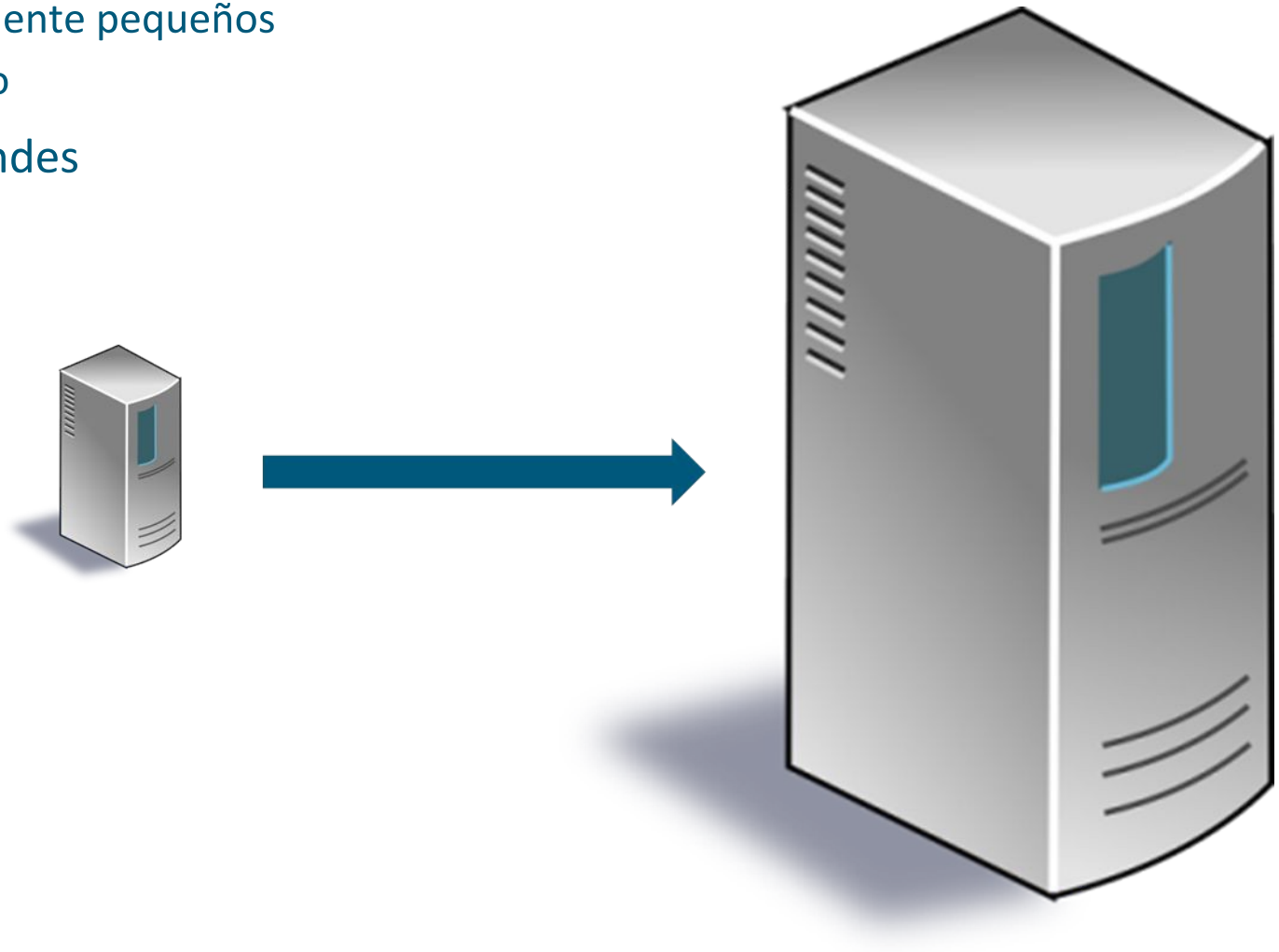
HDFS

Hadoop Distributed FileSystem

- Introducción
- ¿Por qué Hadoop?
- Conceptos básicos y HDFS

Sistemas tradicionales

- Sistemas de computación tradicional
  - Conjuntos de datos relativamente pequeños
  - Procesamiento muy complejo
- Solución: Ordenadores más grandes
  - Procesadores más rápidos
  - Más memoria

