Big Data Aplicado

Josep Garcia

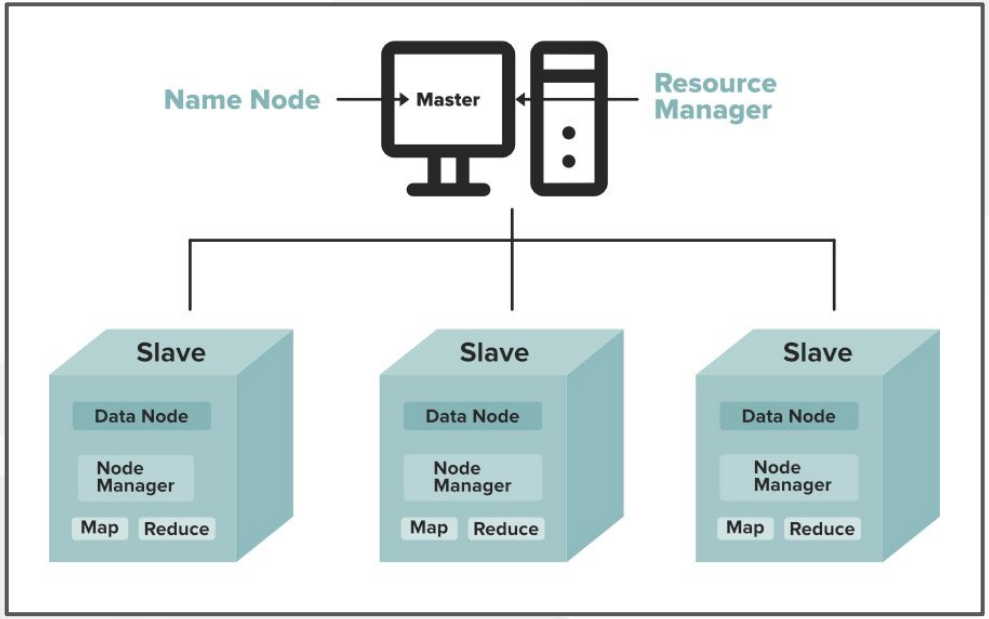
IES EDUARDO PRIMO MARQUÉS

INTRODUCCIÓN A HDFS

- HDFS es un sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de hardware sin perder datos.
- Es la parte de almacenamiento de Datos de Hadoop

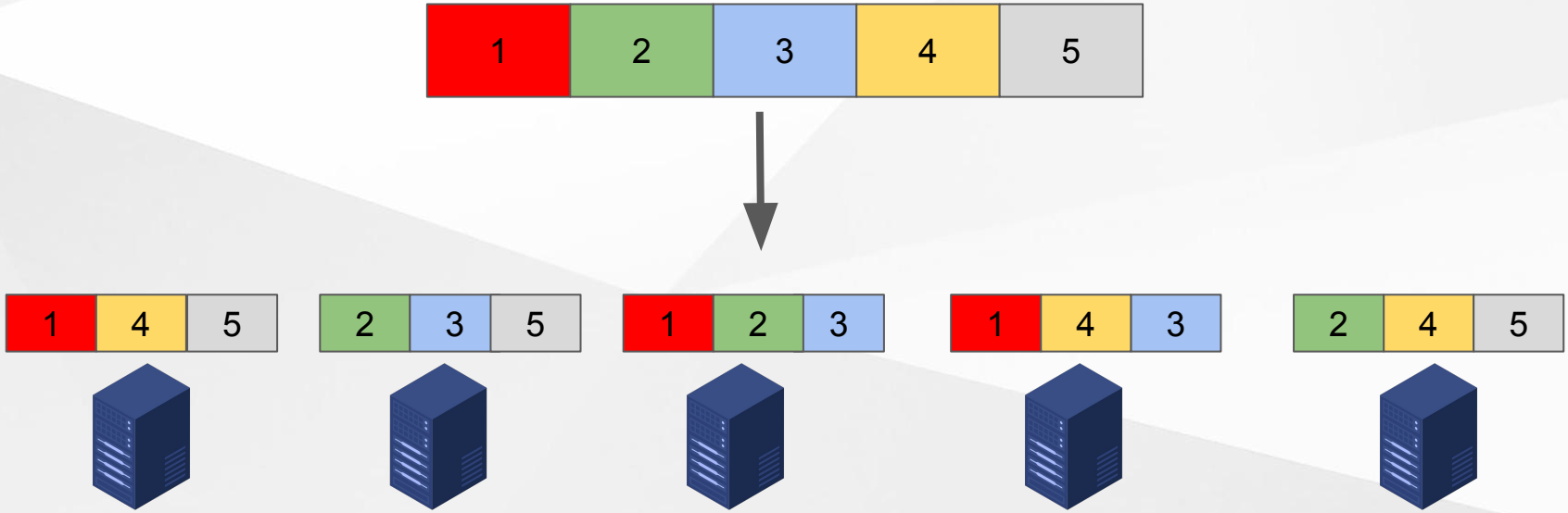
INTRODUCCIÓN A HDFS

- HDFS gestiona el almacenamiento en el cluster, dividiendo los ficheros en bloques y almacenando copias duplicadas a través de los nodos.
- Por defecto se replican en 3 nodos distintos.



INTRODUCCIÓN A HDFS

Ejemplo de guardado de un fichero en HDFS.



Nodos Hadoop

INTRODUCCIÓN A HDFS

- Al crear un cluster hadoop hay:
- Un nodo que actúa como maestro de datos. Sólo tienen metadatos
- El resto de nodos son esclavos. Contiene los datos propiamente dichos.

Maestro

Esclavos

INTRODUCCIÓN A HDFS

ESTRUCTURA DE LOS METADATOS DENTRO DE HDFS

HDFS dispone de unos ficheros que gestionan los cambios que se producen en el Cluster de HDFS.

- edits_000xxxx: Cambios que se van produciendo en la base de datos.
- edits_inprogress_xxxx: Datos que se están escribiendo en el momento.
- fsimage_00000xxxx: Copia (foto) del sistema de ficheros en un momento concreto.

INTRODUCCIÓN A HDFS

¿Cómo funciona al arrancar HDFS?

Al arrancar HDFS (start-dfs.sh) se carga en memoria el último fichero "**fsimage**" disponible junto con los edits que no hayan sido procesados.

INTRODUCCIÓN A HDFS