

Formación senior en Hadoop

PARTE DE LA FORMACIÓN BIG DATA ENGINEER DE BIG DATA ACADEMY

Concepto

BIG DATA
ACADEMY

Hadoop

Es un framework implementado en Java que permite el **almacenamiento y procesamiento distribuido** de grandes conjuntos de datos estructurados, semi-estructurados y no estructurados. Está **diseñado para trabajar en clústers** con miles de máquinas y tiene una alta tolerancia a fallas.



Objetivo fundamental

Objetivo fundamental de Hadoop

1. Almacenar archivos de manera distribuida [HDFS]
2. Procesar archivos de manera distribuida [YARN + MAPREDUCE]

Naturaleza de funcionamiento

¿Qué es trabajar de manera distribuida?

Significa **trabajar sobre un clúster**. Un clúster es una agrupación de servidores (computadoras) conectadas sobre una red generalmente LAN.



**Servidores
(o nodos del clúster)**



Switch



Rack



Clúster

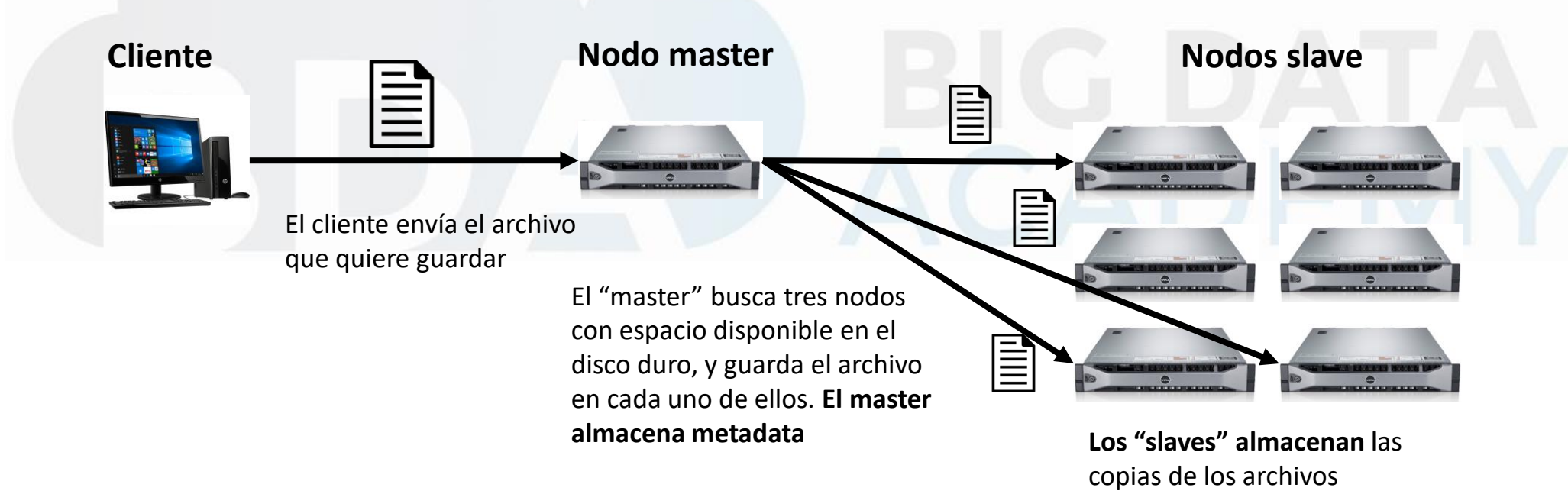
Un clúster Hadoop

Por lo general un clúster Hadoop está **conformado por al menos un nodo llamado “master” y tres o más nodos llamados “esclavos”**. El nodo “master” es el que recibe peticiones de almacenamiento o procesamiento desde algún cliente y delega el trabajo a los nodos “slave”.



