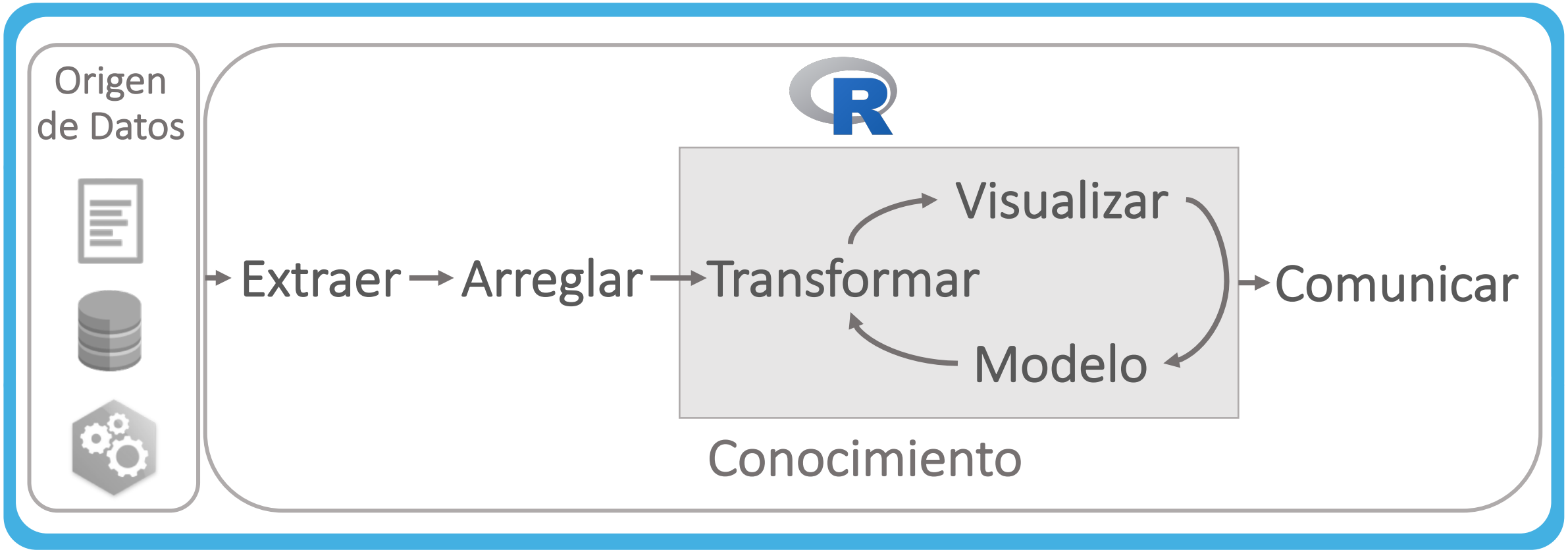
Ciencia de datos en Spark usando R y sparklyr

