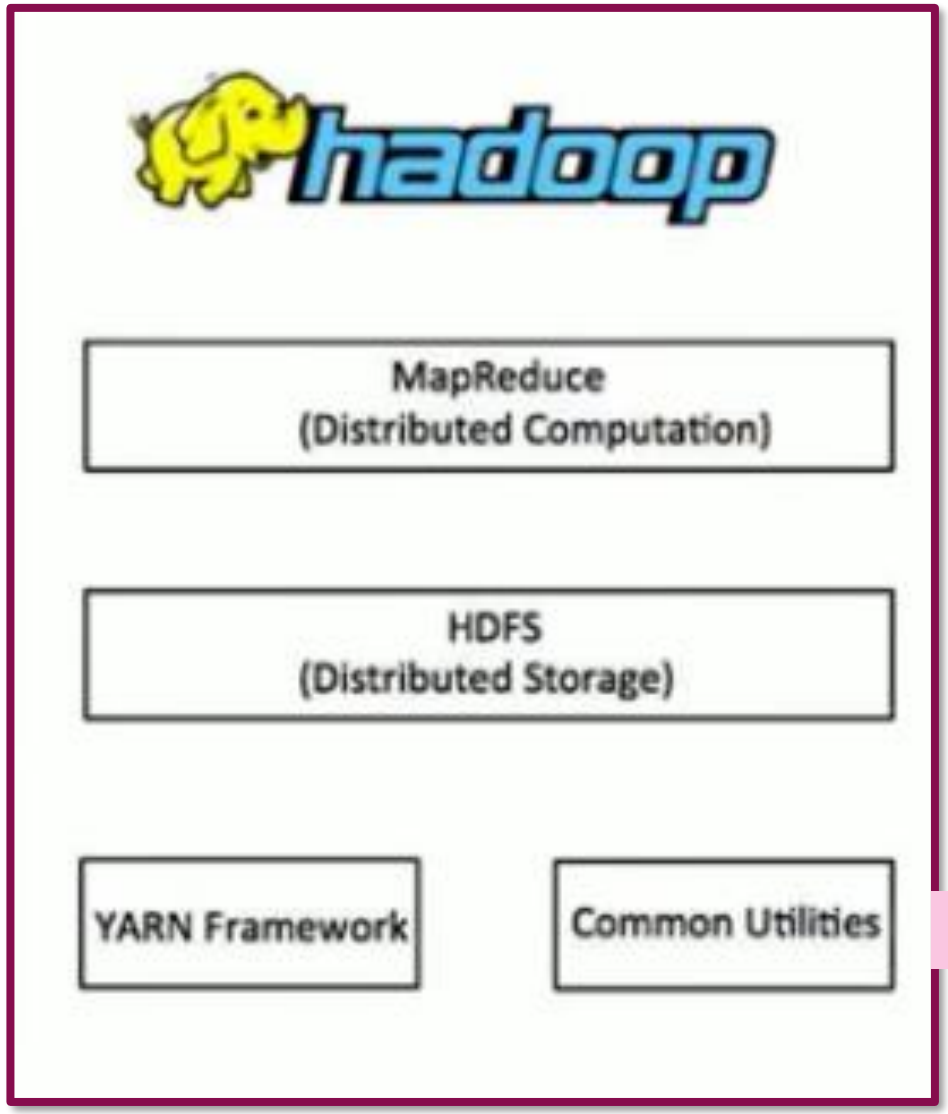
Hadoop es un Framework que permite el procesamiento distribuido de grandes cantidades de datos usando modelos de programación simple sobre un cluster de máquinas

Características Básicas

- Procesamiento Distribuido
- Eficiencia
- Economía
- Escalable
- Tolerante a fallos
- Open source

Arquitectura Básica de Hadoop

- 4 módulos principales
- Crear procesos Map : Grupos de datos
- Crear procesos Reduce: Cálculos sobre esos datos



Paralelizar los datos

Sistema de archivos instalado en c/máquinas. Soporte del MapReduce

Hard y librerías

Gestor de recursos (distribuídos en las distintas máquinas)

