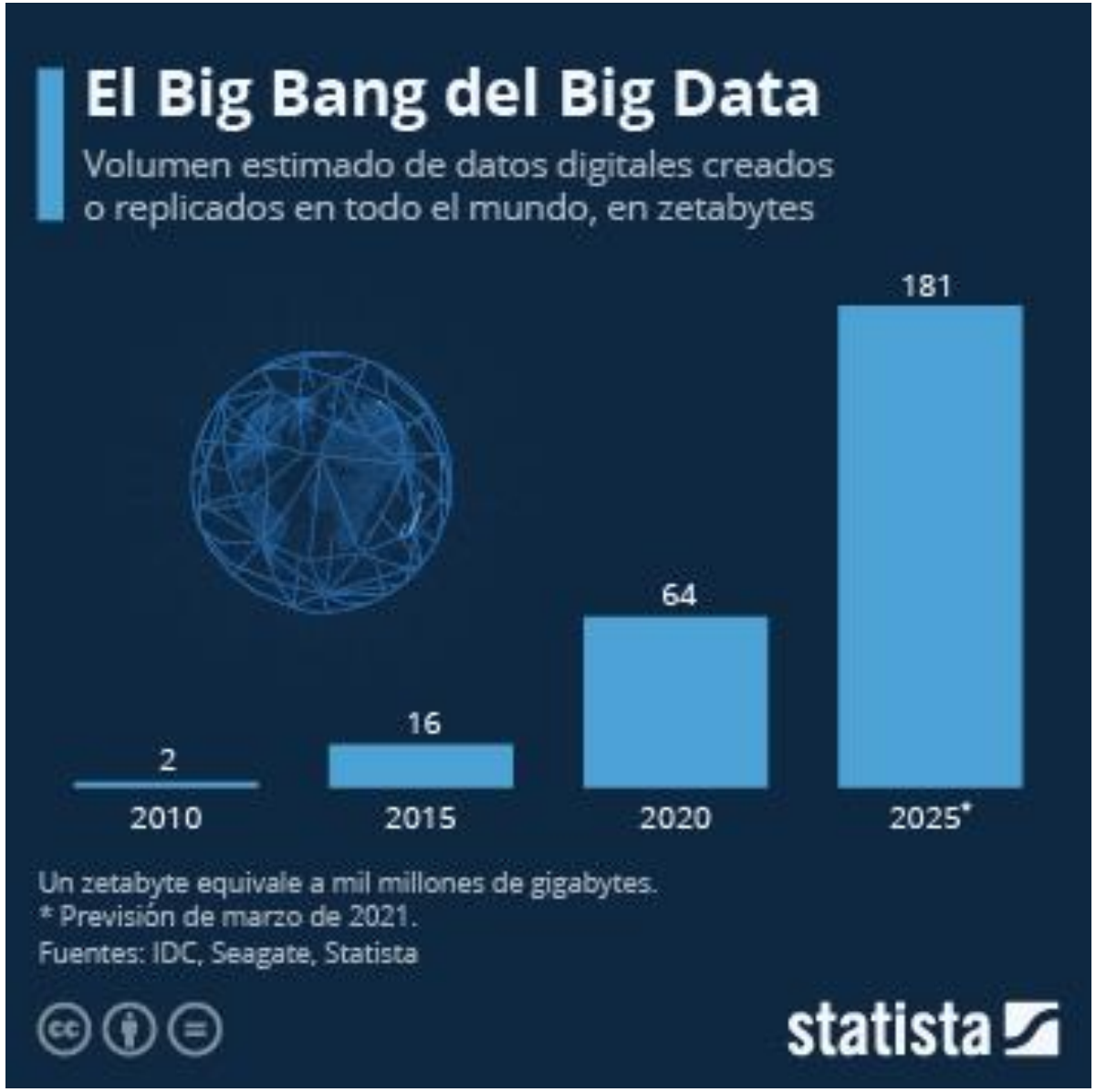
BIG DATA PYTHON

PRESENTACION

Conocimientos previos Python

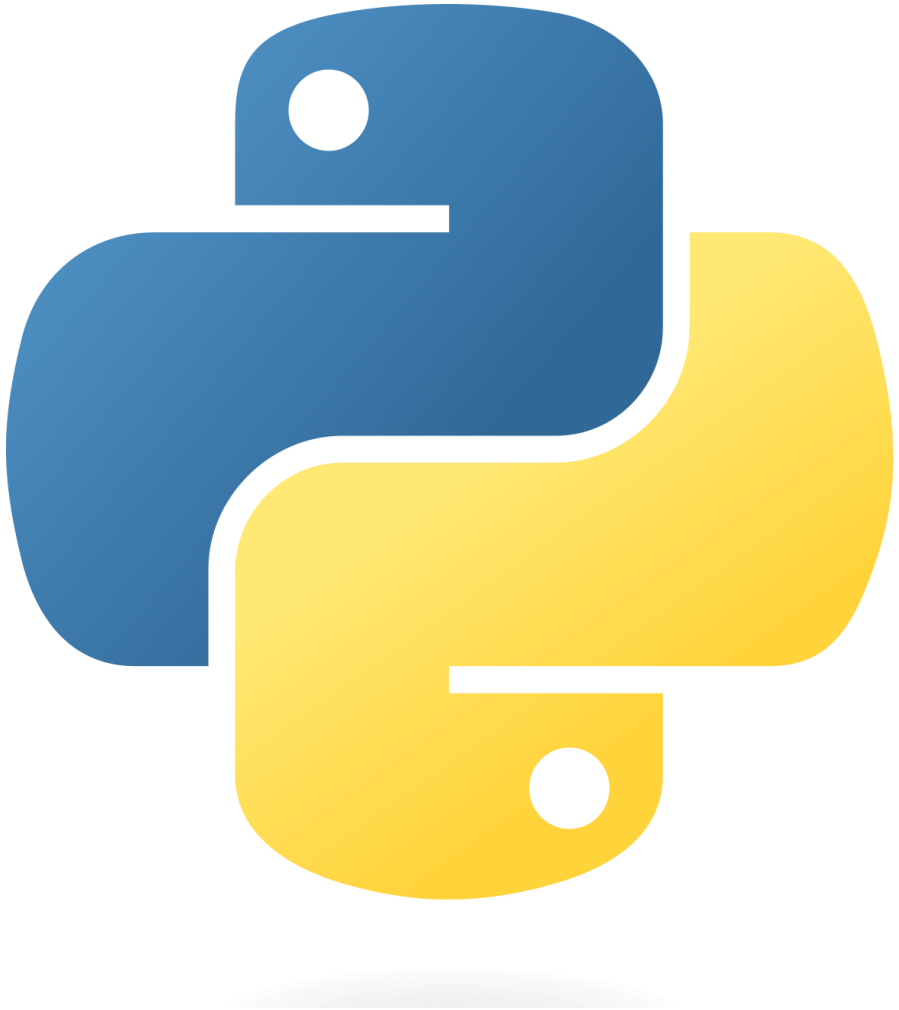
Conocimientos previos Spark

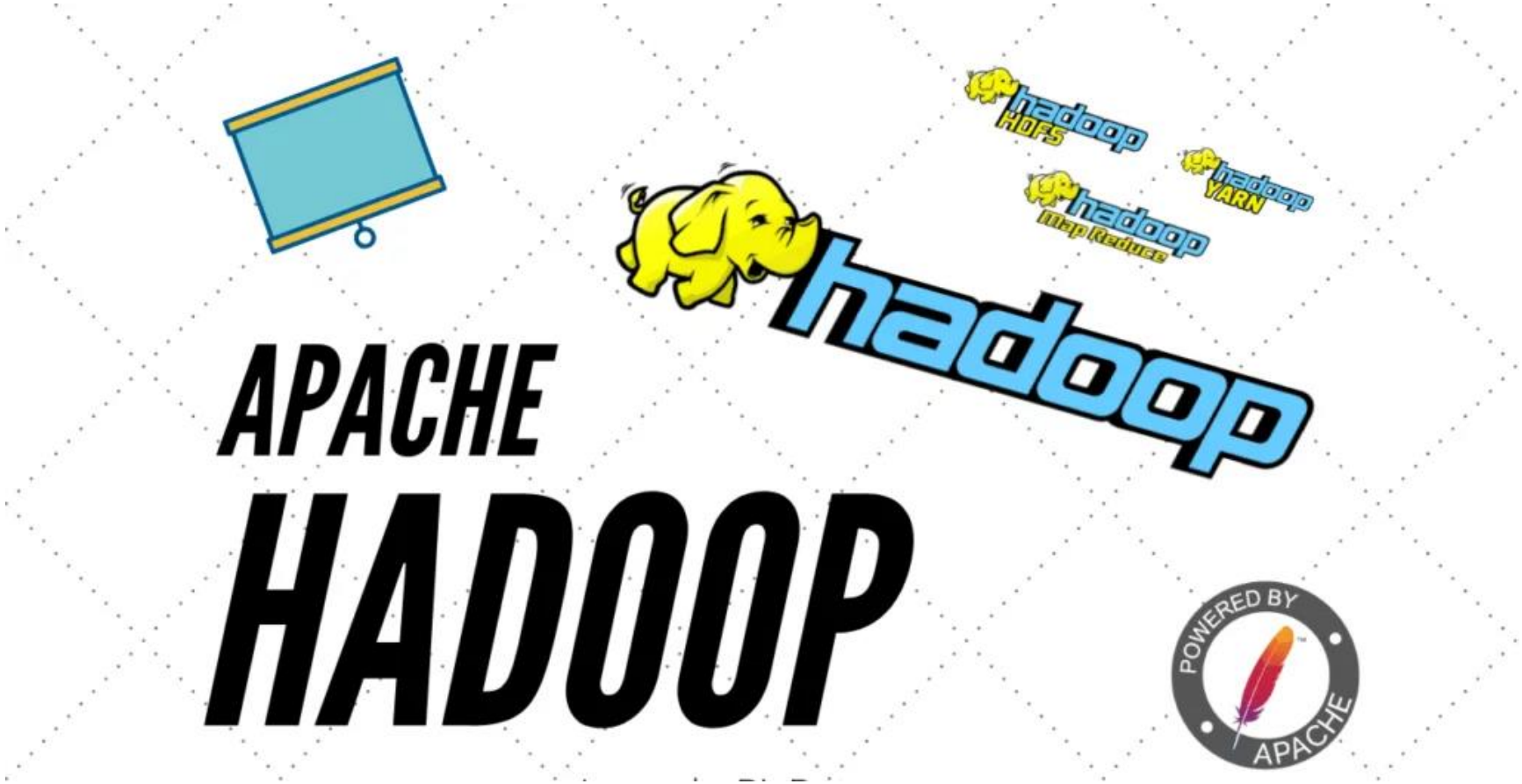
BIG DATA


El Big Bang del Big Data
Volumen estimado de datos digitales creados o replicados en todo el mundo, en zetabytes
2
16
64
181
2010
2015
2020
2025*
Un zetabyte equivale a mil millones de gigabytes.
* Previsión de marzo de 2021.
Fuentes: IDC, Seagate, Statista
statista

