R y Spark para la Ciencia de Datos

Edgar Ruiz

@theotheredgar

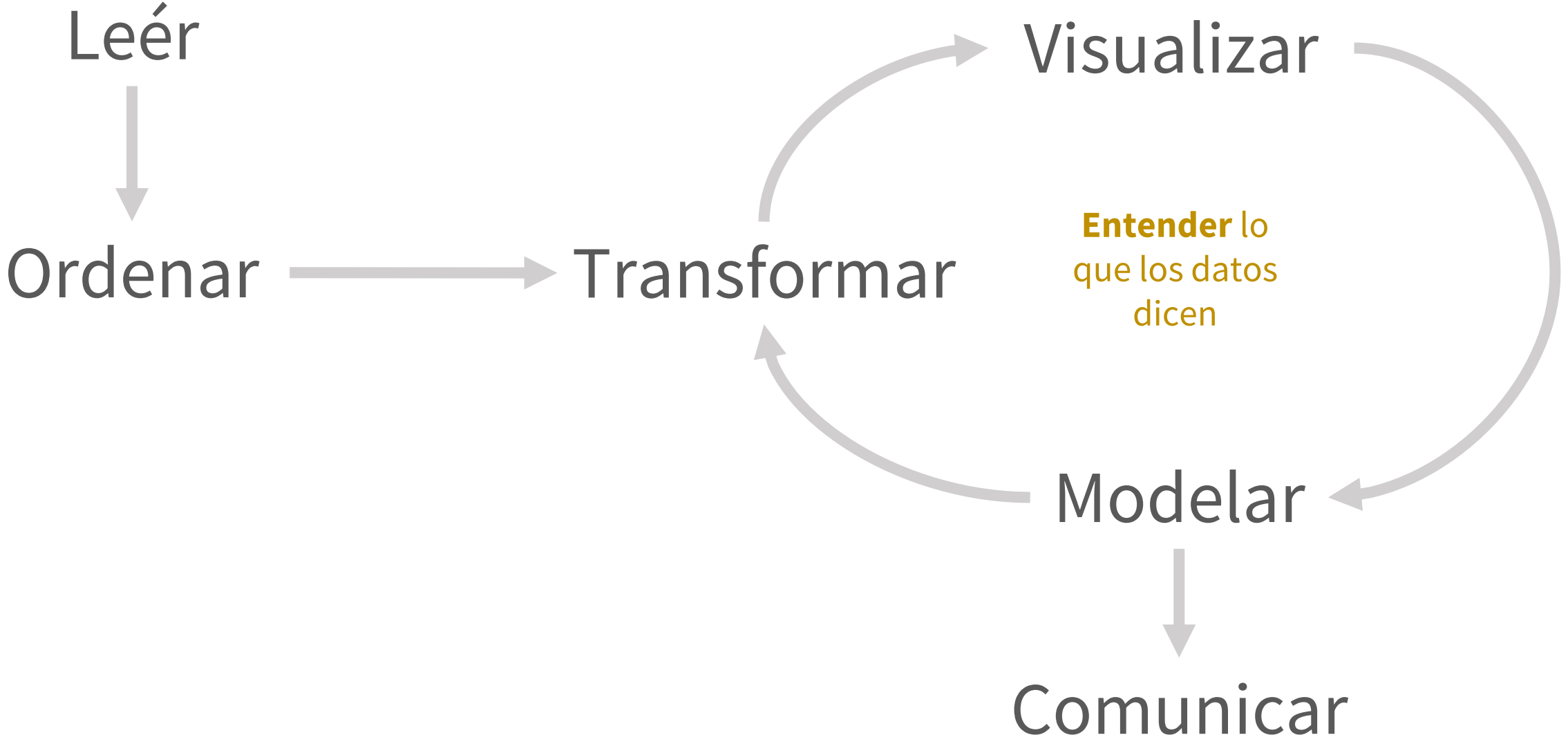
linkedin.com/in/edgararuiz

github.com/edgararuiz

24 de marzo del 2019

RStudio

Ciencia de Datos



Todo se prepara dentro de R

Leér

Ordenar

Transformar

Visualizar

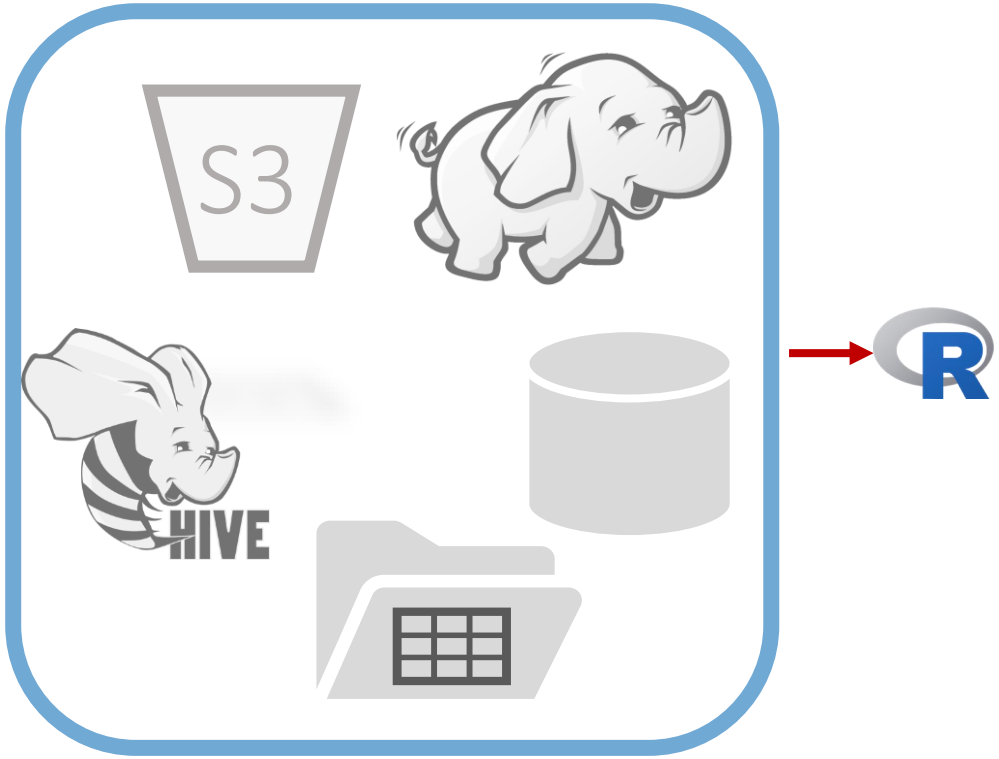
Entender lo que los datos dicen

Modelar

Comunicar

RStudio

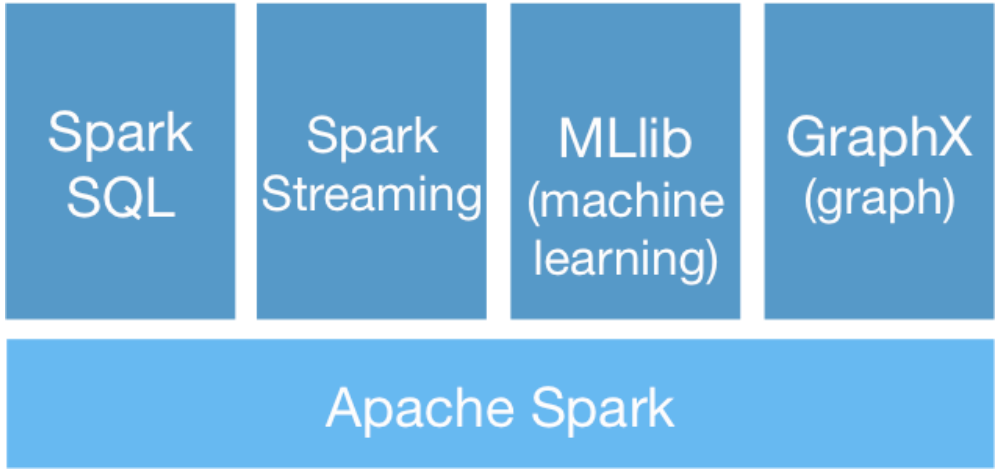
Datos mas grandes que el RAM



Que es Apache Spark?

Motor analítico para procesar datos a gran escala

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



Tipos de análisis disponibles en Spark

- Modelos de regresión
- Modelos de clasificación
- Modelos de agrupación (clustering)
- Modelos de gráficas (GraphX)
- Análisis sobre datos stream (constante flujo)
- Análisis de texto, incluyendo modelos

Que es sparklyr ?

Es una interfaz para R y Spark

- Provee suporte a dplyr dentro de Spark
- Acceso a todo el API de Spark
- …incluyendo Pipelines