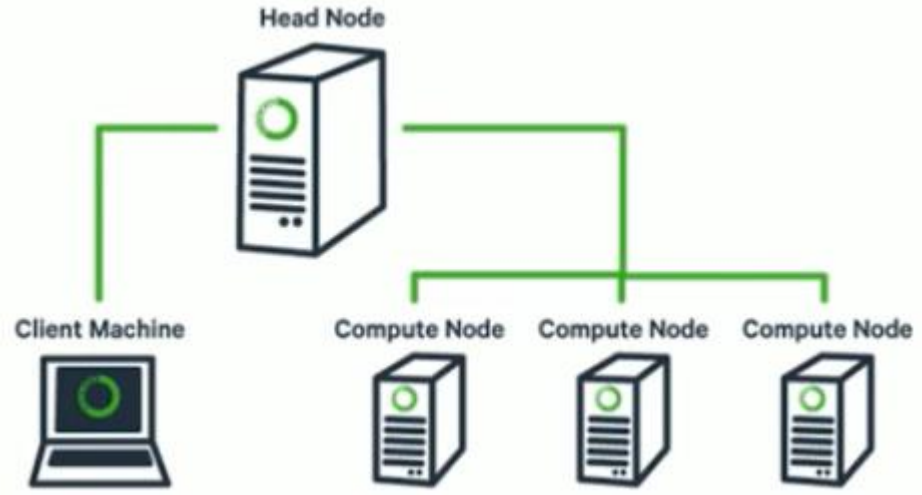
Configuración – Hadoop Cluster

La configuración habitual de Hadoop es tenerlo en un cluster de máquinas.:

1 a n- máquinas maestras

M- máquinas esclavas x cada maestra

La maestra gestiona las tareas y envía a las esclavas para el procesamiento de los datos. Luego estas últimas devuelven el resultado a la Maestra



Opciones para trabajar con Hadoop

Microsoft Azure
https://
azure.microsoft.com/
es-es/services/
hdinsight/

Servicio de Microsoft

Permite tener las máquinas en la nube. Se paga según la cantidad y características

Hortonworks
https://
es.hortonworks.com/
downloads/#data-
platform

Crea maquinas virtuales

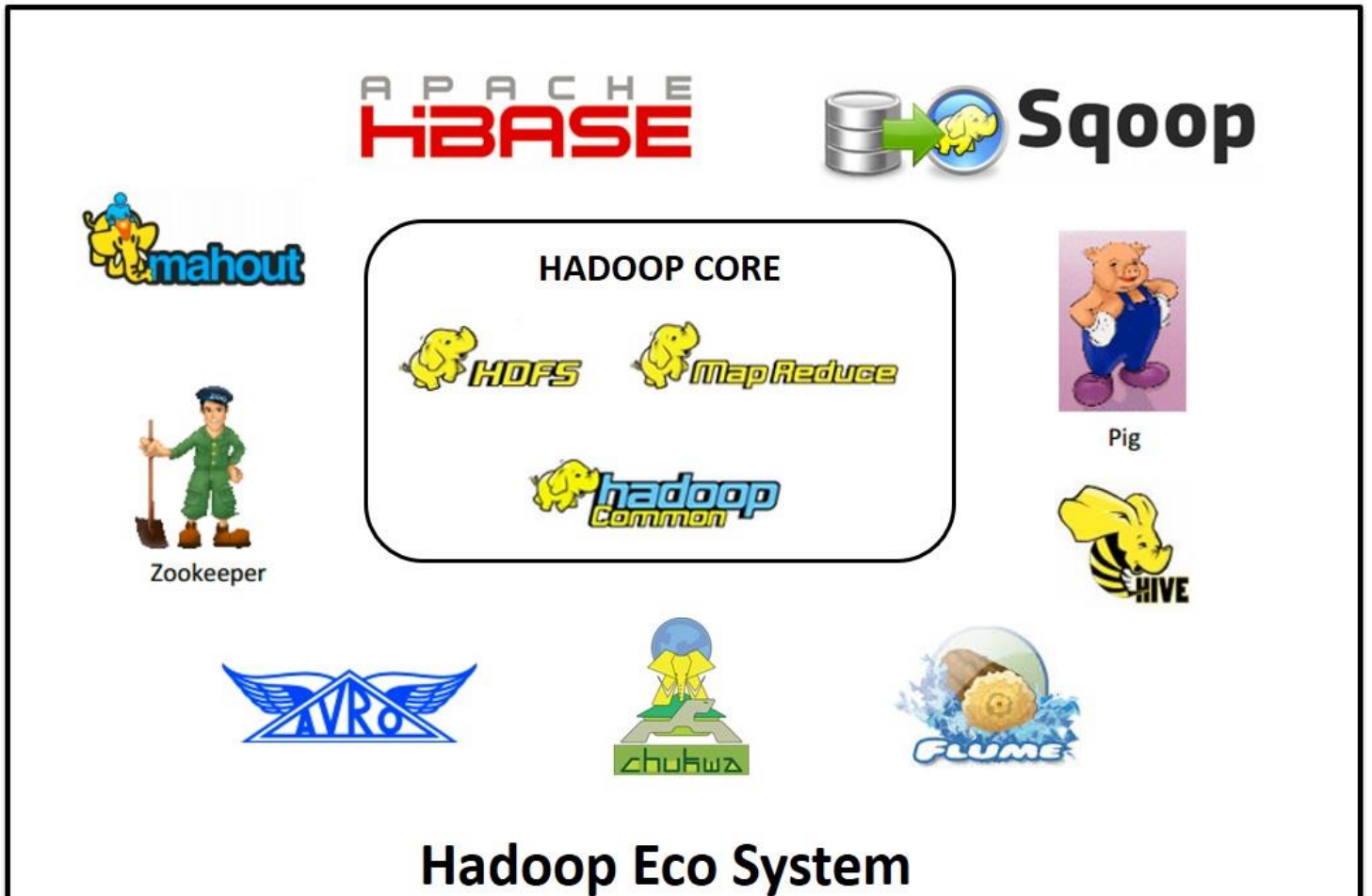
Cloudera
https://
www.cloudera.com/
downloads.html

