

HDFS has a master/slave architecture.

* **NameNode：**An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes.
* **DataNodes：**In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The DataNodes are responsible for serving read and write requests from the file system’s clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.
* HDFS exposes a file system **namespace** and allows user data to be stored in files.

File Formats：

* **TextFileFlat** file with data in comma-, tab-, or space-separated value format or JSON notation.hive, hive:text【行】
* **ParquetCompressed** columnar data representation.hive【列】
* **SequenceFileFlat** file consisting of binary key/value pairs.hive
* **RCFileRecord** columnar data consisting of binary key/value pairs; high row compression rate.hive, hive:rc
* **ORC**：Optimized Row Columnar 【优先推荐】

运维目标：NameNode的RPC高并发、DataNode的数据均匀分布