Big Data

Data Science Bootcamp

The Bridge

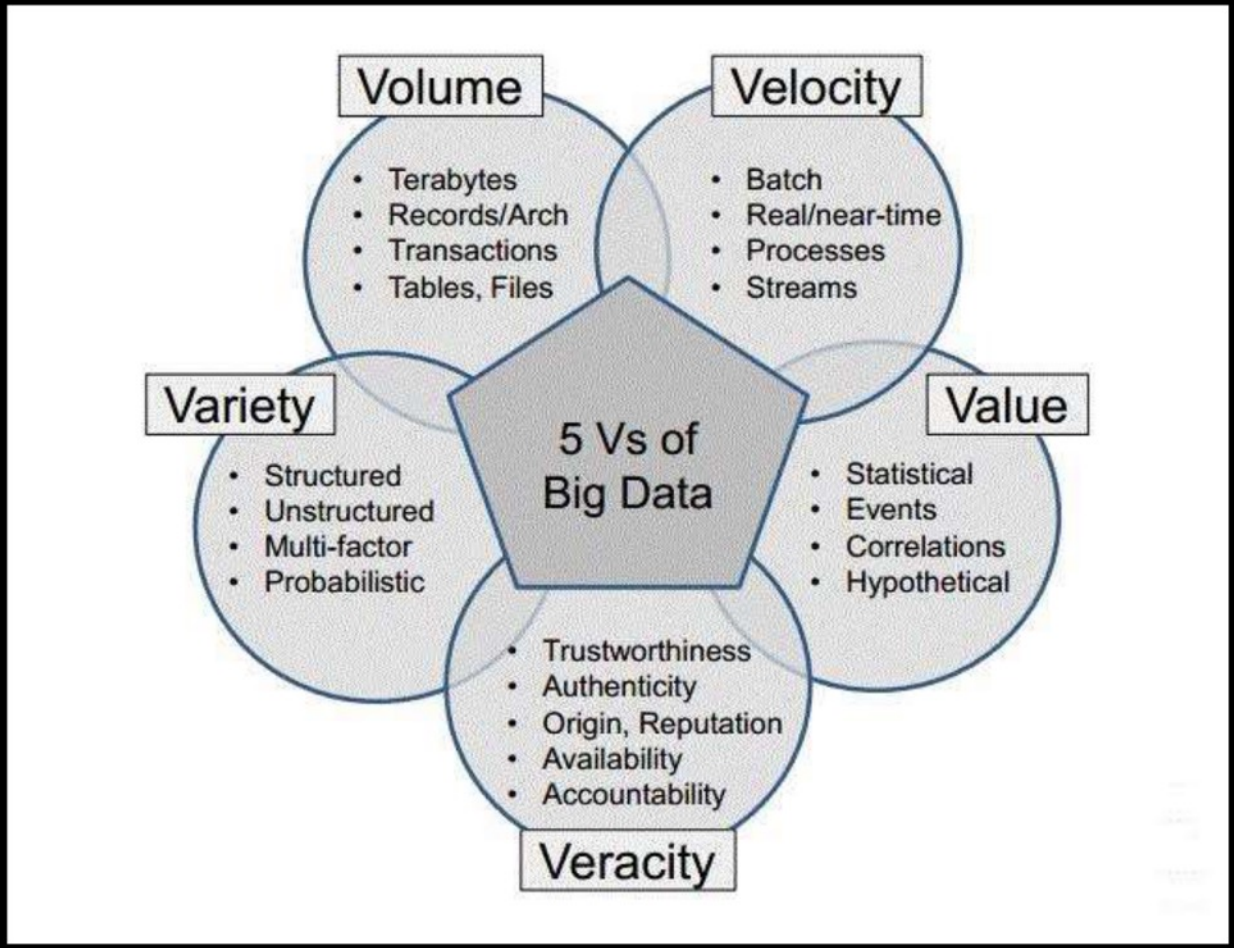
Big Data

Definición

*"Conjuntos de datos o combinaciones de conjuntos de datos cuyo **tamaño (volumen), complejidad (variabilidad)** y **velocidad de crecimiento (velocidad)** dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles."*

Las famosas 3 V’s

Se suman dos más al carro



¿Por qué ha surgido esto ahora?

Internet

Generación de datos

Capacidad de cómputo

Más fuentes de datos

Visión 360 del cliente con mayor cantidad de fuentes de datos

Web
Gran cantidad de info a través de las cookies, logs, navegación de los usuaris

Terceros
Posibilidad de compra de información anonimizada o no, a terceros

Redes Sociales
Redes como Twitter, LinkedIn, Facebook

Internet of Things
Uso masivo de sensores sincronizados con la nube y generando datos en real time