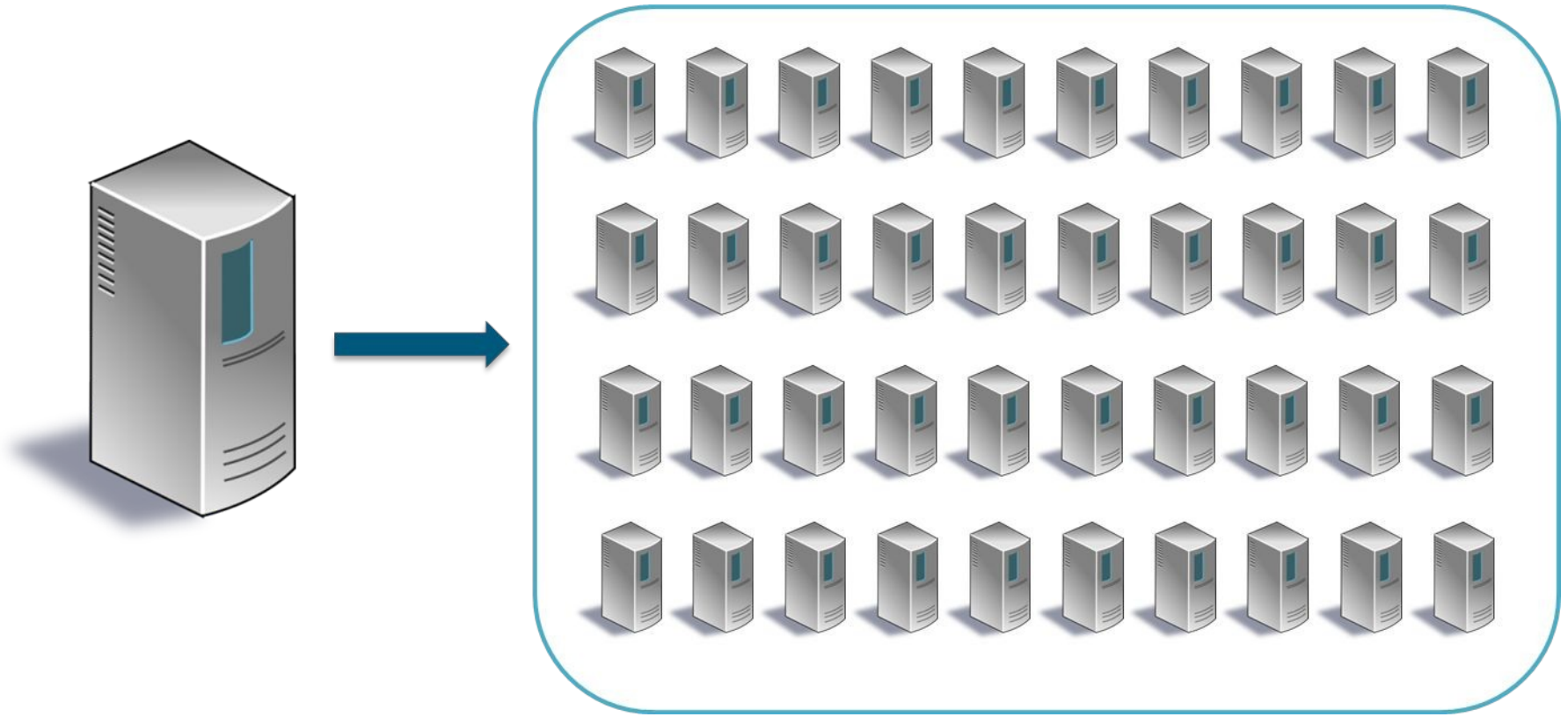


- Introducción
- ¿Por qué Hadoop?
- Conceptos básicos y HDFS

- ¿Por qué Hadoop?
  - Sistemas distribuidos
  - Introducción a Hadoop

Sistemas distribuidos

- ¿Solución?
  - Múltiples ordenadores para realizar el mismo trabajo



- ¿Por qué Hadoop?
  - Sistemas distribuidos
  - Introducción a Hadoop

¿Por qué Hadoop?