Como funciona sparklyr

Función de R

- `dplyr`
- Aprendizaje automático
- Extensiones

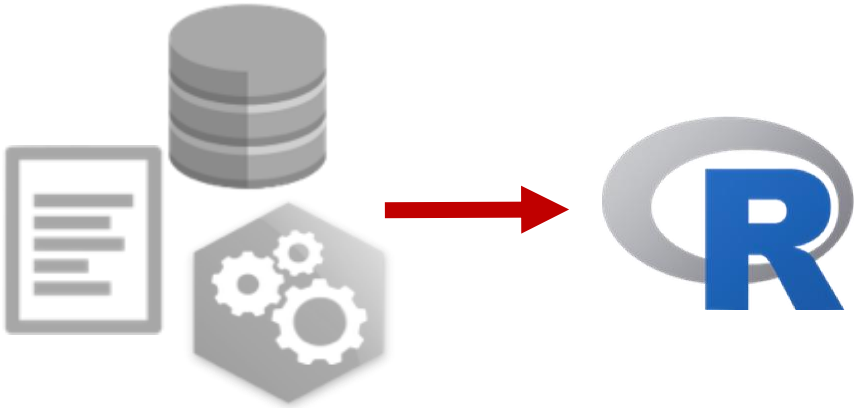
Spark Shell