LAS 7 V'S DEL BIG DATA

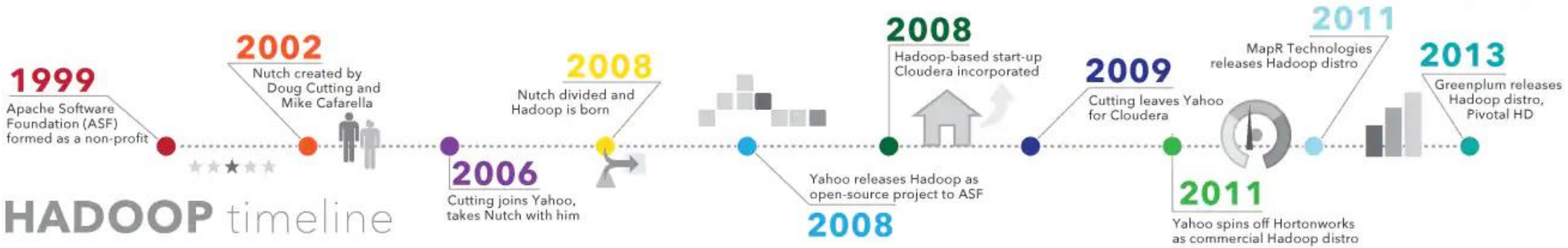
PYTHON



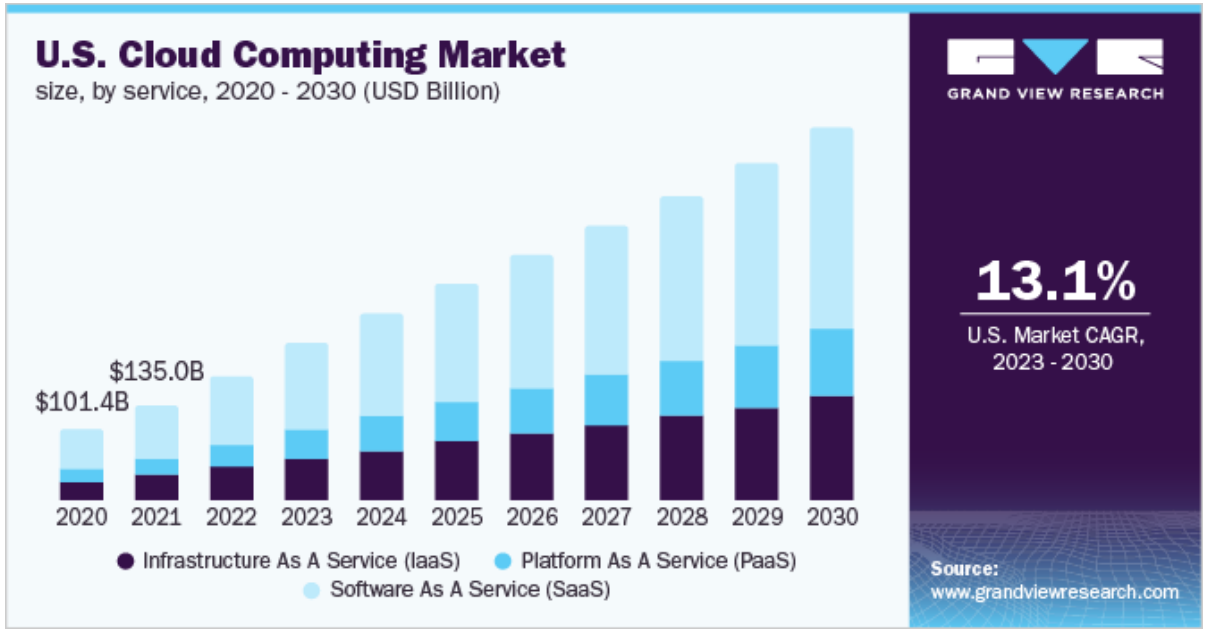

hadoop
HDFS
hadoop
YARN
hadoop
Map Reduce
APACHE
HADOOP
hadoop
POWERED BY
APACHE

HADOOP

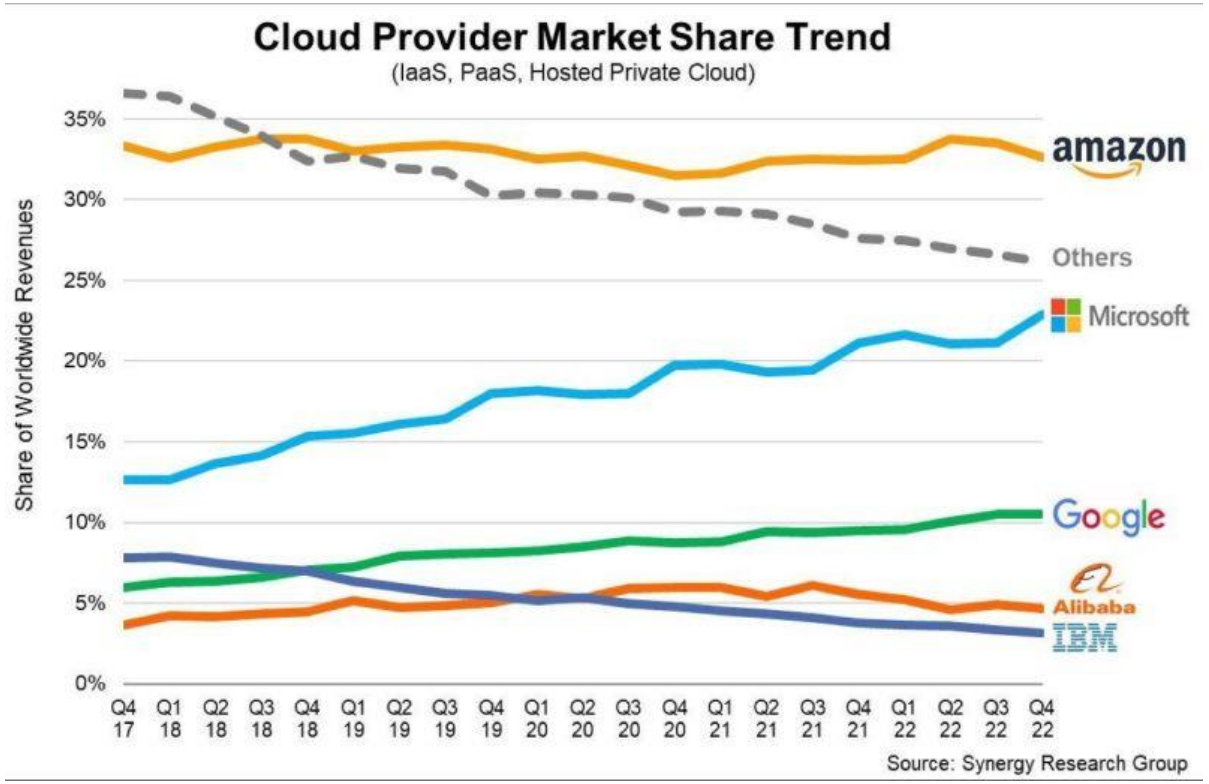
HISTORIA HADOOP



