Usando “Big Data” con R

Data Days 2022

Edgar Ruiz

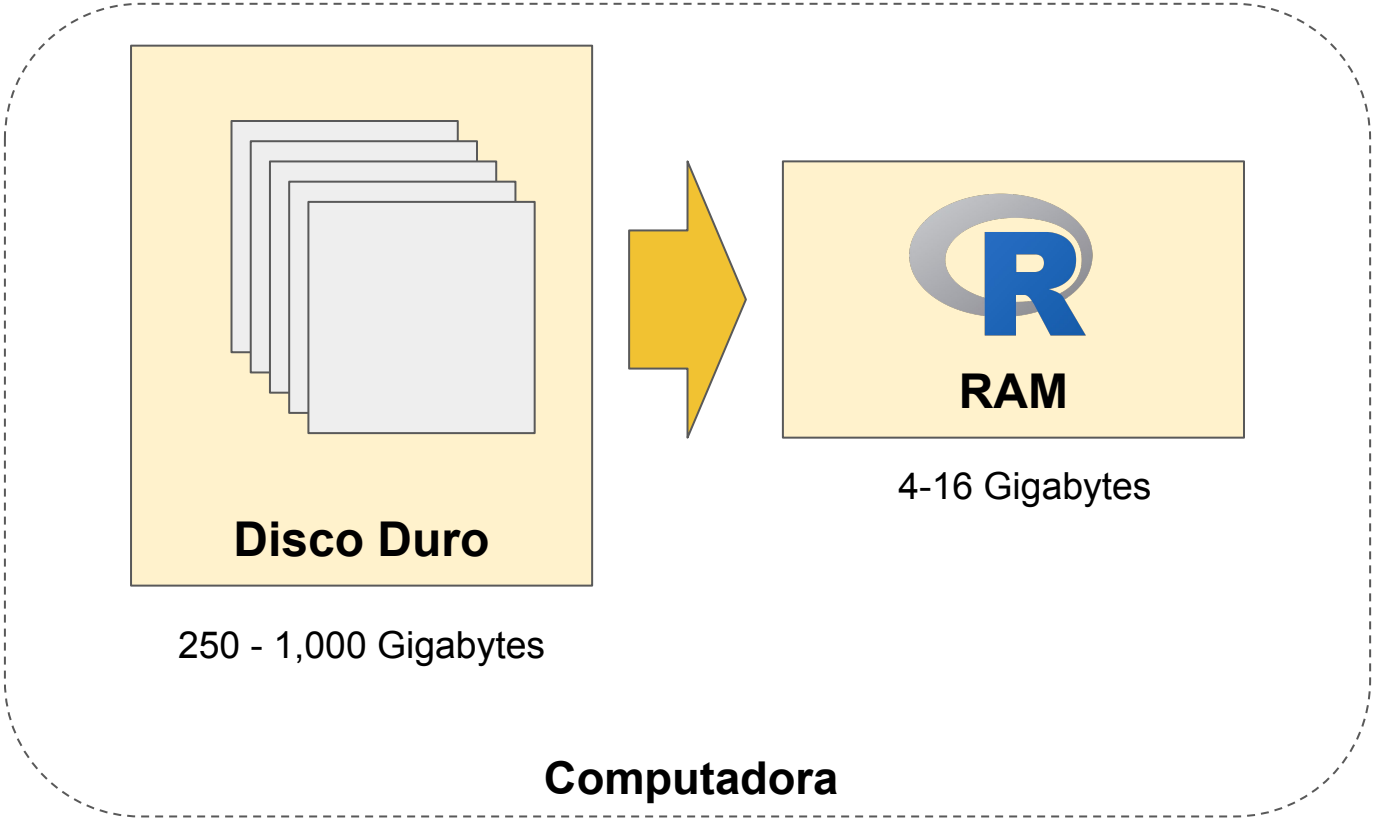
Twitter: **@theotheredgar**

linkedin.com/in/edgararuiz

Elementos físicos

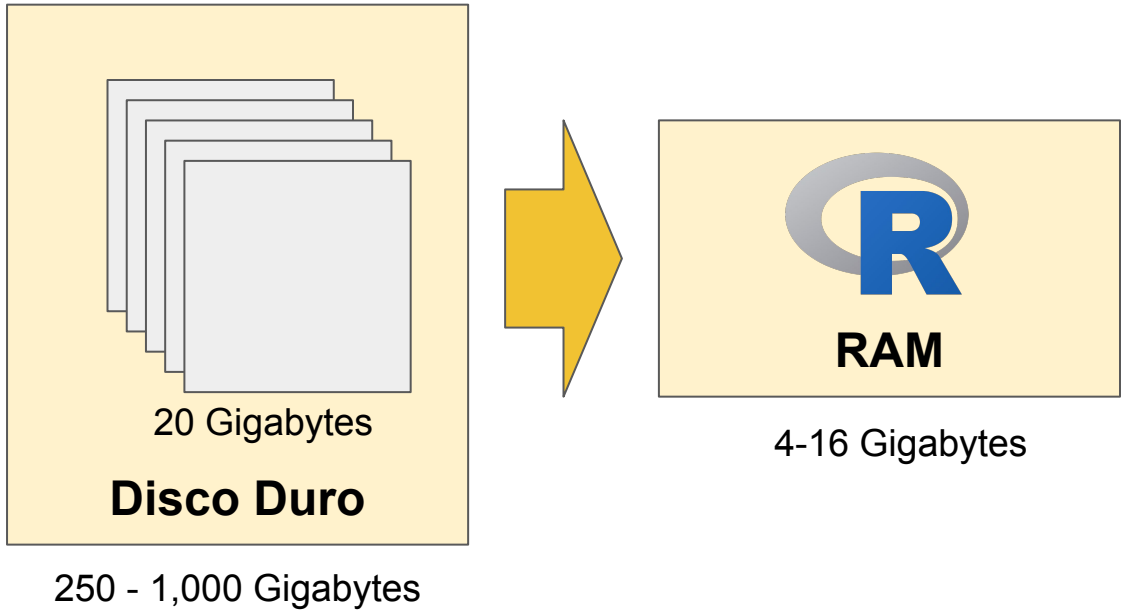
- **Disco duro** - Donde grabamos nuestros archivos. Todo lo que grabamos se queda ahí al menos que lo borramos.
- **RAM** (Random Access Memory) - La “memoria” de la computadora. Estos son “chips” dentro de la computadora que contienen información transitoria, retiene en lo que estamos trabajando en ese momento. *La vasta mayoría de los procesos de R dependen en tener los datos en RAM*

“Quiero leer mis archivos en R”