Descarga maquina virtual que se abre con un cliente de virtualización y el cloudera manager que administra todo el cluster

Se fusionaron hace un año mejorando las prestaciones y servicios,

Las empresas hoy prefieren soluciones en la nube

Ecosistema o Zoo Hadoop



Conveniencia de usar o no Hadoop

MALAS PRÁCTICAS

Es conveniente usar Hadoop cuando:

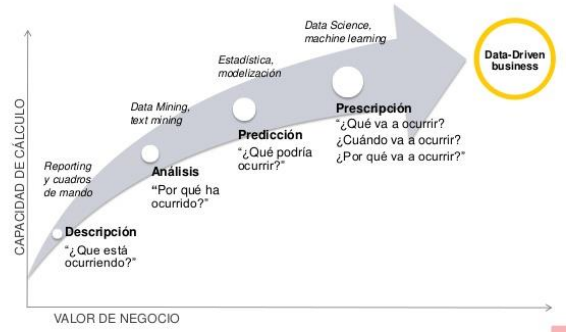
Se deben procesar grandes cantidades de datos

- Cantidades de datos muy grandes, de diferentes fuentes y tipos
- Terabyte, petabyte, hexabyte, zetabytes

Crecimiento exponencial a futuro o decrecimiento temporal

- Constante crecimiento o decrecimiento
- Facil Escalado: añadiendo nuevos nodos al cluster de máauinas

Las plataformas escalables en la nube permiten procesar BIG DATA más rápidamente que con soluciones tradicionales



