The Hadoop Distributed File System (HDFS):

A Storage System for Big Data

HDFS = foundation for Hadoop ecosystem

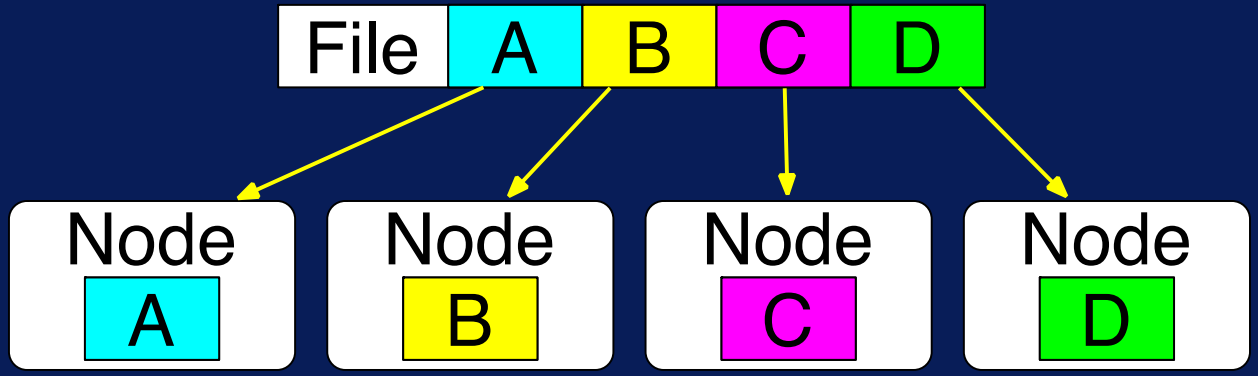
Scalability

Reliability

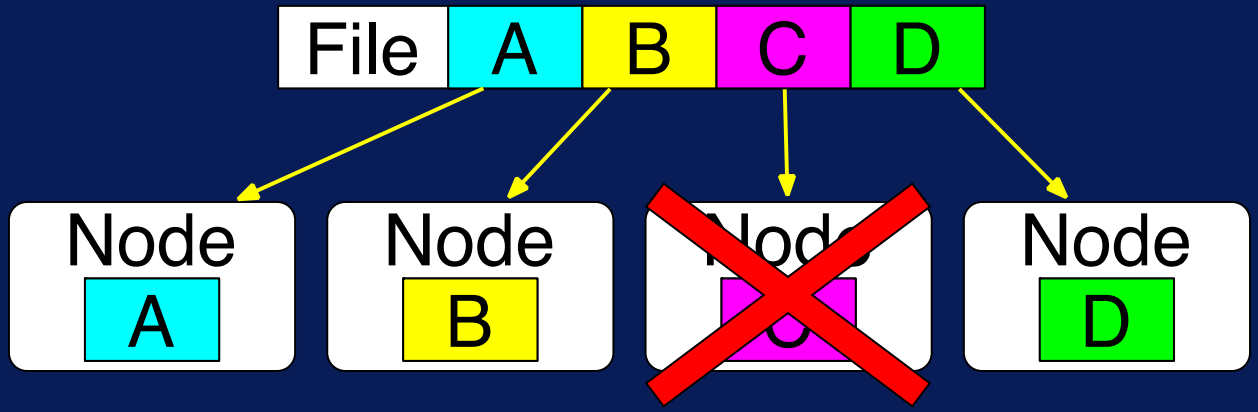
Store massively large data sets

up to 200 Petabytes,
4500 servers,
1 billion files and blocks!