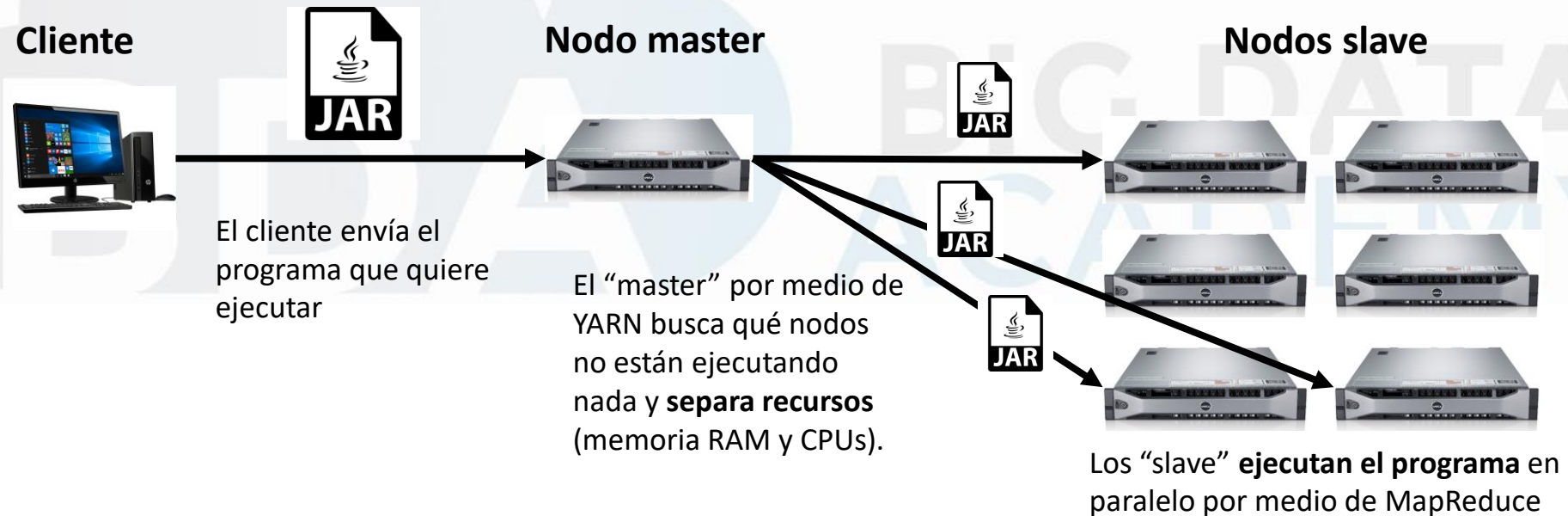
HDFS: Almacenamiento

En Hadoop el módulo que se encarga del almacenamiento y manipulación de archivos es conocido como HDFS (Hadoop Distributed File System). Es módulo se encarga de recibir desde un cliente peticiones de lectura y escritura de archivos y **almacenar los archivos en los nodos "slave"**.



YARN + MapReduce: Procesamiento

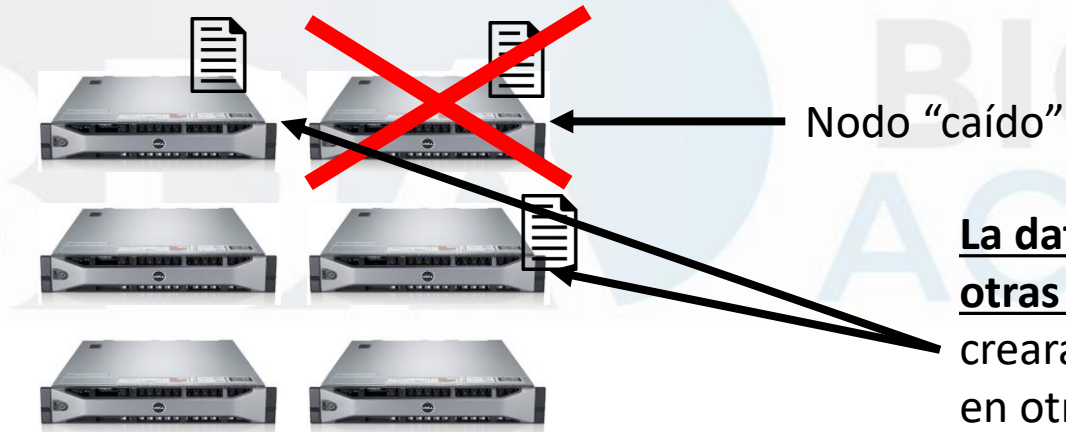
En Hadoop los módulos que se encargan del procesamiento de archivos son YARN (Yet Another Resource Negotiator) y MapReduce. El módulo de YARN verifica los nodos “slave” que están libres y los selecciona para el procesamiento. El módulo MapReduce ejecuta el procesamiento.



¿Por qué Hadoop almacena tres copias de un mismo archivo?