- El Big Data surge para solucionar problemas:
  - Cómo almacenar y trabajar con grandes volúmenes de datos
  - Cómo analizar estos datos de naturaleza diversa
- Hadoop es un framework que permite el procesamiento de grandes volúmenes de datos
- Utiliza una arquitectura maestro/esclavo con su propio sistema de ficheros para almacenamiento (Hadoop Distributed File System, HDFS)
- Se ejecuta el código donde están almacenados los datos

- Introducción
- ¿Por qué Hadoop?
- Conceptos básicos y HDFS

- Conceptos básicos y HDFS
  - El proyecto Hadoop y sus componentes
  - Hadoop Distributed File System (HDFS)

Componentes de Hadoop

Clúster simple

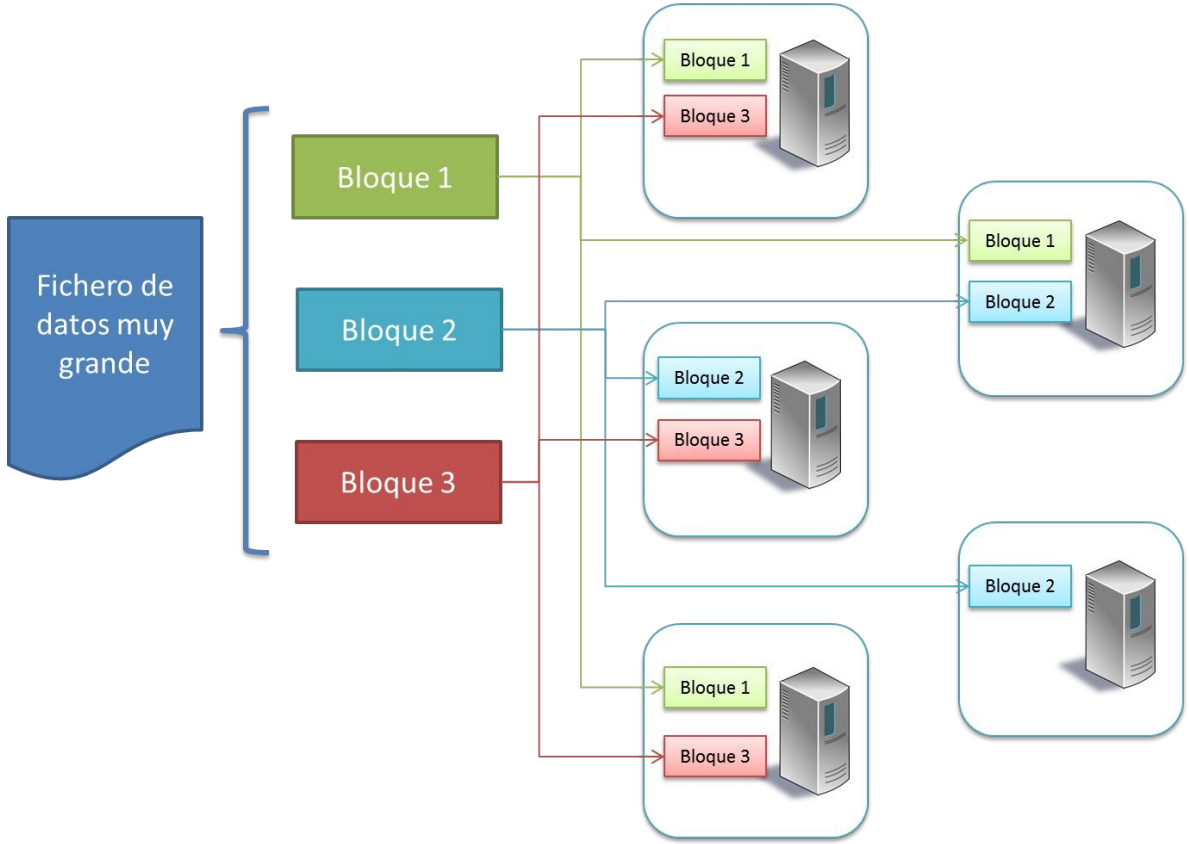
- Un clúster Hadoop es un grupo de ordenadores trabajando juntos para procesar los datos
- Hadoop tiene un arquitectura maestro esclavo:
  - Dos nodos maestros:
    - NameNode: Gestiona HDFS
    - ResourceManager: Gestiona las aplicaciones
  - Muchos nodos esclavos:
    - Almacenan los datos en HDFS
    - Procesan los datos

- Conceptos básicos y HDFS
  - El proyecto Hadoop y sus componentes de Hadoop
  - Hadoop Distributed File System (HDFS)

- Sistema de ficheros escrito en Java
- Se crea sobre el sistema de ficheros nativo (ext3, ext4 o xfs)
- Almacena cantidades masivas de datos
- Proporciona replicación de datos