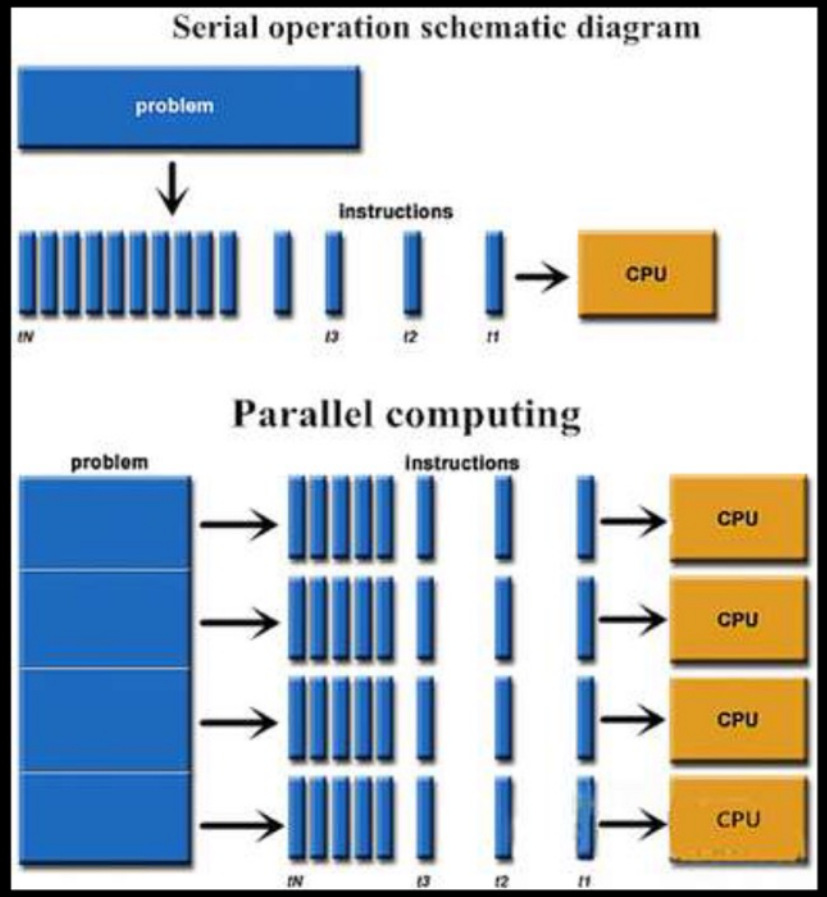
Info no estructurada
Datos no estructurados como imágenes, HTML, XML, voz.

Tecnologías

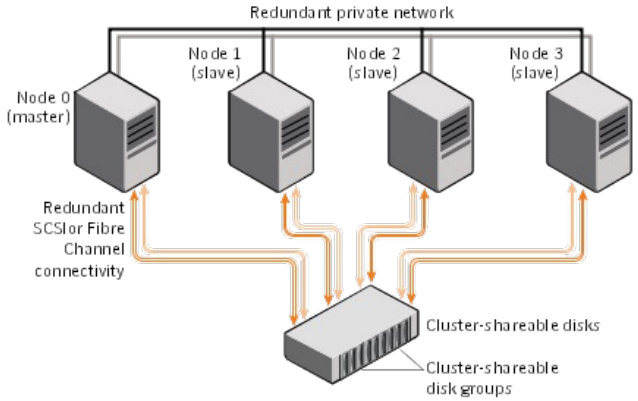
Sistemas distribuidos

Cuando la CPU no da más de sí

Computación distribuida en un ordenador



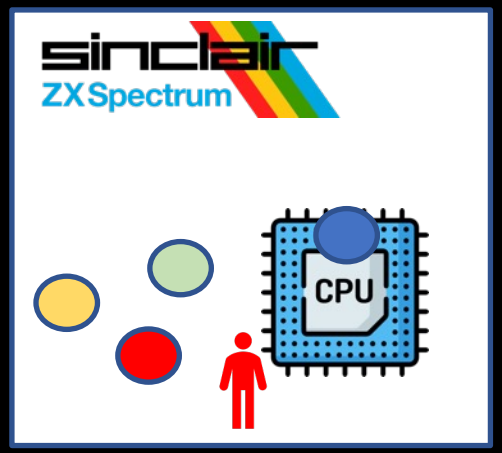
Computación distribuida en varios ordenadores



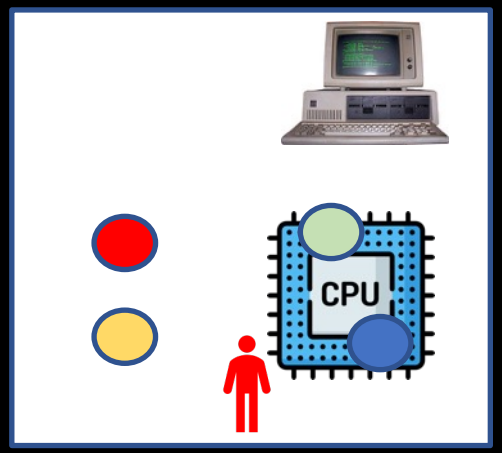
A cluster, in the context of servers, is a group of computers that are connected with each other and operate closely to act as a single computer. Speedy local area networks enhance a cluster of computers' abilities to operate at an exceptionally rapid pace.

Un poco de historia para entender mejor

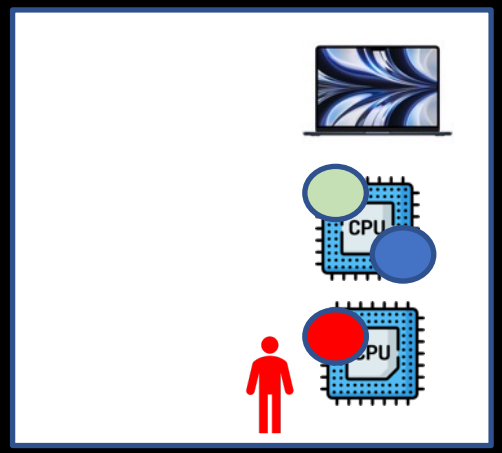
De mi Spectrum a AWS



Una CPU, un solo programa a la vez



Una CPU, varios programas a la vez (se les da un poquito de tiempo y se les echa)



Varias CPUS(núcleos), varios programas a la vez (se les da un poquito de tiempo de cada CPU y se les echa)