Hadoop almacena tres copias del mismo archivo por dos razones:

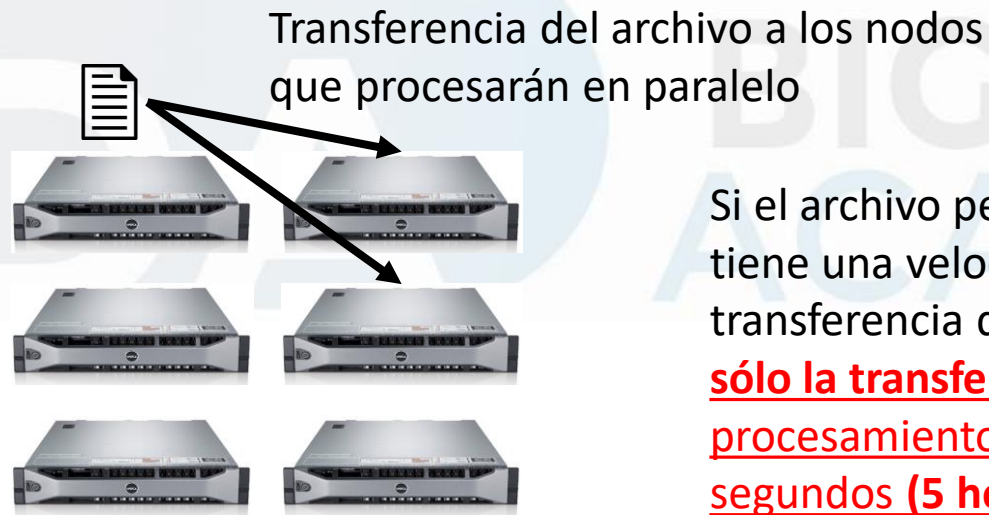
Si uno de los nodos “cae” (por ejemplo el disco duro se malogra), aún tenemos dos copias en otros dos nodos y no perdemos los datos.



La data aún es accesible por las otras dos copias. Es más el clúster creará una nueva copia del archivo en otro nodo para mantener las tres copias.

¿Por qué Hadoop almacena tres copias de un mismo archivo?

Si sólo hubiera una copia, el resto de nodos que procesará el archivo debería recibir una copia de este antes de procesarlo, esto significa un tiempo de transferencia de archivo en la red LAN.











Si el archivo pesa 10TB y la red tiene una velocidad de transferencia de 1GBPS, entonces sólo la transferencia previa al procesamiento demoraría 20000 segundos (5 horas y media)

La capacidad física de un clúster

La capacidad física de un clúster está **definida por la suma de los recursos computacionales que posea cada uno de sus nodos**. Los recursos computacionales son la memoria RAM, la cantidad de discos duros, la capacidad de cada disco duro y el número de CPUs.

La siguiente tabla muestra el cálculo de la capacidad física de un clúster según sus nodos.

	NODO	NÚMERO DE DISCOS DUROS	CAPACIDAD POR DISCO (TB)	ESPACIO TOTAL (GB)	RAM (GB)	NÚMERO DE CPUS
	1	10	1 TB	10 TB	128 GB	8
	2	12	1 TB	12 TB	128 GB	8
	3	4	4 TB	16 TB	128 GB	16
	4	4	4 TB	16 TB	256 GB	8
	5	5	2 TB	10 TB	64 GB	4
	6	7	2 TB	14 TB	64 GB	8
	7	9	1 TB	9 TB	128 GB	16
	8	10	1 TB	10 TB	64 GB	4
			CLÚSTER	97 TB	960 GB	72

¿Por qué Hadoop está hecho sobre Java?

