ESCUELA NACIONAL DE ESTUDIOS SUPERIORES UNIDAD MORELIA

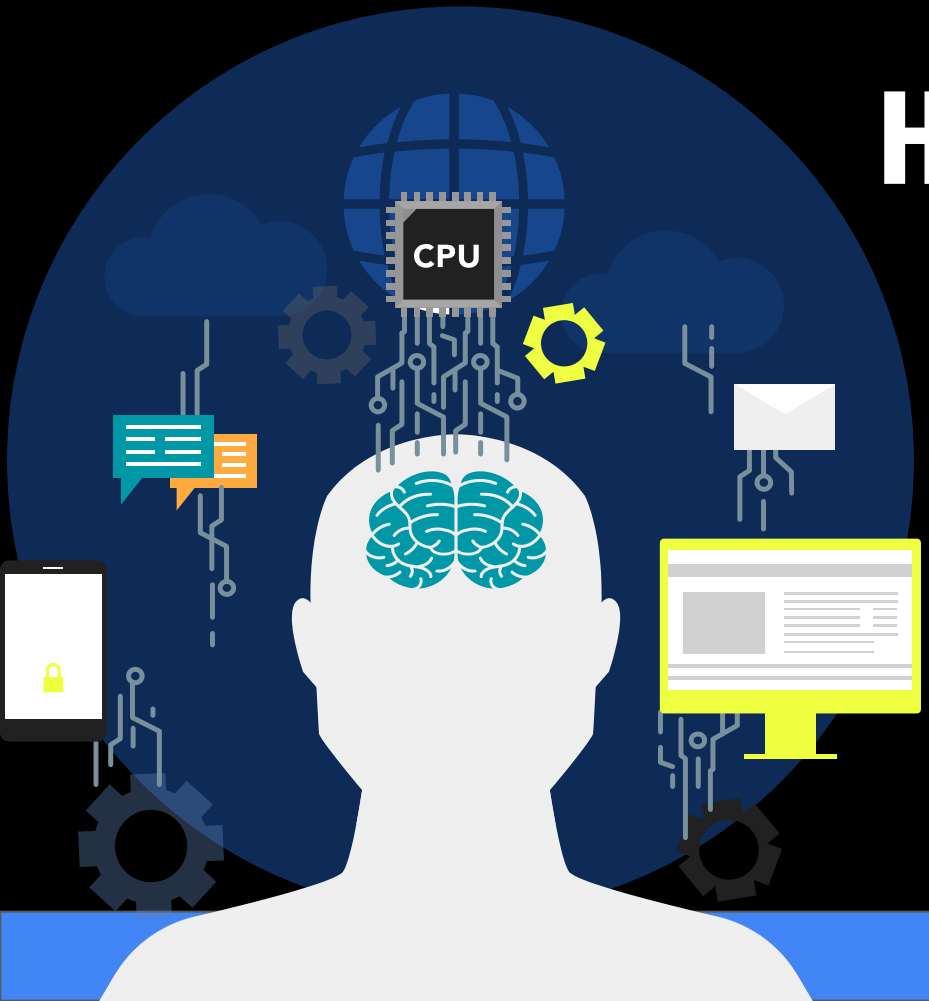
iies UNAM

UTM Universidad Tecnológica de Morelia

Diplomado en ML y DL aplicado a grandes volúmenes de datos

PAPIME PE103124

Módulo I: ML (Marzo 2024)



Hadoop Distributed File System (HDFS)

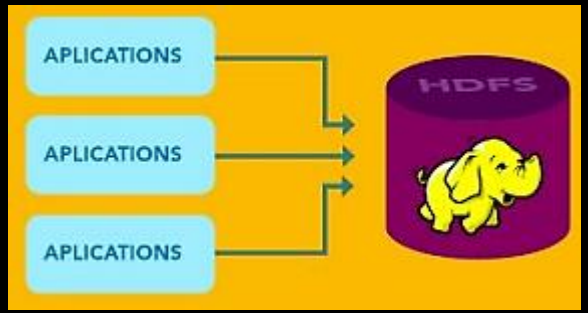
Sergio Rogelio Tinoco Martínez
Heberto Ferreira Medina
José Luis Cendejas Valdez

Contenido

- ¿Qué es HDFS?
- ¿Cómo funciona?
- Ventajas

¿Qué es HDFS?

HDFS (*Hadoop Distributed File System*) es una tecnología de almacenamiento distribuido de información. Proporciona a las aplicaciones la capacidad de acceder a los datos en los lugares donde estén almacenados.

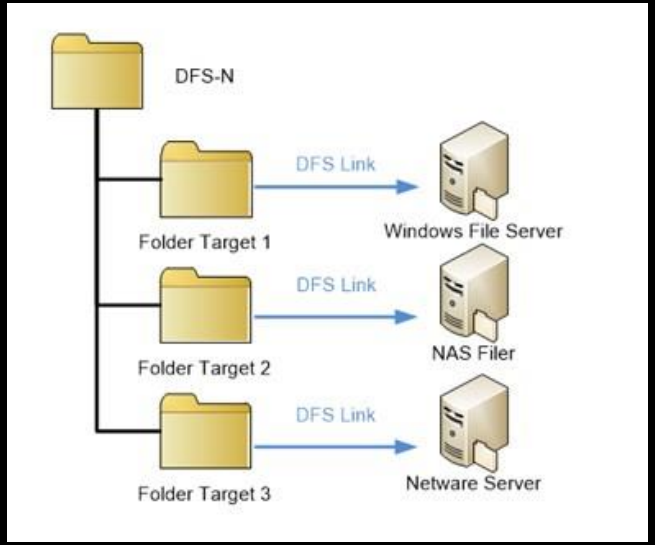


¿Qué es HDFS? – 2 –

HDFS permite el almacenamiento de datos estructurados, no estructurados y semiestructurados, en grandes cantidades y distribuidos en múltiples nodos.