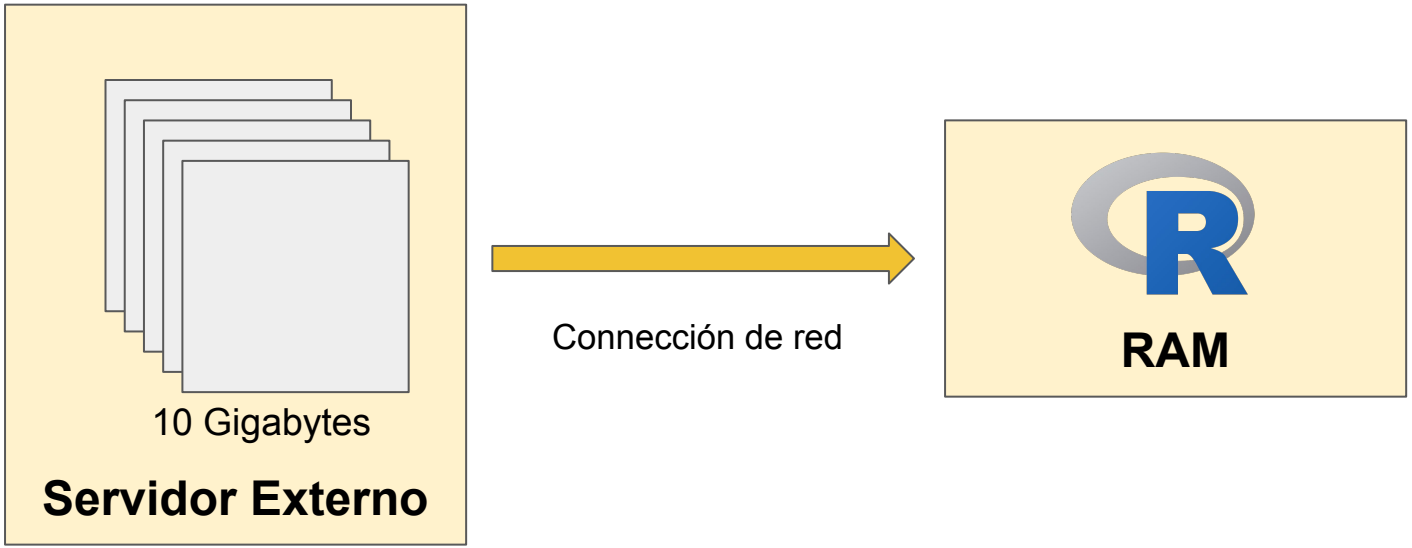
“Big Data” o Datos Masivos

Datos más grandes que el RAM de nuestra computadora



“Big Data” o Datos Masivos

Datos que residen en otra computadora o servidor y la transmisión completa (download) no es ideal.

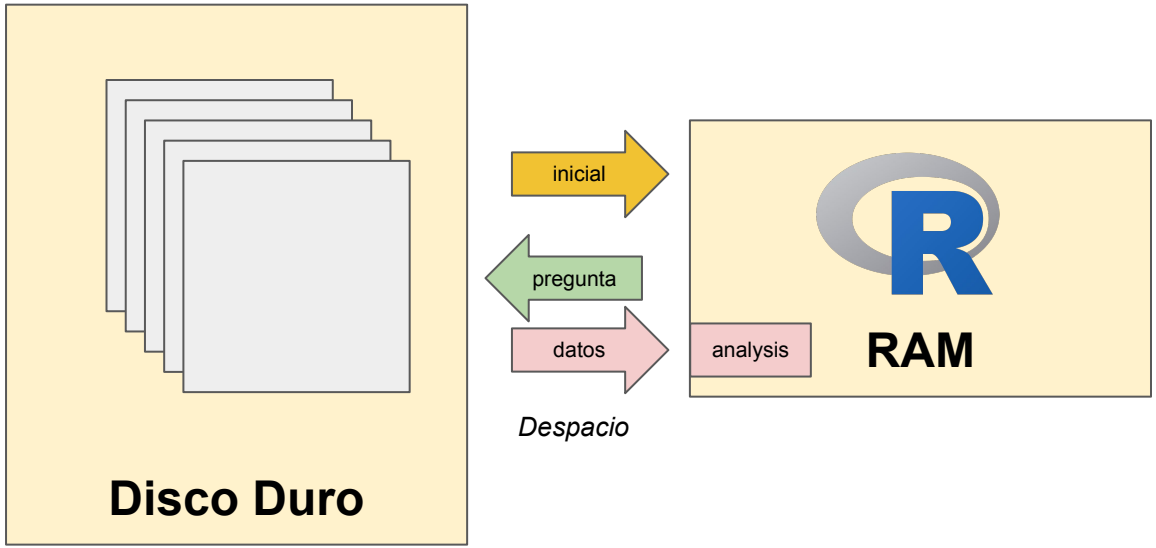


Resumen

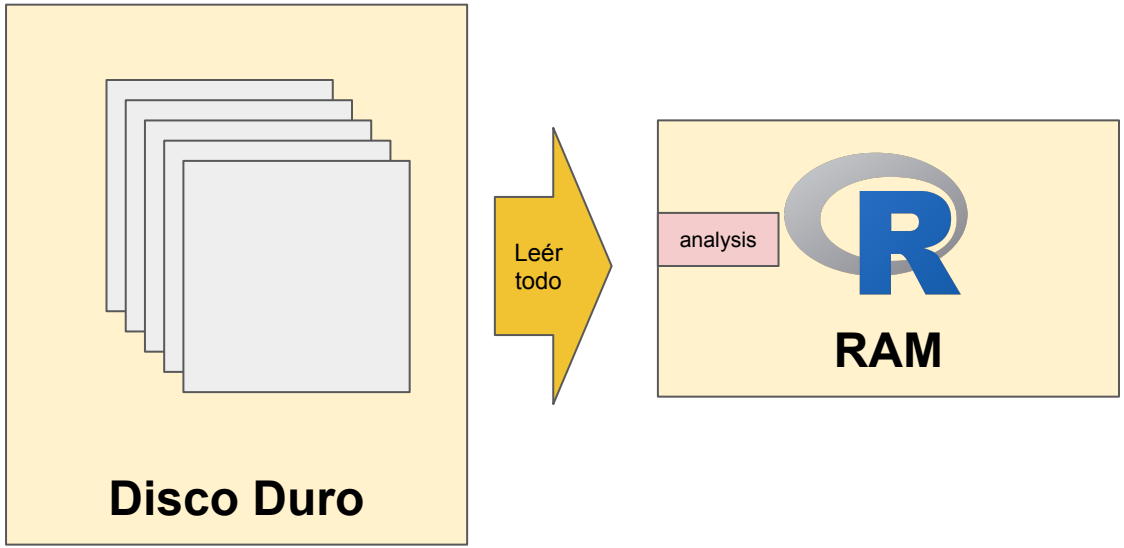
Método	Ventajas / Desventajas
`tidyverse`	• Mas lento para analysis • Modelos y visualizaciones son posibles
`data.table`	• Rápido para analysis • Modelos y visualizaciones son posibles • Tiene que cargar todos los datos
Arrow	• Rápido analysis • No carga los datos en la memoria • Trabaja mejor con Parquet • Modelos y visualizaciones no son posibles, al menos que uno convierta los datos
Bases de Datos	• Podemos utilizar la base de datos como el “motor” de los analysis • Modelos y visualizaciones no son posibles, al menos que uno convierta los datos
Spark	• Rápido analysis • No carga los datos en la memoria • Gran cantidad de modelos • Visualizaciones no son posibles, al menos que uno convierta los datos

