

Seminario de Estadística I
Introducción a la Ciencia de Datos y Machine Learning

HDFS

Universidad Nacional Autónoma de México
Facultad de Ciencias



Contenido

1 Hadoop

- ¿Qué es Hadoop
- Módulos

2 Modos de Ejecución

- Modos

3 Sistema de Ficheros HDFS

- HDFS

4 Arquitectura

- DataNode y NameNode
- Comunicación

5 Interacción con HDFS via comandos

- Comandos



Sobre Hadoop

- Hadoop es framework que permite procesamiento distribuido de grandes cantidades de datos entre clusters de computadoras.
- Esta diseñado para escalarse de un único servidor a miles de máquinas.
- Esta diseñado para detectar y manejar fallos, ofreciendo alta disponibilidad.



Módulos

Módulos de Hadoop

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce:
- Hadoop Ozone:
- Hadoop Submarine



Modo Standalone, Pseudo-Distribuidos)

- **Modo Standalone** Modo no distribuido en el que hadoop corre como un único proceso Java.
- **Modo Pseudo-Distribuido** Hadoop corre en un único nodo donde cada demonio Hadoop corre en un proceso Java separado.

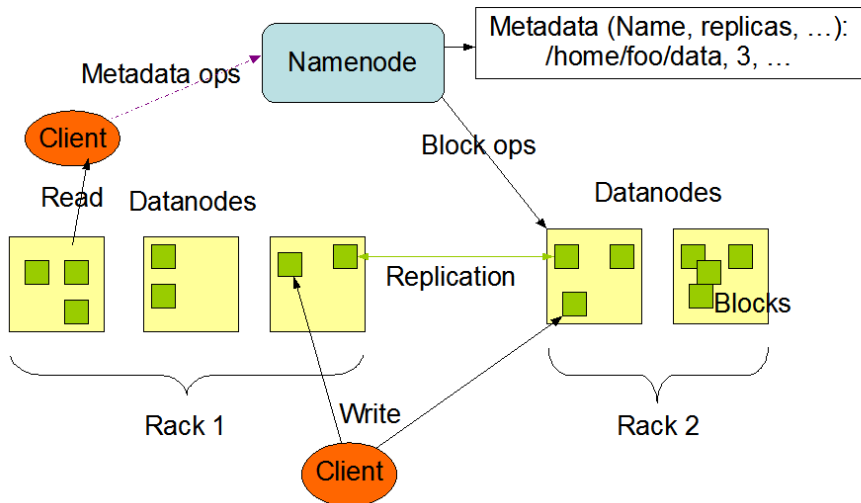


Introducción a HDFS

- El Hadoop Distributed File System (HDFS), es un sistema de ficheros distribuido altamente tolerante a fallos y esta diseñado para implementarse en software de bajo costo.
- Proporciona acceso de alto rendimiento a datos de la aplicación y adecuado para grandes volúmenes de datos.



HDFS Architecture



HDFS web-ui

Para interactuar con HDFS lo primero es “invocar” a los demonios NameNode, DataNode y el SecondaryNamenode con **start-dfs.sh**. Verificamos que estén corriendo escribiendo `jps` en la terminal o bien con la interfaz web en nuestro navegador con la dirección **`http://localhost:9870`**, recordemos que 9870 es el puerto por default en hadoop 3 para ver el NameNode. Una vez lanzado hdfs podemos utilizar linea de comando o un cliente hdfs en python para realizar acciones sobre el sistema de ficheros. A continuación se mostraran los comandos básicos de interacción.



Algunos comandos básicos

Comandos HDFS

- **hdfs dfs -mkdir path**
Crea un directorio en la ruta indicada
- **hdfs dfs -ls path**
Lista el contenido dentro del directorio indicado en la ruta
- **hdfs dfs -du path**
Muestra el uso en disco
- **hdfs dfs -mv pathorigen pathdestino**
Mueve un directorio indicado por el path origen al pathdestino en hdfs.
- **hdfs dfs -rmr path**
Elimina de manera recursiva, es decir elimina directorios y sus subcarpetas.
- **hdfs dfs -put pathlocal pathenhdfs**
Subir archivos del path local a hdfs
- **hdfs dfs -get pathhdfs destino**
Obtener archivos de hdfs a local

