

CIENCIA & DATOS

CIENCIA DE DATOS PARA TODOS

INTRODUCCIÓN AL MUNDO DE SPARK CON R

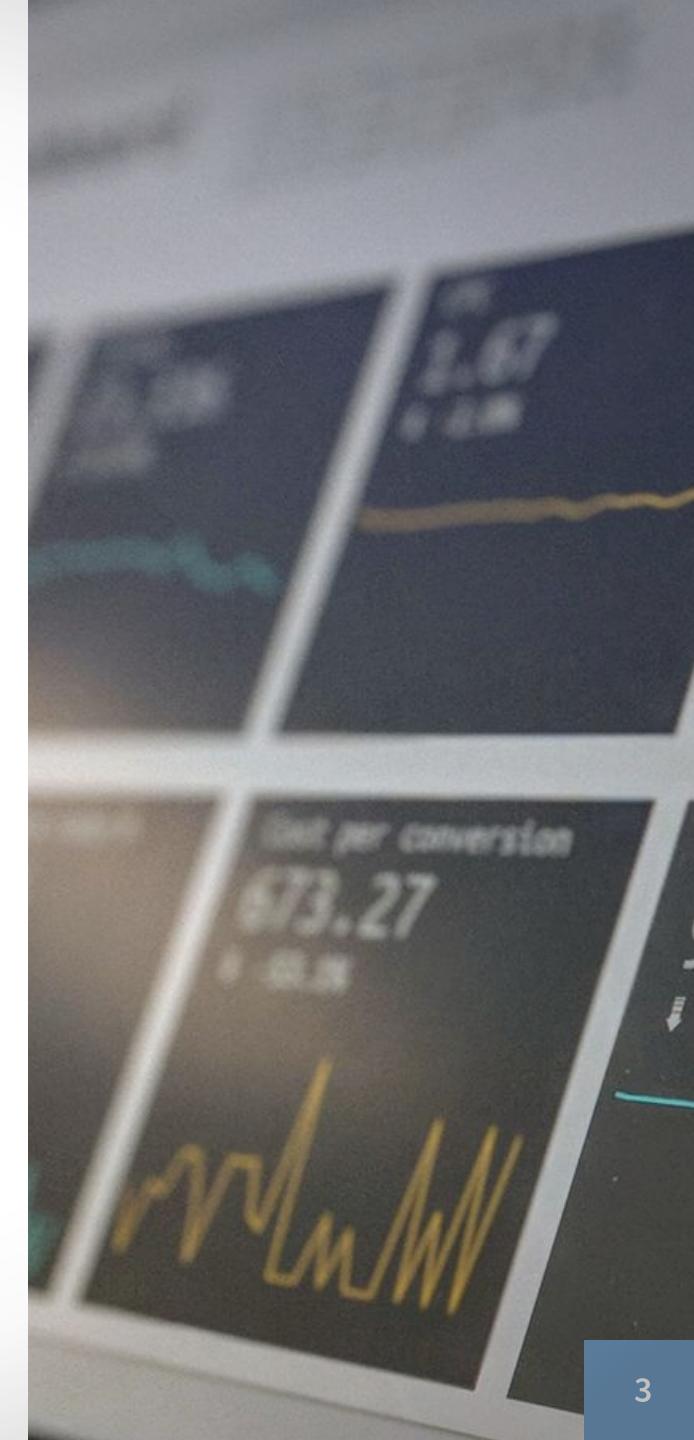
FAVIO VÁZQUEZ
Data Scientist

LO QUE APRENDRÁS

- 1 ¿QUÉ ES R?**
- 2 LO BÁSICO DE R PARA CIENCIA DE DATOS**
- 3 INTRODUCCIÓN A SPARK**
- 4 SPARKLYR**
- 5 PRÓXIMOS PASOS**

R LANGUAGE

- R ES UN **LENGUAJE Y ENTORNO PARA COMPUTACIÓN Y GRÁFICOS ESTADÍSTICOS**. ES UN PROYECTO DE GNU QUE ES SIMILAR AL LENGUAJE Y ENTORNO **S** QUE SE DESARROLLÓ EN BELL LABORATORIES.



```
message", m => {
  = m.split(" ")
  (a[0]){
    "connect":
  (a[1]){
    if(clients.has(a[1])){
      ws.send("connected")
      ws.id = a[1];
    }else{
      ws.id = a[1]
      clients.set(a[1], {
        ws.send("connected")
      })
    }else{
      let id = Math.random()
      ws.id = id;
      clients.set(id, {cli
```

PROGRAMACIÓN PARA DS

- **ES UNO DE LOS PILARES DE LA CIENCIA DE DATOS**
- **ES NECESARIA PARA CONSTRUIR SOLUCIONES DIGITALES**
- **NO NECESITAS SER UN EXPERTO PERO SE REQUIERE DE UN BUEN NIVEL**
- **ESTÁ ENFOCADA EN LA MANIPULACIÓN DE DATOS**
- **DEBES MANEJAR HERRAMIENTAS DE VISUALIZACIÓN**
- **LA PROGRAMACIÓN PARA BIG DATA ES IMPORTANTE, Y FÁCIL CON SPARK + R**
- **SE APRENDE HACIENDO**

LA LIBRERÍAS QUE DEBES SABER





¿QUÉ ES APACHE SPARK?

