# DataNode

A DataNode stores data in the [HadoopFileSystem]. A functional filesystem has more than one DataNode, with data replicated across them.

On startup, a DataNode connects to the NameNode; spinning until that service comes up. It then responds to requests from the NameNode for filesystem operations.

Client applications can talk directly to a DataNode, once the NameNode has provided the location of the data. Similarly, MapReduce operations farmed out to TaskTracker instances near a DataNode, talk directly to the DataNode to access the files. TaskTracker instances can, indeed should, be deployed on the same servers that host DataNode instances, so that MapReduce operations are performed close to the data.

DataNode instances can talk to each other, which is what they do when they are replicating data.

- There is usually no need to use RAID storage for DataNode data, because data is designed to be replicated across multiple servers, rather than multiple disks on the same server.
- An ideal configuration is for a server to have a DataNode, a TaskTracker, and then physical disks one TaskTracker slot per CPU. This will allow every TaskTracker 100% of a CPU, and separate disks to read and write data.
- Avoid using NFS for data storage in production system.