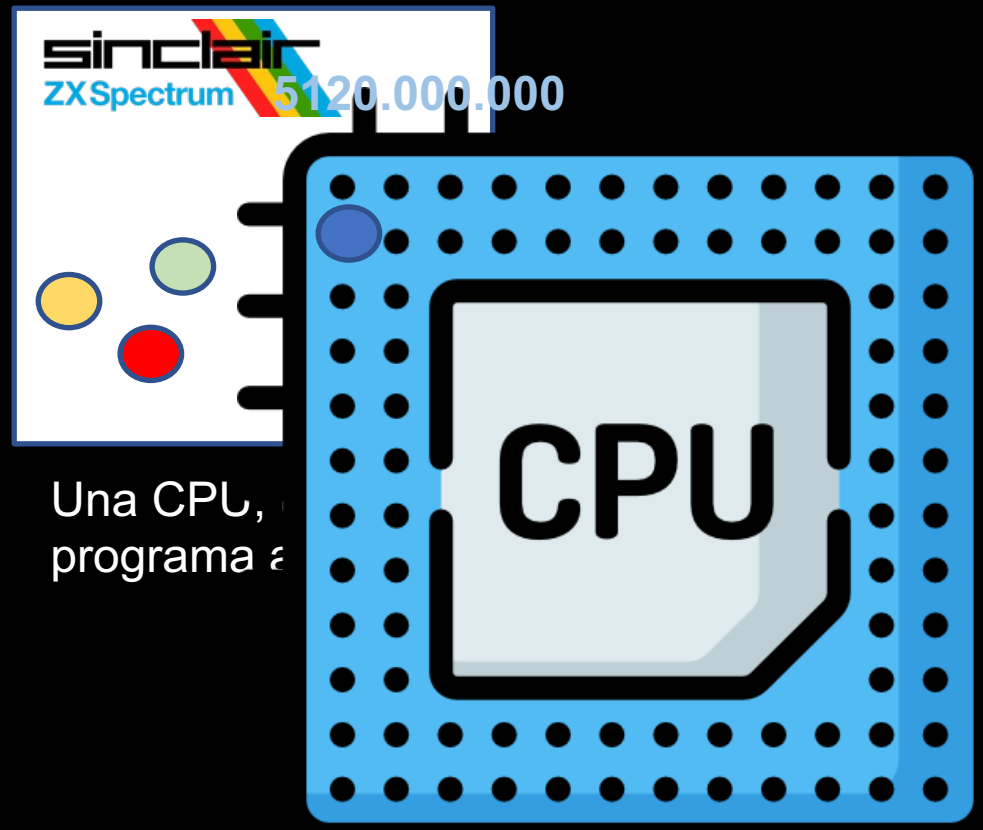
Varios ordenadores, varias CPUs, programas distribuidos manualmente

SISTEMA OPERATIVO

CLOUD/HOSTING SW

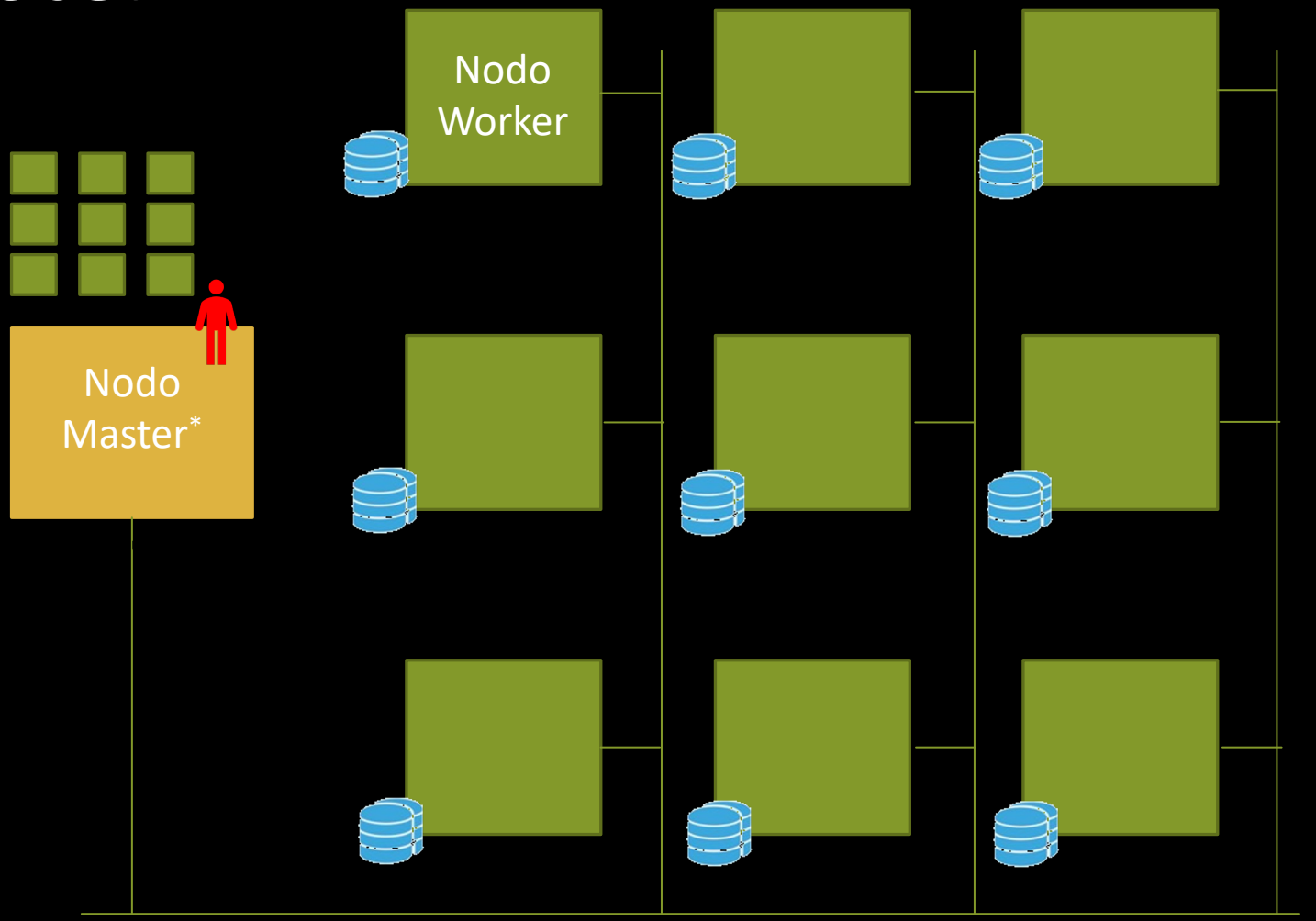
Un clúster = ordenadores como si fueran uno

