Big Data con R: del laptop a la nube, escalando tu análisis de datos

STATDATA-Hub 2024

José Fernando Zea

¿Qué hecho en datos?

EL TIEMPO
CASA EDITORIAL

Chief Data Scientist

Casa Editorial El Tiempo (CEET)

Data Science

5 años, DNP, Colpatria, UIAF, CEET

Estadístico

DANE, Naciones Unidas, Alcaldía de Bogotá, consultoras (*datos*)

BOGOTAR
Users Group

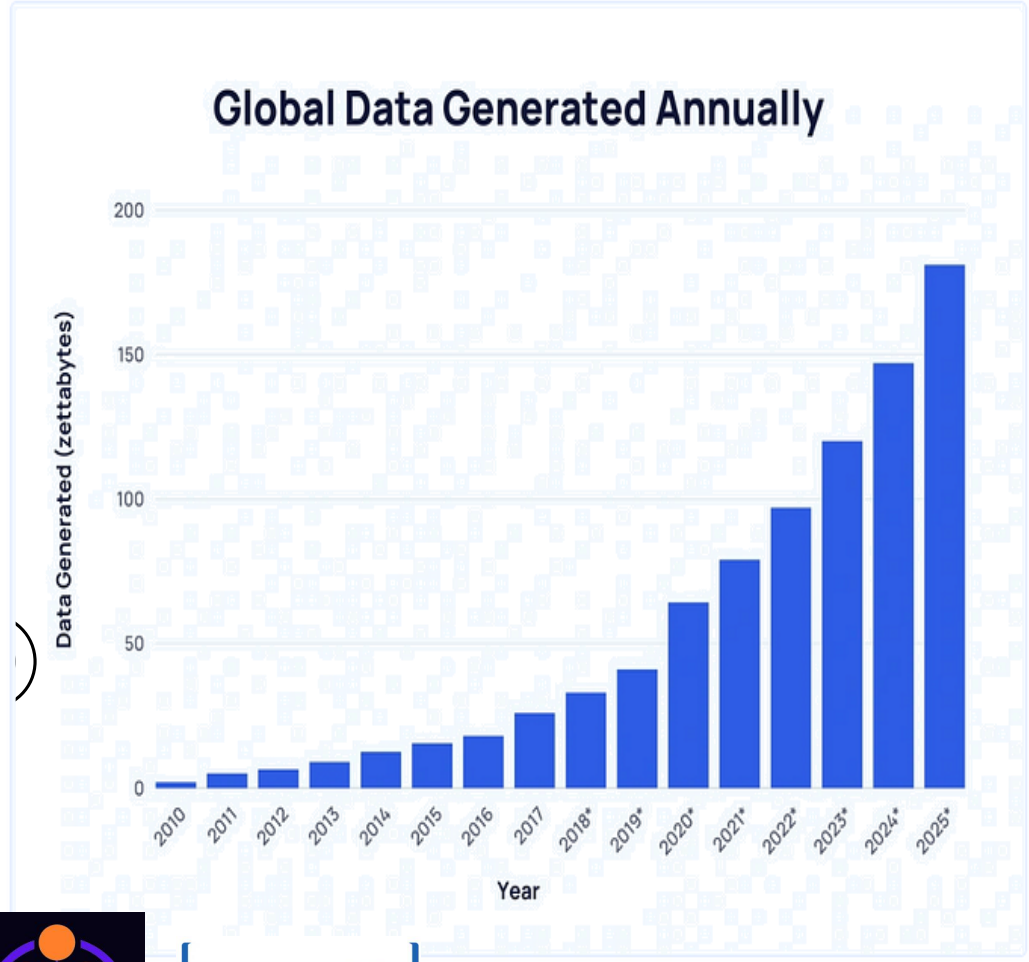
Bogota R User Group - BRUG

- ✓ 1938 miembros a hoy

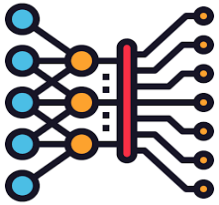
https://www.meetup.com/bogota-r-users-group/