Almacenamiento de la información en HDFS

- Los ficheros se dividen en bloques (El tamaño de bloque por defecto es 128MB)
- Cada bloque es replicado en múltiples nodos (por defecto 3 réplicas)
- El NameNode almacena los metadatos
- Cliente comunica directamente con Data Nodes



Comandos HDFS

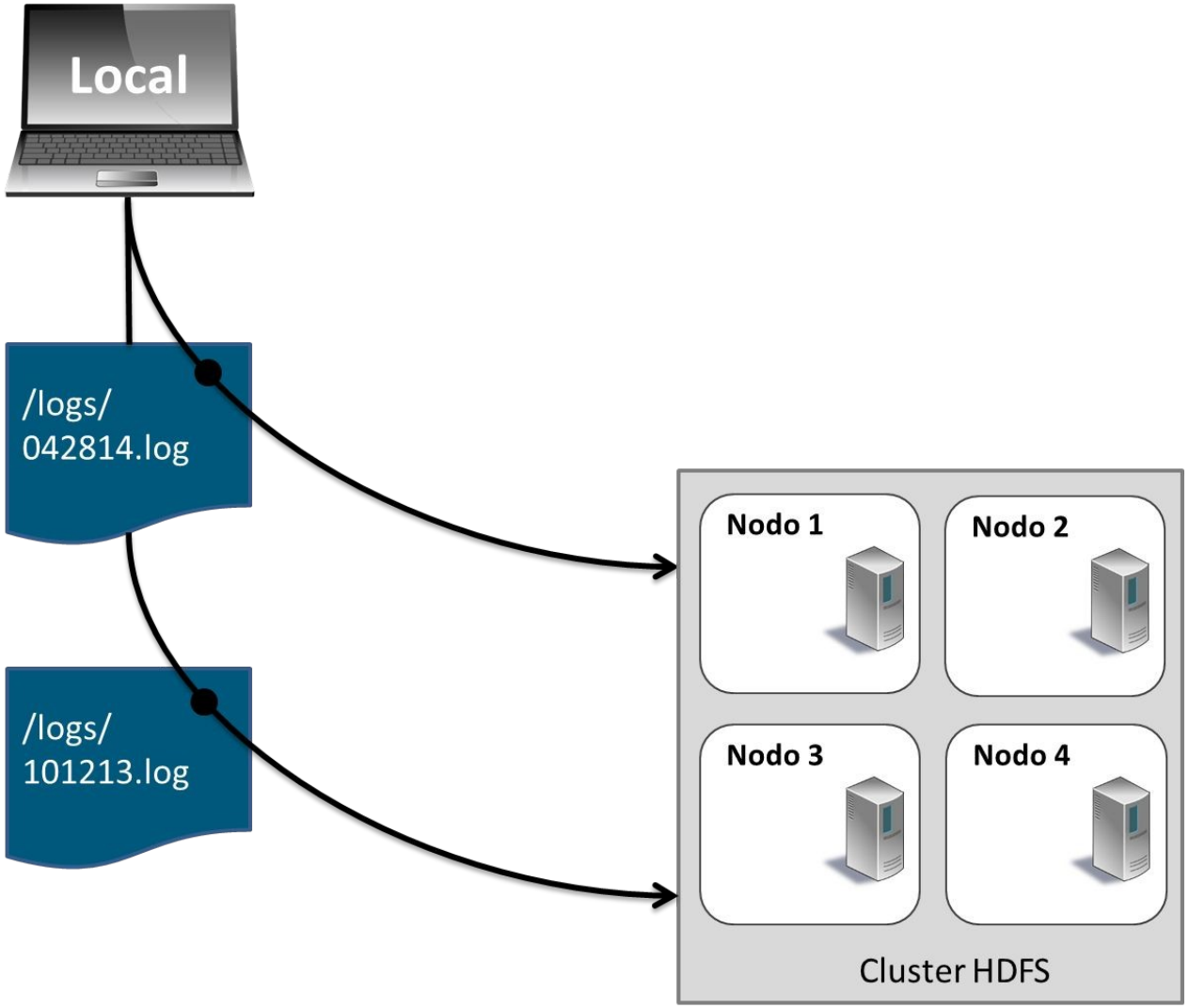
- La aplicación *hadoop* es un cliente de Hadoop que permite ejecutar comandos en línea de comandos. Tiene la sintaxis:

```
hdfs dfs –command <args>
```