Varios ordenadores, varias CPUs, programas distribuidos manualmente



Una CPU, programa a

Un clúster



Los Workes reciben otros nombres como Ejecutores (executors) y hace mucho mucho tiempo slaves

(*)Tanto Master como Workers son ORDENADORES (Instancias)

Es el SW de gestión del clúster

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen
Sirve para almacenar grandes volúmenes de información

Backups
Guarda copias de la información en diferentes nodos

Tolerancia a fallos
En caso de que se caiga un nodo, cuenta con otros para mantener el servicio

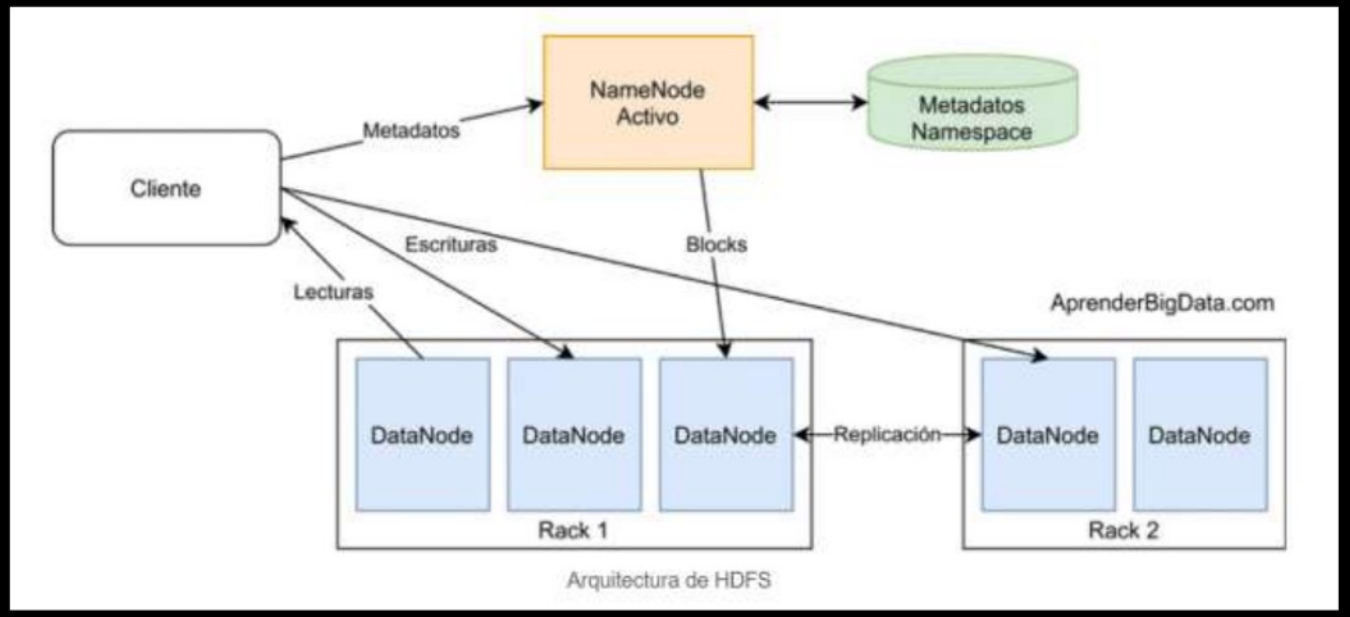
YARN
Gestor de recursos de Hadoop

Escalabilidad
Es cuestión de añadir nuevos nodos, de hardware económico

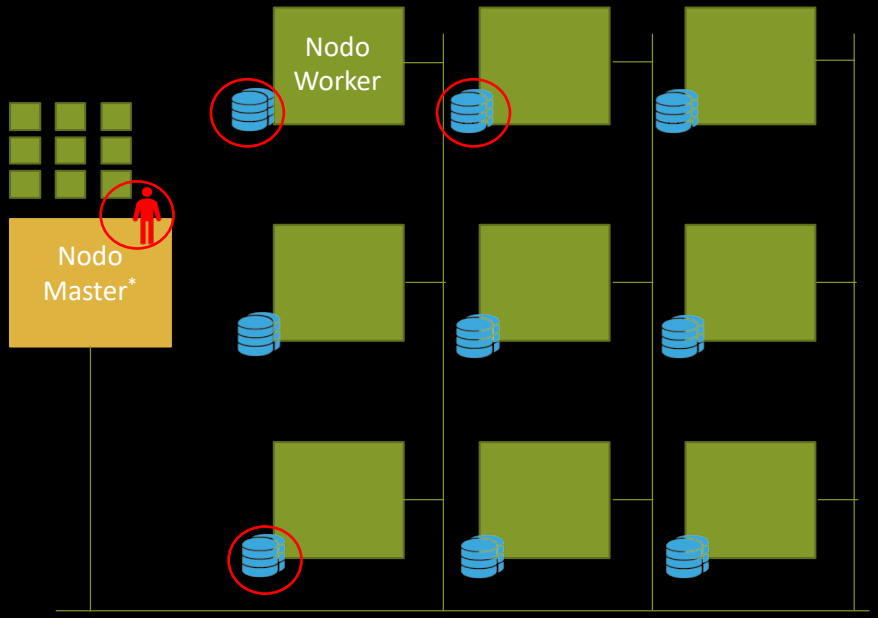
HDFS

High Distributed File System

"Sistema de ficheros distribuidos de Hadoop. Sirve para el almacenamiento masivo de información, tanto para datos estructurados, semi-estructurados y no estructurados."