EVOLUCIÓN CLOUD COMPUTING

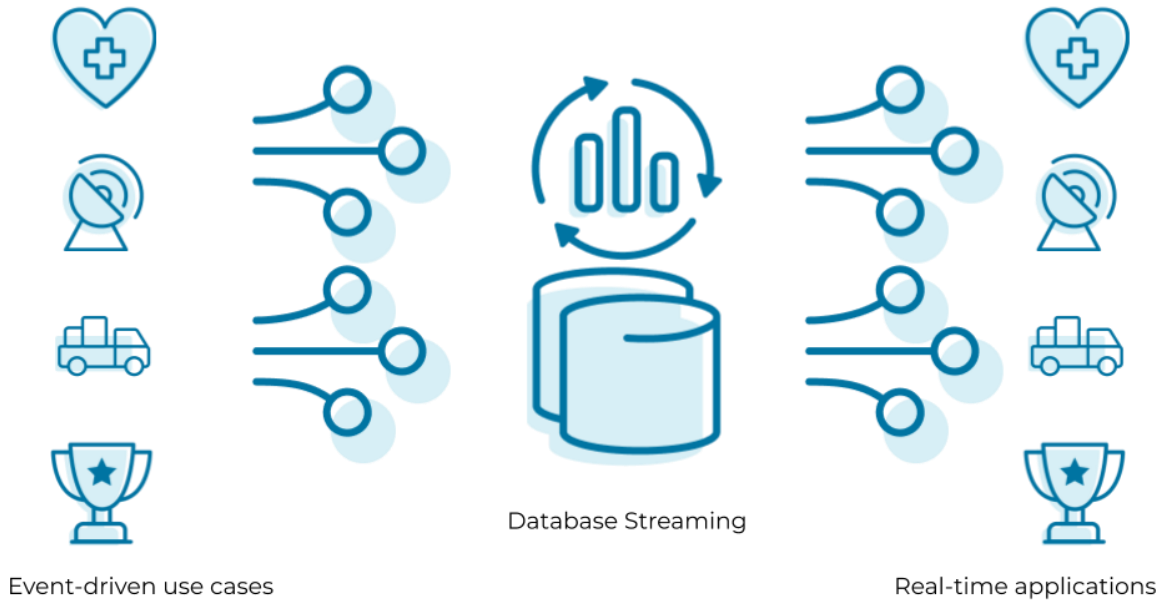


PRINCIPALES PROOVEDORES

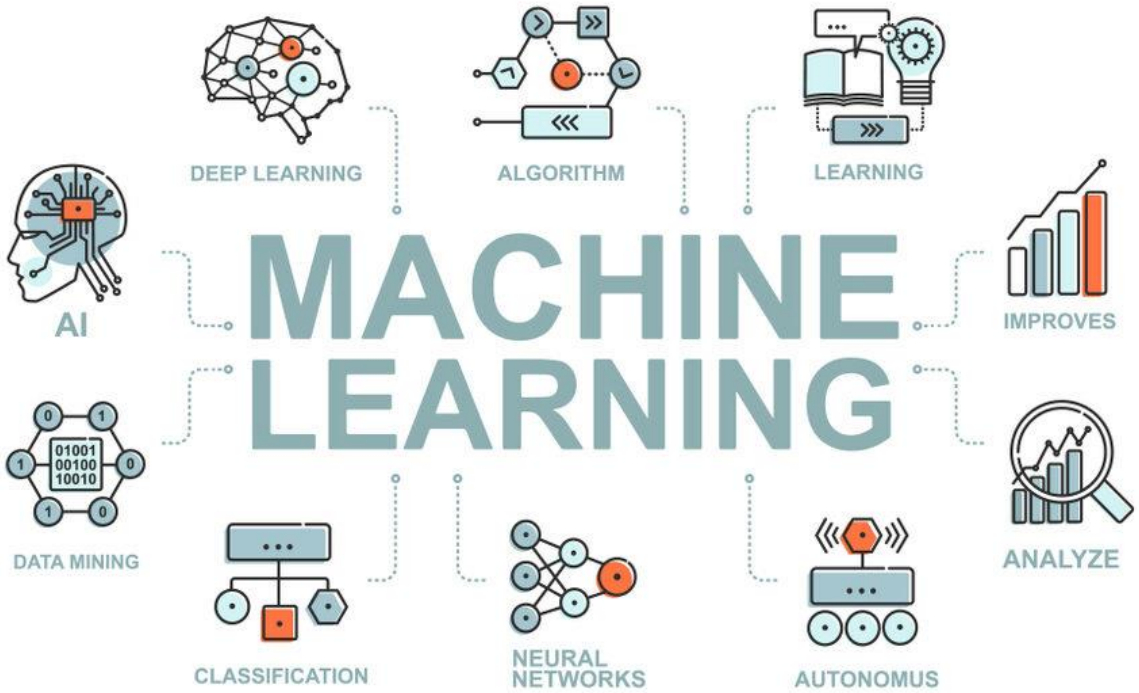


VENTAJAS DE CLOUD COMPUTING

STREAMING DATA

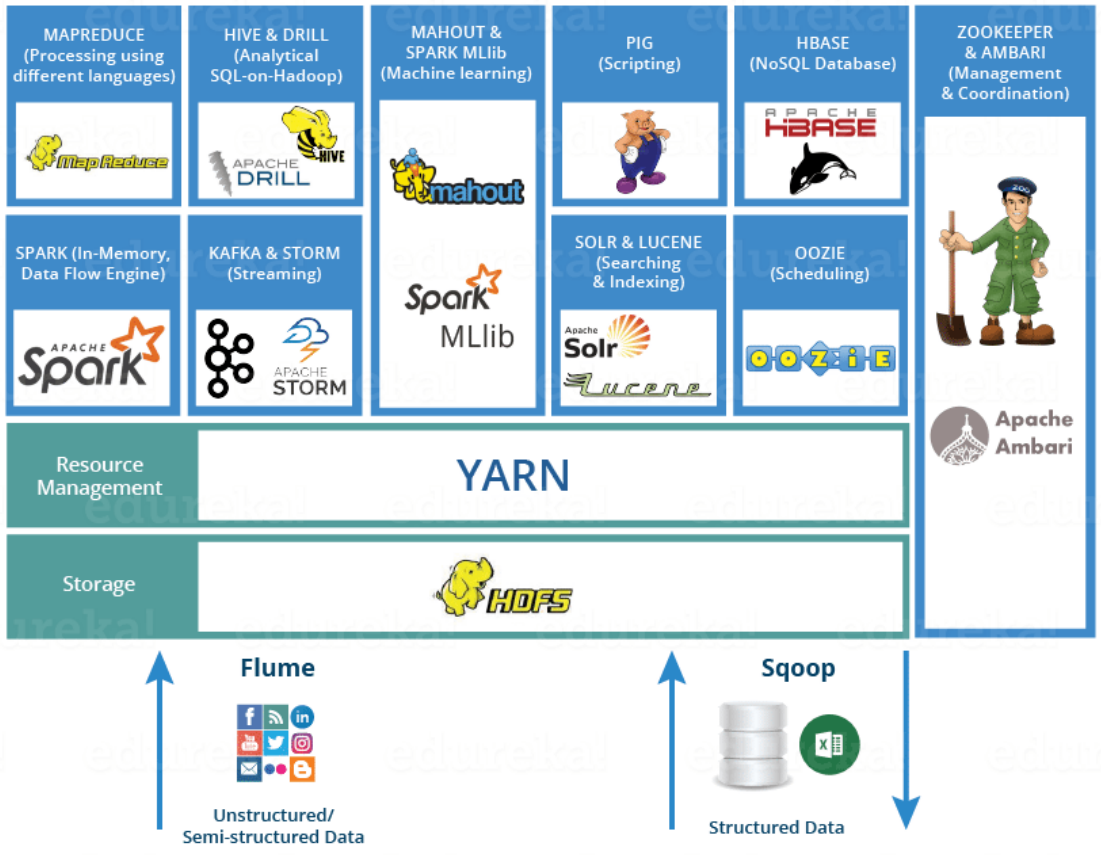


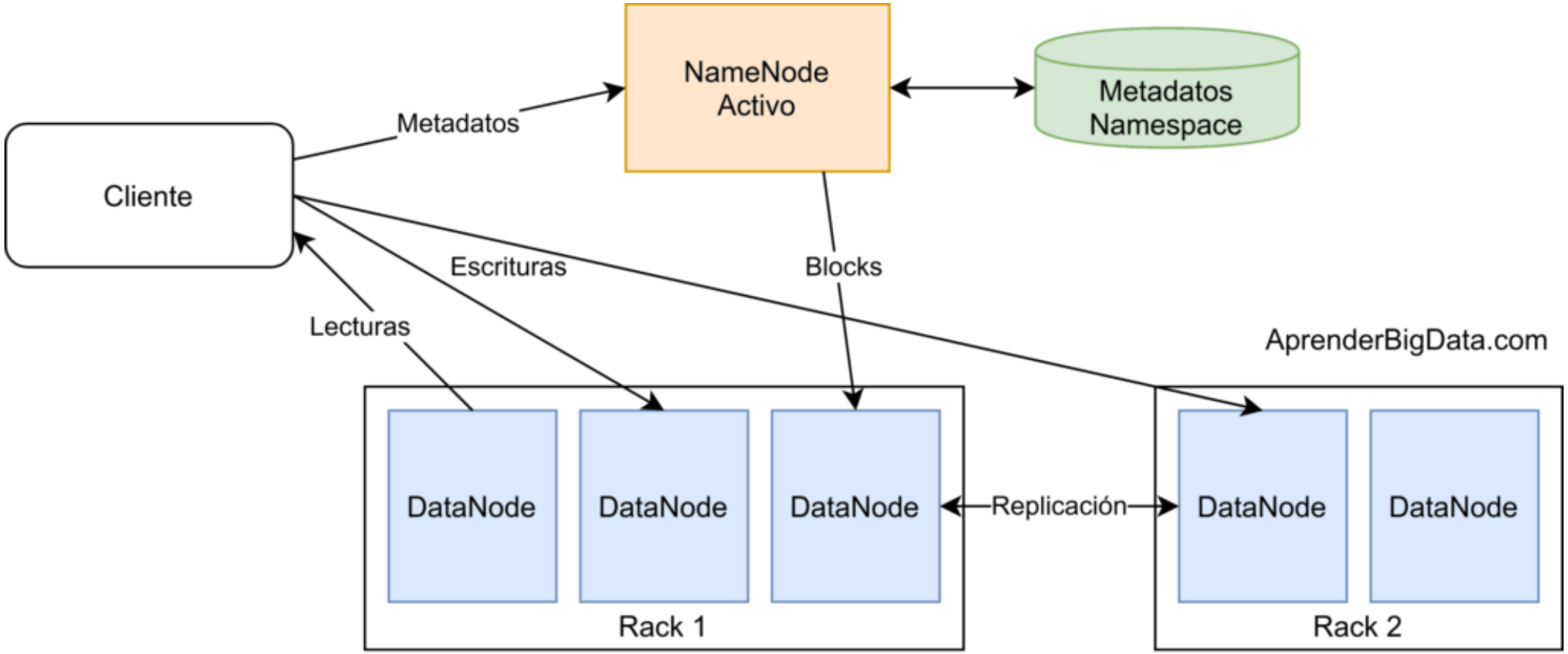