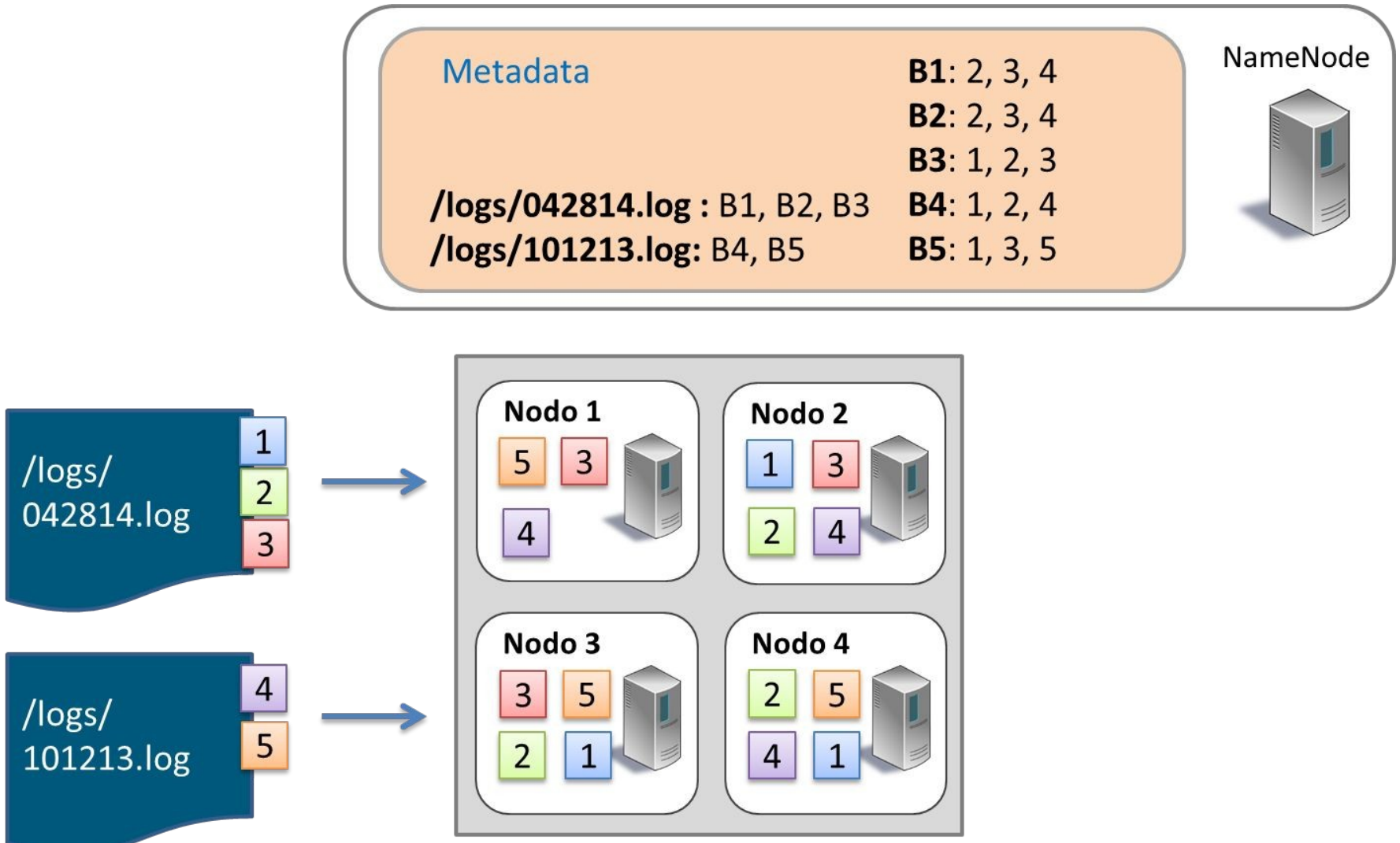
- Algunos comandos hadoop:

Comando	Acción
ls	Lista el contenido de los directorios.
count	Cuenta el numero de directorios, ficheros y bytes en un directorio.
chgrp, chown y chmod	Cambia los permisos de ficheros y directorios.
help	Muestra la ayuda.
cat y text	Muestra el contenido de los ficheros.
tail	Muestra el último 1KB del contenido del fichero.
get y copyToLocal	Copia un fichero o directorio desde HDFS al sistema de ficheros local.
put y copyFromLocal	Copia un fichero o directorio desde el sistema de ficheros local a HDFS.
getMerge	Obtiene una mezcla de dos ficheros en único fichero y lo almacena en el sistema de ficheros local.
mv	Mueve ficheros y directorios en HDFS.
cp	Copia ficheros y directorios en HDFS.
mkdir	Crea directorios nuevos en HDFS.
rm	Borra un fichero en HDFS (mueve a la papelera).
rm -r	Borra un directorio recursivamente en HDFS (mueve a la papelera).