APACHE Spark™

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

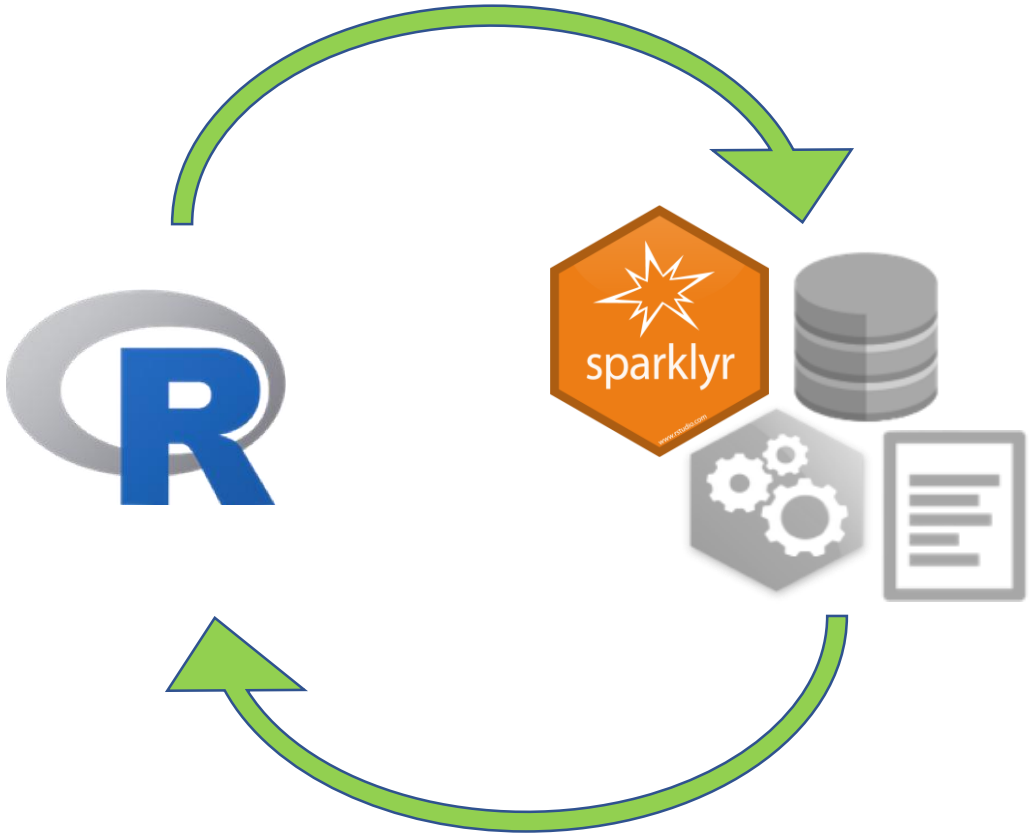
Paquete de Spark

La idea principal



