Big Data Aplicado

Josep Garcia

IES EDUARDO PRIMO MARQUÉS

¿QUÉ ES BIG DATA?

Una gran cantidad de información que tengo que procesar para extraer algún tipo de dato y que por métodos tradicionales no se puede o es complicado de realizar.

HADOOP

Hadoop es casi un sinónimo de “Big Data”.

Es un entorno distribuido de:

- Datos
- Procesos

Es un sistema tipo cluster que se puede escalar horizontalmente con hardware relativamente “barato” (commodity hardware).

* Barato para empresas: 64/128Gb RAM, varios cores…

HADOOP

Hadoop implementa **procesamiento en paralelo** a través de nodos de datos en un **sistema de ficheros distribuidos**.

**“Si una máquina lo hace en 10 minutos,
con 10 máquinas se hará en 1 minuto”**

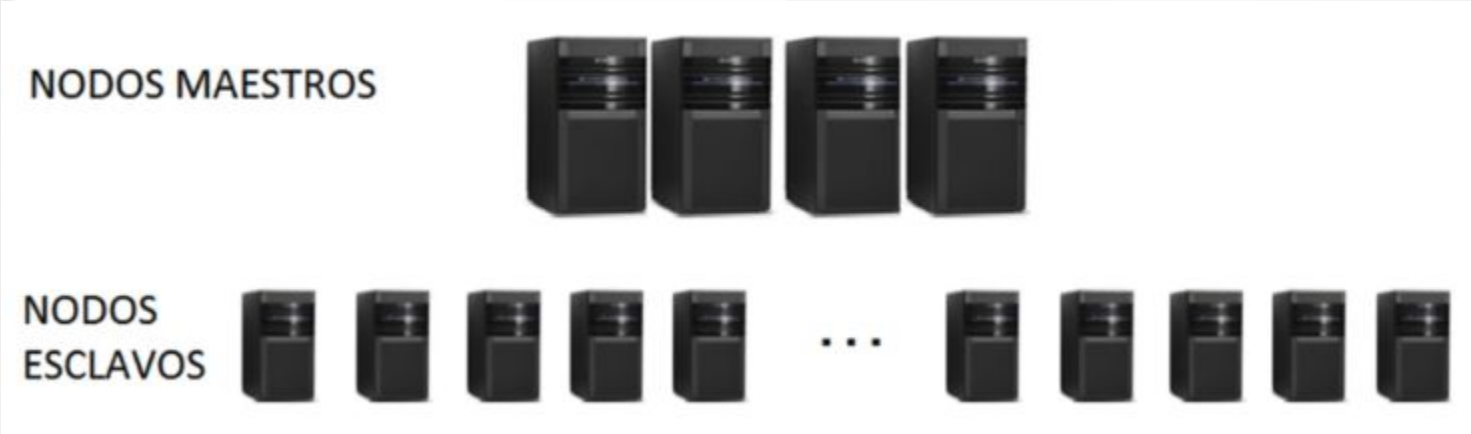
(No es exáctamente así pero es similar)

Divide y vencerás: Dividimos nuestros procesos en varias máquinas para que cada una se encargue de uno.

HADOOP

Nodos maestros: Gobiernan a los nodos esclavos.

Nodos esclavos: Realizan el procesamiento de la información.



HADOOP

- Uno de los puntos fuertes es que Hadoop está diseñado para ejecutarse en servidores de bajo coste y que dispone de una **gran tolerancia a fallos**.

- Los fallos **se tratan como una regla** y no como una excepción. Presupone que siempre va a haber una máquina que se va a estropear (siempre hay disponibilidad en otro nodo para realizar el proceso).

- Si no se dispone de infraestructura física, se puede utilizar la nube.

HADOOP

Principales componentes

Hadoop Common	Librerías comunes
MapReduce (Yarn)	Procesos
HDFS (Hadoop Distributed File System)	Sistema de archivos

HADOOP

Está diseñado para escalar desde unos pocos nodos a miles de máquinas, cada una de ellas ofreciendo la lógica de negocio y el almacenamiento a nivel local.

HADOOP

El core de Hadoop está formado por dos componentes básicos.

DATOS

PROCESOS

HADOOP DATOS

HDFS

DATOS

- Sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de HW sin perder datos.
- Si un nodo falla, el cluster puede continuar trabajando sin perder datos o interrumpir el trabajo, sencillamente redistribuye el trabajo entre los nodos restantes
  del cluster.

HADOOP PROCESOS

PROCESOS

- Formas de procesamiento:
  - Map Reduce V1
  - Map Reduce V2 - Yarn
- De forma general son algoritmos de procesamiento de datos que implementan procesos en paralelo.
- Es decir, distribuye las tareas a través de los nodos de un cluster.

HADOOP ECOSISTEMA

- **HBase**: orientada a valores/claves.
- **Hive**: permite lanzar comandos sql sobre hadoop.
- **Pig**: Lenguaje scripting de alto nivel.
- **Zookeeper**: mantener información de configuración, gestión de nombre, y para facilitar la sincronización de servicios.
- **Sqoop**: Herramienta diseñada para transferir datos masivos desde Hadoop a otros entornos como Bases de Datos relacionales.
- **Spark**: Motor muy eficiente de procesamiento de datos a gran escala (puede trabajar independientemente de Hadoop). Implementa MapReduce en tiempo real (al contrario de hadoop).

DISTRIBUCIONES

Hay distintas empresas que ofrecen soluciones “empaquetadas” para Hadoop.

- Cloudera.
- Hortonworks -> Cloudera.
- IBM Open Platform
- …