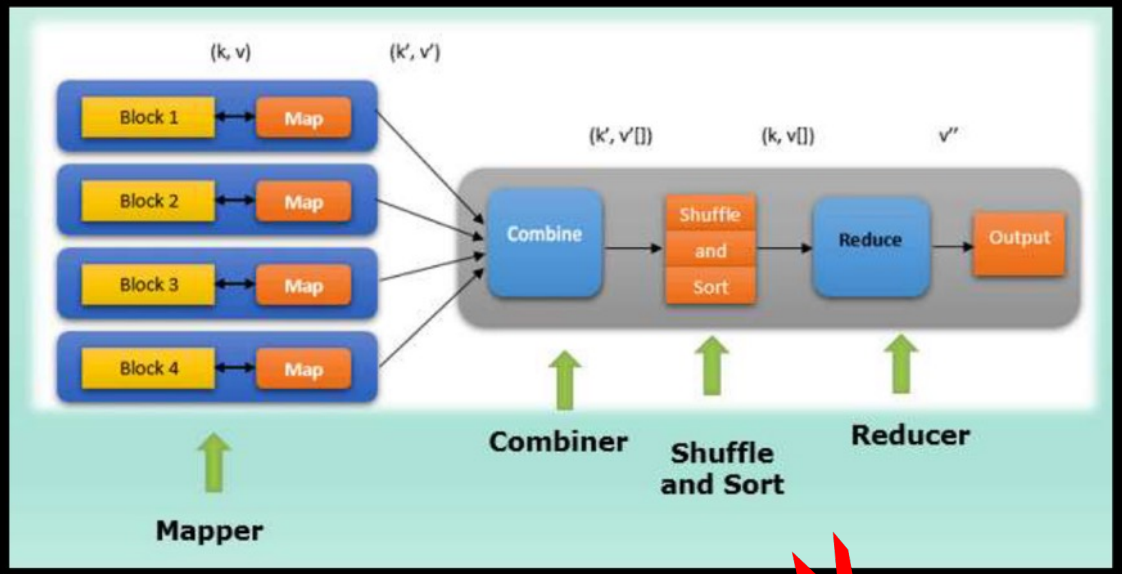
Arquitectura de HDFS

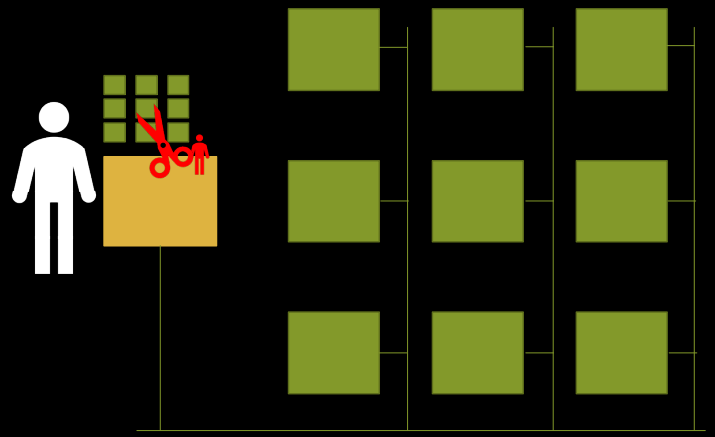


Map Reduce

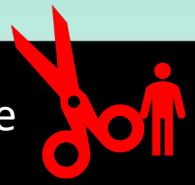
Paradigma de programación

"MapReduce es una técnica de procesamiento y un programa modelo de computación distribuida basada en java. Mediante el Map se generan pares clave-valor y en el Reduce se produce la agregación."





La "primera" versión, junto con YARN, de

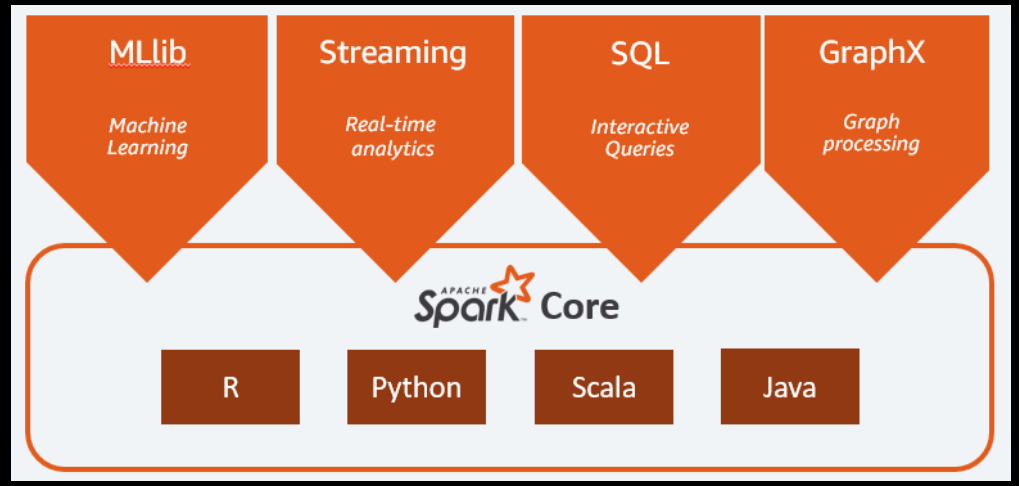


… y llega Spark

La gracia de Spark es que hace que programar sobre un cluster se parezca más a programar sobre un único ordenador (casi te puedes olvidar de que estás en un cluster, casi…)

¿Qué es Apache Spark? Google Cloud

Apache Spark es un motor unificado de analíticas para procesar datos a gran escala que integra módulos para SQL, streaming, aprendizaje automático y procesamiento de grafos.



Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.

aws

Spark

Nociones básicas

Transformaciones y acciones

Una transformación es cualquier modificación que hagamos sobre los datos, como por ejemplo, un filtrado. Mientras que en una acción necesitamos ejecutar todas las transformaciones ya que estamos pidiendo un resultado, como un count o un show. **Es el trigger de la ejecución.**

Lazy evaluation

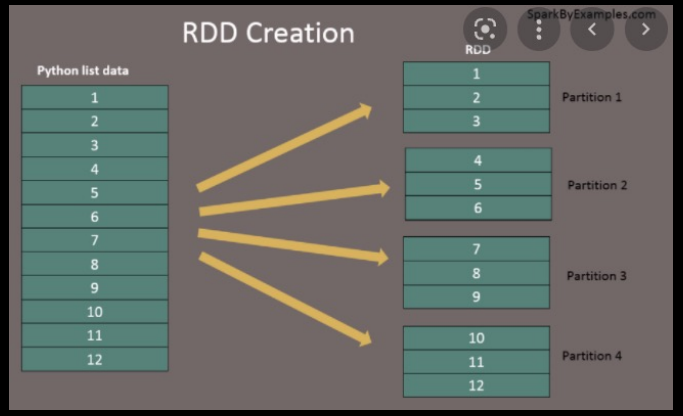
Spark no ejecuta todas las operaciones hasta que no se ve en la necesidad de mostrar datos con una acción. Evalúa todas las operaciones de la ejecución y las ordena como crea más conveniente para que la ejecución sea óptima

Para más: http://ashkrit.blogspot.com/2018/09/anatomy-of-apache-spark-job.html

Spark

Nociones básicas

RDD = Estructura básica de datos (particionada y que se puede mantener en memoria)



Driver Program = Básicamente mi programa (del que saldrán “hijos” a ejecutarse en los Workers/Executors y que luego concentrará y agrupará las salidas de estos)

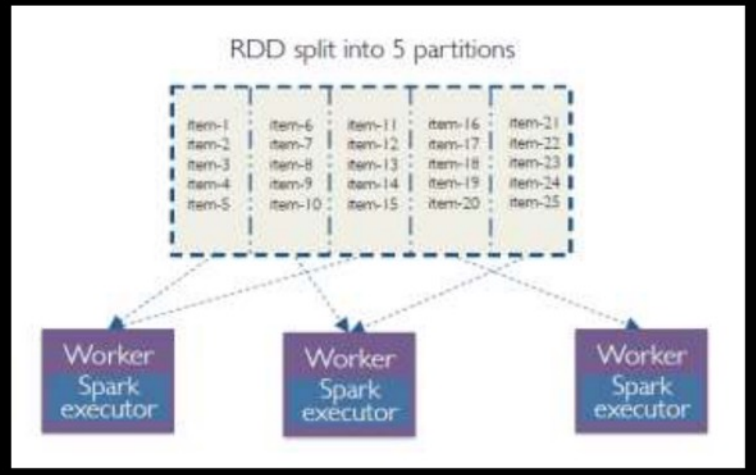
- At a high level, every Spark application consists of a *driver program* that runs the user’s main function and executes various *parallel operations* on a cluster.
- The main abstraction Spark provides is a *resilient distributed dataset* (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel
- RDDs are created by starting with a file in the Hadoop file system (or any other Hadoop-supported file system), or an existing Scala collection in the driver program, and transforming it.
- Users may also ask Spark to *persist* an RDD in memory, allowing it to be reused efficiently across parallel operations. Finally, RDDs automatically recover from node failures.

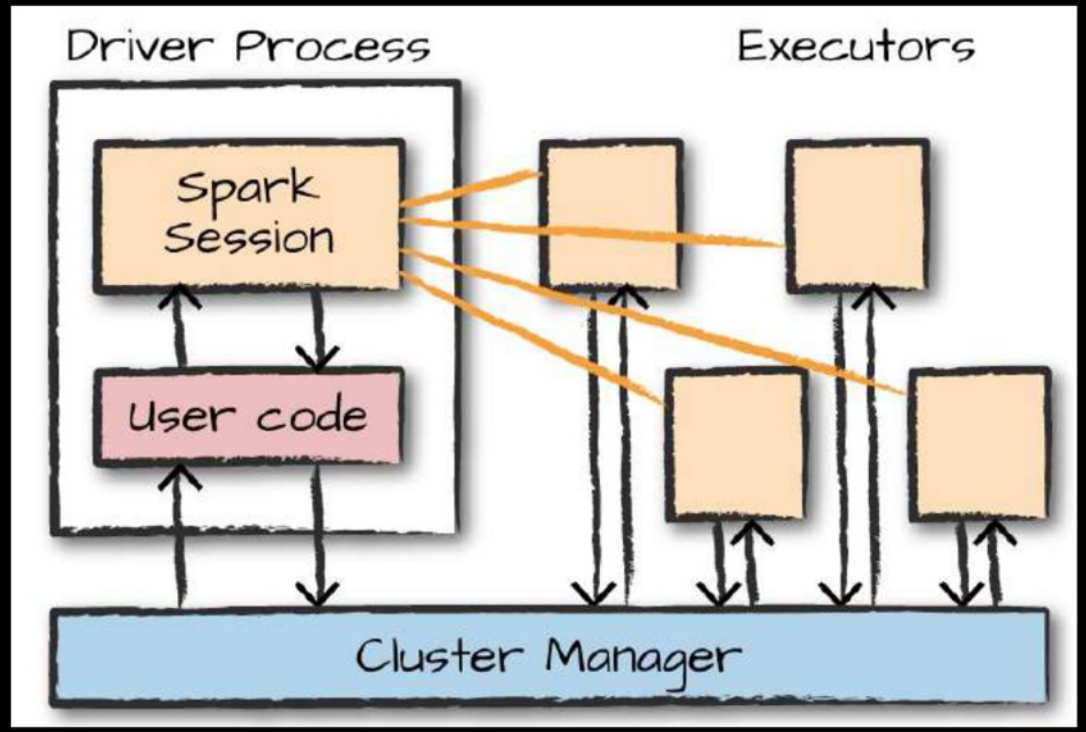
Spark

Nociones básicas

Partición
Una partición es una serie de filas de un DF que se almacenan en una maquina fisica, dentro de un cluster, por ejemplo, partición por fecha.