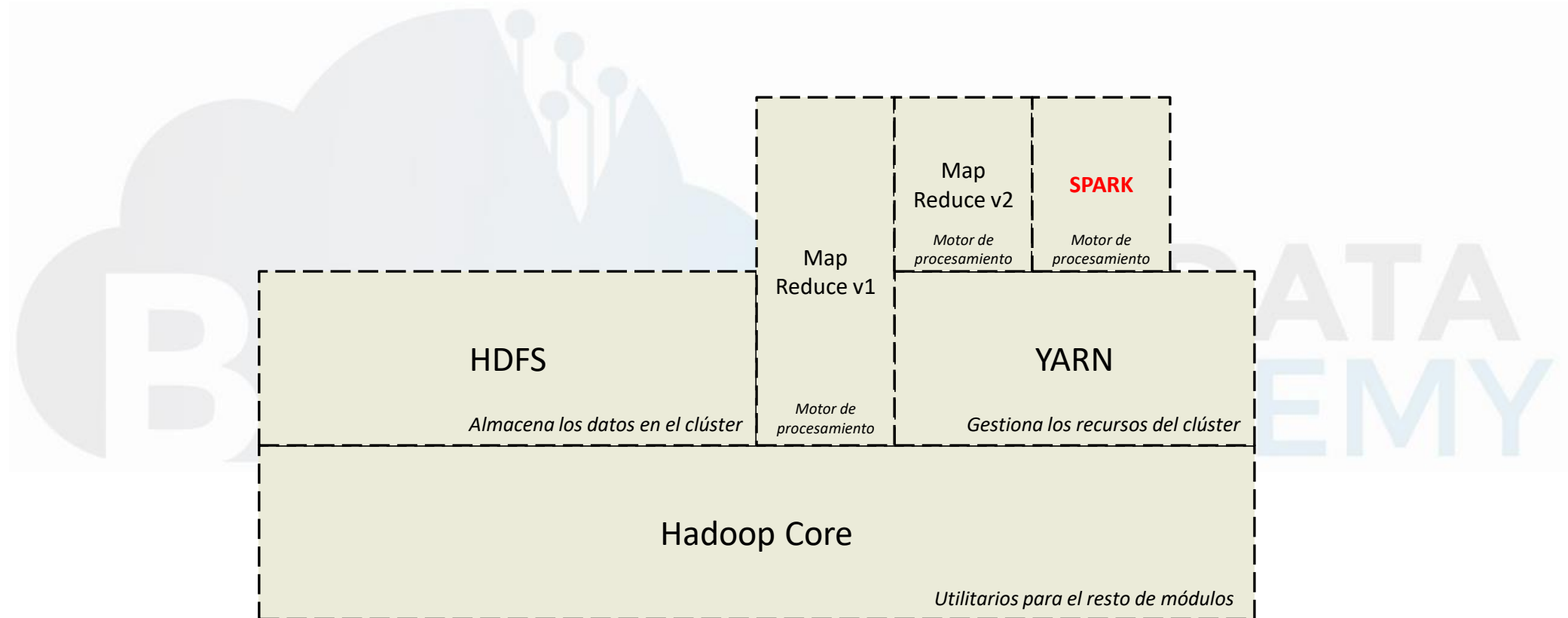
Hadoop está implementado sobre Java ya que este lenguaje de programación es perfecto para ejecutar programas que utilicen grandes cantidades de recursos computacionales. Java está optimizado para ejecutar programas que requieran de cientos de gigas de memoria RAM.