MACHINE LEARNING



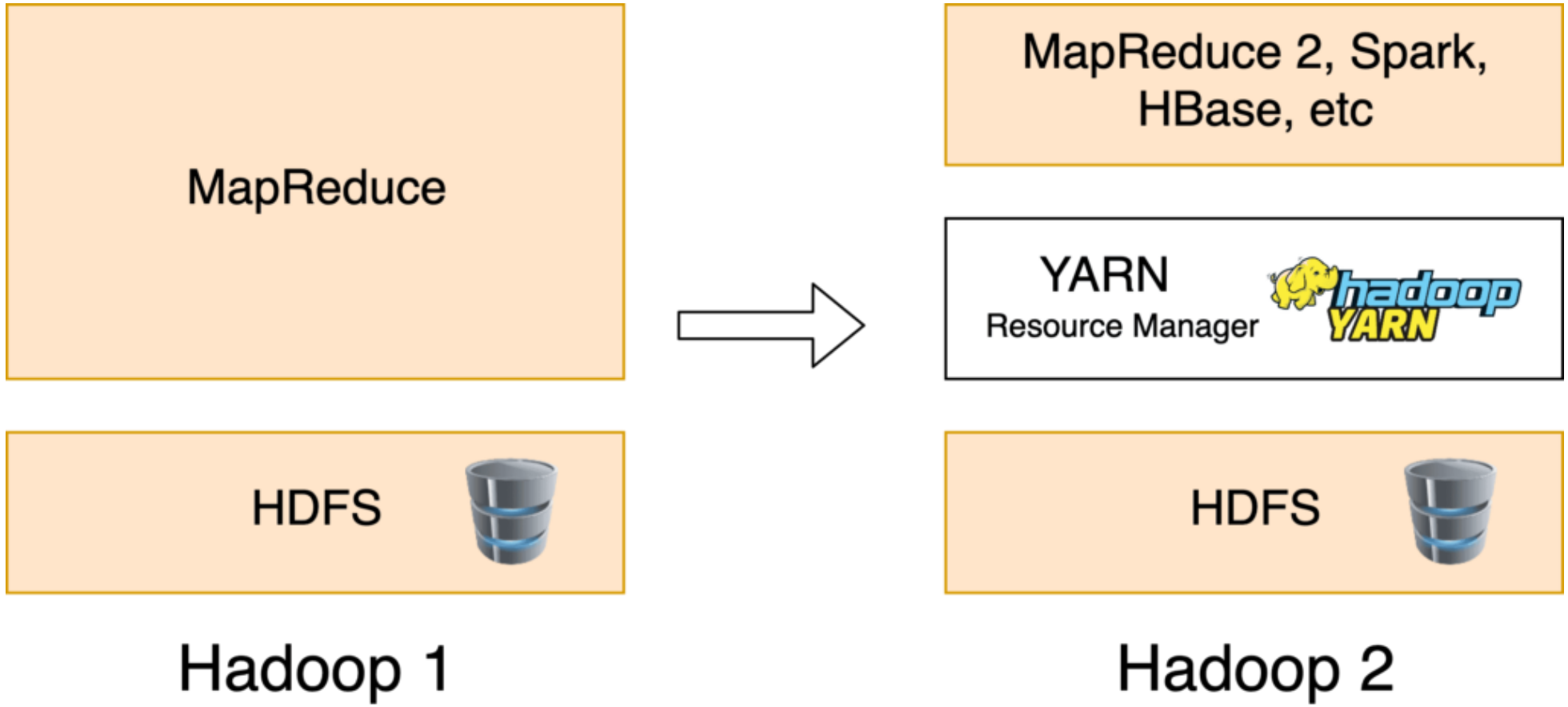
NurPhoto/picture alliance

ECOSISTEMA HADOOP



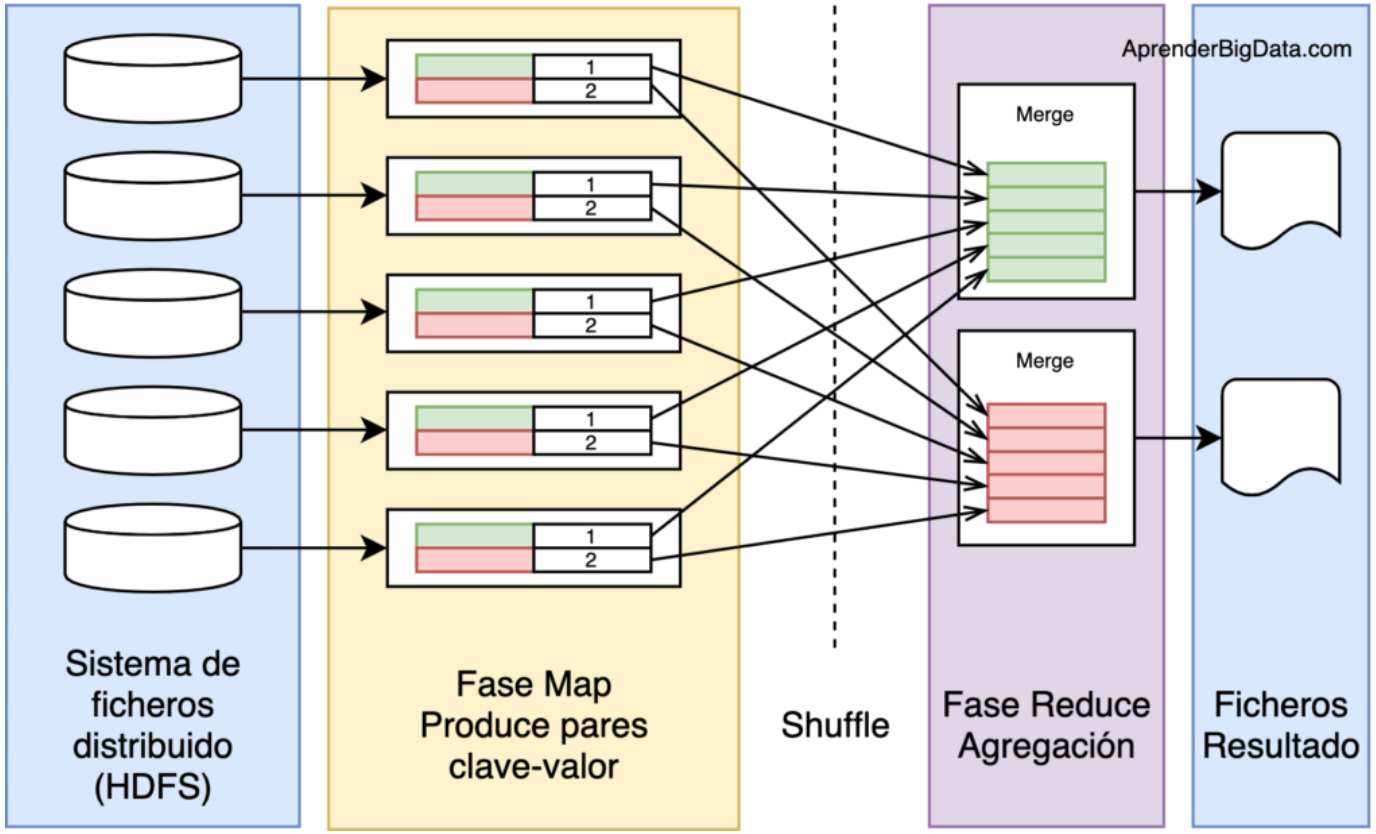


ARQUITECTURA DE HADOOP


MapReduce
HDFS
Hadoop 1
MapReduce 2, Spark,
HBase, etc
YARN
Resource Manager
hadoop
YARN
HDFS
Hadoop 2

