
Unidad 1. Introducción a Apache Hadoop

— BIG DATA APLICADO-CEIABD —

Google File System (octubre 2003), almacena petabytes a bajo coste utilizando un modelo de almacenamiento distribuido.

En 2004 publican como resolver el procesamiento sobre conjuntos de datos voluminosos mediante **MapReduce**.

Surgía por esa época **Apache Nutch** como motor de búsqueda. Poco después (2006) se llamaría **Hadoop**.

Hadoop es una plataforma que permite almacenar y procesar grandes volúmenes de datos.

Apache Hadoop es una plataforma opensource que ofrece la capacidad de almacenar y procesar, a “bajo” coste, grandes volúmenes de datos, sin importar su estructura, en un entorno distribuido, escalable y tolerante a fallos, basado en la utilización de hardware commodity y en un paradigma acercamiento del procesamiento a los datos.

- **El coste es más bajo que otros sistemas tradicionales de gestión de datos.**
- **Hadoop es escalable tanto en almacenamiento como en procesamiento.**
- **Hadoop no tiene unos requerimientos de hardware muy específicos.**

Los componentes core principales de Hadoop son HDFS y YARN:

- HDFS**: un sistema de ficheros (capa de almacenamiento) que almacena los datos en una estructura basada en espacios de nombres (directorios, subdirectorios, etc.).
- YARN**: un gestor de recursos (capa de procesamiento) que permite ejecutar aplicaciones sobre los datos almacenados en HDFS.
- MapReduce**: un sistema de procesamiento masivo de datos que se puede utilizar directamente, programando sobre su API, o indirectamente, con aplicaciones que lo utilizan de forma transparente.

Sin embargo, normalmente se identifica el nombre Hadoop con todo el ecosistema de componentes independientes que suelen incluirse para dotar a Hadoop de funcionalidades necesarias en proyectos Big Data empresariales, como puede ser la ingesta de información, el acceso a datos con lenguajes estándar, o las capacidades de administración y monitorización.

Los principales componentes o proyectos asociados al ecosistema Hadoop son los siguientes:

Nombre	Descripción
Apache Hive	Permite acceder a ficheros de datos estructurados o semiestructurados que están en HDFS como si fueran una tabla de una base de datos relacional, utilizando un lenguaje similar a SQL.
Apache Pig	Utilidad para definir flujos de datos de transformación o consulta mediante un lenguaje de scripting.
Apache HBase	Base de datos NoSQL de tipo columnar que permite el acceso aleatorio, atómico y con operaciones de edición de datos.
Apache Flume	Componente para ingestar streams de datos procedentes de sistemas real-time en Hadoop.

Apache Sqoop	Componente para importar o exportar datos estructurados desde bases de datos relacionales a Hadoop y viceversa.
Apache Oozie	Herramienta que permite definir flujos de trabajo en Hadoop así como su orquestación y planificación.
Apache ZooKeeper	Herramienta técnica que permite sincronizar el estado de los diferentes servicios distribuidos de Hadoop.
Apache Storm	Sistema de procesamiento real-time de eventos con baja latencia.
Apache Spark	Aunque habitualmente no se asocia al ecosistema Hadoop, Apache Spark ha sido el mejor complemento de Hadoop en los últimos años. Apache Spark es un motor de procesamiento masivo de datos muy eficiente que ofrece funcionalidades para ingeniería de datos, machine learning, grafos, etc.

Apache Kafka	Sistema de mensajería que permite recoger eventos en tiempo real así como su procesamiento.
Apache Atlas	Herramienta de gobierno de datos de Hadoop.
Apache Accumulo	Base de datos NoSQL que ofrece funcionalidades de acceso aleatorio y atómico.
Apache Mahout	Conjunto de librerías para desarrollo y ejecución de modelos de machine learning utilizando las capacidades de computación de Hadoop.
Apache Phoenix	Capa que permite acceder a los datos de HBase mediante interfaz SQL .
Apache Zeppelin	Aplicación web de <u>notebooks</u> que permite a los Data Scientists realizar análisis y evaluar código de forma sencilla, así como la colaboración entre equipos.
Apache Impala	Herramienta con funcionalidad similar a Hive (tratamiento de los datos de HDFS mediante SQL) pero con un rendimiento elevado (tiempos de respuesta menores).