- **ES UN MOTOR GENERAL Y MUY RÁPIDO PARA EL PROCESAMIENTO EN PARALELO DE DATOS EN GRAN ESCALA.**
- **TIENE UNA API PARA SCALA, JAVA, PYTHON AND R**

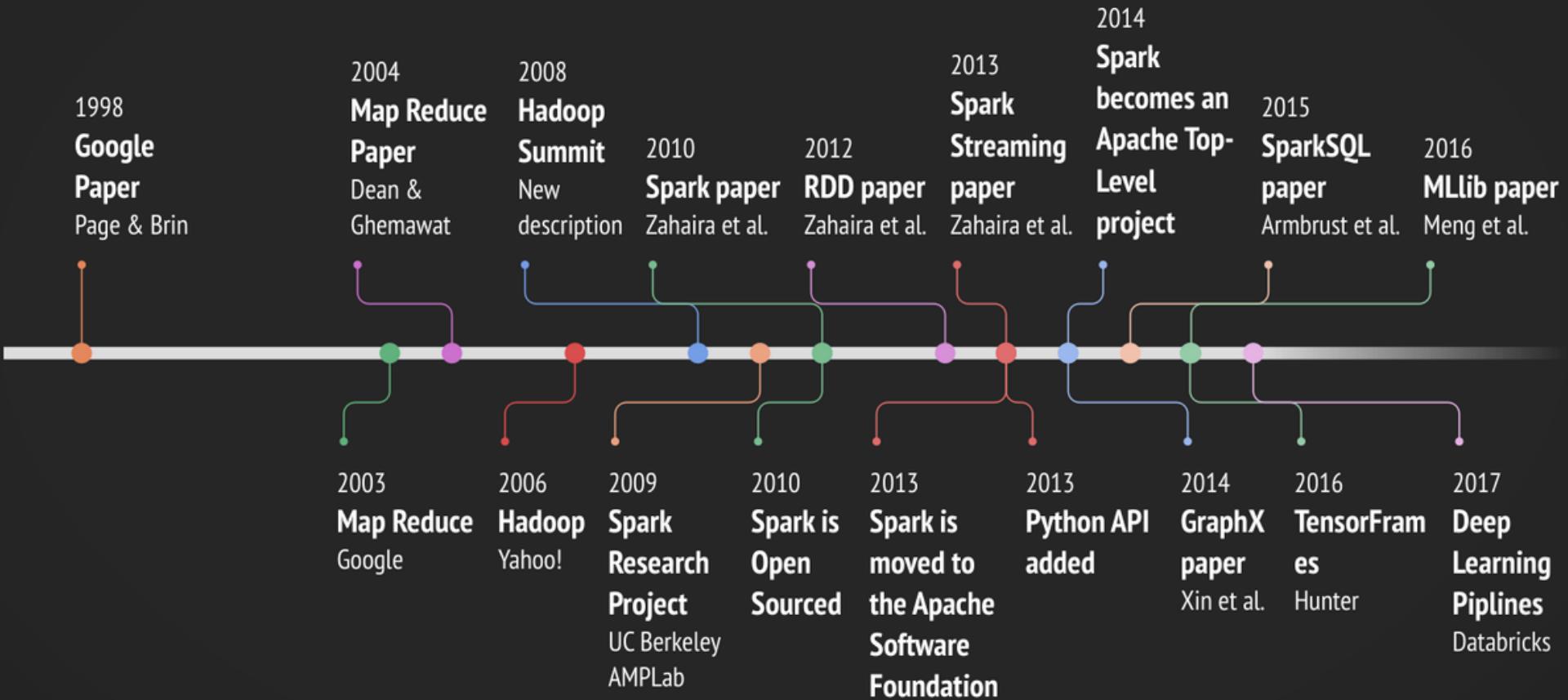


APACHE SPARK



- SE CONECTA FÁCILMENTE A CASI TODAS LAS BASES DE DATOS, SISTEMAS DISTRIBUIDOS Y MÁS
- NORMALMENTE SE EJECUTA EN ENTORNOS CLOUDERA, USANDO YARN Y DESARROLLÁNDOSE EN NOTEBOOKS JUPYTER.

Apache Spark Timeline



Spark
SQL

Spark
Streaming

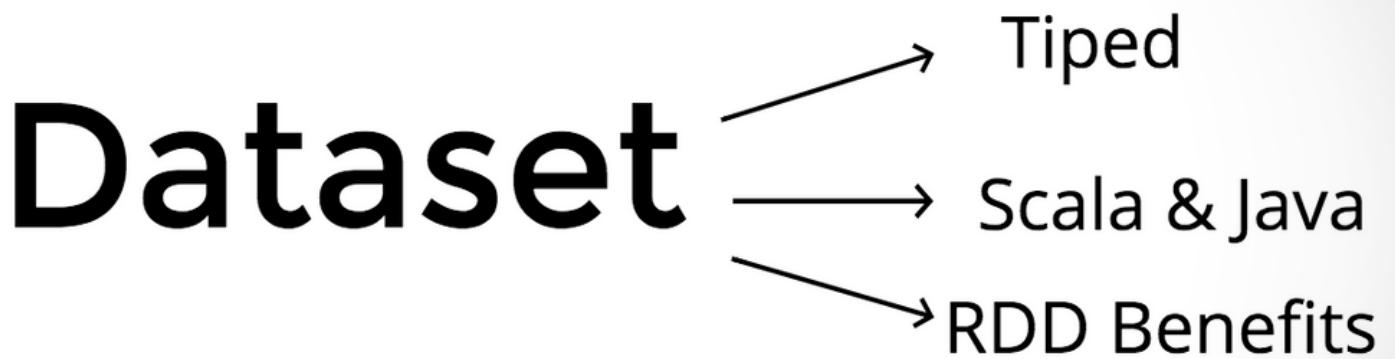
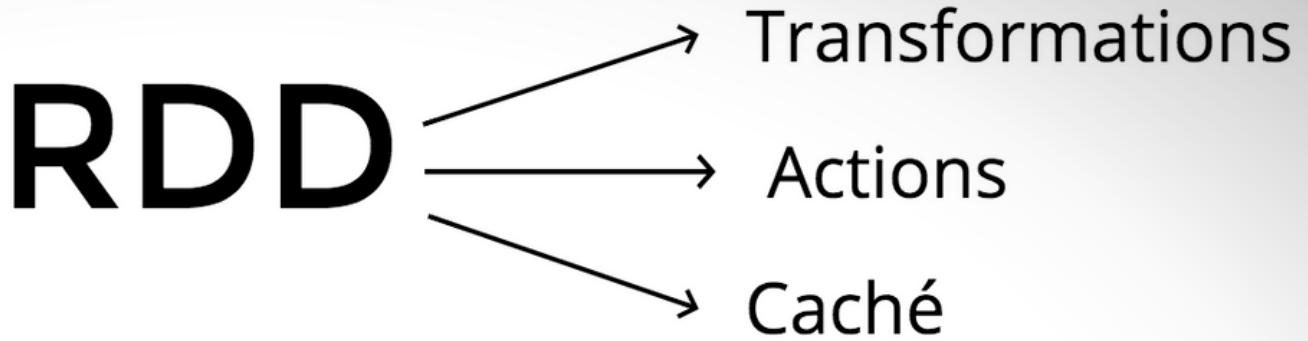
MLlib
(machine
learning)