BOGOTÁR Users Group

Navegamos un mar de datos



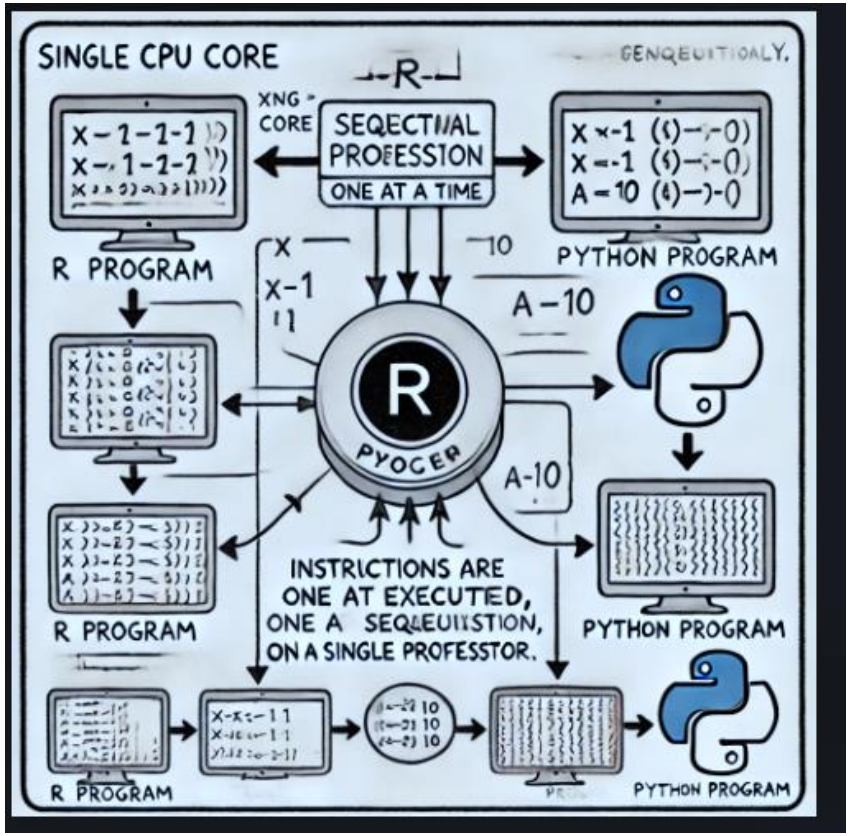
"We expect the data universe to grow more than 10 times from 2020 to 2030, reaching 660 zettabytes—equivalent to 610 iPhones (128GB) per person".
USB Wealth Management



BOGOTAR Users Group

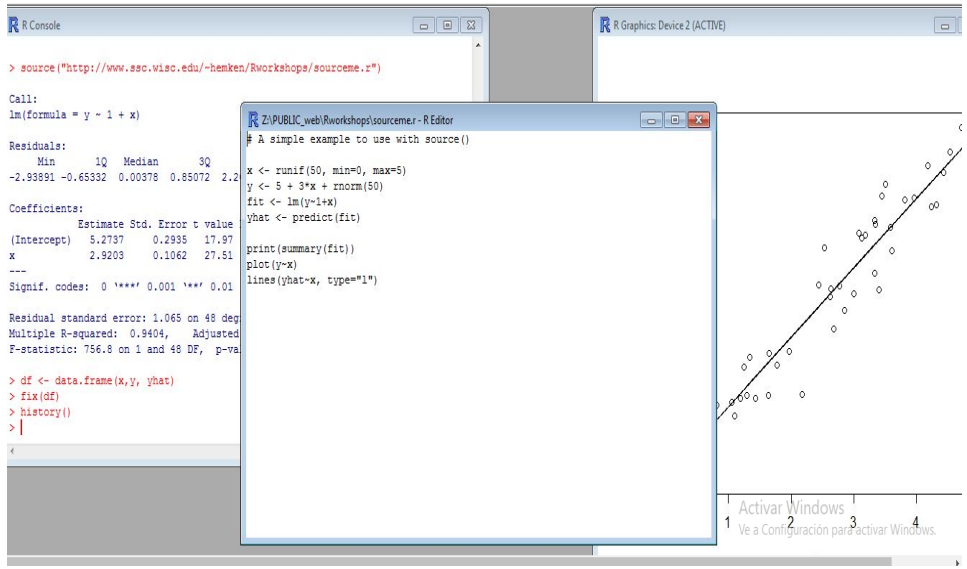
¿Por qué Big Data con R?