YARN

MAP REDUCE

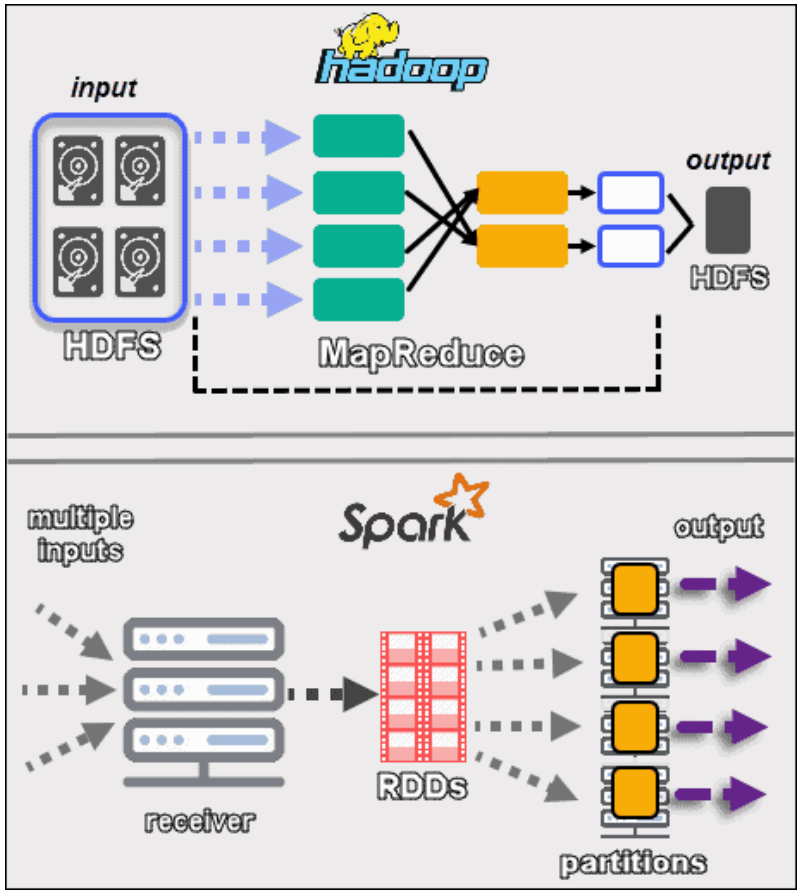

AprenderBigData.com
1
2
Merge
Merge
Sistema de ficheros distribuido (HDFS)
Fase Map
Produce pares clave-valor
Shuffle
Fase Reduce
Agregación
Ficheros Resultado

SPARK

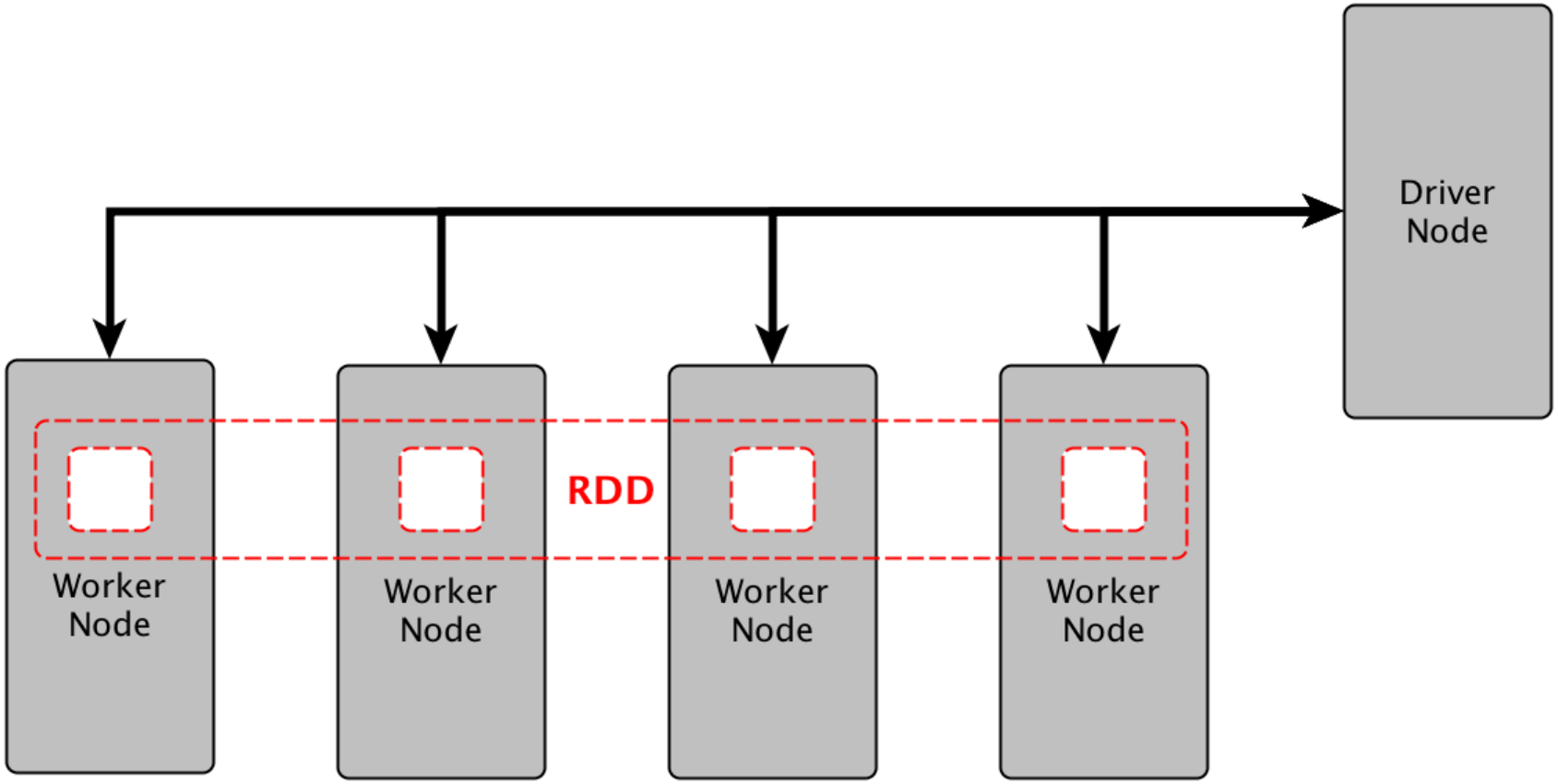
- “MapReduce va en camino de convertirse en un legado, tenemos Spark” dice el hombre detrás de Apache Hadoop, *Doug Cutting* .