GraphX
(graph)

Apache Spark

**ESTRUCTURA
DE SPARK**

ESTRUCTURA DE DATOS EN SPARK



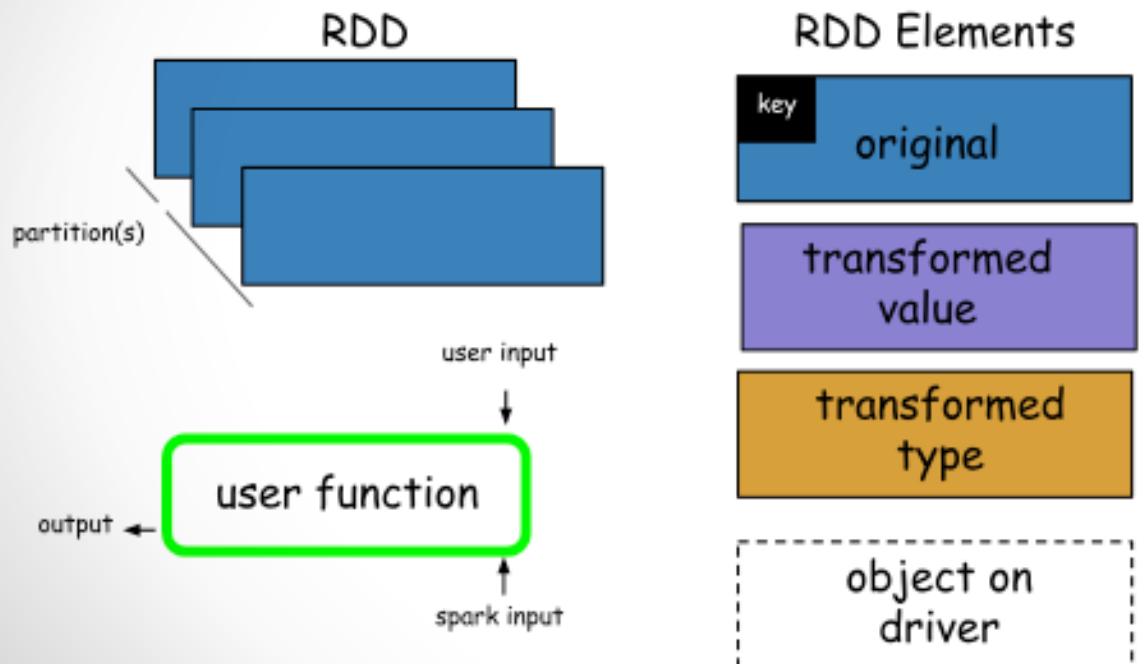
LA HISTORIA DEL DF DE SPARK

RDD

DATAFRAMES

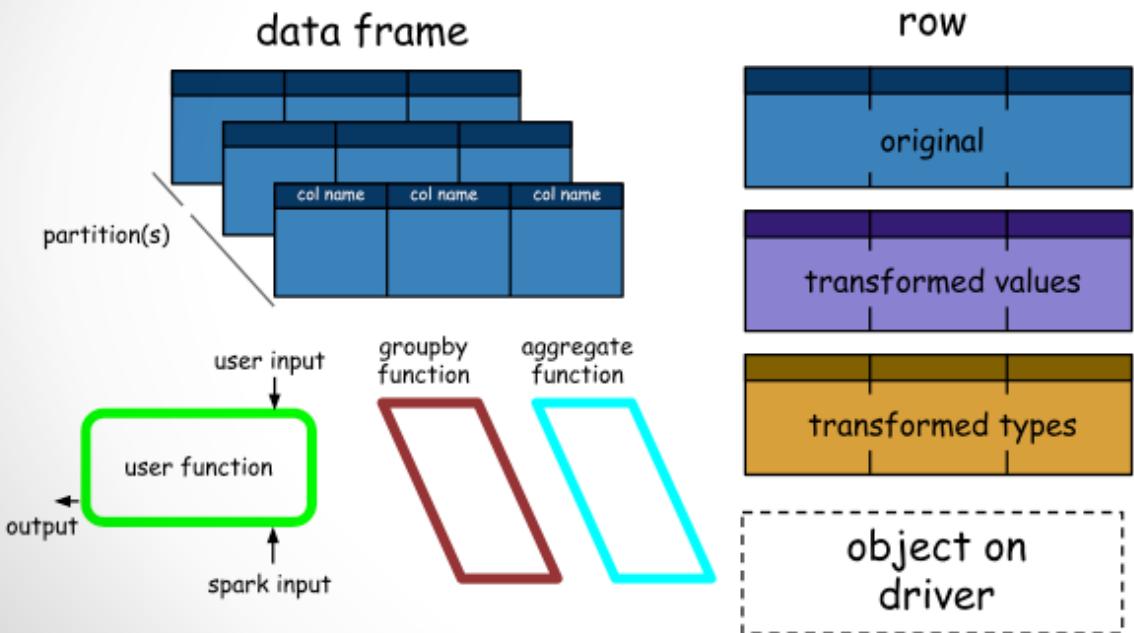
DATASETS

RESILIENT DISTRIBUTED DATASETS



- **UN RDD ES LA ABSTRACCIÓN DE DATOS BÁSICOS DE SPARK.**
- **SON UNA COLECCIÓN DISTRIBUIDA DE OBJETOS JVM INMUTABLES QUE PERMITEN REALIZAR CÁLCULOS MUY RÁPIDAMENTE, Y SON LA COLUMNA VERTEBRAL DE APACHE SPARK.**

