

---

# Spark como Motor de Procesamiento Big Data

---

---

# Concepto

---

BIG DATA  
ACADEMY

# Spark

Es un motor de procesamiento distribuido paralelo in-memory. Proporciona apis en Java, Scala, Python y R. Spark mantiene la escalabilidad lineal y la tolerancia a fallos de MapReduce, pero amplía sus bondades gracias a varias funcionalidades.



---

# Objetivo fundamental

---

A large, faint, light blue watermark of the BDA logo and the text "BIG DATA ACADEMY" is centered in the background of the slide, behind the main title.

# Objetivo fundamental de Spark

---

**Ejecutar procesos  
lo más rápido  
posible**

---

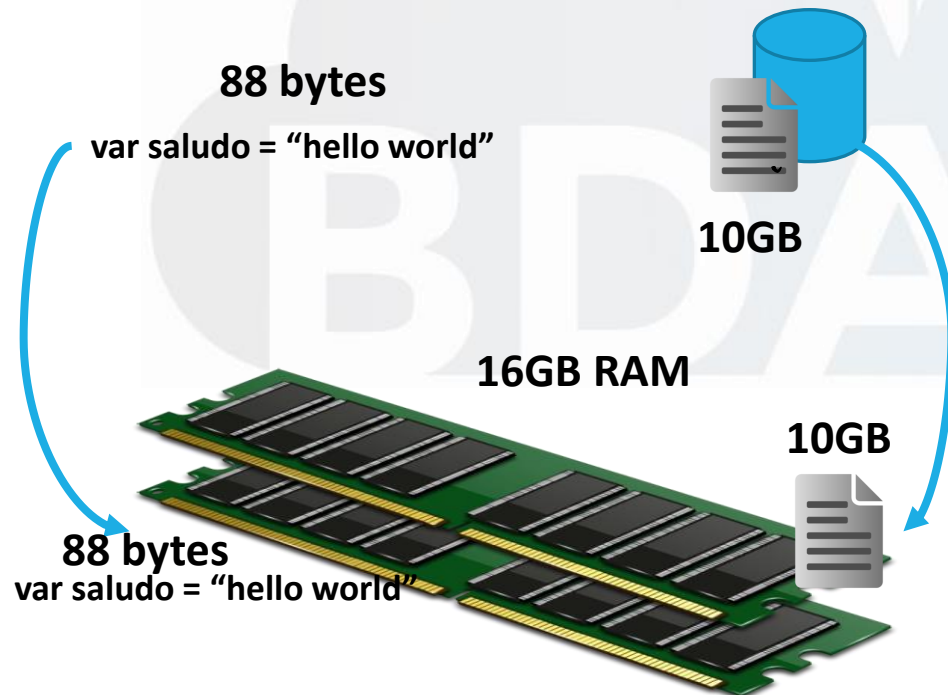
# Variables in-memory

---

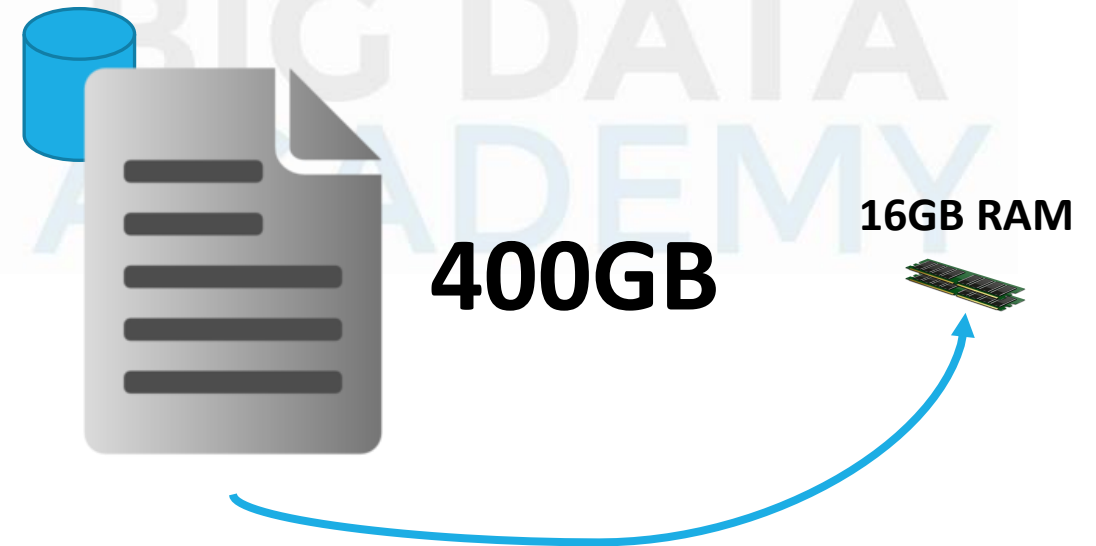
A large, faint, light-blue watermark is centered on the slide. It consists of a stylized cloud icon with a mountain peak inside, followed by the text "BDA" in a large, bold, sans-serif font, and "BIG DATA ACADEMY" in a smaller, all-caps, sans-serif font below it.

# Variables en memoria

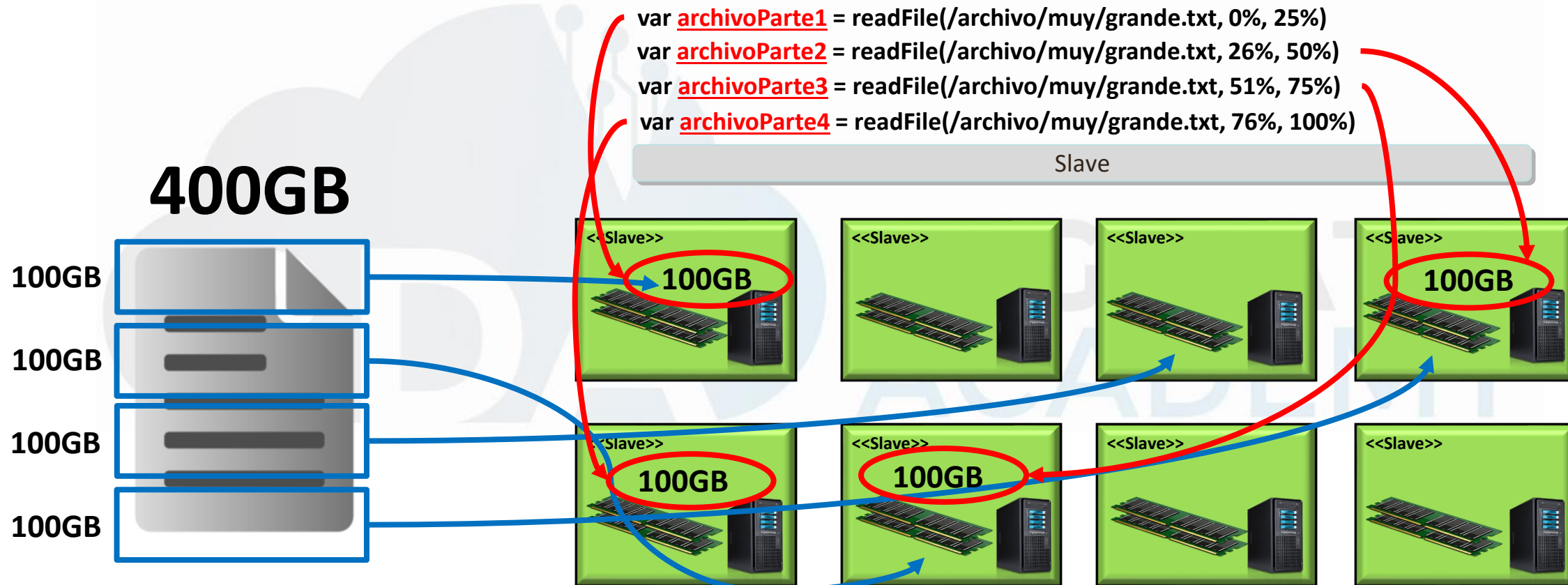
¿Cómo se crea una variable en memoria?