databricks

DATABRICKS

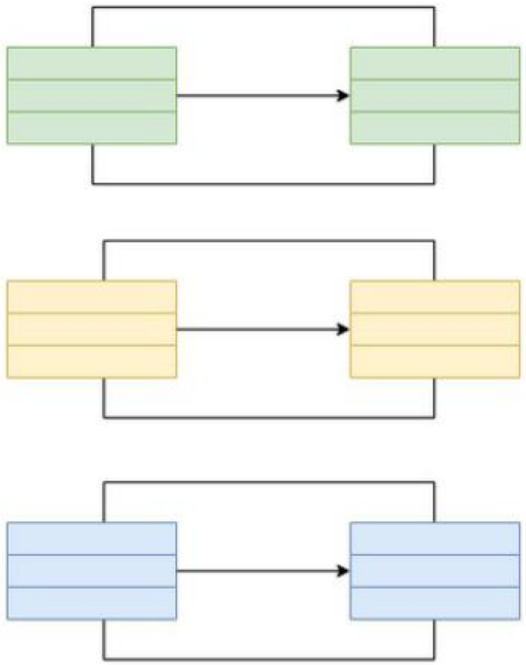


SPARK VS MAP REDUCE

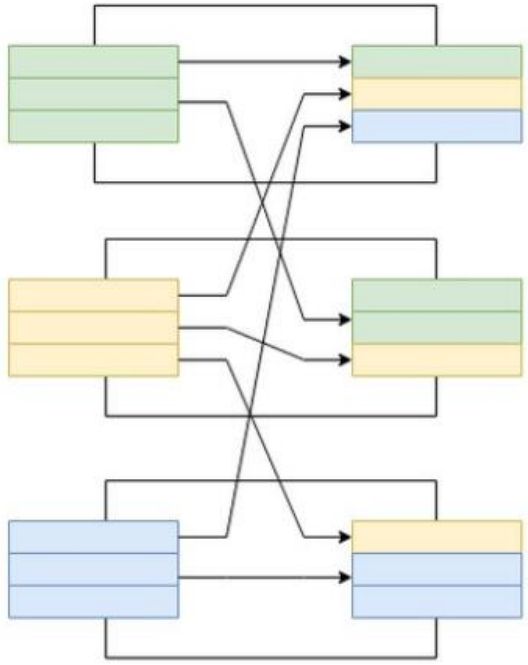

Driver
Node
RDD
Worker
Node
Worker
Node
Worker
Node
Worker
Node

RDD’S

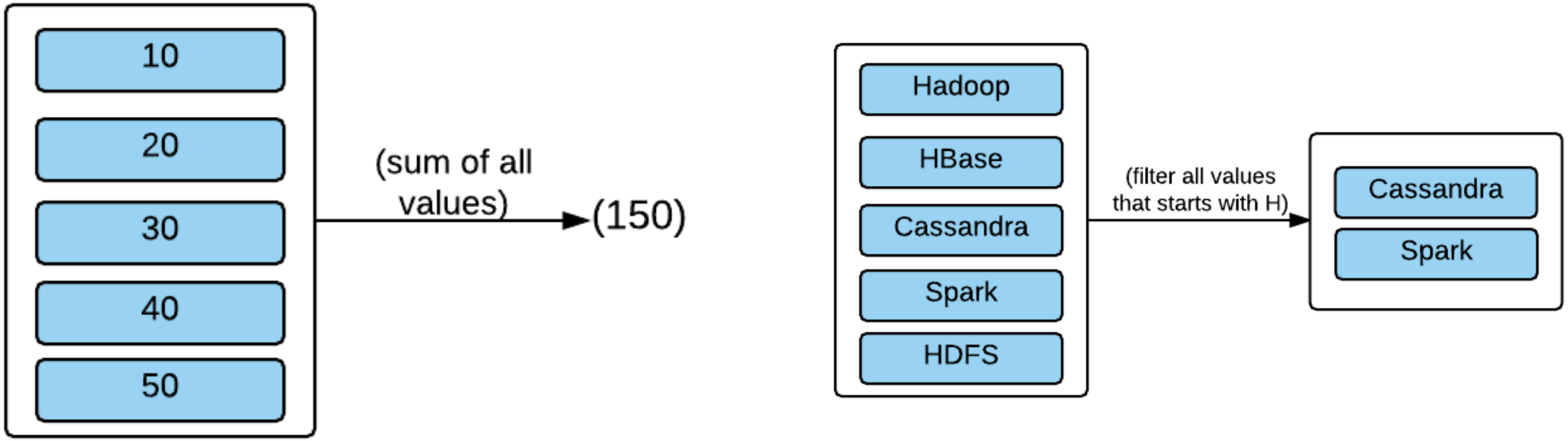
TIPOS TRANSFORMACIONES



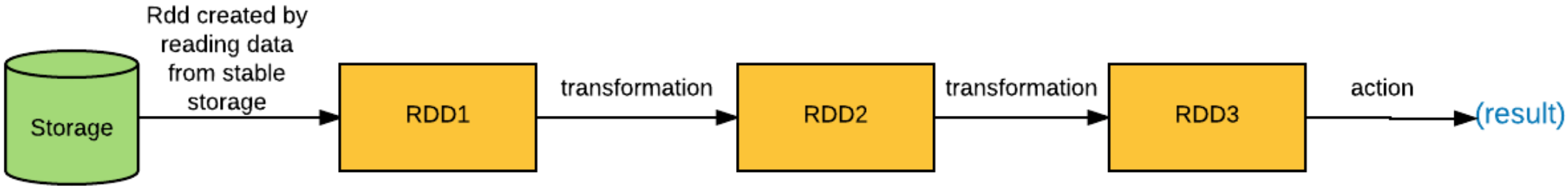
Transformación narrow



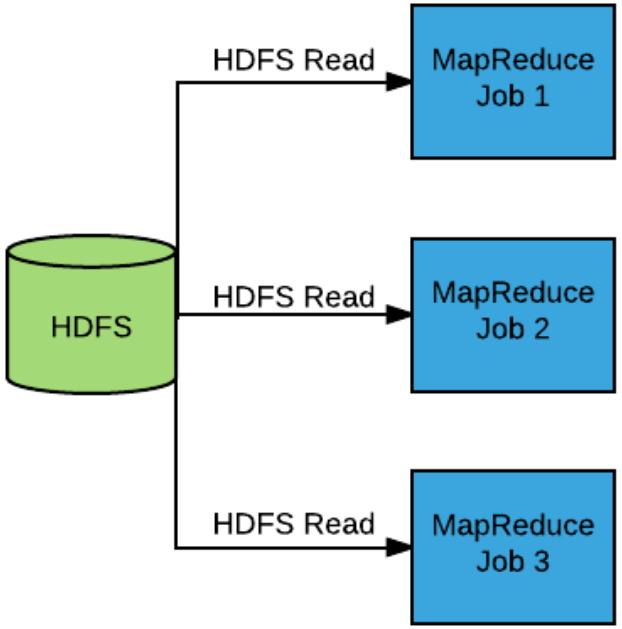
