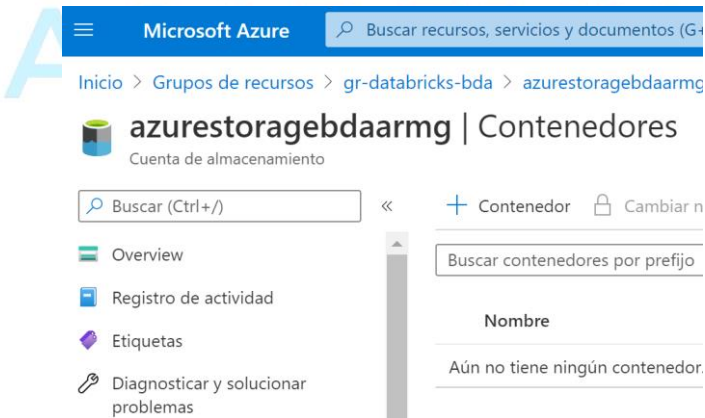
### 4.3. En “Información general” seleccionamos nuestro storage “azurestoragebdaarmg”



4.4. Seleccionamos “Contenedores”



4.5. Crearemos un contenedor. Seleccionamos “+ Contenedores”



4.6. Configuramos

Nombre	dataset
Nivel de acceso público	Privada

Damos clic en “Crear”

## Nuevo contenedor ×

Nombre \*

dataset ✓

Nivel de acceso público ⓘ

Privada (sin acceso anónimo) ▼

▼ Avanzado

### 4.7. Hacemos clic en el contenedor “dataset”



### 4.8. Crearemos dos directorios. Damos clic en “Agregar directorio”

⬆ Cargar + Agregar directorio (

**Método de autenticación:** Clave de acceso

**Ubicación:** dataset

Creamos los directorios:




input
output

### 4.9. Damos clic sobre el directorio “input”

Nombre	Modificado
<input type="checkbox"/>  input	
<input type="checkbox"/>  output	




## 4.10. Damos clic en “Cargar”

 Cargar
  Agregar directorio
  Actualizar

**Método de autenticación:** Clave de acceso ([Cambiar a](#))  
**Ubicación:** [dataset](#) / [input](#)

Buscar blobs por prefijo (distingue mayúsculas de mir


Nombre	Modificado
 [-]	

## 4.11. Seleccionamos el archivo “persona.data” y damos clic en “Cargar”

**Cargar blob** ×

dataset/input/

Files ⓘ



☐ Sobrescribir los archivos si ya existen

☒ Avanzado

**Cargar**

## 5. OBTENER CLAVES DE ACCESO AL STORAGE




## 5.1. Buscamos el servicio

Grupos de recursos




## 5.2. Seleccionamos el recurso

gr-databricks-bda

## 5.3. En “Información general” seleccionamos nuestro storage “azurestoragebdaarmg”

 **gr-databricks-bda**  

Grupo de recursos

<<  Crear  Editar columnas  Eliminar

☒ Información general  
☐ Registro de actividad  
☐ Control de acceso (IAM)  
☐ Etiquetas  
☐ Visualizador de recursos  
☐ Eventos



**Configuración**  
☐ Implementaciones  
☐ Seguridad  
☐ Directivas  
☐ Propiedades  
☐ Bloqueos

**Información esencial**  
 Suscripción (Mover) : [Pay-as-you-go](#)  
 Id. de suscripción : 51844055-0202-427d-857  
 Etiquetas (Editar) : [Haga clic aquí para agreg](#)

**Recursos** Recomendaciones

**Tipo == todo**

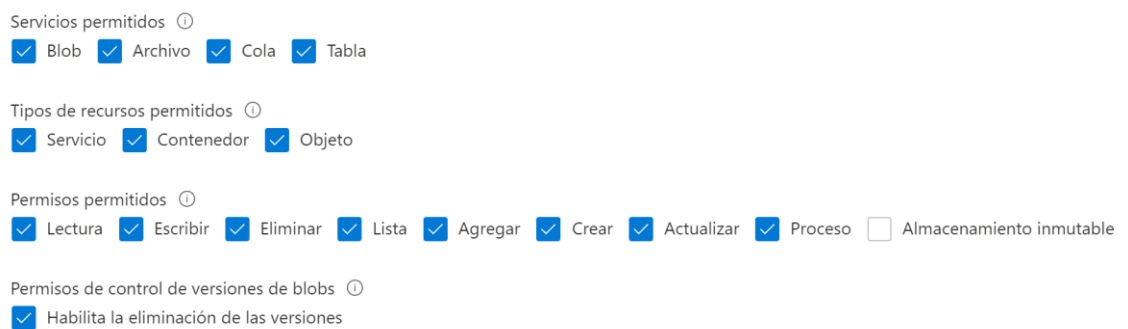
Mostrando de 1 a 2 de 2 registros. ☐ Mostr