DATAFRAMES



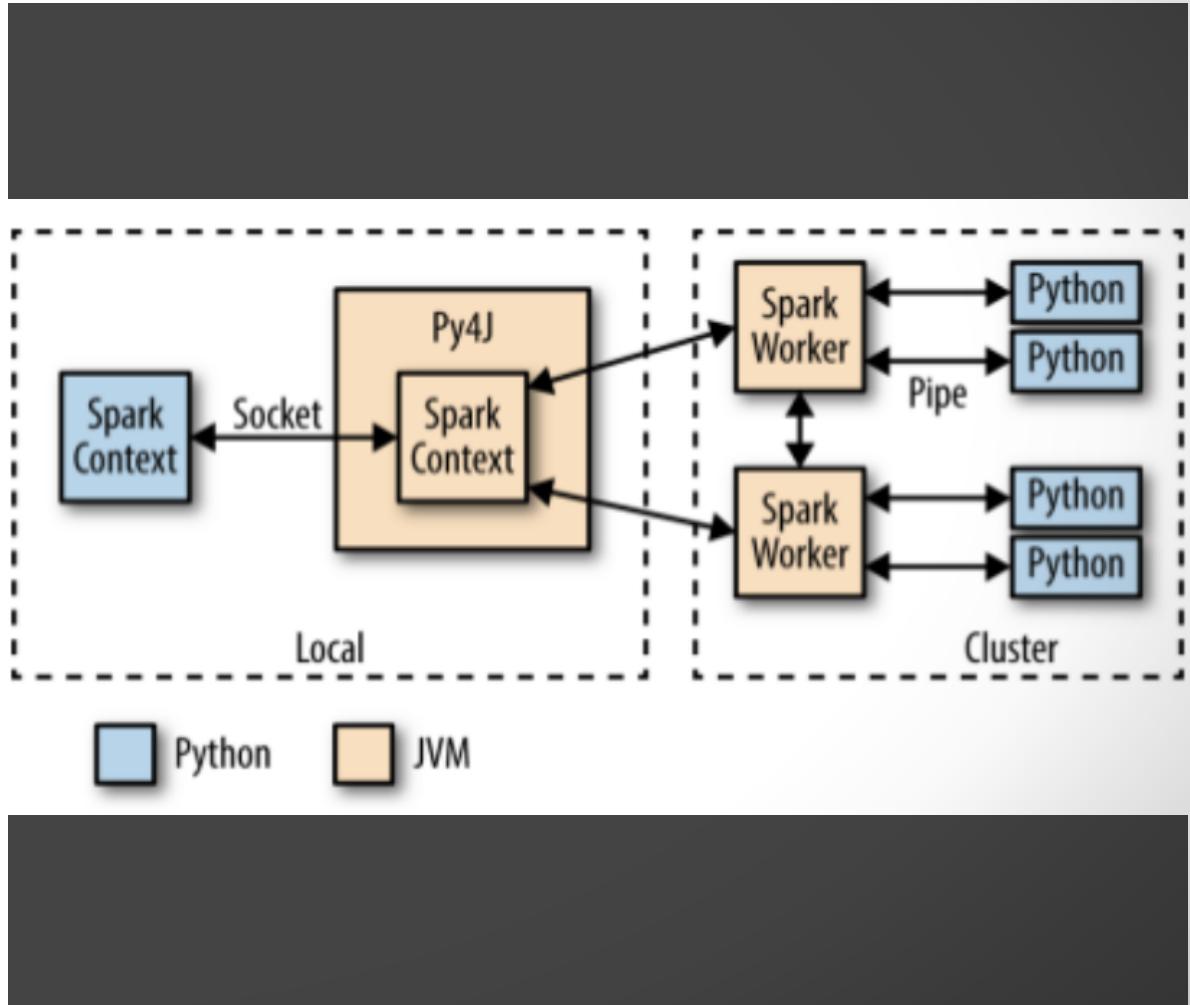
- **LO QUE VEREMOS Y UTILIZAREMOS TODOS LOS DÍAS. MUY SIMILAR A LOS DATAFRAMES DE PANDAS O R**
- **SE PUEDEN OPTIMIZAR, LO QUE SIGNIFICA QUE DESPUÉS DE HACER CLIC EN EJECUTAR, SPARK HARÁ LO MEJOR PARA EJECUTAR EL CÓDIGO LO MÁS RÁPIDO QUE PUEDA**

