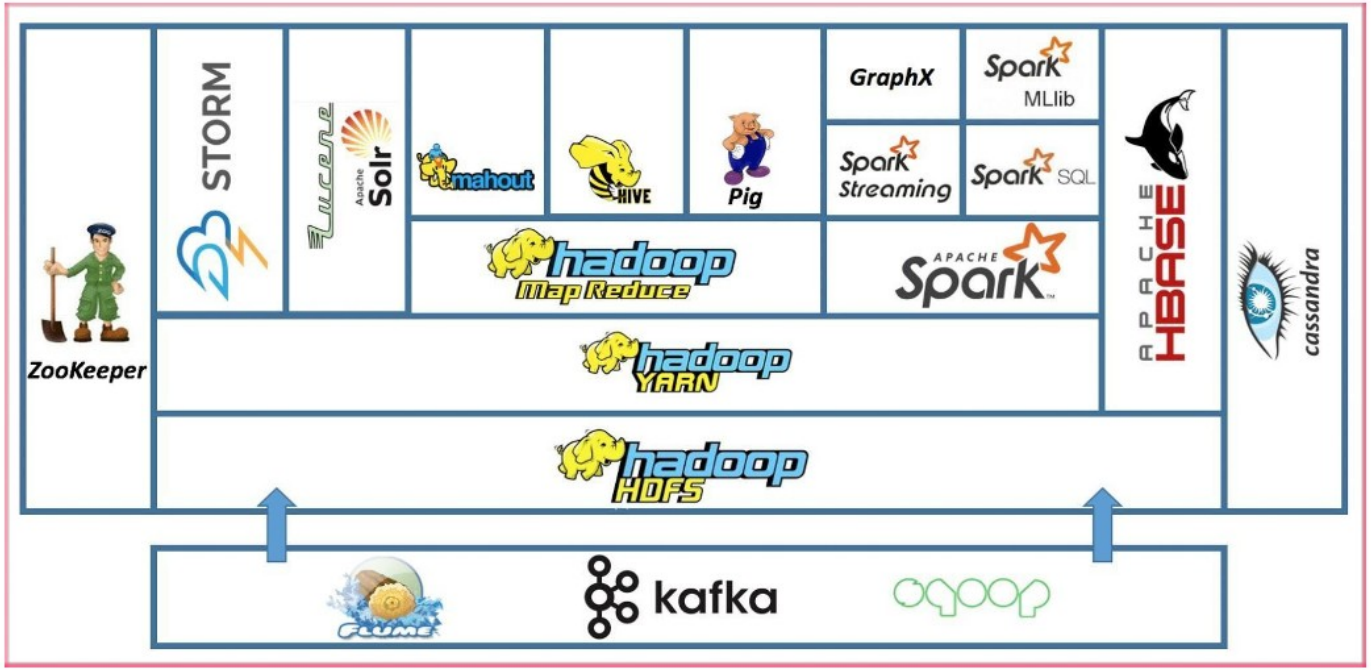
David Granados Zafra

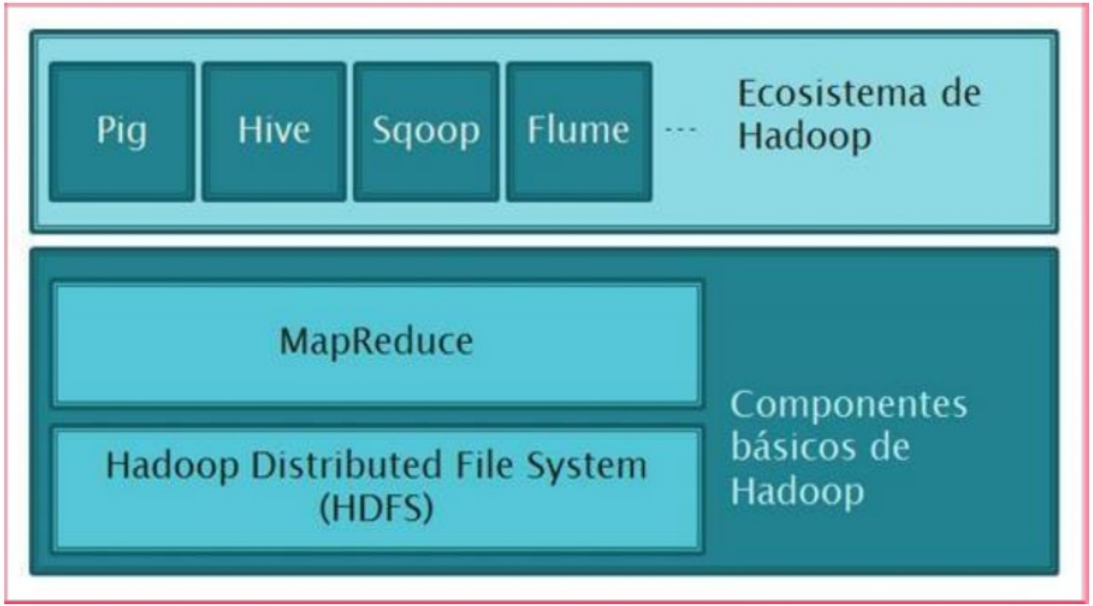
Conceptos Básicos

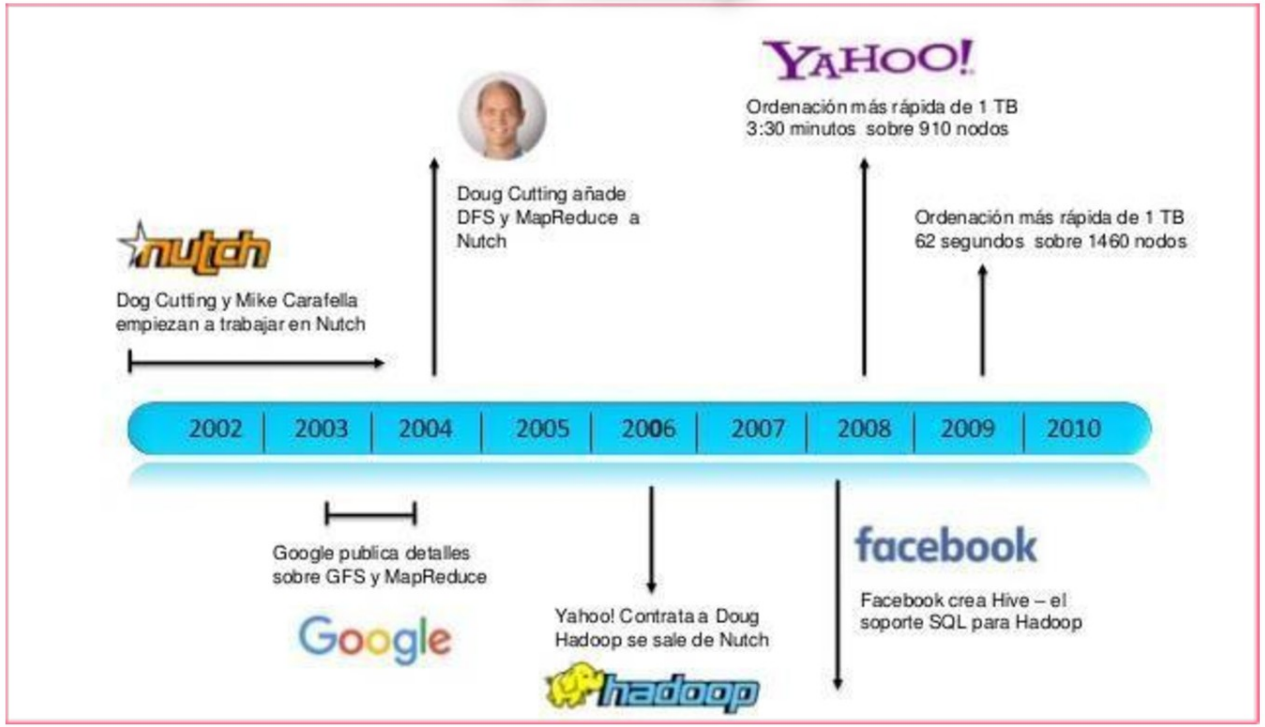
- Apache Hadoop es un conjunto de tecnologías de código abierto.
- Se emplea para el almacenamiento y procesamiento distribuido de datos a gran escala
-

