In [ ]:

*## #######################################################################################################*  
*##*  
*## @copyright Big Data Academy [info@bigdataacademy.org]*  
*## @professor Alonso Melgarejo [alonsoraulmgs@gmail.com]*  
*##*  
*## #######################################################################################################*

1. Tamaño de clúster[¶](#gjdgxs)

In [ ]:

*#Definimos el tamaño del clúster*  
*#*  
*# Cantidad de servidores: 2*  
**%**number\_of\_workers 2  
*# Tipo de servidores: G 1X (16 GB RAM | 4 vCPU)*  
**%**worker\_type G.1X

2. Lectura de datos[¶](#30j0zll)

In [ ]:

*#Leemos los datos*  
*#IMPORTANTE: Reemplazamos "XXX" por nuestras iniciales*  
*#IMPORTANTE: La primera vez que ejecutamos una porción de código, el clúster se crea (TIEMPO: 1 MINUTO)*  
dfPersona **=** spark**.**read**.**format("parquet")**.**load("s3://deltalakebdaXXX/silver/persona/")

In [ ]:

*#Verificamos*  
dfPersona**.**show()

3. Procesamiento[¶](#1fob9te)

In [ ]:

*#Filtramos a los que tienen más de 60 años*  
dfReporte **=** dfPersona**.**filter(dfPersona["EDAD"] **>** 60)

In [ ]:

*#Verificamos*  
dfReporte**.**show()

4. Almacenamiento[¶](#3znysh7)

In [ ]:

*#Escribimos el resultado final*  
*#IMPORTANTE: Reemplazamos "XXX" por nuestras iniciales*  
dfReporte**.**write**.**format("parquet")**.**mode("overwrite")**.**save("s3://deltalakebdaXXX/gold/reporte/")

In [ ]:

*#Verificamos la escritura*  
*#IMPORTANTE: Reemplazamos "XXX" por nuestras iniciales*  
dfReporteRecuperado **=** spark**.**read**.**format("parquet")**.**load("s3://deltalakebdaXXX/gold/reporte/")  
  
*#Verificamos*  
dfReporteRecuperado**.**show()