Edgar Ruiz – Marzo 2017

Ciencia de datos con R

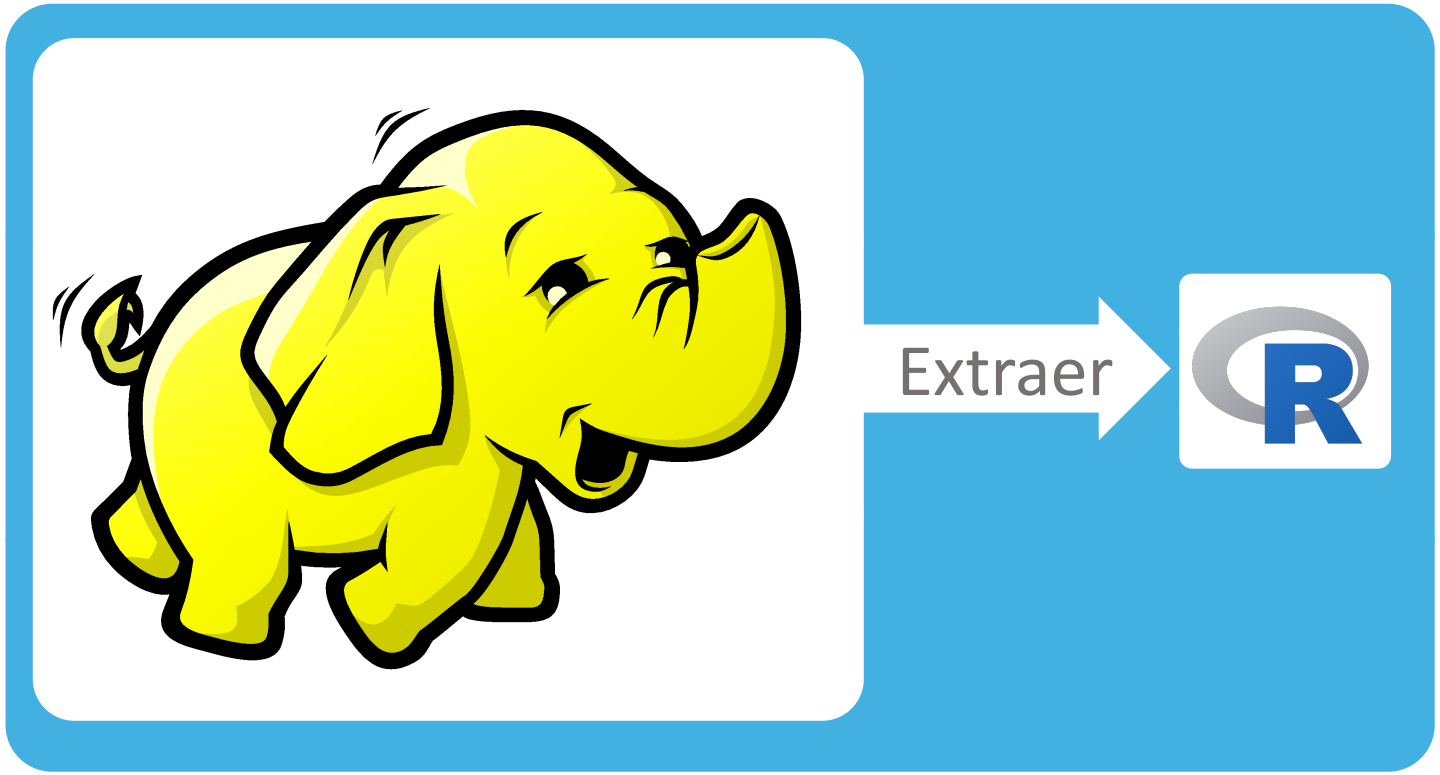


R for Data Science, Wickham & Grolemund

Practicando Ciencia de Datos en un Lago de Datos