Gran variedad de datos

Tipos diferentes

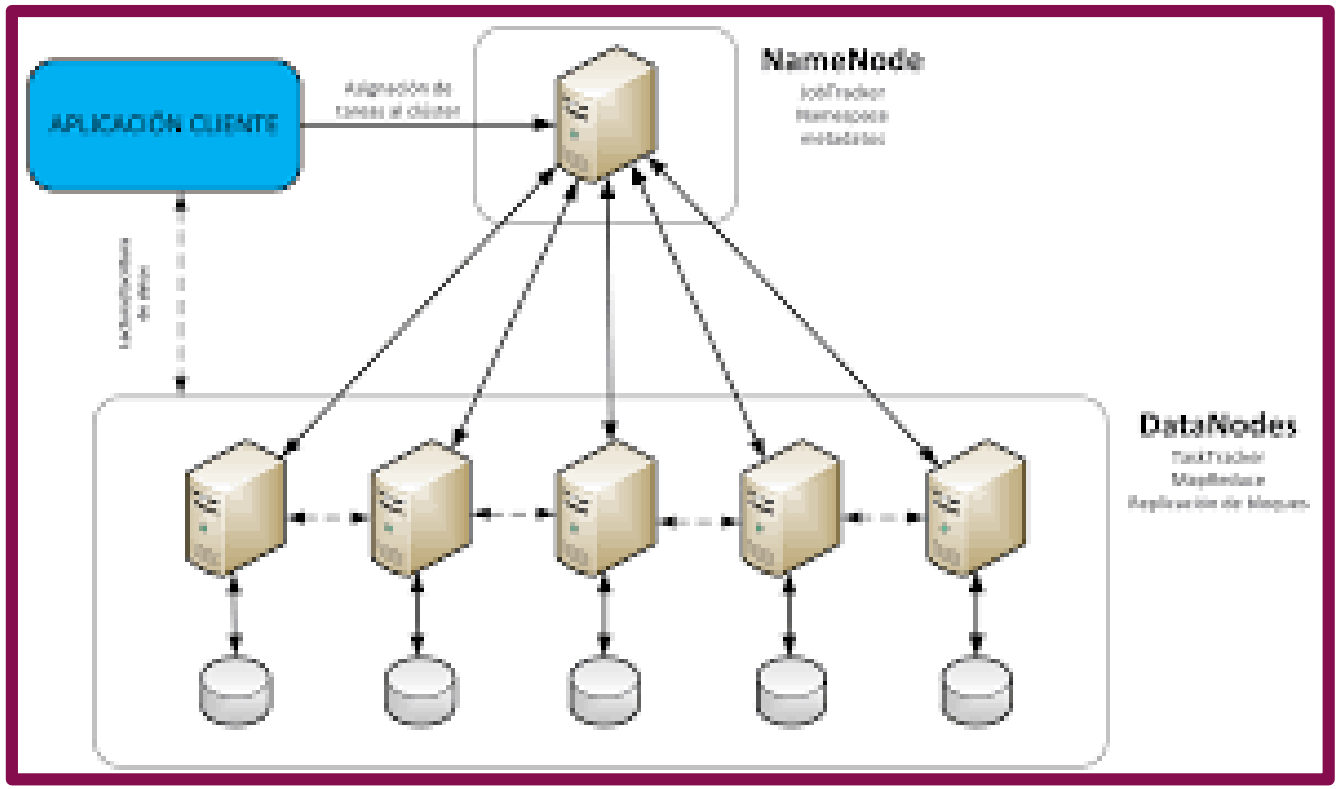
Fuentes distintas

Imagen texto sql sensores

Es conveniente usar Hadoop cuando:

Se necesita paralelizar los procesos

- Cantidades datos hacen necesaria la paralelizar los datos.
- Apache Hadoop tiene grandes ficheros divididos en grupos que se trataran paralelamente ganando en velocidad.



No es conveniente usar Hadoop cuando:

Se deben analizar datos en tiempo real

- Los Procesos Map Reduce llevan su tiempo
- Horas días semanas porque trabaja en disco
- Se debería usar **Apache Spark** que trabaja en memoria y tiene una latencia mucho más baja

Para sistemas con Bases de datos relacionales

- Modelo de datos muy complejo
- Join, unión, filter, etc
- Se debería usar Apache Hive, Permite lanzar consultas SQL sobre el HSFS.
- Así tendremos los ficheros alojados en el HDFS al que le crearemos con el APACHE HIVE una peueña estructura de esos datos para ser consultados con el lenguaje HivsQl similar al Sql.