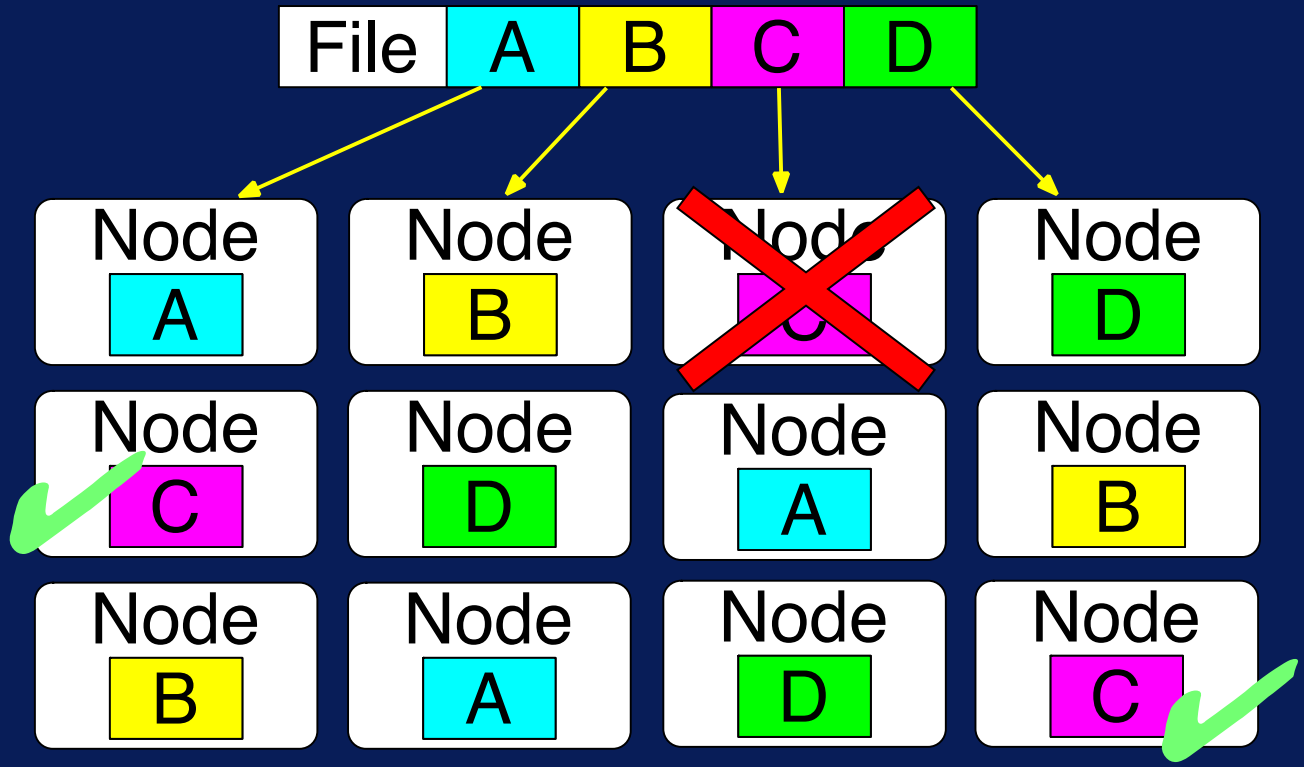
HDFS splits files across nodes for parallel access



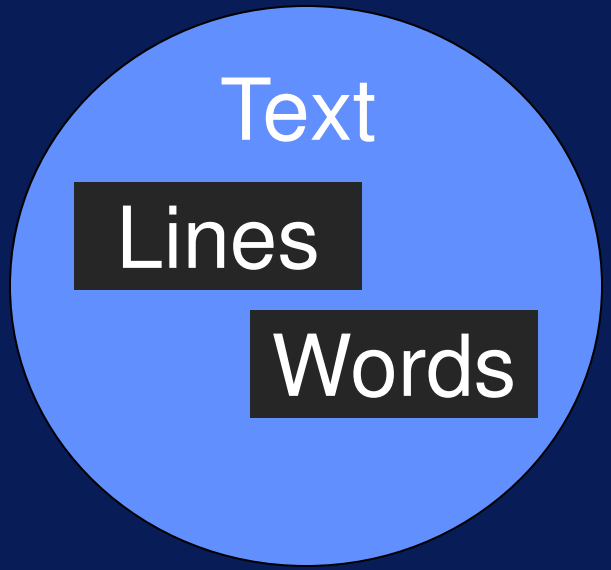
What happens if node fails?