Hadoop como el origen de datos

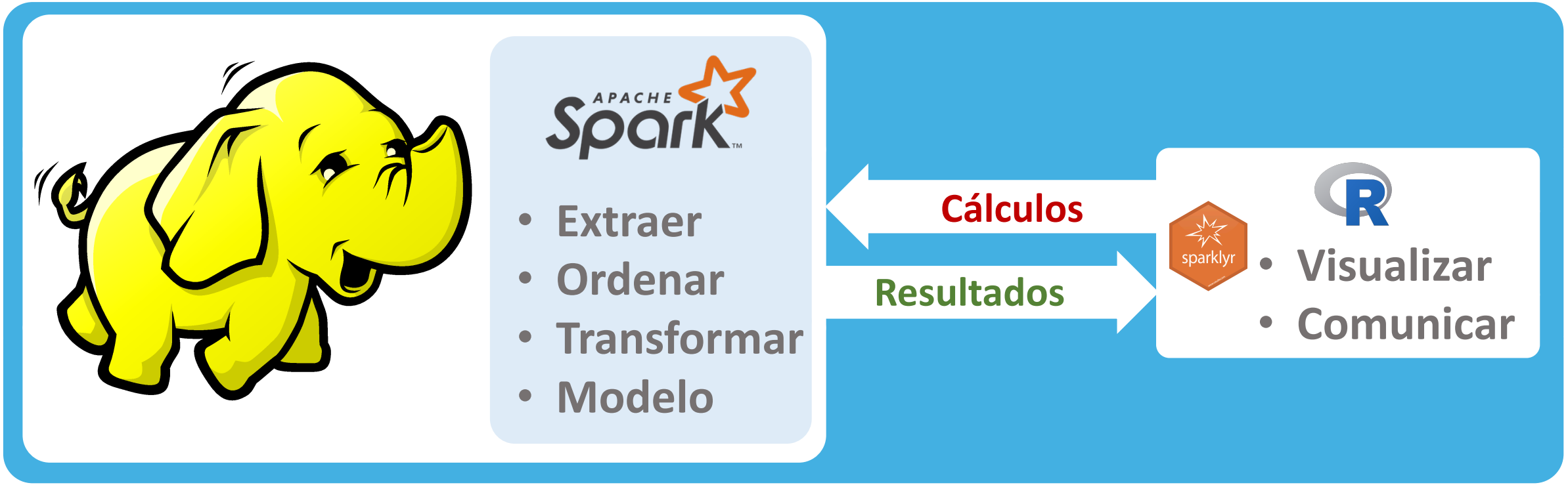
Problema: Muchos datos para analizar en RAM con R



Alternativa: Tomar una muestra pequeñísima o usar cuantos datos posibles dentro de R