Componentes



Componentes




YAHOO!
Ordenación más rápida de 1 TB
3:30 minutos sobre 910 nodos
Doug Cutting añade
DFS y MapReduce a
Nutch
Ordenación más rápida de 1 TB
62 segundos sobre 1460 nodos
nutch
Dog Cutting y Mike Carafella
empiezan a trabajar en Nutch
2002
2003
2004
2005
2006
2007
2008
2009
2010
Google publica detalles
sobre GFS y MapReduce
Google
Yahoo! Contrata a Doug
Hadoop se sale de Nutch
hadoop
facebook
Facebook crea Hive – el
soporte SQL para Hadoop

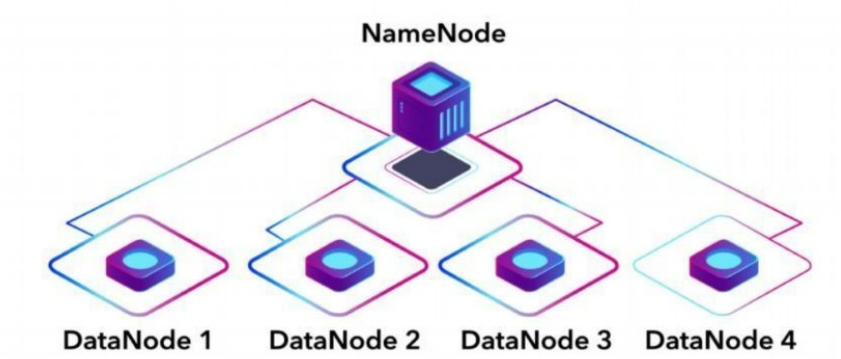
Características

- Procesamiento Distribuido
  - Utiliza HDFS (Hadoop Distributed File System) para almacenar grandes cantidades de datos
  - HDFS divide los datos en bloques y los distribuye en múltiples nodos en un cluster para proporcionar redundancia y tolerancia a fallos.

Cluster

- Un cluster lo podríamos definir como:
  - Conjunto de computadoras que trabajan juntas para resolver una tarea.
  - Suelen estar formadas por equipo de “Bajo coste
- Tienen las siguientes características:
  - Escalabilidad
  - Confiabilidad
  - Eficiencia

Cluster



Procesamiento Distribuido

- La capa de procesamiento en Hadoop se realiza mediante el framework de procesamiento de Map Reduce.
- MapReduce divide las tareas de procesamiento en etapas de “map” y “reduce” (mapeo y reducción)
- Las distribuye en varios nodos del cluster lo cual permite el procesamiento en paralelo.

Escalabilidad

- Hadoop está diseñado para escalar de manera horizontal
- Se basa en usar dispositivos de “bajo coste”
- Podemos ir añadiendo más dispositivos según demanda.
- Podemos paralelizar de forma automática así como conseguir una gran tolerancia a fallos.
-

Java

- Tradicionalmente se programa en Java
- En la actualidad también podemos hacer uso de otros lenguajes como Python y R

Ecosistema Hadoop

- Hadoop es un ecosistema, lo que quiere decir que es una colección de componentes.
-

Modelo de programación declarativo.

- MapReduce es el modelo original de programación para Hadoop
- Con el tiempo se necesitaron alternativas más eficientes, con lo que surgieron otros lenguajes como Hive o Pig
-

Apache Yarn

- YARN (Yet Another Resource Negotiator)
- Es un administrador de recursos que permite la gestión eficiente de recursos en el cluster
- Esto permite usar alternativas a Map Reduce, como Spark
-