Los más utilizados son: **Apache Spark, Apache Hive y Apache Kafka**, además de los componentes core: **HDFS y YARN**.

Para solventar las dos dificultades de instalación, surgen las **distribuciones comerciales de Hadoop**, que contienen en un único paquete la mayor parte de componentes del ecosistema, resolviendo dependencias, añadiendo incluso utilidades, e incorporando la posibilidad de contratar soporte empresarial 24x7. Es decir, una distribución comercial ofrece:

- Un "instalador" de toda la plataforma, simplificando enormemente el proceso de instalación y despliegue de la plataforma.
- Un servicio de soporte 24x7 para resolver todas las incidencias que puedan aparecer en la plataforma en producción.
- Documentación más completa que la que se puede encontrar en los proyectos Apache.

Las principales distribuciones que aparecieron son:

- Cloudera**: fue la primera distribución en salir al mercado (2009) y la que ha tenido un mayor número de clientes. Utiliza la mayor parte de componentes de Apache, en algún aso realizando algunas modificaciones, y añade algún componente propietario (Cloudera Manager, Cloudera Navigator, etc.).
- Hortonworks**: surgió en 2012 y es una distribución que contiene, sin ninguna modificación, los componentes originales de Apache. Se fusionó con Cloudera en 2018.
- MAPR**: rehízo la mayor parte de componentes utilizando los mismos interfaces pero reimplementando el core para ofrecer un mayor rendimiento. Cerró en 2019.

Además de las distribuciones mencionadas, es necesario añadir las soluciones Hadoop-as-a-Service de los proveedores de cloud:

- Amazon Elastic Map Reduce (EMR).
- Microsoft Azure HDInsight (y evoluciones).
- Google Dataproc.

Estas soluciones aportan algunas ventajas muy interesantes:

-Reducen considerablemente el tiempo de aprovisionamiento (instalación, configuración y despliegue) de infraestructuras Hadoop, de meses en el caso de instalaciones en la propia infraestructura de las empresas, a minutos en un proveedor cloud. Las empresas se encuentran inmersas en procesos de transformación digital donde prima lo que se conoce como el time-to-market, es decir, la rapidez para lanzar nuevas soluciones.

-Ofrecen **elasticidad**, es decir, cuando lanzas una plataforma Hadoop en la nube, si necesitas más capacidad o potencia, el proceso de escalar o incrementar el tamaño de la infraestructura es muy sencilla, y lo mismo ocurre si deseas reducir el tamaño de la plataforma.

-Ofrecen **pago por uso**: el coste suele ser en número de servidores por las horas que están levantados, por lo que por un lado no requiere una inversión inicial importante y por otro, se paga sólo por el tamaño de la plataforma, que como hemos visto, puede adecuarse a la necesidad real en cada momento (elasticidad).

En resumen, las principales ventajas son una reducción del riesgo (no hay inversión inicial) y un incremento de la agilidad.

Sin embargo, estas soluciones cloud presentan algunas **desventajas**:

- Se produce un efecto que se denomina vendor lock-in, es decir, la barrera para salir de una solución cloud a otra de otro fabricante cloud o a un Hadoop propio, es elevada. Por ejemplo, los proveedores cloud aplican un cargo por sacar los datos fuera de su entorno.

- Las soluciones que ofrecen no suelen ser estándar, sino adaptaciones de Hadoop que han realizado los proveedores.
- El coste puede ser mucho más elevado y de hecho, difícilmente se conoce a priori al utilizar fórmulas de cálculo de los costes que añaden a veces variables que no se pueden estimar (por ejemplo, el consumo de CPU que vamos a tener).

Al **conjunto de servidores** que trabajan en conjunto para implementar las funcionalidades de Apache Hadoop se le denomina **clúster**, y a cada uno de los servidores que forman parte del clúster se le denomina **nodo**.