Usando Spark para los cálculos

Solución: Analice los datos con Spark y después recaude los resultados en R



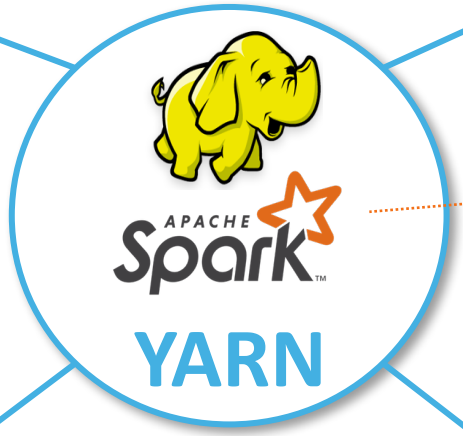
Diseño del grupo de servidores

Use un ”web browser” para entrar a RStudio

Name node

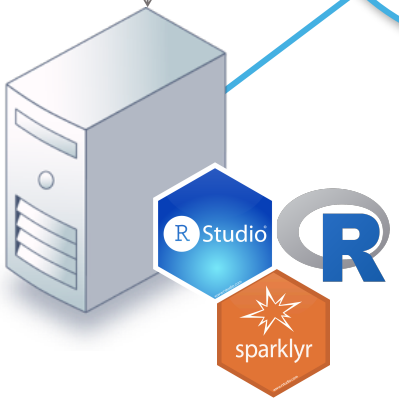
Data node

Todos los servidores necesitan **YARN** y **Spark Gateway**



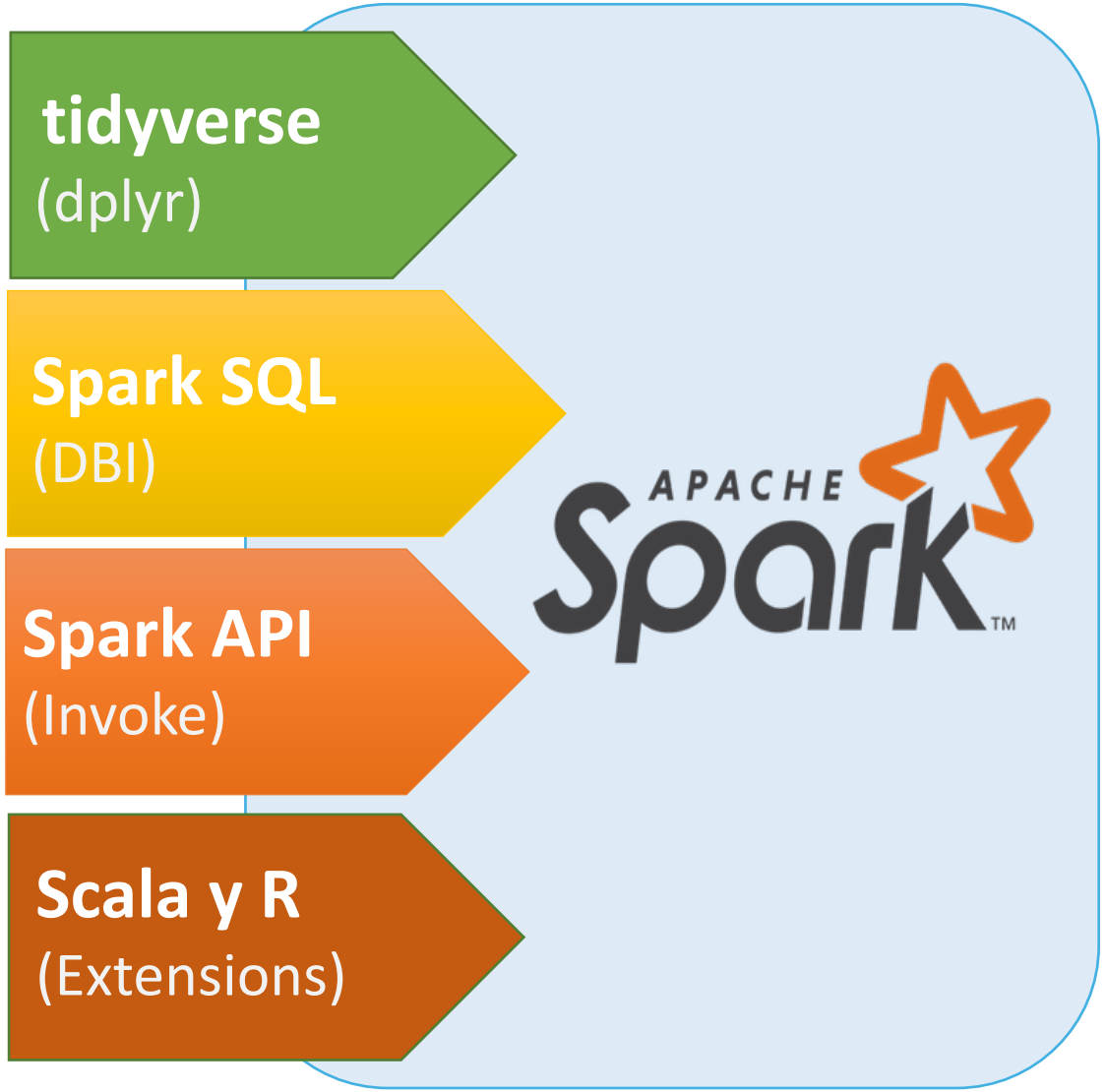
R, RStudio & sparklyr en 1 servidor

Edge node



Data node

Los cuatro niveles de acceso



Por que usar sparklyr?