FUNDAMENTOS DE SPARK

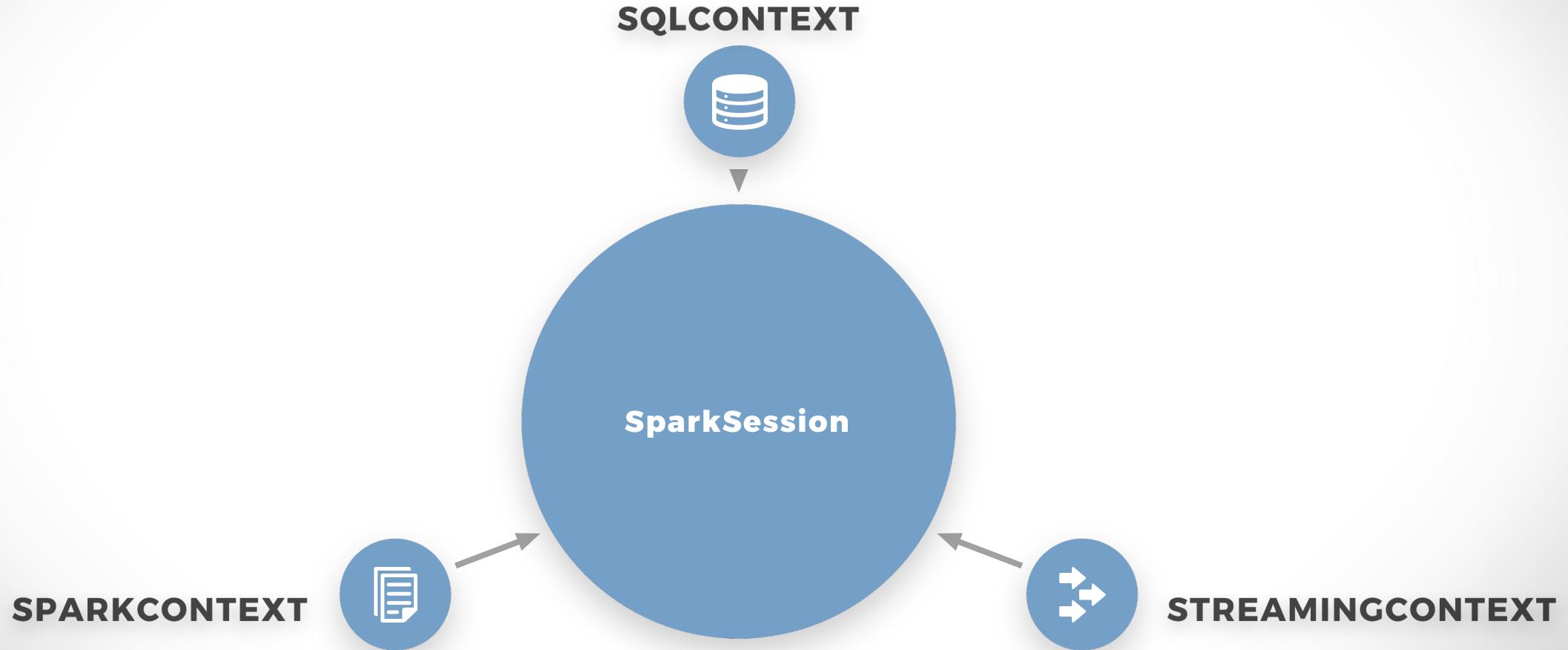


EL SPARKCONTEXT

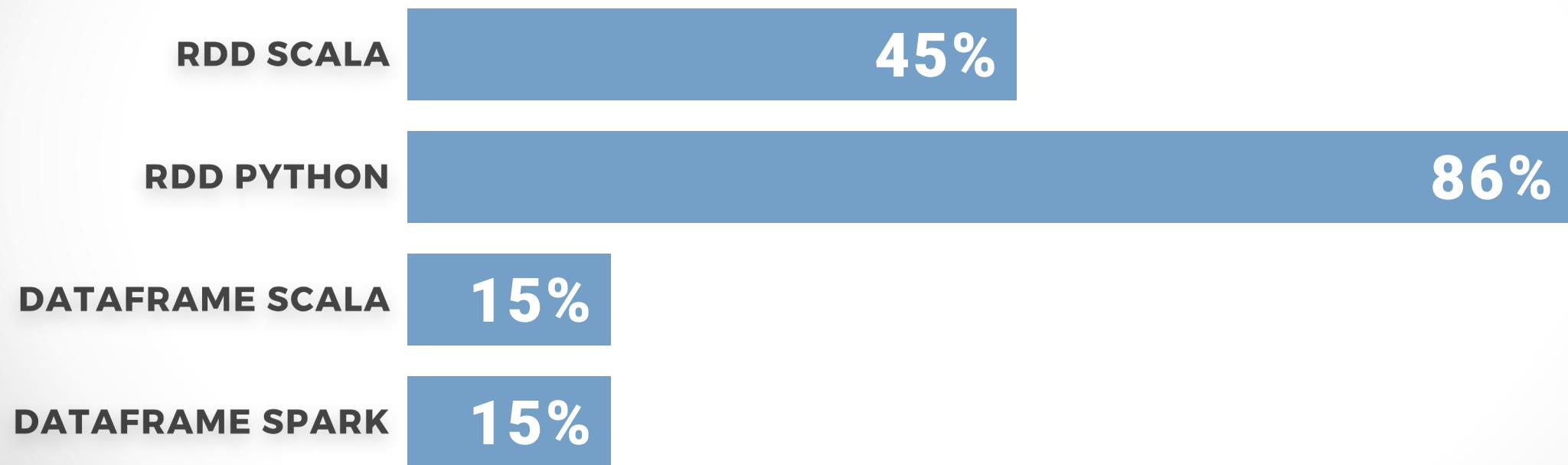
- UN PROGRAMA SPARK PRIMERO CREA UN OBJETO SPARKCONTEXT QUE LE DICE A SPARK CÓMO Y DÓNDE ACCEDER A UN CLÚSTER.
- EL SHELL SPARK CREA AUTOMÁTICAMENTE LA VARIABLE SC.
- EN RSTUDIO Y OTROS PROGRAMAS SE DEBE USAR UN CONSTRUCTOR PARA CREAR UN NUEVO SPARKCONTEXT