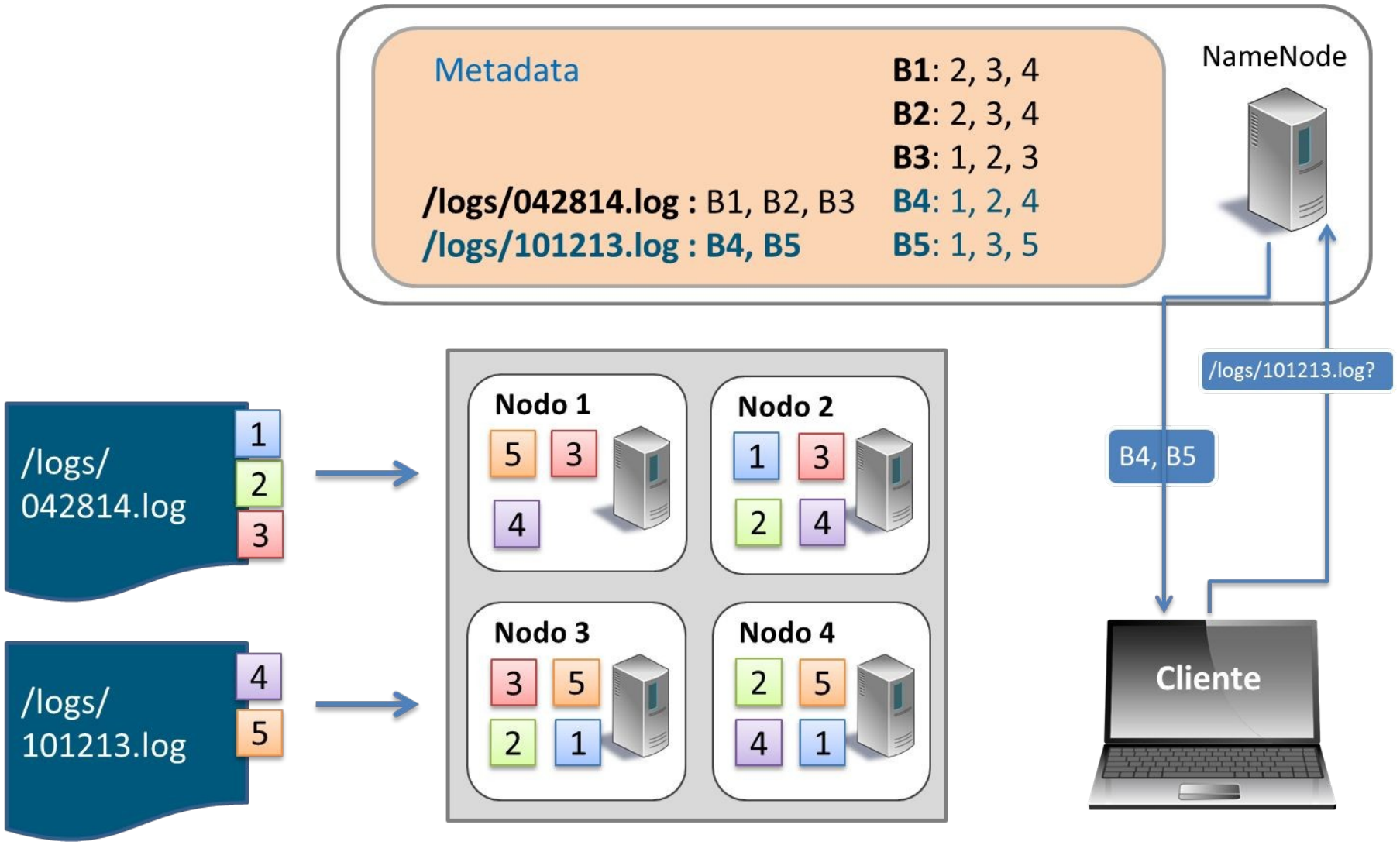
Ejemplo lectura y escritura en HDFS I



Ejemplo lectura y escritura en HDFS II



Ejemplo lectura y escritura en HDFS III



Ejemplo lectura y escritura en HDFS IV