1. R, RStudio & sparklyr solo se necesitan en un servidor

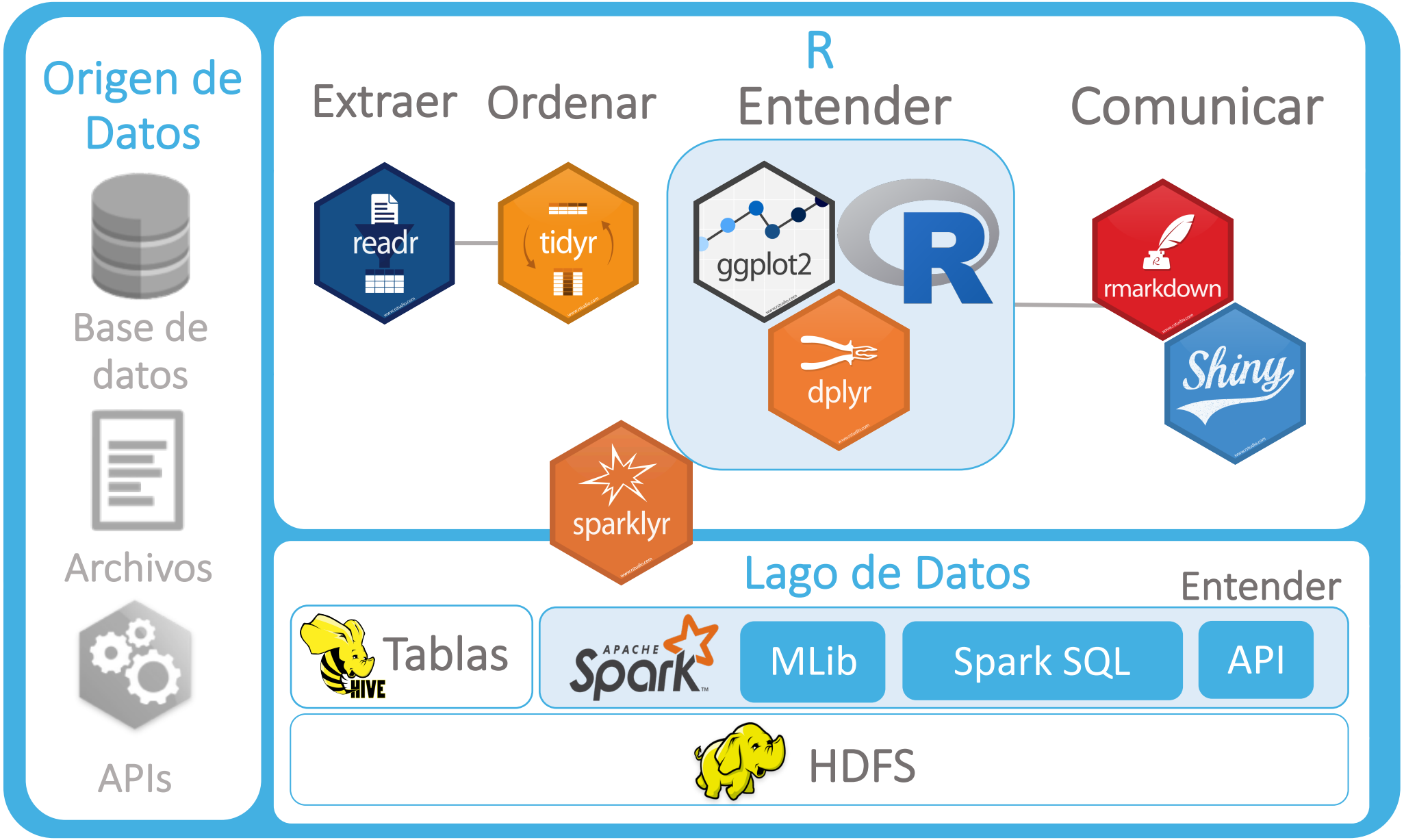
2. Librerías de aprendizaje automático de Spark