<input type="checkbox"/> Nombre ↑↓
<input type="checkbox"/>  azurestoragebdaarmg
<input type="checkbox"/>  ws-databricks-bda

#### 5.4. Seleccionamos “Firma de acceso compartido”



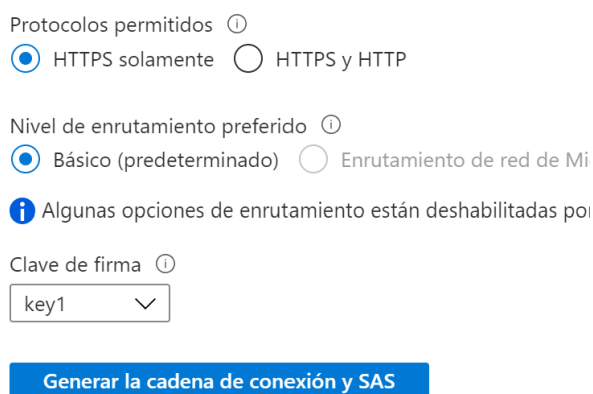
#### 5.5. En tipos de recursos permitidos:



Estarán activadas casi todas las casillas, activamos también:

Servicio
Contenedor
Objeto

#### 5.6. Navegamos hasta al final de la página y seleccionamos “Generar la cadena de conexión y SAS”



### 5.7. De todas las cadenas de conexión generadas, copiaremos el “Token de SAS”

Cadena de conexión  
 BlobEndpoint=https://azurestoragebdaarmg.blob.core.windows.net/;QueueEndpoint=https://azurestoragebdaarmg.queue.core.windows.net/;FileEndpoint=https://azurestorag...

Token de SAS ⓘ  
 ?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-11-15T09:12:45Z&st=2021-11-15T01:12:45Z&spr=https&sig=ayGO077ZqeO00moEm0y%2FyuY5ZPNiPTiTpLtiH...

URL de SAS de Blob service  
 https://azurestoragebdaarmg.blob.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-11-15T09:12:45Z&st=2021-11-15T01:12:45Z&spr=https&si...

URL de SAS del servicio Archivo  
 https://azurestoragebdaarmg.file.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-11-15T09:12:45Z&st=2021-11-15T01:12:45Z&spr=https&sig...

URL de SAS del servicio Cola  
 https://azurestoragebdaarmg.queue.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-11-15T09:12:45Z&st=2021-11-15T01:12:45Z&spr=https&...

URL de SAS de Table service  
 https://azurestoragebdaarmg.table.core.windows.net/?sv=2020-08-04&ss=bfqt&srt=sco&sp=rwdlacupx&se=2021-11-15T09:12:45Z&st=2021-11-15T01:12:45Z&spr=https&si...

## 6. CREAR CLÚSTER DATABRICKS

### 6.1. En el buscador de servicios de Azure, buscamos el servicio:

Databricks

Hacemos clic sobre él



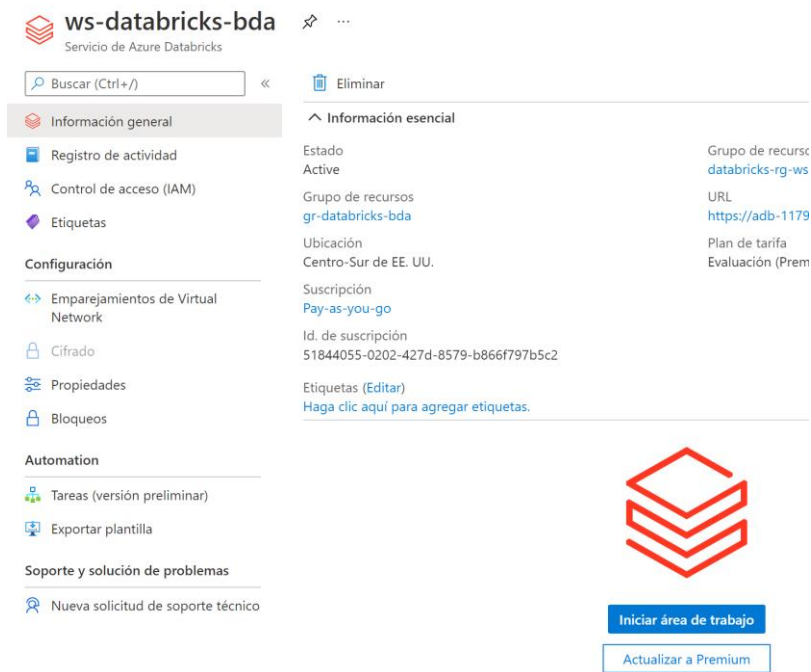
### 6.2. Seleccionamos nuestro workspace “ws-databricks-bda”

Mostrando de 1 a 1 de 1 registros.