Todos funcionan usando `dplyr !!`

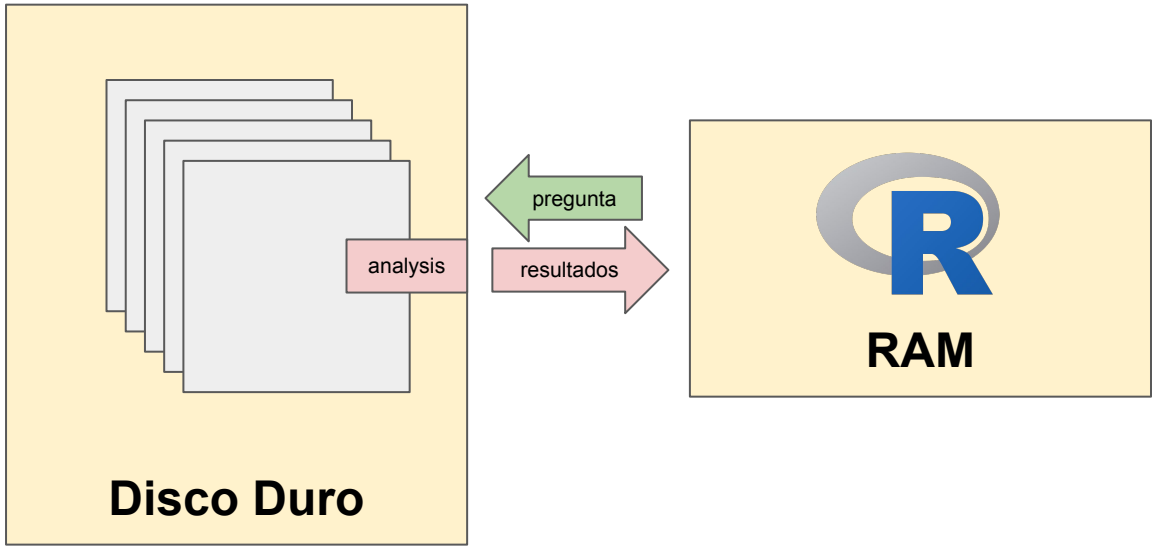
Usando “lazy” read en `readr`



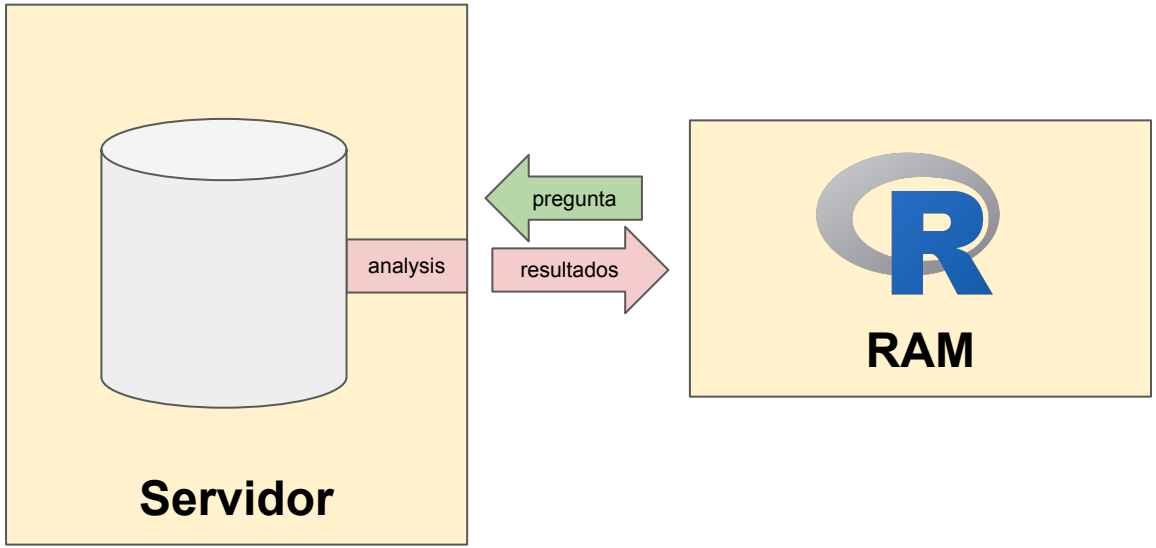
data.table



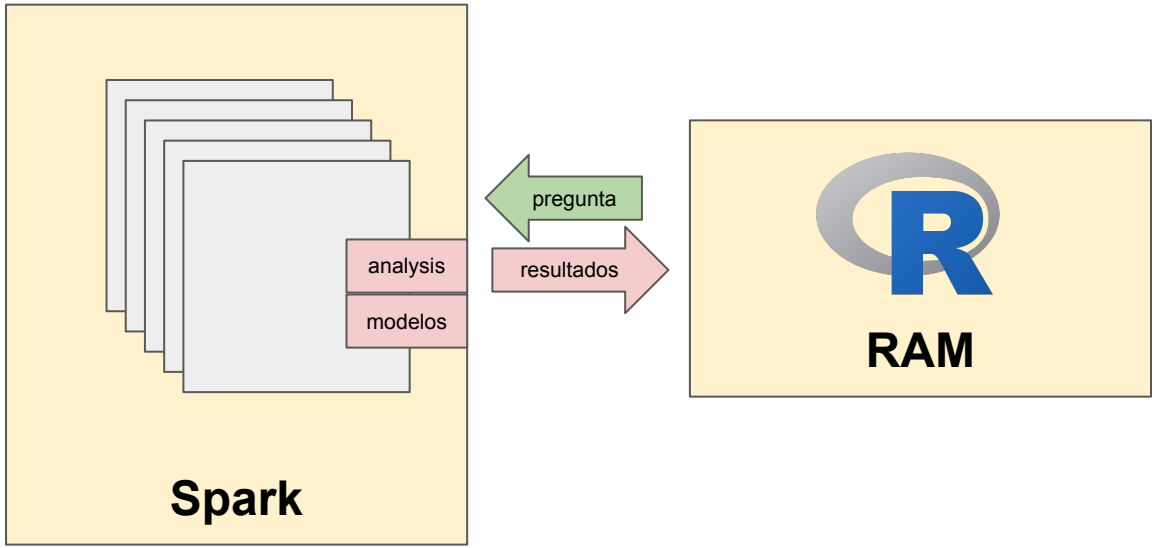
Arrow



Bases de datos (`dbplyr`)



Spark



Resumen

Método	Ventajas / Desventajas
`tidyverse`	• Mas lento para analysis • Modelos y visualizaciones son posibles
`data.table`	• Rápido para analysis • Modelos y visualizaciones son posibles • Tiene que cargar todos los datos
Arrow	• Rápido analysis • No carga los datos en la memoria • Trabaja mejor con Parquet • Modelos y visualizaciones no son posibles, al menos que uno convierta los datos
Bases de Datos	• Podemos utilizar la base de datos como el “motor” de los analysis • Modelos y visualizaciones no son posibles, al menos que uno convierta los datos
Spark	• Rápido analysis • No carga los datos en la memoria • Gran cantidad de modelos • Visualizaciones no son posibles, al menos que uno convierta los datos

Todos funcionan usando `dplyr !!`

Gracias!