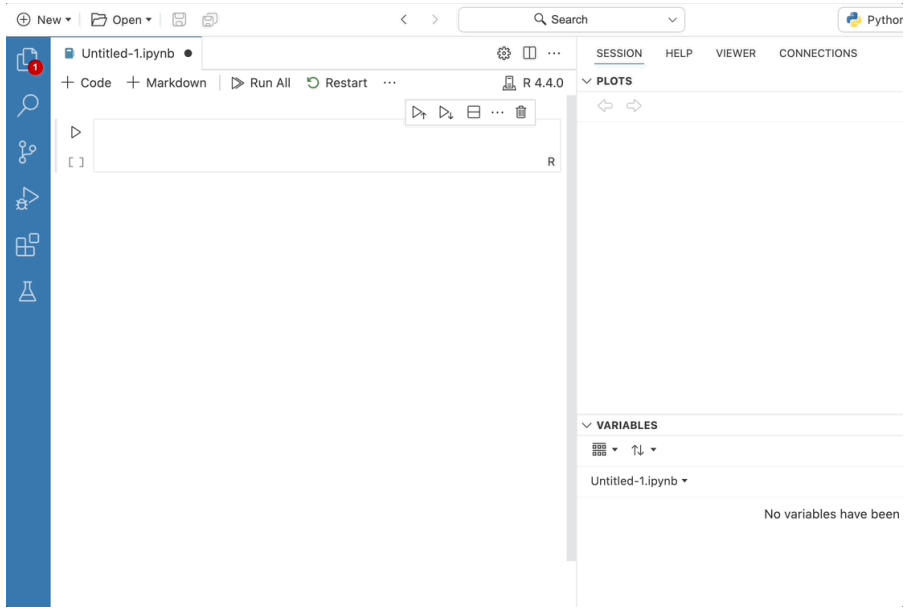
Python		R	
Paquetes	**Sintaxis**	**Paquetes**	**Sintaxis**
pandas 2	pandas	dtplyr	dplyr
polars	polars	tidypolars	dplyr
pyspark	pyspark / SQL	sparklyr	dplyr / SQL
duckdb	SQL	duckdb	dplyr /SQL
arrow	arrow	arrow	dplyr
pyodbc	SQL	DBI / dbplyr	dplyr /SQL



Datawarehouse: Google Bigquery, AWS Redshift,, Azure Databrics, Snowflake, …
DB: Oracle, PostgreSQL, MySQL, ..

R se sigue enseñando como hace 10 años





data <- starwars[starwars$species == "Human" & starwars$height > 185,]

data <- starwars %>% filter(species == "Human", height > 185)

r/statistics • 10 yr. ago
[deleted]

How well does R handle big data nowadays?

Lately I've been doing some short term contracts involving SAS and god dam do I miss R and Python so much. I work with pretty big data and most online sources that say that R can't handle large data due to ram are all pretty dated e.g http://stats.stackexchange.com/questions/33780/r-vs-sas-why-is-sas-preferred-by-private-companies. I don't really keep up to date with these things so I was wondering if anyone has an update on the situation?

DuckDB

APACHE ARROW

sparklyr

Estudio de caso

README

Proyecto de Big Data con R 🚀

¡Bienvenido a este emocionante proyecto de Big Data utilizando R! 🎉 Aquí exploraremos cómo R se puede integrar con herramientas poderosas como Google BigQuery y varios paquetes diseñados para manejar grandes volúmenes de datos.

¿Por qué R para Big Data? 🤔

R no solo es un lenguaje estadístico, sino que también cuenta con un ecosistema robusto para el análisis de Big Data. Entre sus características más destacadas están:

- **Integración con BigQuery**: R puede conectarse fácilmente a Google BigQuery, permitiendo realizar consultas sobre grandes conjuntos de datos sin necesidad de descargarlos localmente. 🌐
- **Paquete** `arrow` : Este paquete permite una eficiente lectura y escritura de datos en formato Apache Arrow, lo que facilita el manejo de grandes volúmenes de datos. 📊
- **Paquete** `duckdb` : DuckDB es un motor de base de datos en memoria que permite realizar consultas SQL sobre grandes conjuntos de datos de manera eficiente. 🦆

Datos 📊

Pueden encontrar los datos en:
https://www.dropbox.com/scl/fo/8v04ytxh4g0pj32ir4c3l/AA3N_RMBtV916MXXoe_4stA/hogares?dl=0&rlkey=l5upmb0deodayh8lzzzrzfihm&subfolder_nav_tracking=1

https://microdatos.dane.gov.co/index.php/catalog/643/data-dictionary

Configuración del Entorno 🛠️

https://github.com/josezea/bigdataR_2024/

Algunas ideas para BRUG

- ✓R en producción (Docker, Kubernetes)
- ✓Integración de R con ambientes productivos (Quarto / Shiny)
- ✓Creación de *API's* (Plumber, *Valvet, faucet entre otros*)
- ✓Creación de páginas web, apps y de dashboards en ambientes productivos
- ✓Deep Learning: torch, tensorflow para R
- ✓Machine Learning en serio: tidymodels, mlr3