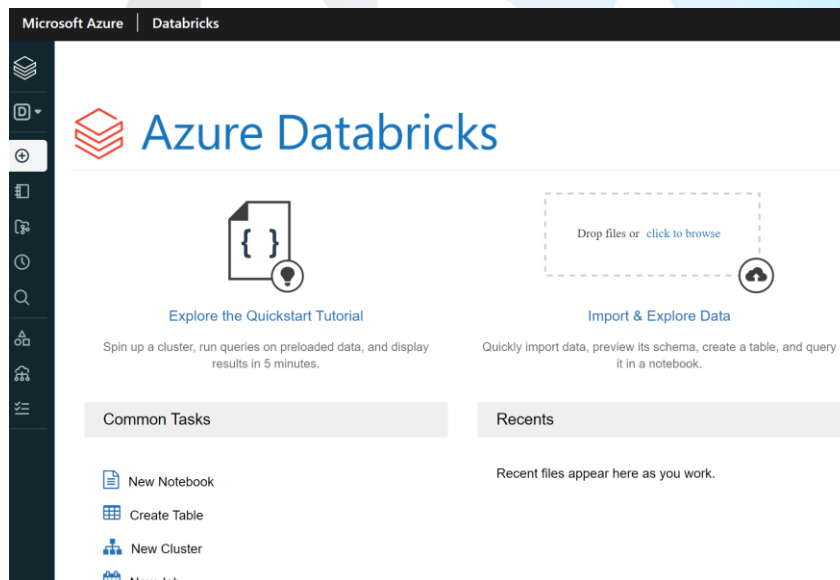
☐ Nombre ↑↓

☐  ws-databricks-bda

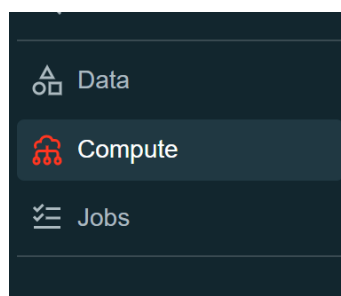
### 6.3. En “Información general” damos clic en “Iniciar área de trabajo”



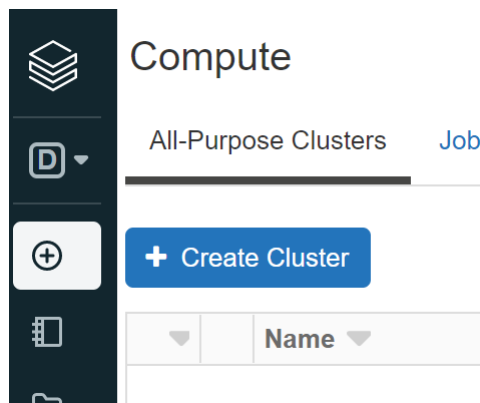
#### 6.4. Nos encontraremos en el Workspace de Databricks



#### 6.5. Seleccionamos la opción "Compute"



## 6.6. Seleccionamos la opción "Create Cluster"



## 6.7. Configuramos


<b>Cluster Name</b>	BIG_DATA_ACADEMY_AZURE
<b>Cluster Mode</b>	Standard
<b>Databricks Runtime Version</b>	Runtime: 9.1
<b>Enable autoscaling</b>	Desmarcar
<b>Terminate after</b>	120
<b>Worker Type</b>	Standard DS3_v2
<b>Workers</b>	1
<b>Driver Type</b>	Same as worker

Hacemos clic en "Create Cluster"

## Create Cluster

## New Cluster



 DBU / hour: 1.5 

Cluster Name

Cluster Mode 



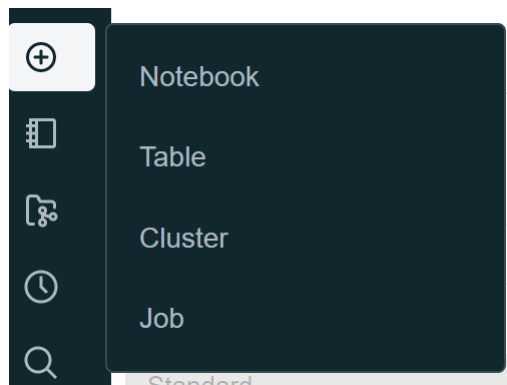
El clúster comenzará a desplegarse **[TIEMPO: 5 MINUTOS]**

