BIG DATA

Introducción a Hadoop



- Hadoop es casi sinónimo del término "Big Data".
- Es un entorno distribuido de:
 - Datos
 - Procesos
- Hadoop es un entorno de tipo High Performance Super Computer que se puede escalar horizontalmente con hardware relativamente barato "commodity hardware"

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

NODOS MAESTROS



NODOS























- Uno de los puntos fuertes de Hadoop es que está diseñado para ejecutarse en servidores de bajo coste y que dispone de una gran tolerencia a fallos
- De hecho, en Hadoop, los fallos de hardware se tratan como una regla y no como una excepción.
- Por tanto se puede montar un cluster de servidores X86 a un precio muy razonable si lo comparamos con grandes servidores
- Y si no, nos queda la nube



- Hadoop en un entorno que suministra librerías open source para la computación distribuida usando varios componentes, Los principales son:
 - Hadoop Common
 - MapReduce
 - Hadoop Distributed File System (HDFS).
- Está diseñado para escalar desde unos pocos nodos a miles de máquinas, cada uno de ellas ofreciendo la lógica de negocio y el almacenamiento a nivel local.

Versiones

- Un poco enrevesadas al mantener varias líneas de trabajo
- En el momento de escribir este manual
 - 1.2.X current stable version, 1.2 release
 - 2.7.1 stable 2.x version
 - 0.23.X similar to 2.X.X but missing NN HA.

 El "core" de Hadoop está formado por dos componentes básicos:

DATOS

PROCESAMIENTO

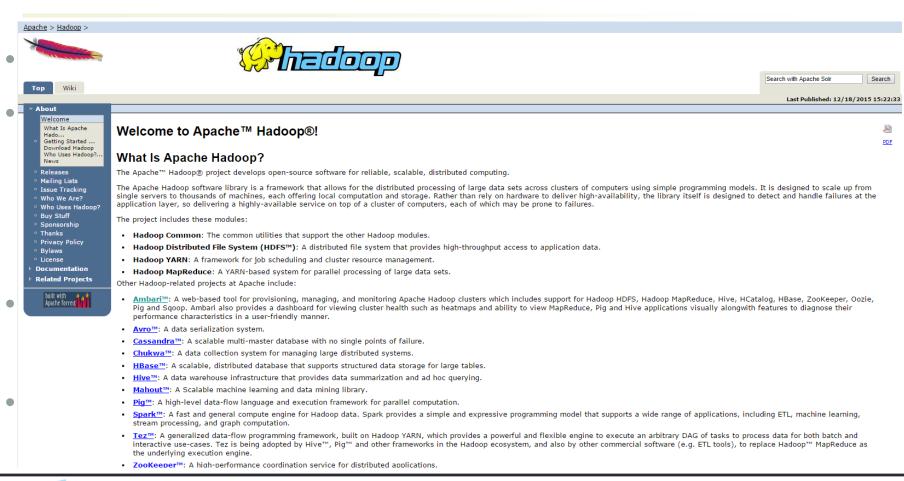
HDFS

- HDFS es un sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de hardware 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.



- Procesos
- En la actualidad existen dos formas de procesamiento distintos
 - Map Reduce V1
 - Map Reduce V2- YARN
- De forma general son algoritmos de procesamientos de datos que implementan procesos en paralelo
- Es decir distribuye las tareas a través de los nodos de un cluster







- Otros proyectos implicados en Hadoop
 - HBase: Una base de datos orientada a valores/claves que se ejecuta sobre HDFS
 - Hive : sistema de funciones que soportan agregación de datos y coinsultas ad hoc sobre MapReduce
 - Pig: Lenguaje de alto nivel para gestionar flujos de datos y ejecución de aplicaciones sobre Hadoop
 - Mahout: entorno de aprendizaje de máquinas implementado en hadoop
 - Zookeeper : servicio centralizado para 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