DESDE SPARK 2.0

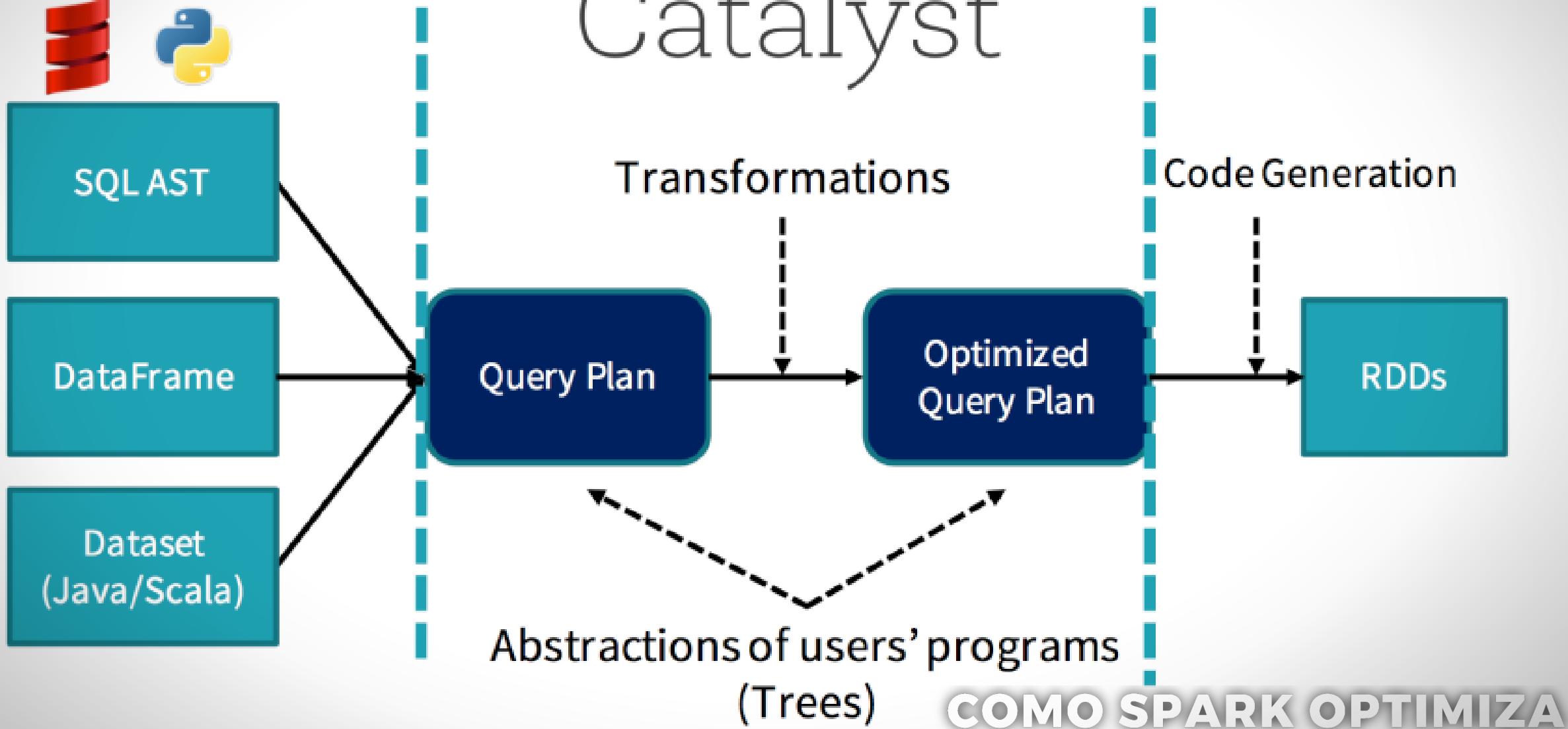


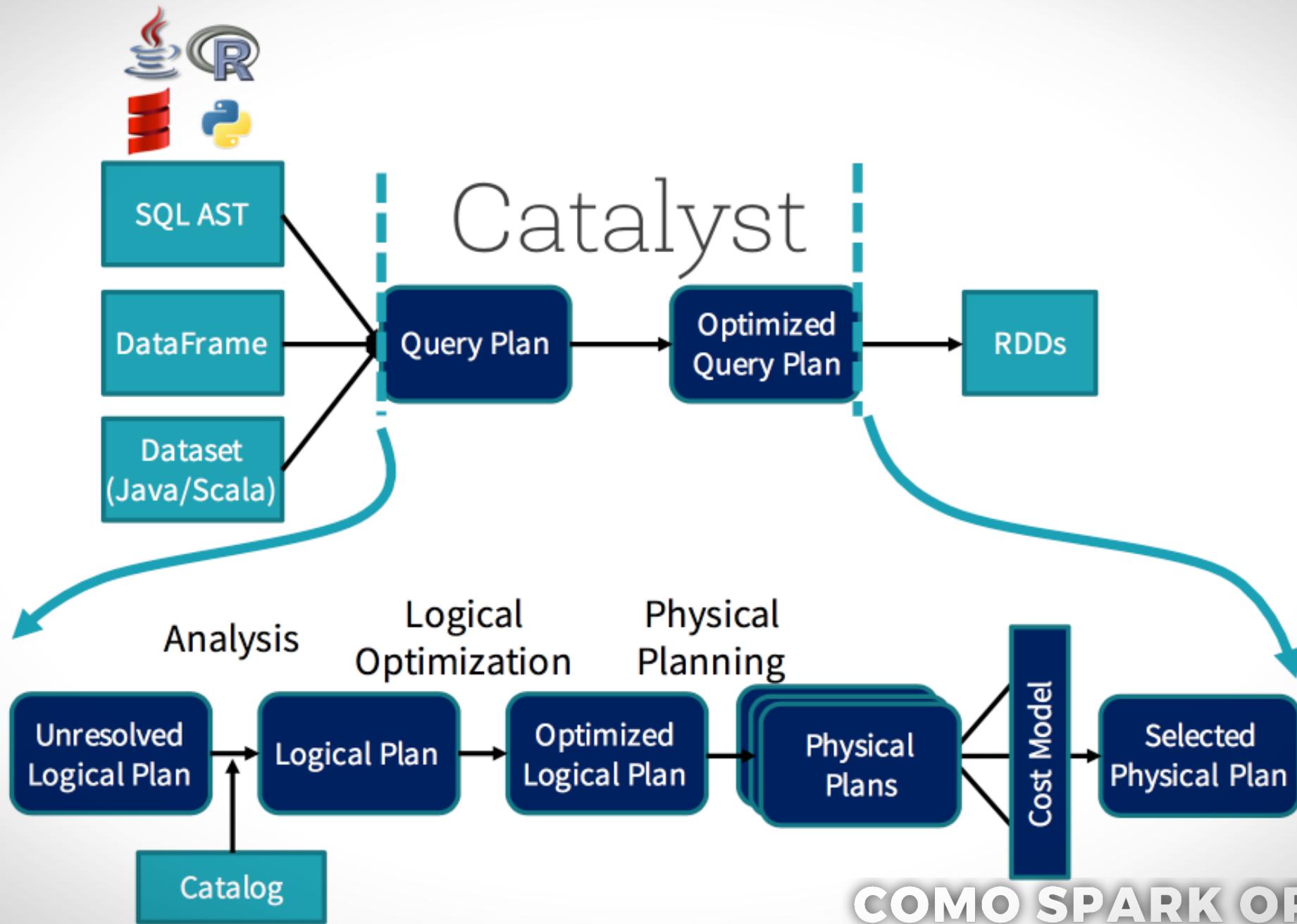
OPTIMIZACIÓN





Catalyst







sparklyr

www.rstudio.com



sparklyr

dplyr

GraphX
(graphframes)