3. Acceso al API de Spark

4. Acceso a las tablas y funciones de Hive

5. Use comandos familiares de dplyr para interactuar con los datos en Spark

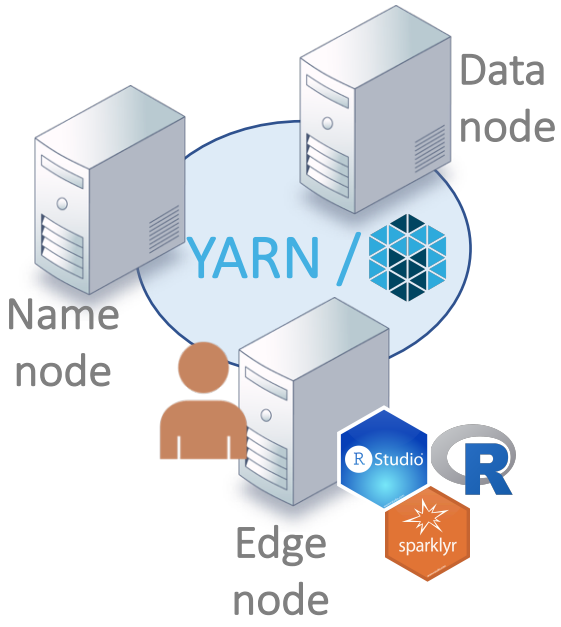
Using Spark & R for Data Science



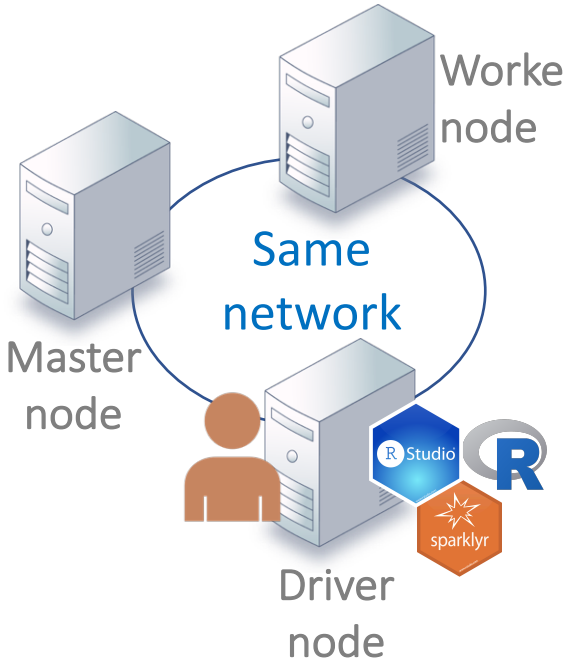
Demonstración

Opciones de instalación

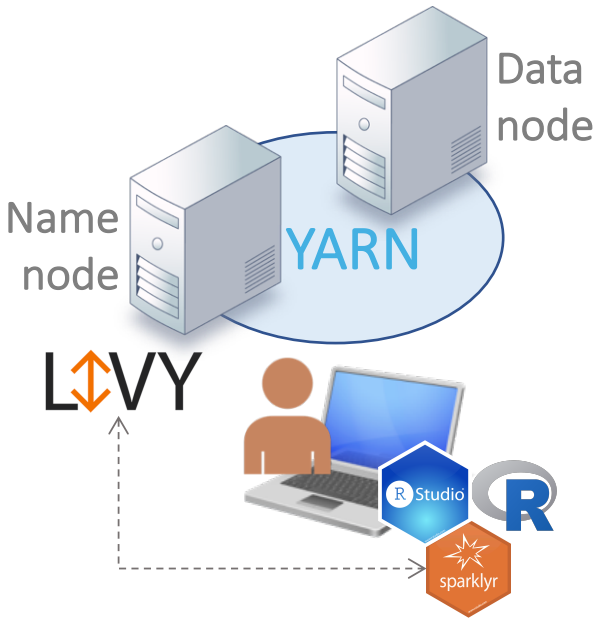
Hadoop



Solo Spark



Livy



Local

Demonstración

R for Data Science Toolchain with Spark

Origen de Datos

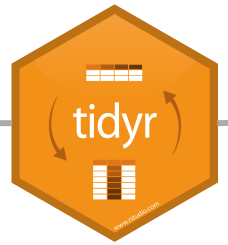
Base de Datos

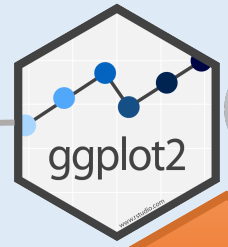
Archivos

APIs

R

Extraer Arreglar Entender Comunicar





RStudio Connect

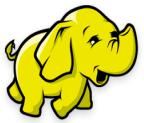
Publicar

Lago de Datos

Entender

Tablas

MLib Spark SQL API



HDFS

RStudio

Recursos

Official Website	spark.rstudio.com
Data Science	http://r4ds.had.co.nz/
GitHub Repository	github.com/rstudio/sparklyr
Cheatsheet	spark.rstudio.com/images/sparklyr-cheatsheet.pdf