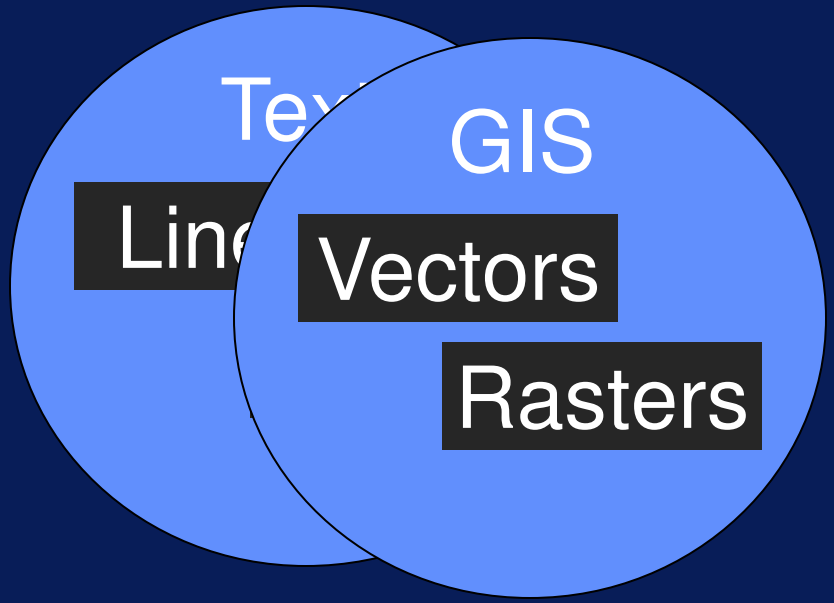
Replication for fault tolerance



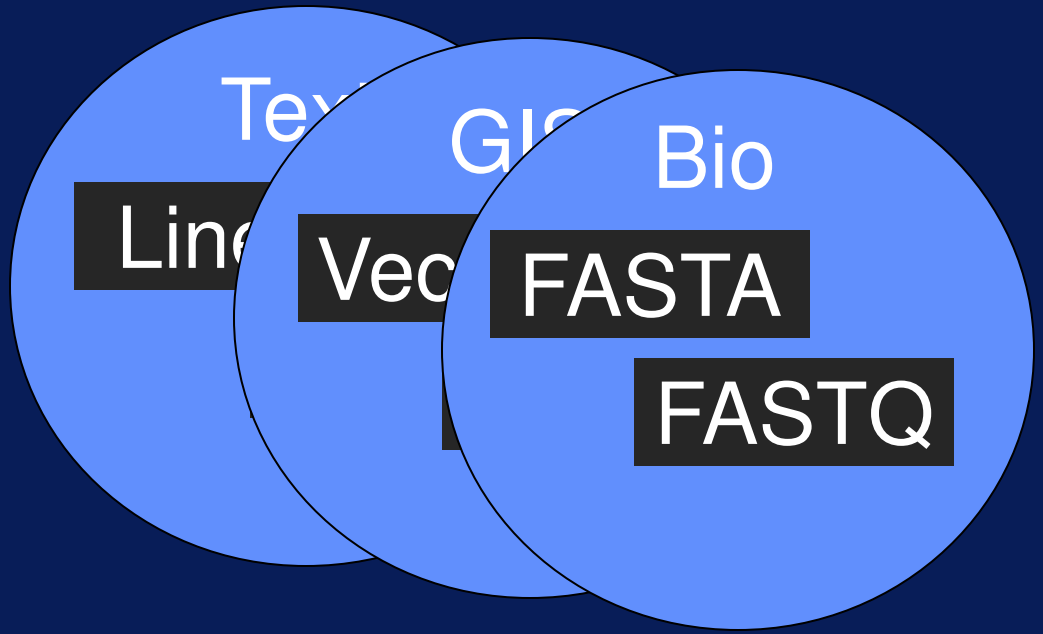
Customized reading to handle *variety* of file types



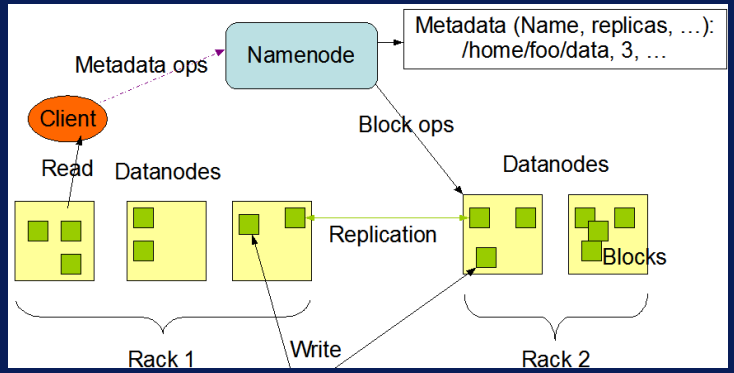
Customized reading to handle *variety* of file types



Customized reading to handle *variety* of file types



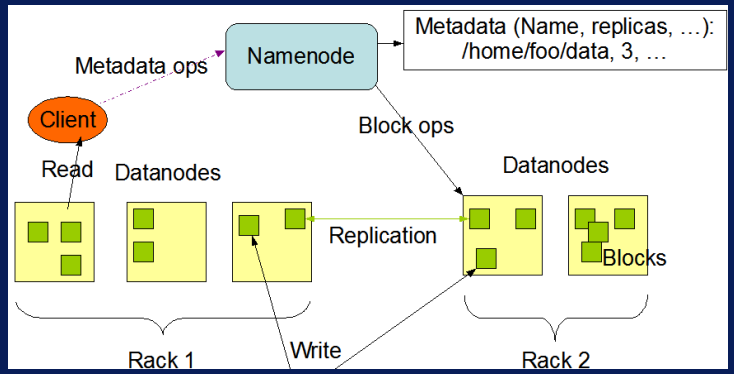
Two key components of HDFS



1. NameNode for metadata

2. DataNode for block storage

Two key components of HDFS



1. NameNode for metadata

Usually one per cluster

2. DataNode for block storage

Usually one per machine

The NameNode coordinates operations

Keeps track of file name, location in directory, etc.