Arquitectura de componentes



Arquitectura de componentes de Hadoop



Arquitectura de servicios

Servicios de HDFS

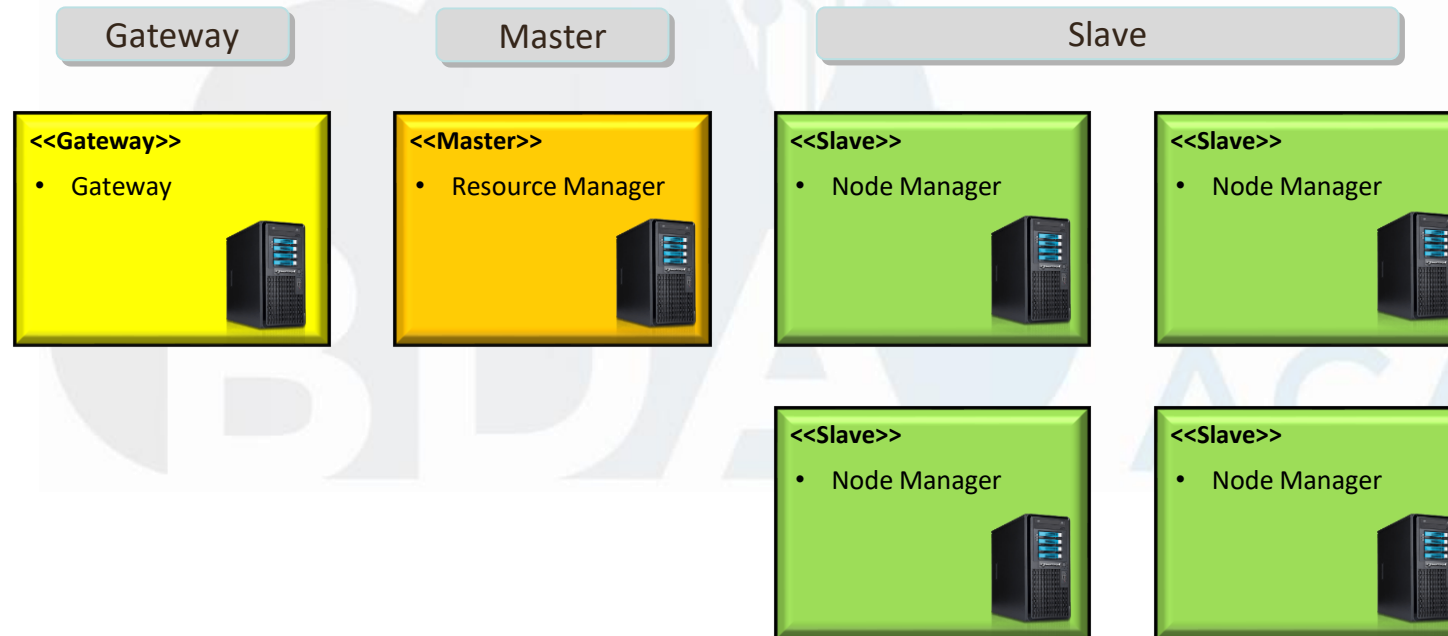


El Gateway es el nodo que tiene instalado las shells desde donde podemos ejecutar nuestros código, es un cliente

DESCRIPCIÓN

- **Gateway:** Nodo con la configuración de HDFS y acceso a las shells
- **HttpFs:** Permite enviar comandos HDFS por medio de un API REST
- **Namenode:** Almacena los metadatos (estructuras de carpetas, tamaño de bloques)
- **Datanode:** Almacena los datos subidos al clúster

Servicios de YARN (MapReduce v2 incluido)



DESCRIPCIÓN

- **Gateway:** Nodo con la configuración de YARN y acceso a las shells
- **Resource Manager:** Decide sobre que nodos slaves separar los recursos.
- **Node Manager:** Ejecuta las decisiones tomadas por el Resource Manager y separa sobre el nodo los recursos solicitados.

Ventajas

Ventajas de Hadoop

1. Abstrae al desarrollador de la distribución de la carga de trabajo, es decir, el desarrollador sólo se preocupa en construir su programa y decirle al clúster que lo ejecute con cierta cantidad de RAM y CPUs, el clúster se encargará de elegir a los servidores que en ese momento estén disponibles
2. Funciona sobre hardware commodity, es decir no necesita de un servidor especial tipo IBM que puede llegar a costar mucho dinero.
3. Permite la escalabilidad lineal, es decir si un proceso se ejecuta en 10 minutos con 5GB de RAM y 3 CPUs, entonces con 10GB de RAM y 6 CPUS deberá ejecutarse en 5 minutos.
4. Permite aumentar la potencia del clúster fácilmente, agregando más nodos al clúster.

Ejercicios teóricos



Ejercicios teóricos [continua...]

1. ¿Qué es Hadoop?
2. ¿Cuáles son los objetivos de hadoop?
3. ¿Qué es un clúster?
4. ¿Cuál es la función de un nodo “gateway”?
5. ¿Cuál es la función de un nodo “master”?
6. ¿Cuál es la función de un nodo “slave”?
7. ¿Cuál es el componente de Hadoop que permite almacenar archivos?
8. ¿Cuál es el componente de Hadoop que permite procesar archivos?
9. ¿Por qué Hadoop realiza tres copias del mismo archivo?
10. Si hay un clúster con 10 nodos, cada nodo tiene 1 disco de 5TB, 32 GB de RAM y 8 CPUs, ¿cuál es la potencia física del clúster?
11. Enumere algunas ventajas de Hadoop.

Ejercicios teóricos [continuación...]

11. ¿Por qué Hadoop está implementado en Java?
12. ¿Cuáles son los componentes de Hadoop? Describa cada uno
13. ¿Cuáles son los servicios de HDFS? Describa cada uno
14. ¿Cuáles son los servicios de YARN? Describa cada uno

Programación básica

Programación básica

Codifiquemos...



Programación avanzada

Programación avanzada

Codifiquemos...



Resumen

BIG DATA
ACADEMY

Resumen

Hablemos...