Extraer datos

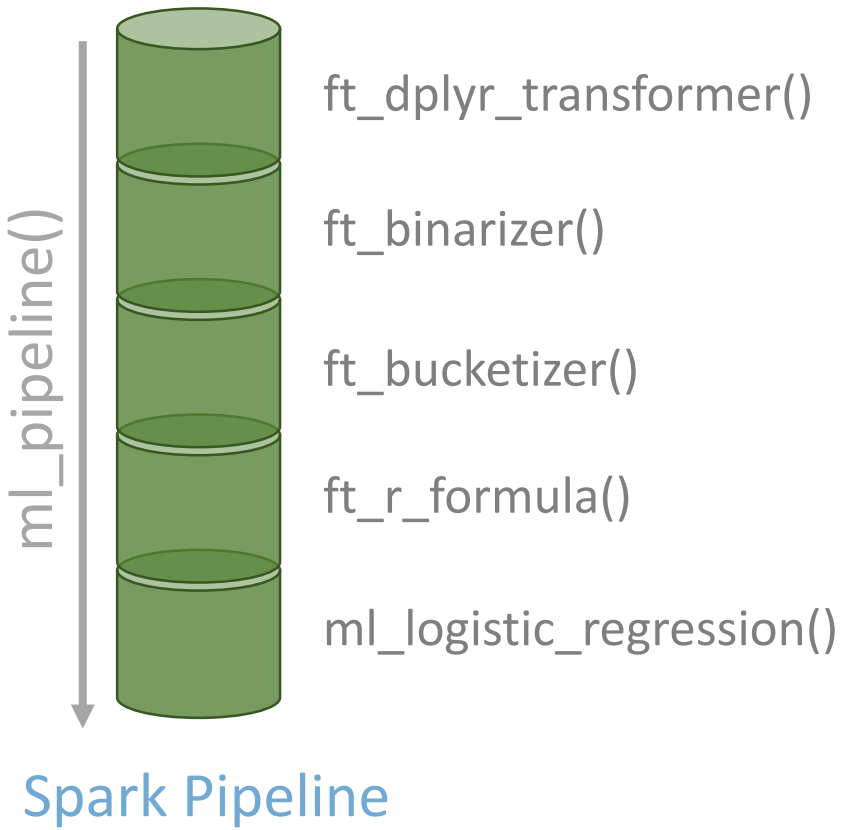
Cálculos



Resultados

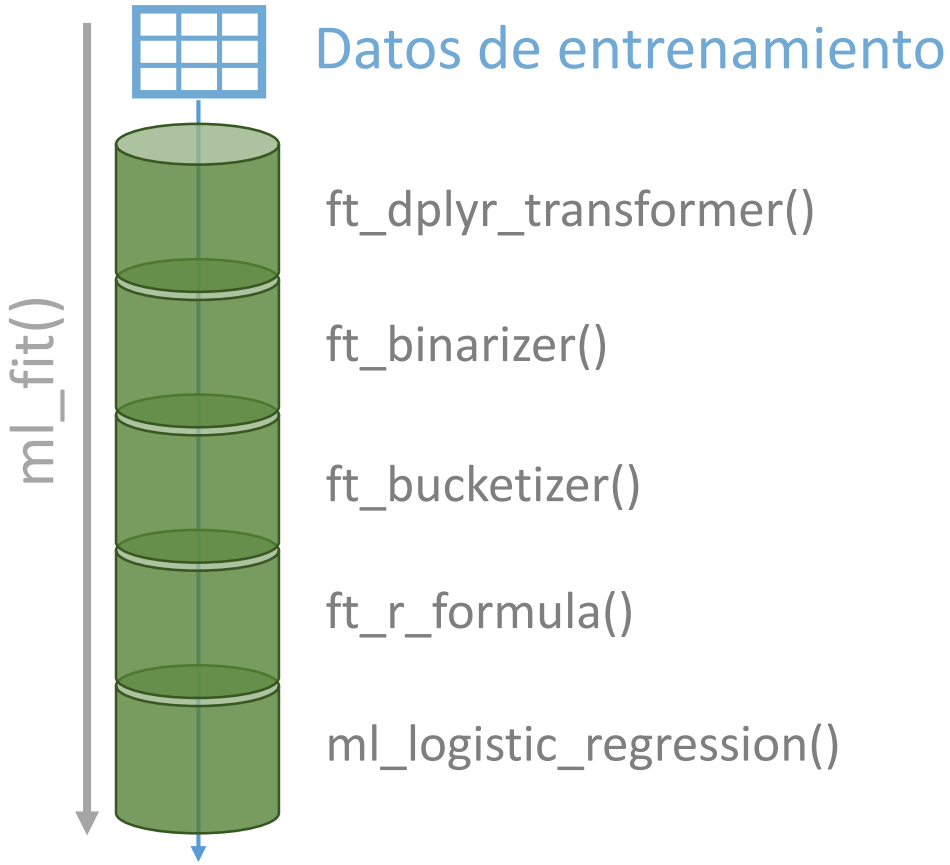
Spark Pipelines (Tubería)