## 7. MONTAR EL STORAGE EN DATABRICKS

## 7.1. Necesitaremos los siguientes datos del storage

<b>Cuenta de almacenamiento</b>	azurestoragebdaarmg
<b>Contenedor</b>	dataset
<b>Token de SAS</b>	Revisar el paso 6.7

## 7.2. Creamos un notebook



## 7.3. Configuramos:

<b>Name</b>	AZURE
<b>Default Language</b>	Scala
<b>Cluster</b>	BIG_DATA_ACADEMY_AZURE

Damos clic en "Create"

### Create Notebook

×

Name

AZURE

Default Language

Scala

▼

Cluster

BIG\_DATA\_ACADEMY\_AZURE

▼

Cancel

Create

## 7.4. Sobre el notebook escribimos

### CONFIGURAMOS LOS PARÁMETROS DE MONTADO

```
var storage = "azurestoragebdaarmg"  
var container = "dataset"  
var tokenSas = "xxxxxx"
```

**REALIZAMOS EL MONTADO**

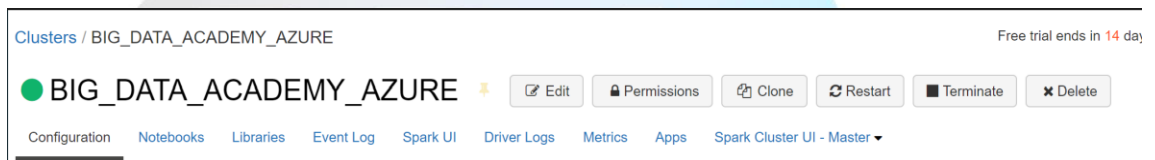
```
var config = "fs.azure.sas." + container + "." + storage + ".blob.core.windows.net"
dbutils.fs.mount(
  source = "wasbs://" + container + "@" + storage + ".blob.core.windows.net",
  mountPoint = "/mnt/" + container,
  extraConfigs = Map(config -> tokenSas)
)
```

**VERIFICAMOS EL MONTADO**

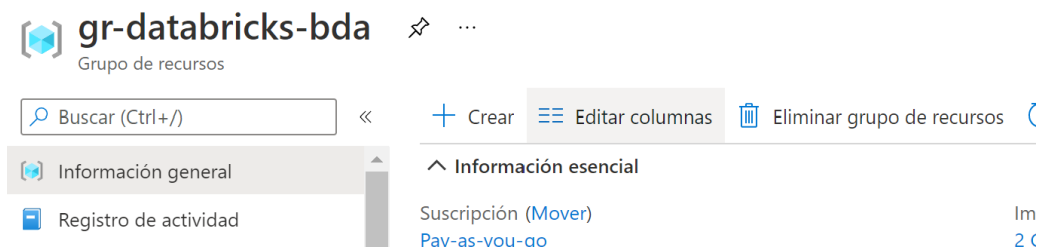
```
%fs ls /mnt/dataset
```

**LEEMOS LOS DATOS**

```
var df = spark.read.format("csv").option("header", "true").option("delimiter",
"|").load("/mnt/dataset/input/persona.data")
df.show()
```

**8. DESTRUIR RECURSOS****8.1. Destruimos el clúster desde la opción "Delete"****8.2. Buscamos el servicio**

**8.3. Seleccionamos el recurso**

**8.4. Seleccionamos "Eliminar grupo de recursos"****8.5. Se nos pedirá confirmar la eliminación, escribimos el nombre del grupo**

Y damos clic en “Eliminar”

¿Está seguro de que desea eliminar 'gr-...'





Advertencia: La eliminación del grupo de recursos 'gr-databricks-bda' no se puede revertir. La acción que va a realizar no se puede deshacer. Si continúa, se eliminará este grupo de recursos y todos los recursos que contiene de forma permanente.

ESCRIBA EL NOMBRE DEL GRUPO DE RECURSOS:

gr-databricks-bda

RECURSOS AFECTADOS

Se eliminarán 2 recursos de este grupo de recursos.

Nombre	Tipo	Ubicación
 azurestoragebdaarmg	Cuenta de almacen...	Centro-Sur de EE. U...
 ws-databricks-bda	Servicio de Azure D...	Centro-Sur de EE. U...

Eliminar

Cancelar

El grupo de recursos y todos los recursos se eliminarán.