Streaming

ML

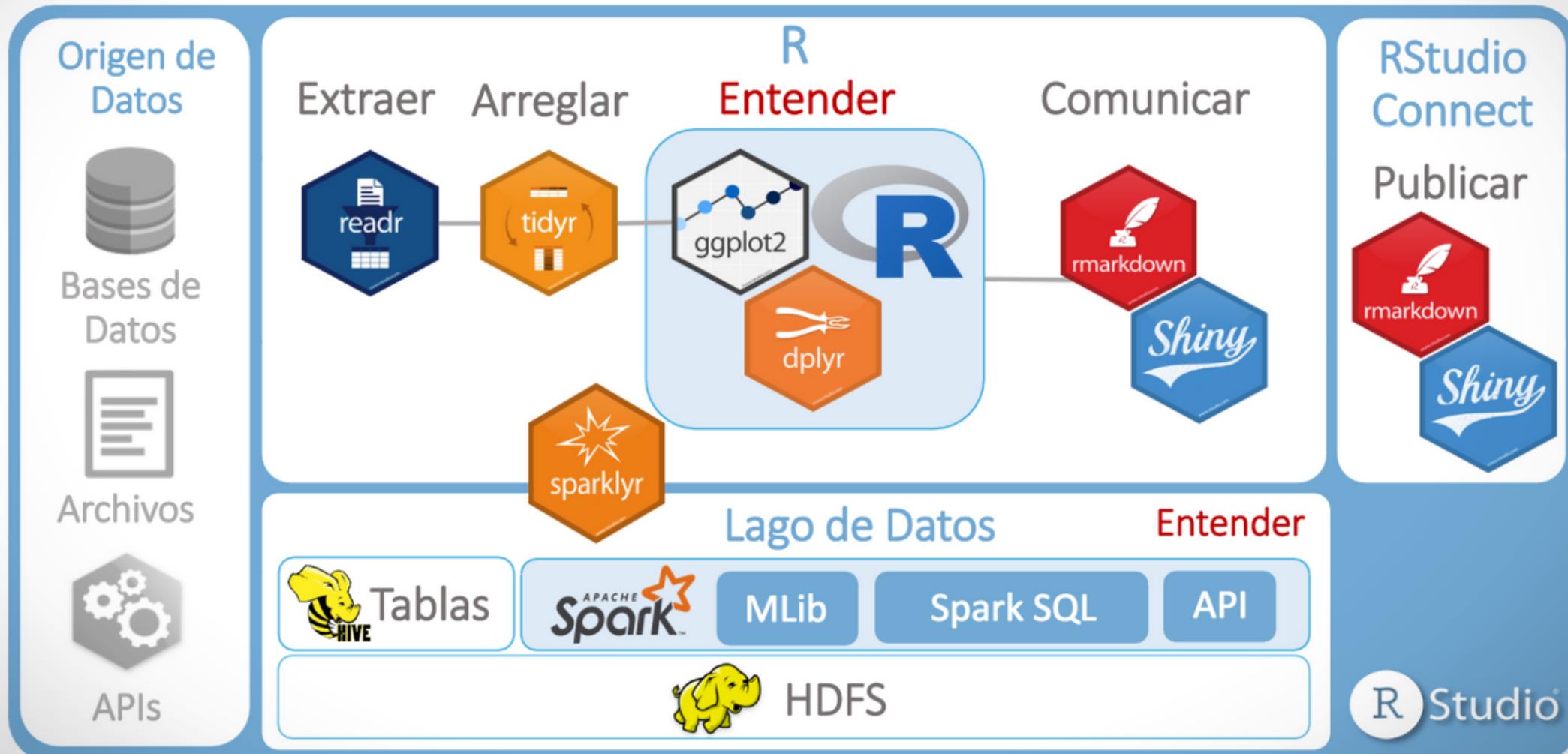
Extensions

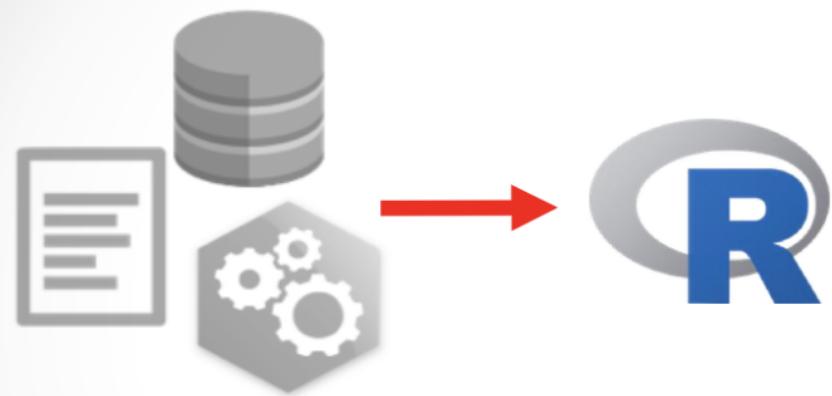


SPARKLYR

- CONECA A SPARK DESDE R. PROPORCIONA UN BACKEND COMPLETO DE DPLYR
- FILTRA Y AGREGA LOS CONJUNTOS DE DATOS DE SPARK Y LUEGO LOS TRAE A R PARA ANÁLISIS Y VISUALIZACIÓN.
- UTILIZA LA BIBLIOTECA DE APRENDIZAJE AUTOMÁTICO DISTRIBUIDO DE SPARK DE R.
- PUEDES CREAR EXTENSIONES QUE LLAMEN A LA API COMPLETA DE SPARK Y PROPORCIONE INTERFACES A LOS PAQUETES DE SPARK.