Estimator



Spark Pipeline

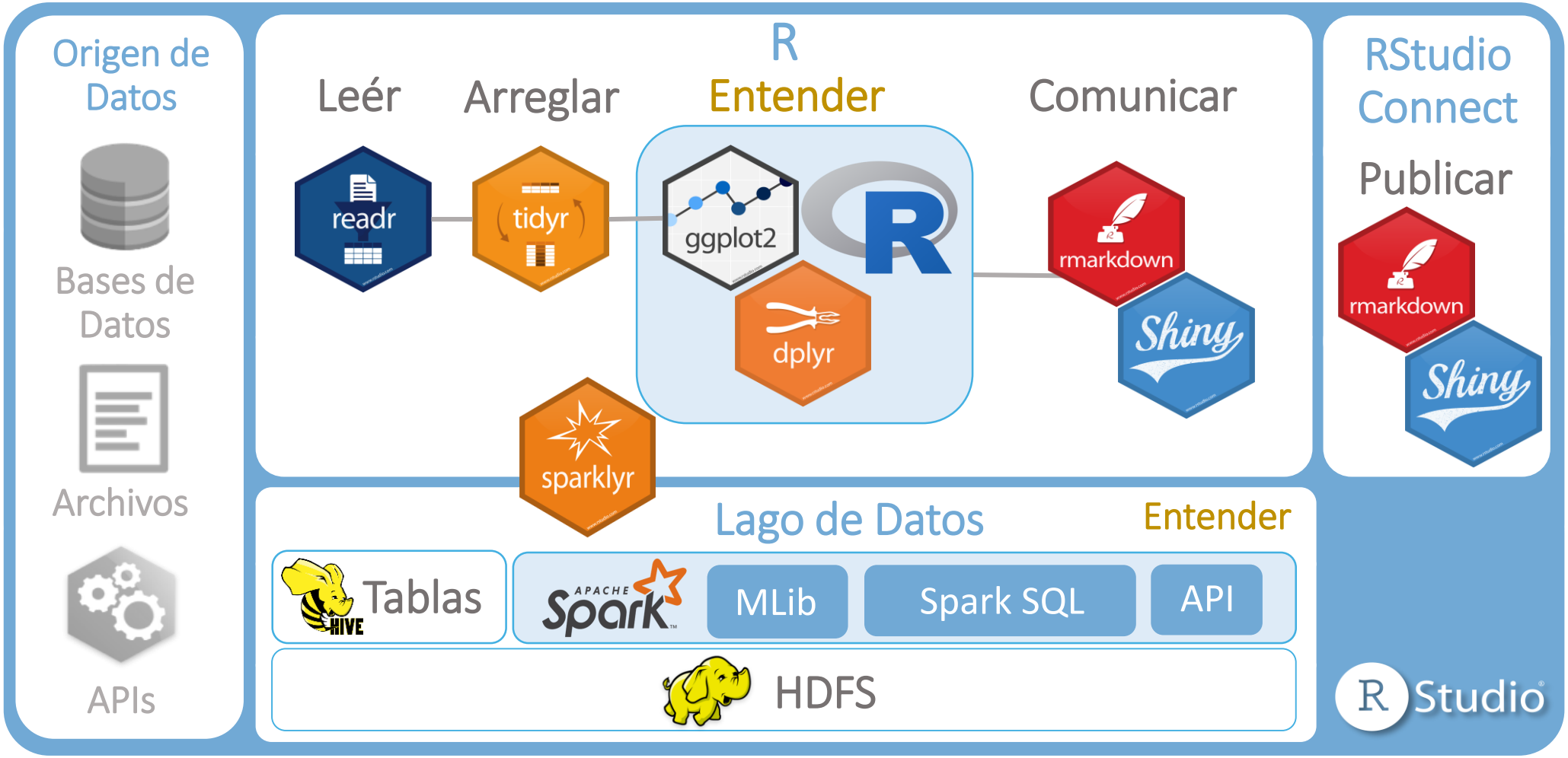
Transformer



Spark PipelineModel

RStudio

Ciencia de Datos con R y Spark



Recursos

Sitio official de sparklyr:

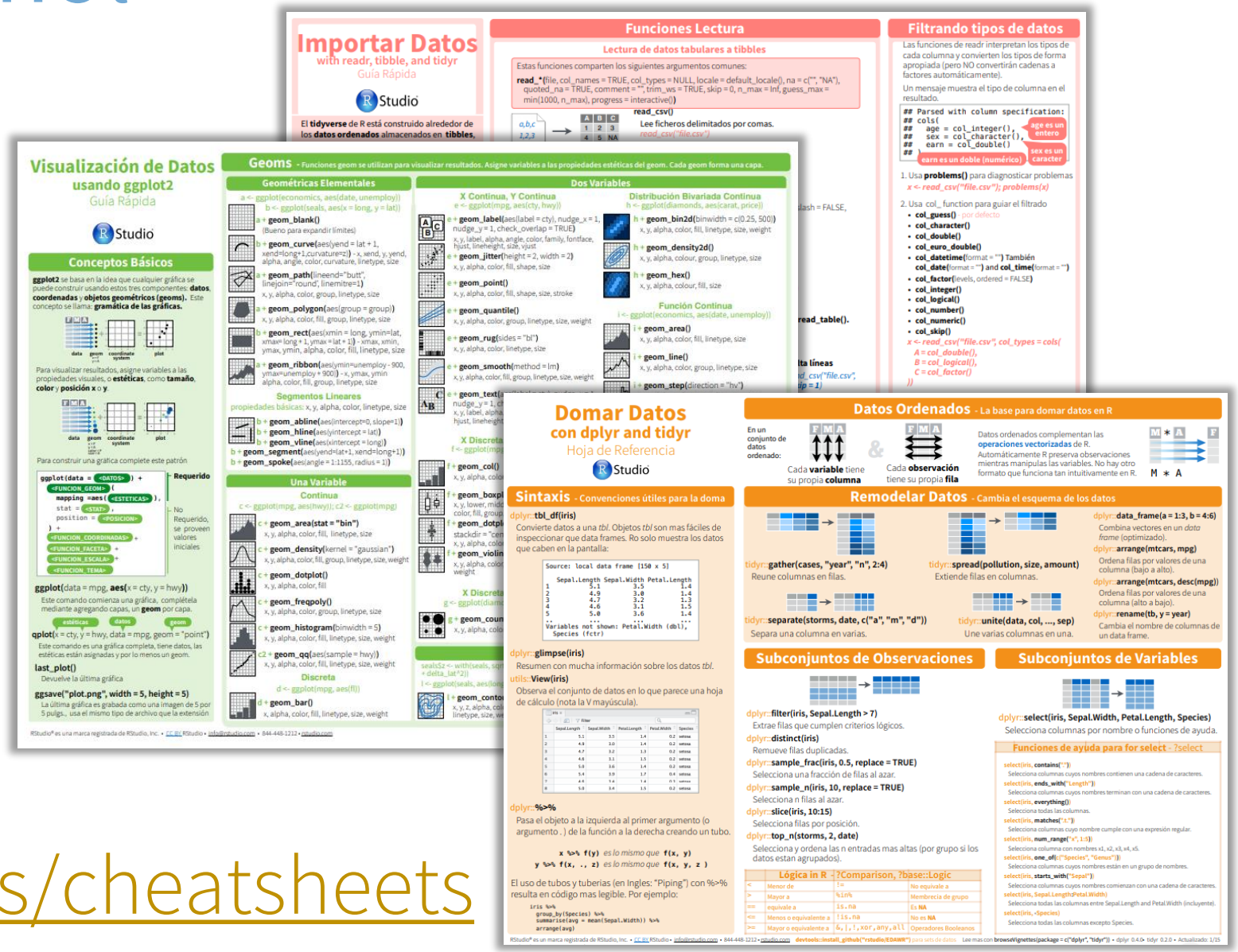
spark.rstudio.com

Sitio official de Spark:

spark.apache.org

Recursos en español

> Para aprender como usar los paquetes en práctica, las Hojas de Referencia, o *Cheatsheets*, son los mejores recursos, no importa el idioma



Cortesía de Frans van Dunné, Carlos Ortega y Santiago Mota aquí:

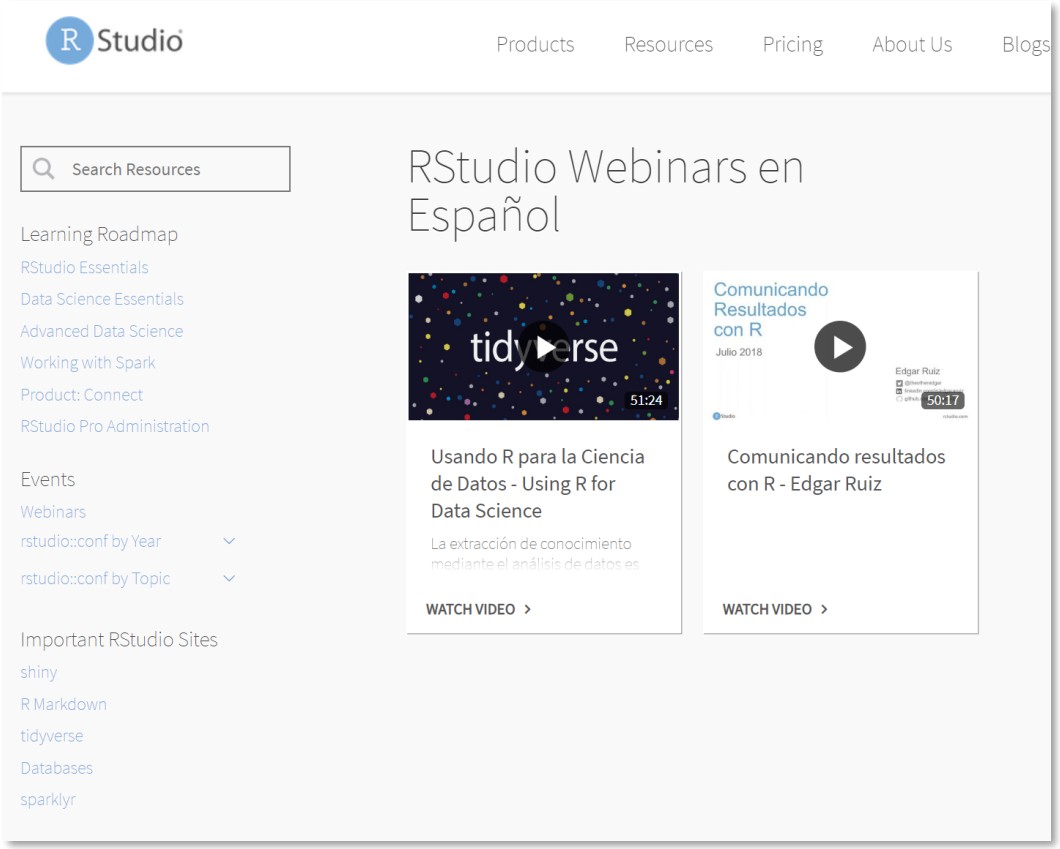
rstudio.com/resources/cheatsheets

RStudio

Recursos en español - Webinars

Una base para aprender a utilizar R de manera efectiva en nuestros análisis. Estos webinars proveen tres cosas:

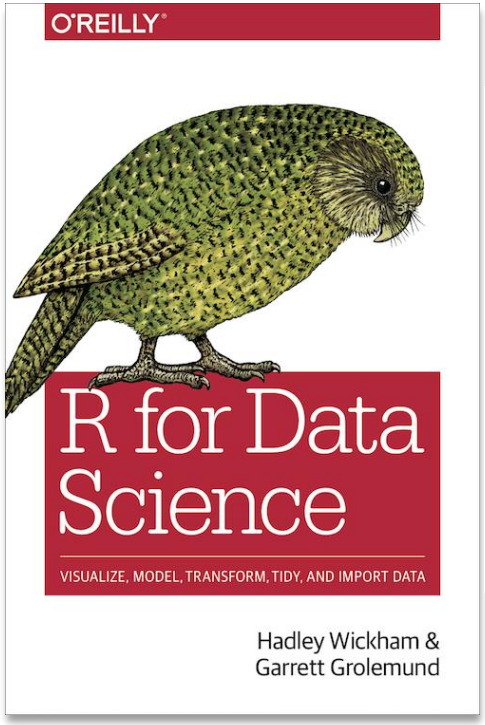
1. Ejemplos de código
2. Presentaciones
3. Video de la sesión



resources.rstudio.com/espanol

RStudio

Muy pronto!



r4ds
en español

github.com/cienciadedatos

RStudio

Materiales

rstd.io/conectar