Taller: Usando sparklyr

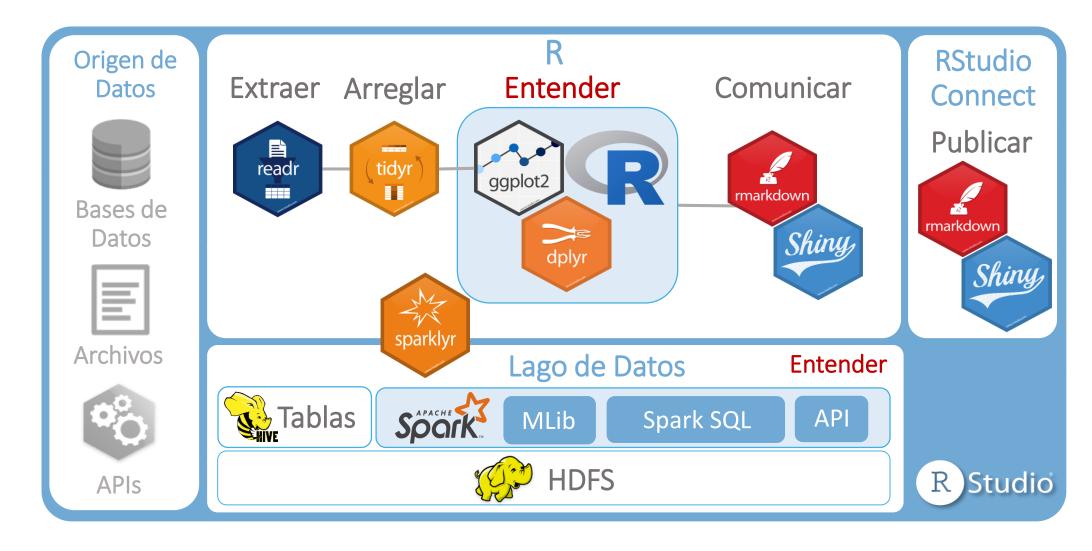
Edgar Ruiz

@theotheredgar
linkedin.com/in/edgararuiz

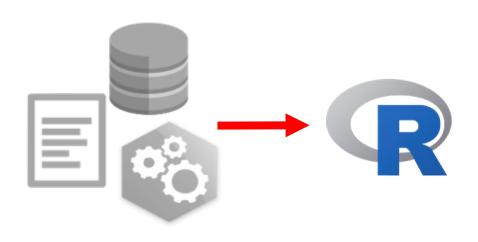
15 de marzo del 2018



Ciencia de Datos con R y Spark



La idea principal



Extraer datos





Como funciona sparklyr





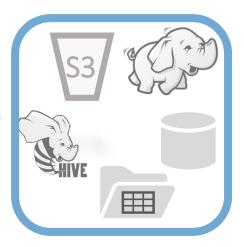
- dplyr
- Aprendizaje automático
- Extensiones





- Informática en clúster
- Aprendizaje automático
- Comunicación usando SQL
- API extensible

Paquete de Spark



Opciones para manejar memoria

Opción 1

Los datos de cada resultado vienen desde el origen.



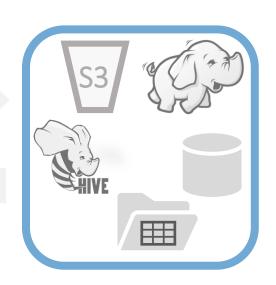
SQL

Resultado



SQL

Resultado



Opción 2

Copiar los datos en Spark. Los resultados están en la memoria.

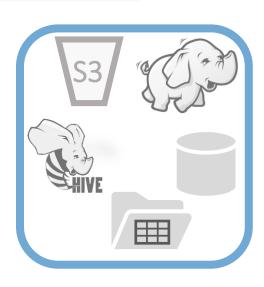


SQL

Resultado



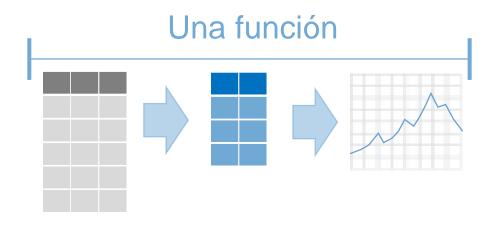
Copiar



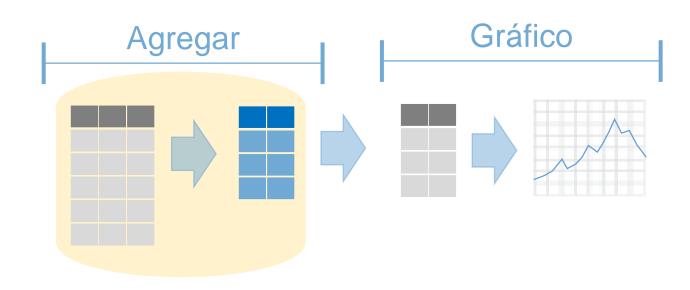


Visualizar datos

Datos locales



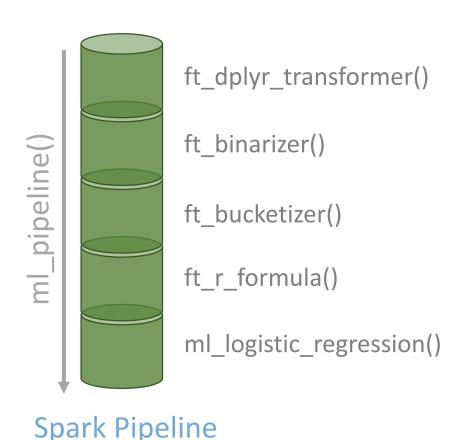
Datos lejanos



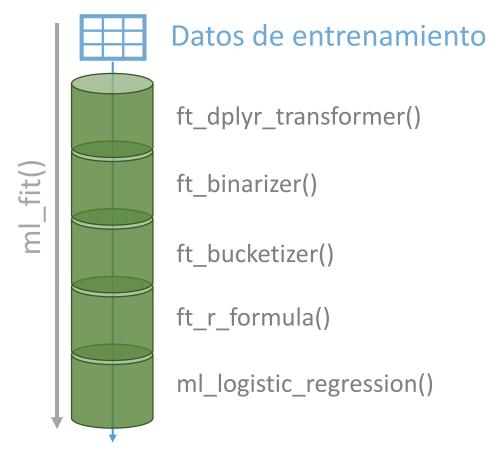


Spark Pipelines (Tubería)

Estimator



Transformer



Spark PipelineModel



Links

Sitio official de sparklyr: http://spark.rstudio.com/

Sitio official de Spark: http://spark.apache.org/