Spark como Motor de Procesamiento Big Data

Spark

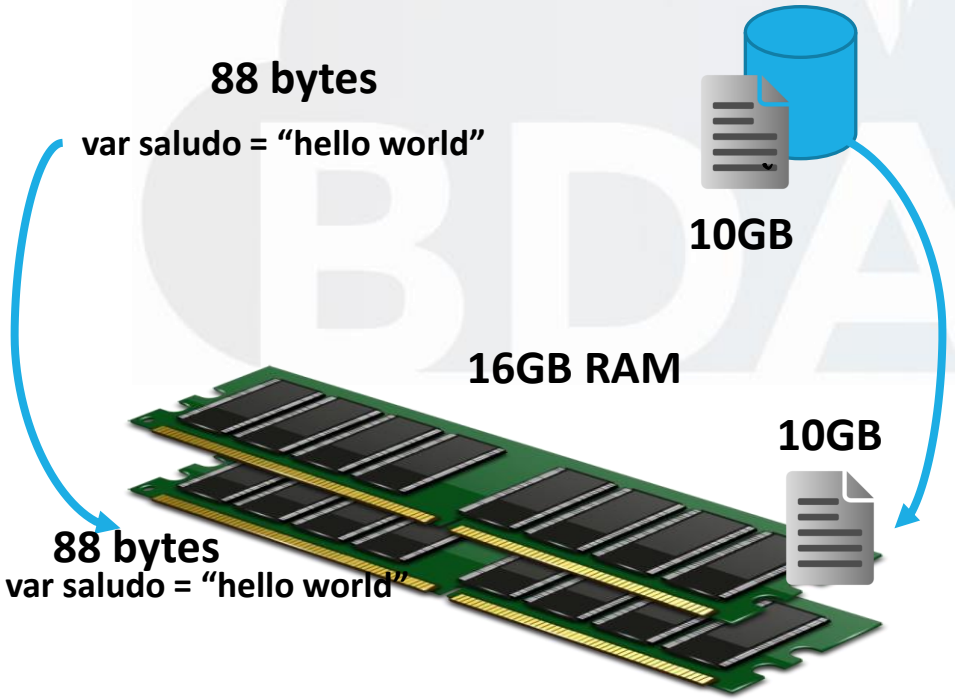
Es un **motor de procesamiento distribuido paralelo in-memory**. Proporciana **apis en Java, Scala, Python y R**. Spark mantiene la **escalabilidad lineal y la tolerancia a fallos** de MapReduce, pero amplía sus bondades gracias a varias funcionalidades.

Objetivo fundamental de Spark

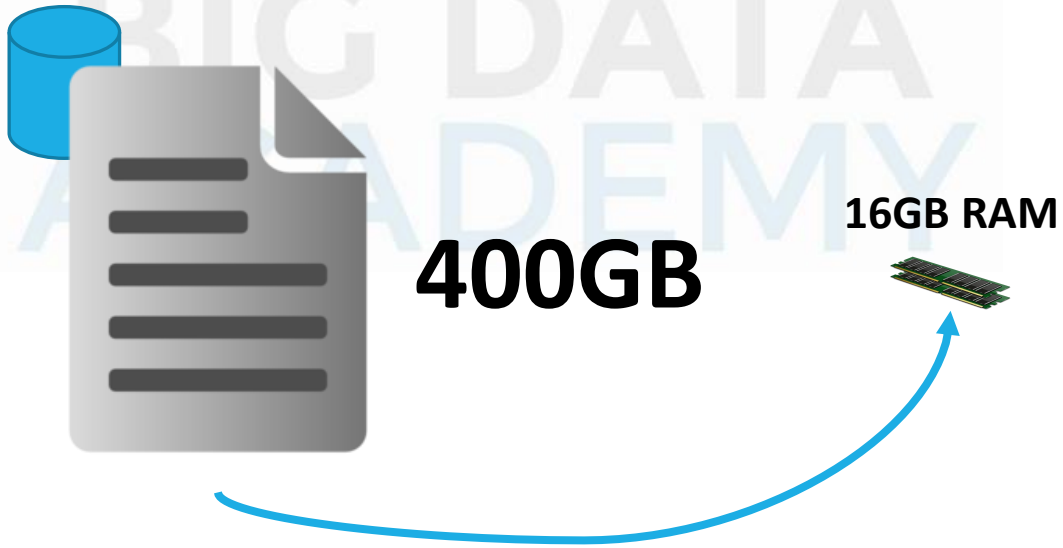
Ejecutar procesos **lo más rápido posible**