A partir de ahora, cuando usemos la palabra "**clúster** de Hadoop" debes pensar en el conjunto de servidores que forman la plataforma que está en ejecución, y cuando usemos la palabra "nodo" debes pensar en cada uno de los servidores que componen el clúster.

Los nodos pueden ser de tres tipos diferentes:

-**Nodos worker**, que realizan los trabajos. Por ejemplo, para el almacenamiento, cada worker se ocupará de almacenar una parte, mientras que para la ejecución de trabajos, cada worker realiza una parte del trabajo.






-**Nodos master**, que controlan la ejecución de los trabajos o el almacenamiento de los datos. Son los nodos que controlan el trabajo que realizan los nodos worker, por ejemplo, asignando a cada worker una parte del proceso o de los datos a almacenar, vigilando que están realizando el trabajo y no están caídos, rebalanceando el trabajo a otros nodos en caso de que un worker tenga problemas, etc.

-Nodos edge o frontera que hacen de puente entre el clúster y la red exterior y proporcionan interfaces, ya que normalmente un clúster Hadoop no tiene conexión con el resto de servidores e infraestructura de la empresa, por lo que toda la comunicación desde el exterior hacia el clúster se canaliza a través de los nodos frontera, que además, ofrecen los APIs para poder invocar a servicios del clúster.

NODOS MASTER

- Deben ser **más fiables** porque gestionan los servicios críticos del clúster.
- No almacenan datos generales, solo información de control.
- Su potencia depende del tamaño del clúster (más nodos = más exigencia).






Configuración típica:

-  **Discos:** 2–4 discos en RAID (1, 5 o 10), de 2–4 TB, para tener copias en caso de fallo.
-  **CPU:** 2 procesadores de 6–8 núcleos cada uno (procesamiento intensivo).
-  **Memoria:** 128–256 GB de RAM de alta calidad.
-  **Red:** conexión rápida, normalmente de 10 Gbps duplicada (hasta 20 Gbps) o redes avanzadas tipo Infiniband (>50 Gbps).
-  **Fuente de alimentación:** redundante para evitar cortes eléctricos.

NODOS WORKER

- Se enfocan en **almacenamiento y procesamiento**, no en resiliencia.
- Se asume que **pueden fallar**, por lo que no se invierte tanto en redundancia (fuentes, discos espejo, etc.).

Configuración típica:

-  **Discos:** Muchos discos grandes (3–4 TB) en **JBOD** (sin replicación), normalmente 10–12 discos por nodo. La replicación la hace HDFS.
-  **CPU:** 2 procesadores de gama media, 6–8 núcleos cada uno.
-  **Memoria:** Mínimo 64 GB, habitual 128–256 GB para tareas de procesamiento.
-  **Red:** Igual que los nodos master, 10–20 Gbps.
-  **Fuente de alimentación:** básica, no redundante, porque el sistema tolera fallos de nodos.

Por lo tanto, el hardware típico donde se ejecuta un cluster Hadoop es el siguiente:

Tipo de nodo	Disco	CPU	Memoria	Red	Coste aproximado
Master	2 HD x 2-3 TB RAID	2 CPU x 8 cores	256 Gb RAM	20 Gbps	5.000 - 15.000 € / nodo
Worker	12 HD x 2-3 TB JBOD	2 CPU x 8 cores	256 Gb RAM		3.000 - 12.000 € / nodo
Edge	2 HD x 2-3 TB RAID	2 CPU x 8 cores	256 Gb RAM		5.000 - 10.000 € / nodo

La implantación de una plataforma Hadoop implica tres tipos de costes principales:

1. **Hardware:** incluye la compra de servidores y equipos de red.
2. **Soporte empresarial:** suele costar entre **5.000 y 15.000 euros por nodo y año**.
3. **Servicios profesionales o consultoría:** varían según la complejidad y tamaño del clúster.

Por ejemplo, un clúster con **50 nodos worker, 4 master y 2 frontera** tendría un coste aproximado de **430.000 euros en hardware y 350.000 euros anuales en soporte**.