¿Qué es HDFS? – 3 –

HDFS está basado en el lenguaje de programación Java, lo cual permite obtener una visión de los recursos como una sola unidad. Crea una capa de abstracción por encima de los datos, como si fuera un sistema de archivos únicos.



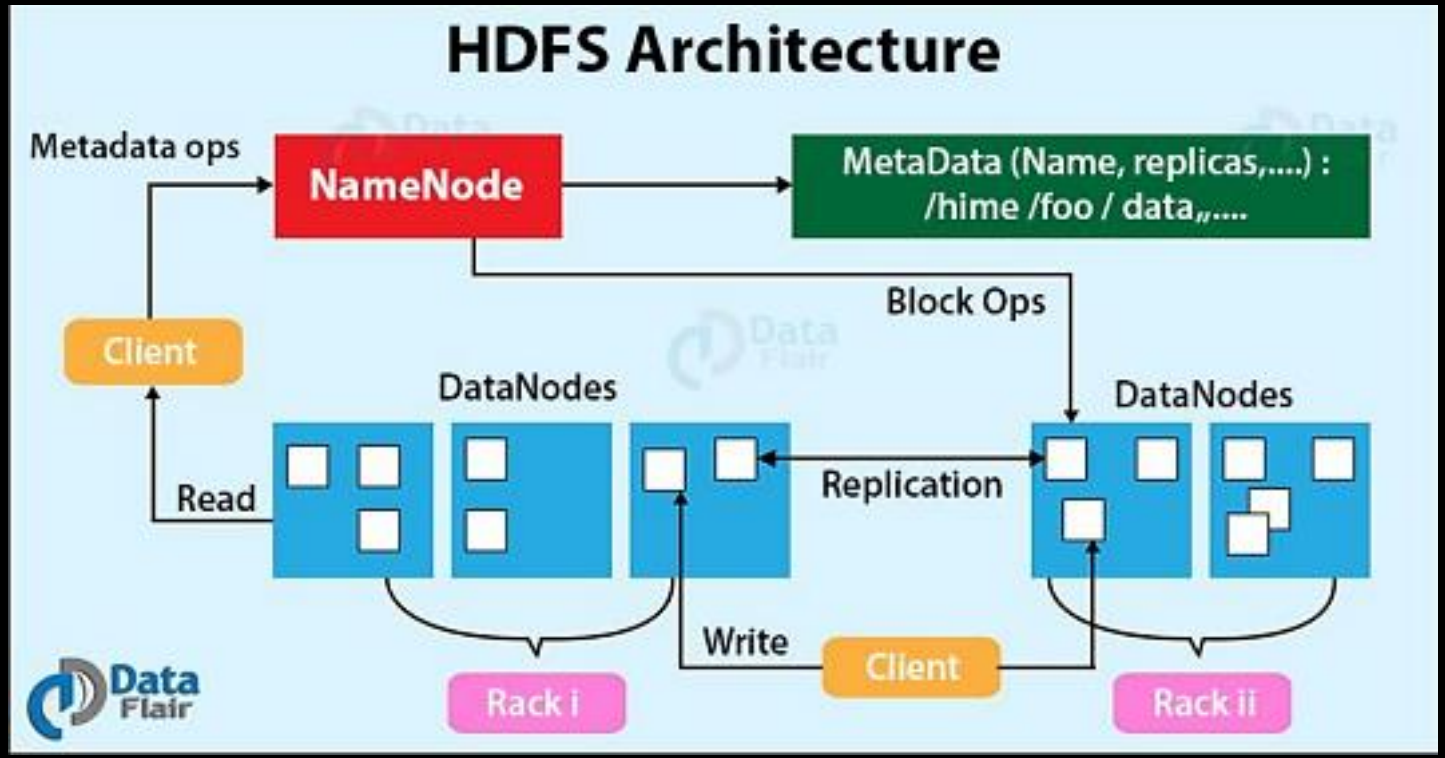
¿Cómo funciona?

- ***NameNode***: es el maestro o nodo principal que almacena los metadatos de los archivos y gestiona el acceso a la información.
- ***DataNodes***: son las computadoras que almacenan la información del clúster y son a los que se les añaden discos duros o se incrementan en número, a fin de aumentar la capacidad de almacenamiento del sistema y conseguir la escalabilidad horizontal.

¿Cómo funciona? – 2 –

- HDFS tiene un modelo *Write-Once Read Many* (***WORM***). Significa que no se pueden editar los archivos almacenados en el HDFS, pero sí se pueden añadir datos.
- Escritura: recibir y replicar la información.
- Lectura: reconstruir la información para desplegarla.

¿Cómo funciona? – 3 –



Ventajas

- Escalabilidad: más discos.
- Eficiencia: WORM.
- Reducción de costos: software LIBRE.
- Gestión de errores: “*Murió pero vivió*”.
- Flexibilidad: *Data Lakes* (repositorios únicos de los datos crudos del sistema, de sensores, sociales, etc. y los datos transformados usados para generar informes, visualización, analítica y ML).