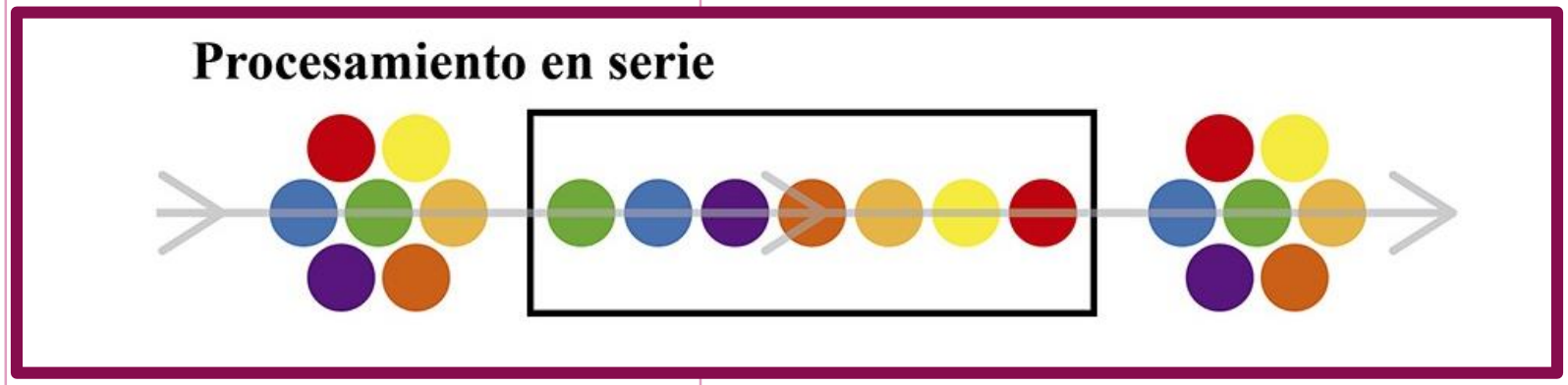
Cuando queremos modificar nuestros datos

- HDFS : Write once ready many
- Una vez escrito solo podemos borrar o agregar al final

No es conveniente usar Hadoop cuando:

No se pueda paralelizar el trabajo

- Si nuestros datos no se pueden paralelizar perdemos la principal característica de Hadoop
- Si los das datos deben ser tratados de manera secuencial



Malas Prácticas

Mejorar el Diseño y análisis

- Cientos de ficheros pequeños en HDFS, cuando está pensado para alojar ficheros muy grandes (tera, peta) almacenándolos en bloques de 128 MB
- Cientos de procesos Map con poca duración (es donde se paraleliza, se distribuyen los datos) Demasiados procesos Map se paralelizó mas de la cuenta, es poco útil.
- Pocos reduce en el procesamiento de grandes datos, pocos Reduce para ficheros muy grandes. Si tenemos ficheros de 2 gigas y tenemos 2 Reduce, se debe paralelizar más.
- Muchas salidas y pequeñas rn los Reduce: Solo sale lo útil. Menos ficheros de más tamaño

Aplicaciones prácticas: casos de éxito

Facebook

- **Facebook termina acumulando cantidades masivas de datos, probablemente más que su organización típica, especialmente teniendo en cuenta la cantidad de medios que consume. Uno de los desafíos que ha enfrentado Facebook desde los primeros días es desarrollar una forma escalable de almacenar y procesar todos estos bytes, ya que el uso de estos datos históricos es una parte muy importante de cómo pueden mejorar la experiencia del usuario en Facebook.**
- **La rápida adopción de Hadoop en Facebook se ha visto favorecida por un par de decisiones clave. En primer lugar, los desarrolladores pueden escribir programas de reducción de mapas en el idioma que elijan. En segundo lugar, Facebook ha adoptado SQL como un paradigma familiar para abordar y operar en grandes conjuntos de datos. La mayoría de los datos almacenados en el sistema de archivos de Hadoop se publican como tablas.**

ebay -- amazon

- **Comenzron con la implementación de hadoop y evolucionar hasta convertise en una de la clou más exitosas**

Oracle

- **Los clientes de Oracle se enfrentan a un problema de big data, y Hadoop se ha convertido en la respuesta inicial, aunque Oracle es reacio a admitirlo.**

- **Hoy no solo utiliza Hadoop para algunas soluciones sino que evolucionó hasta su propia nube**

Yahoo

- **El proyecto Hadoop es una parte integral de Yahoo! infraestructura en la nube, y es el corazón de muchos de los procesos comerciales importantes de Yahoo !. Yahoo gestiona los clústeres de Hadoop más grandes del mundo, trabaja con instituciones académicas y otras grandes corporaciones en investigación avanzada de computación en la nube y sus ingenieros son los principales participantes de la comunidad de Hadoop.**

- **Hadoop con seguridad:**
- **Impide el acceso no autorizado a los datos en clústeres de Hadoop**
- **Autentica a los usuarios que comparten datos confidenciales de negocios**
- **Reduce los costos operativos al consolidar los clústeres de Hadoop**
- **Coloca datos para nuevas clases de aplicaciones**