Ciencia de Datos con R y Spark





Extraer
datos



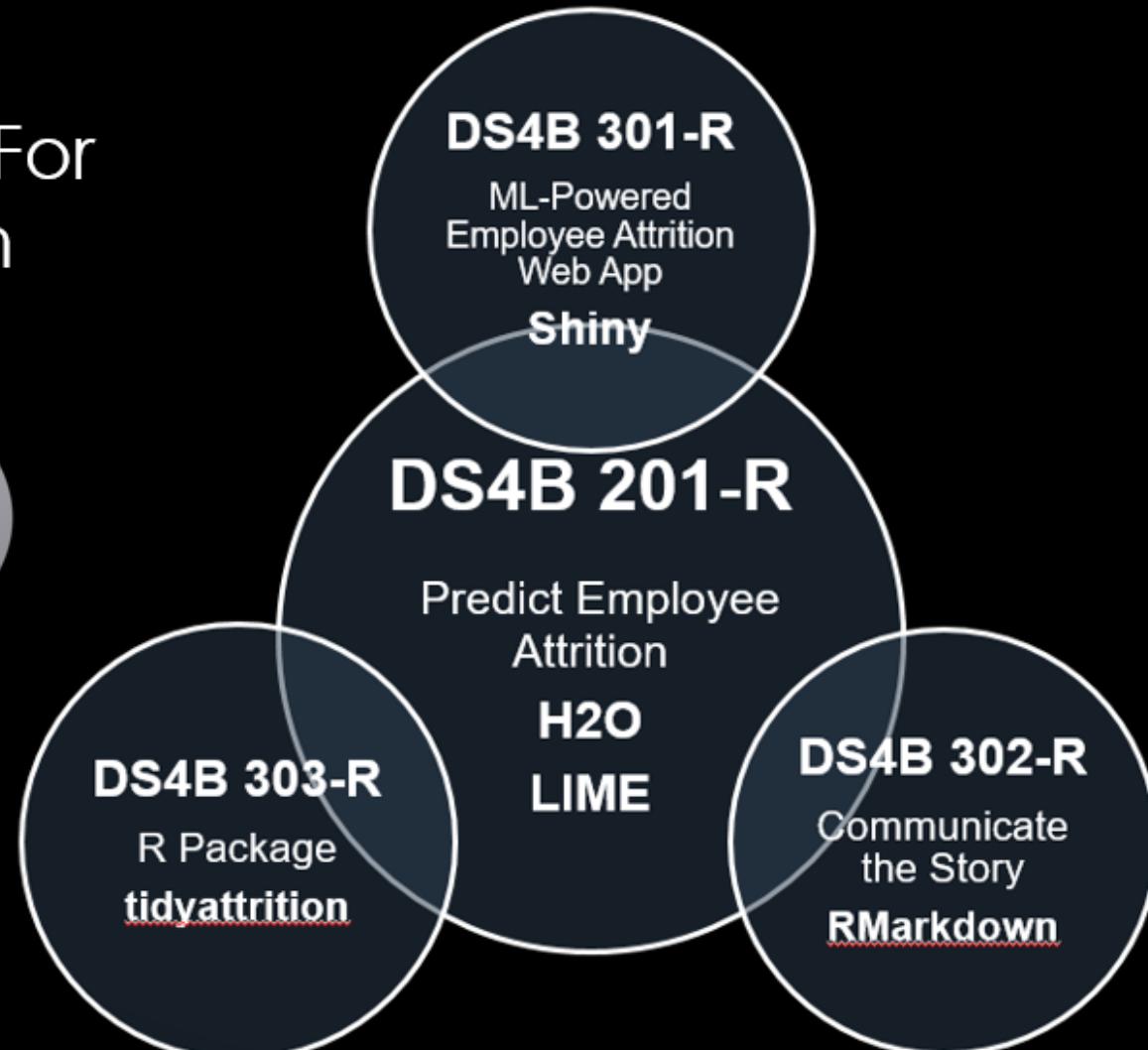
**¿QUIERES
APRENDER MÁS?**

R Track

Data Science For
Business With



Business Science
University



Data Science For Business With Python (DS4B 201-P)

Data Science For
Business With



Business Science
University



university.business-science.io



Favio Vázquez

Founder / Ciencia y Datos

@ favio@cienciaydatos.org

 @faviovaz

 <https://www.linkedin.com/in/faviovazquez/>

 www.cienciaydatos.org