¿Y si tengo un archivo muy grande?



# En un clúster clásico

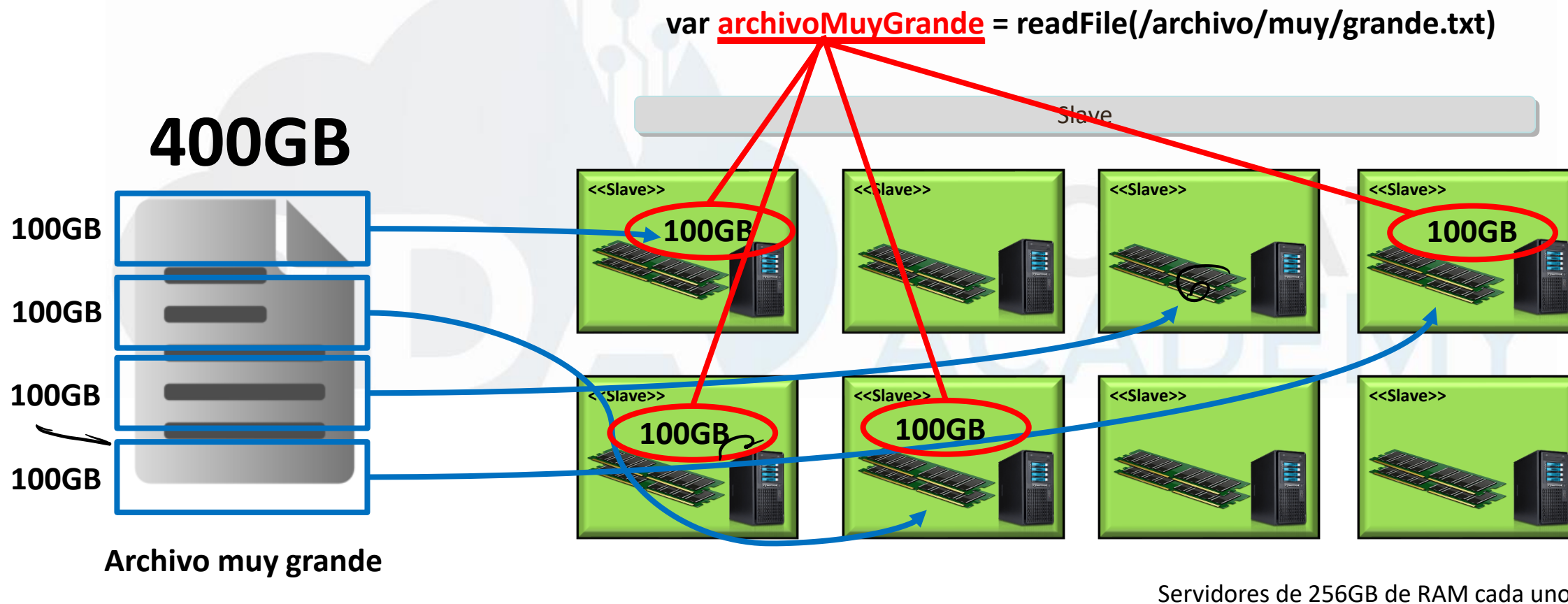


Archivo muy grande

Servidores de 256GB de RAM cada uno



# RDD: Resilient Distributed Dataset



# Agregando estructura a los RDD: Los Dataframes

**RDD**

+

**METADATA**

=

**DATAFRAME**



(CAMPO1 STRING,  
CAMPO2 INT,  
CAMPO3 DOUBLE,  
...)