Transformación wide

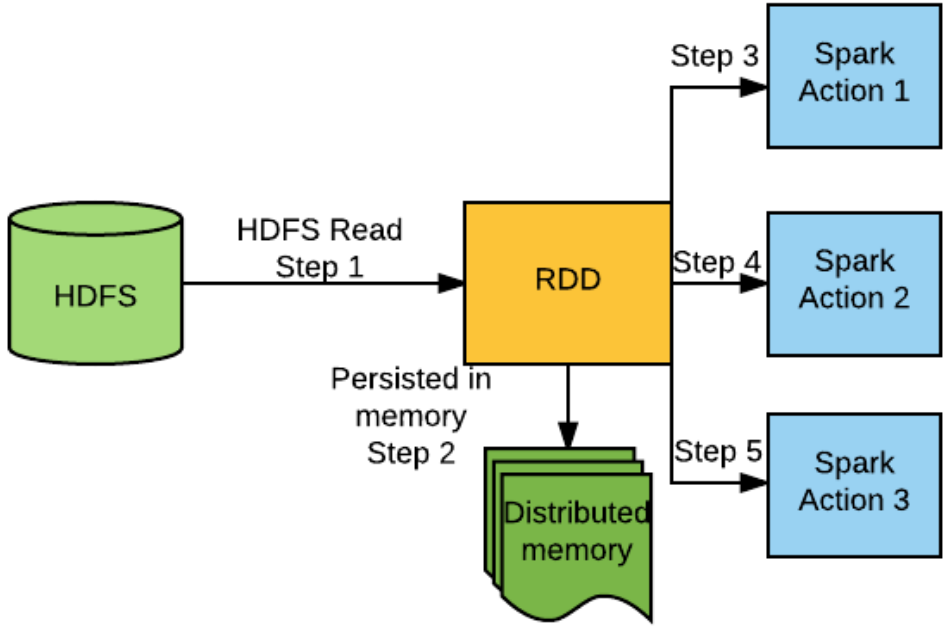


TRANSFORMACIONES Y ACCIONES



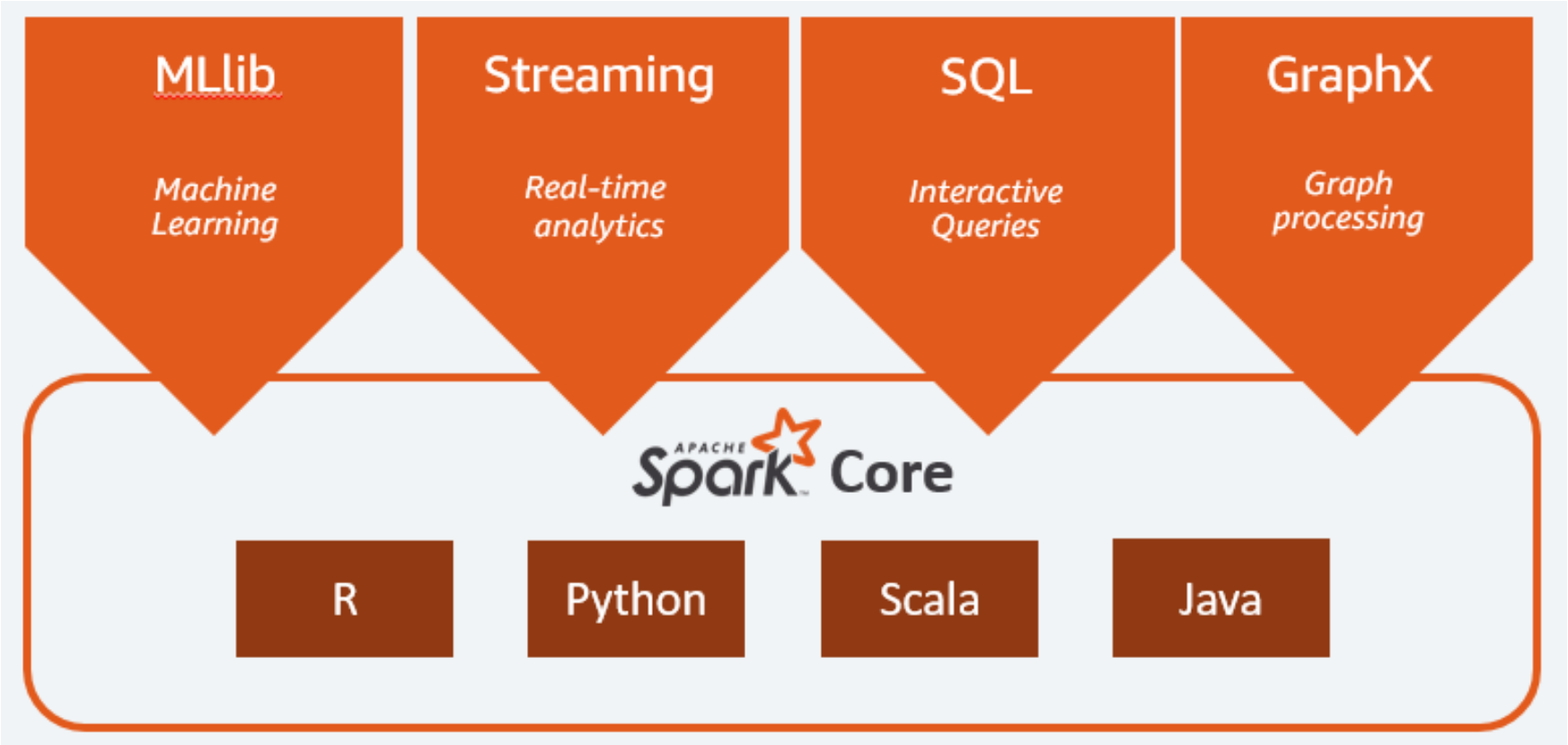
EVALUACIÓN PEREZOSA

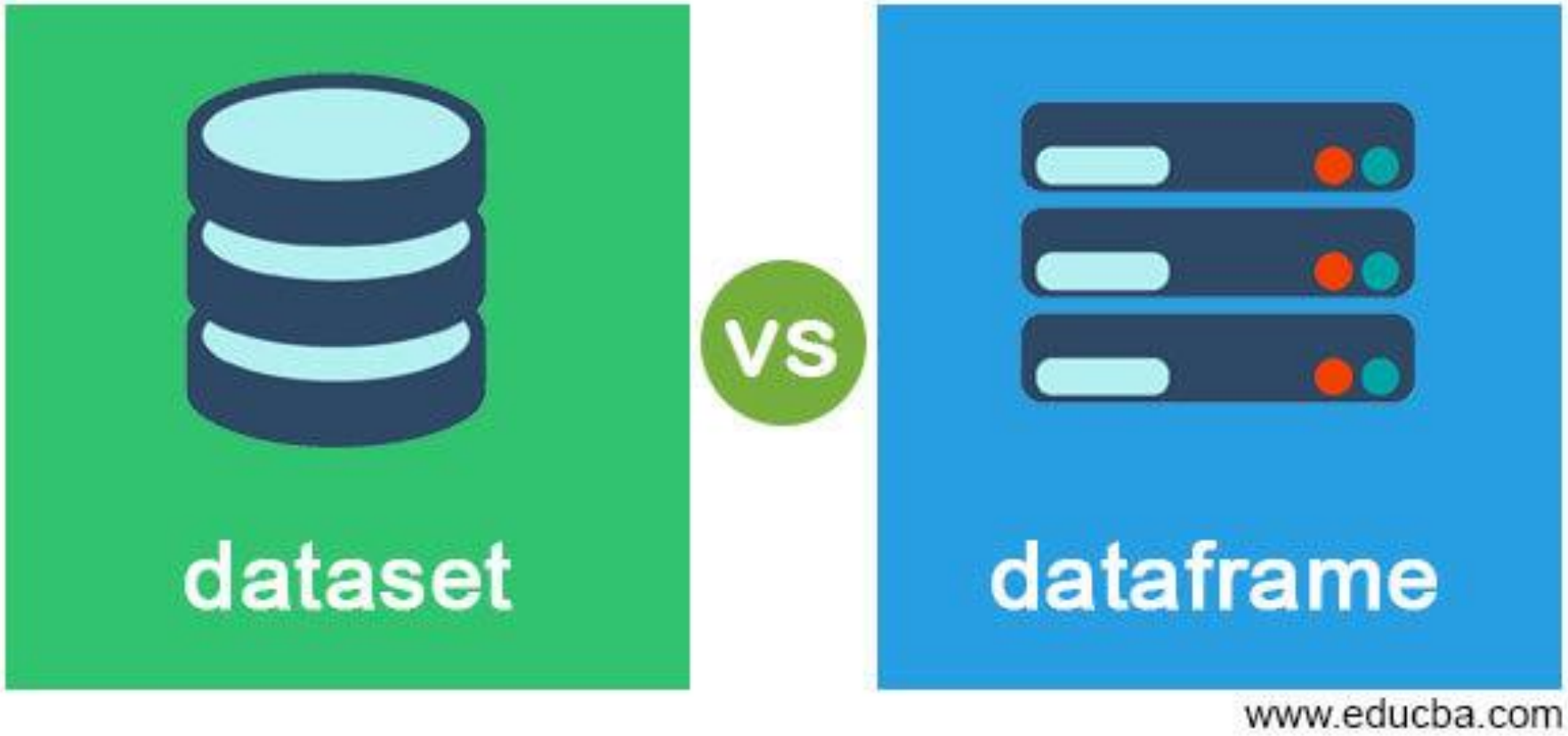




BENEFICIOS RDD

ECOSISTEMA SPARK





DATASETS Y DATAFRAMES

https://www.linkedin.com/pulse/rdd-datarame-datasets-apache-spark-shahzad-aslam/