Trabajamos a alto nivel, no con las particiones

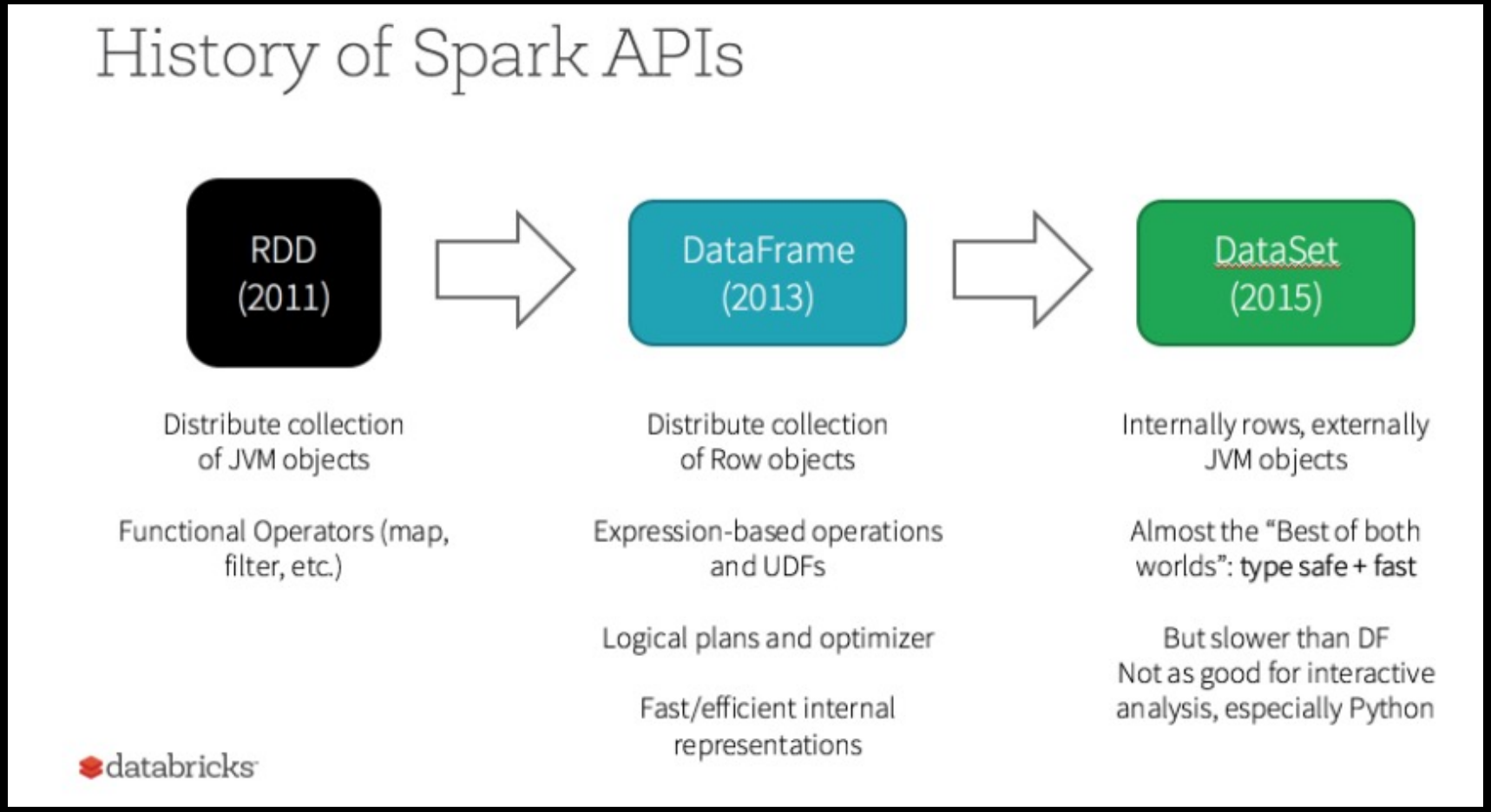




Para más: https://luminousmen.com/post/spark-partitions

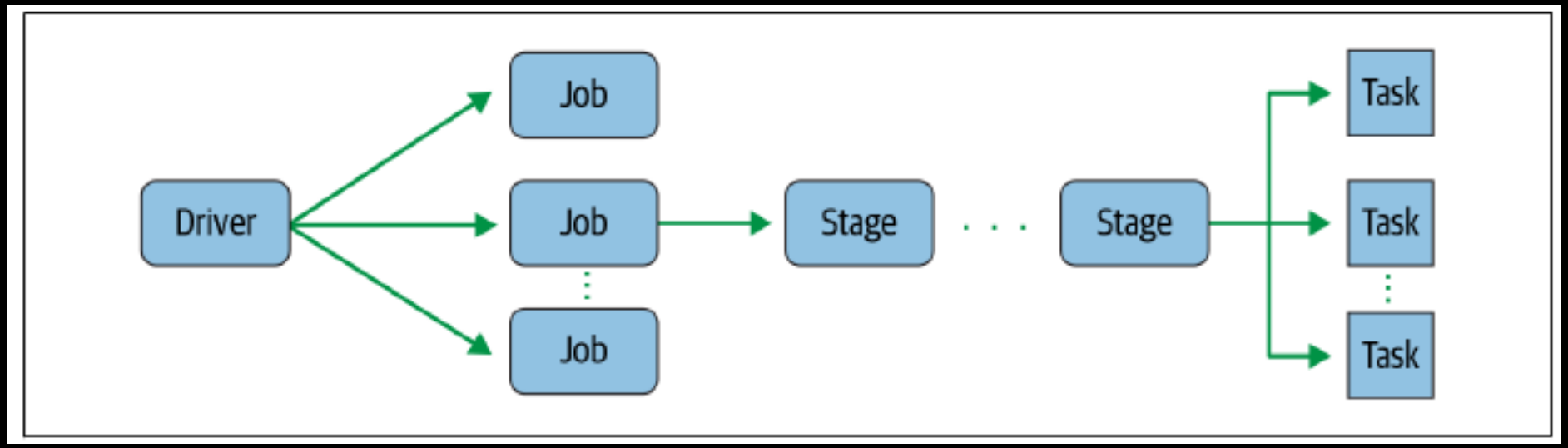
Spark

Estructuras

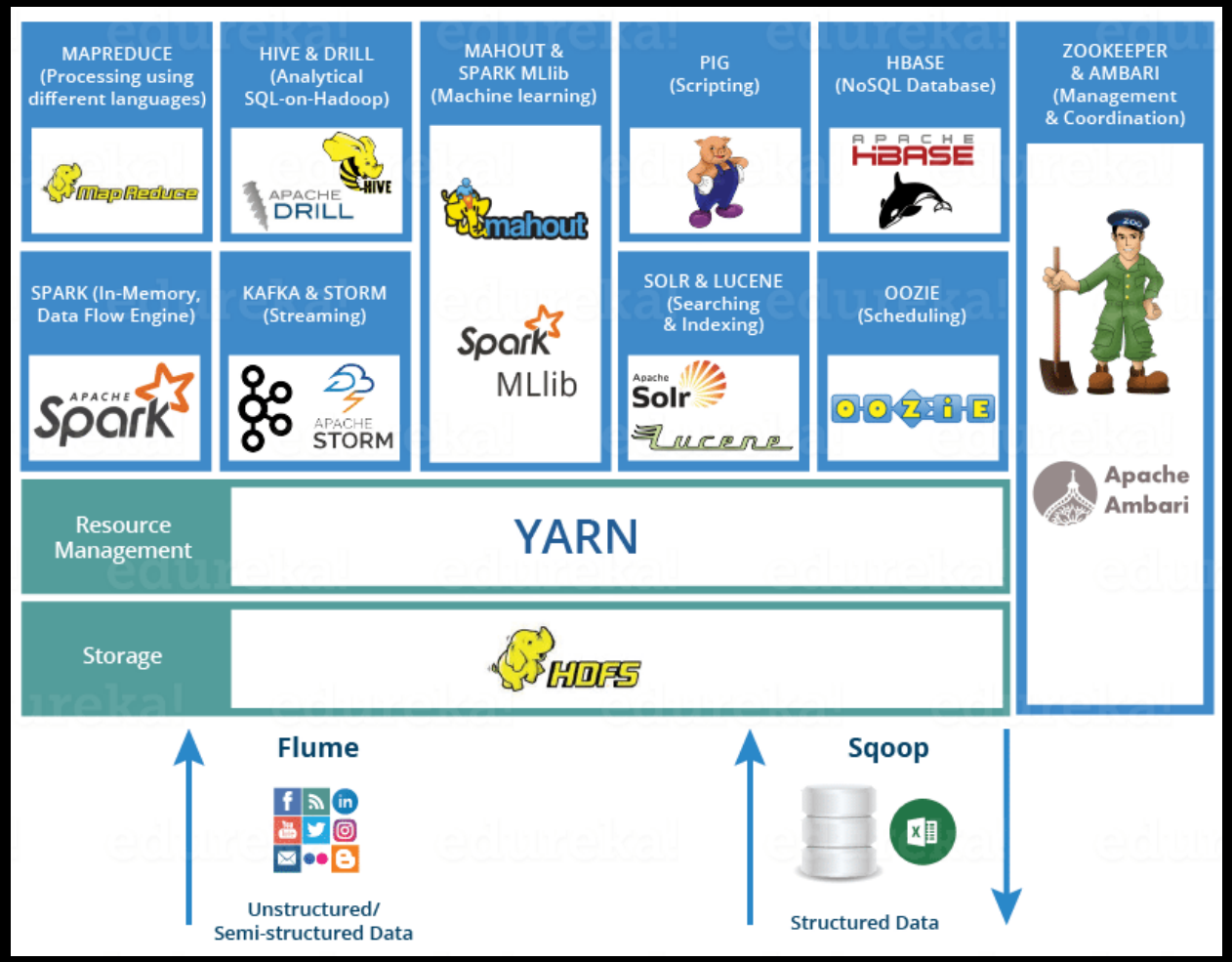


Spark

Orquestación



Hadoop Ecosystem



Los destacados (para mí :-)):

- HDFS
- Yarn
- Spark (Spark Mlib)
- Hive
- Kafka

Para más:
https://www.edureka.co/blog/hadoop-ecosystem

Spark

Qué vamos a hacer

- Instalar Pyspark
- Comandos básicos Pyspark Dataframes y equivalencias con Pandas
- Usar spark mlib para entrenar un modelo

- Si da tiempo:
  - Databricks
  - RDDs
  - UDFs
  - Modelado un poco más complejo