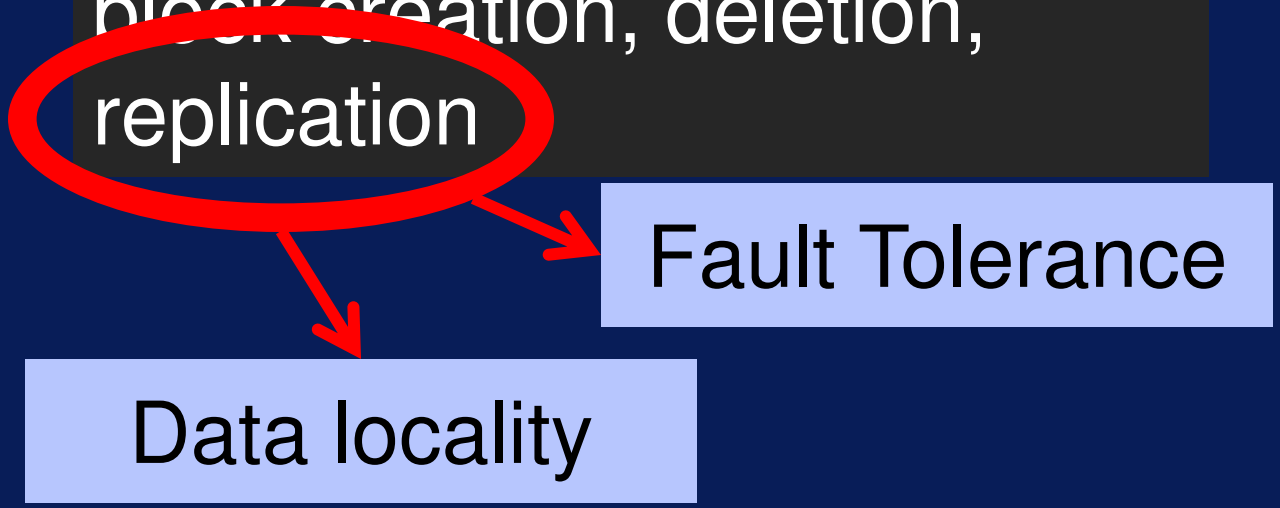
Mapping of contents on DataNode.

DataNode stores file blocks

Listens to NameNode for block creation, deletion, replication

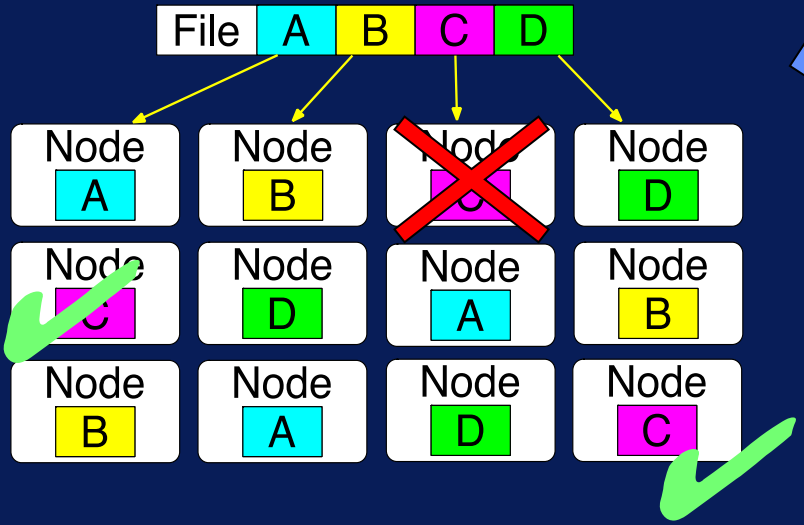
DataNode stores file blocks

Listens to NameNode for block creation, deletion, replication



Data partitioning

Data replication



Scalability

Fault tolerance

Data locality