Variables en memoria

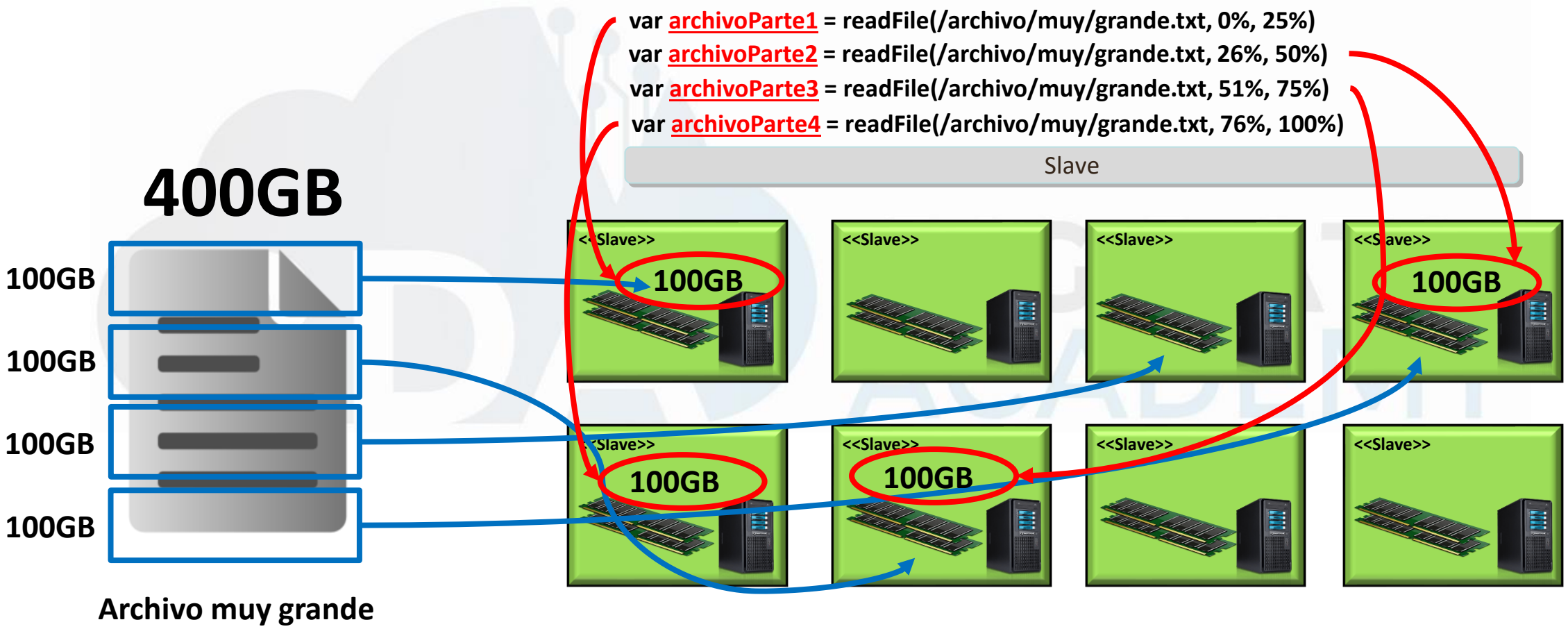
¿Cómo se crea una variable en memoria?



¿Y si tengo un archivo muy grande?

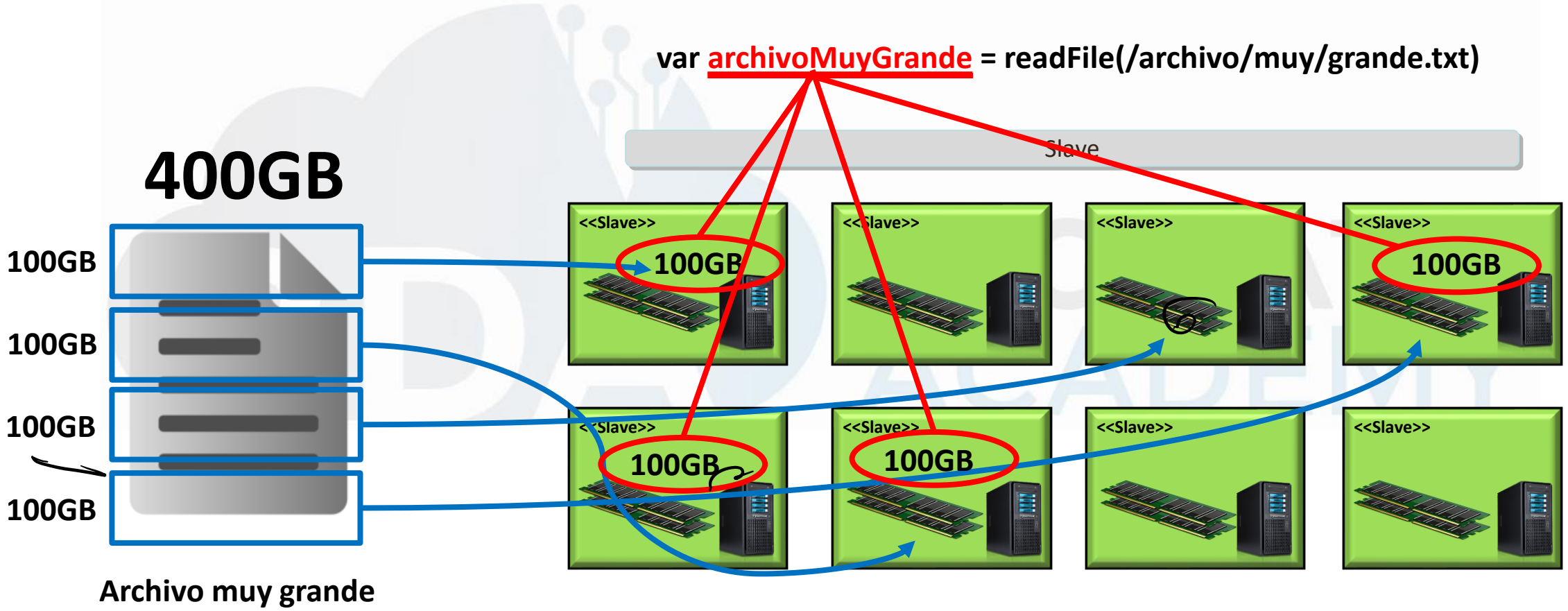


En un clúster clásico



Servidores de 256GB de RAM cada uno

RDD: Resilient Distributed Dataset



Servidores de 256GB de RAM cada uno

Agregando estructura a los RDD: Los Dataframes

RDD + **METADATA** = **DATAFRAME**

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