fault tolerance & failure handling

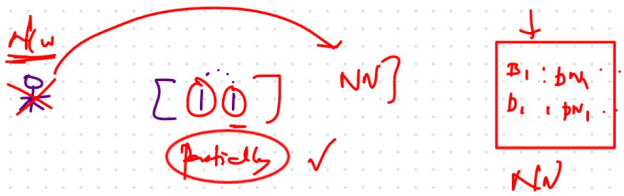
1. DN fails during the write
2. Client fail
3. Network/Rack failure
4. Namenode failure
5. DN failure

Dfs Duplicl
1, 2

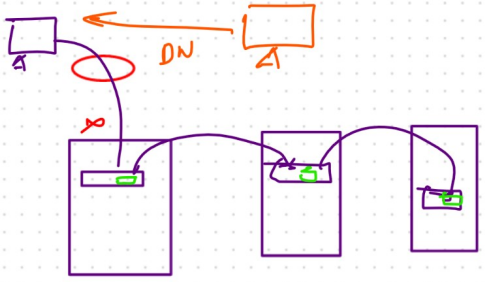
Data Node failure During Write:

- Client receives error report in acknowledge [1 0 1] 1
- Client informs Name node
- Name node marks block as under-replication
- Name node provides Client with a new Data Node address
- Client writes only to the new Data Node to maintain replication factor

Client failure :



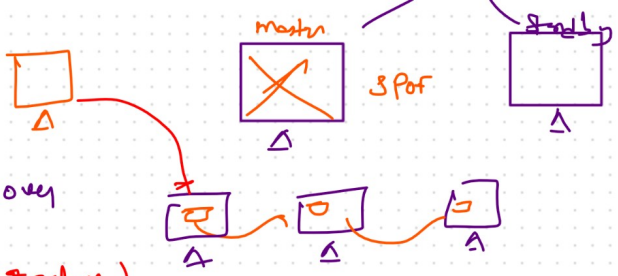
- partially written blocks are marked as corrupt
- Name Node clean up incomplete Blocks during regular maintenance



Network or Rack failure:

- disconnected replicas are marked as temporarily unavailable
- If issue persists, damage is marked as permanent
- Replication begins to maintain to replication factor

Name node failure :



- Standby Name node takes over
  (High available architecture)

H. A cluster

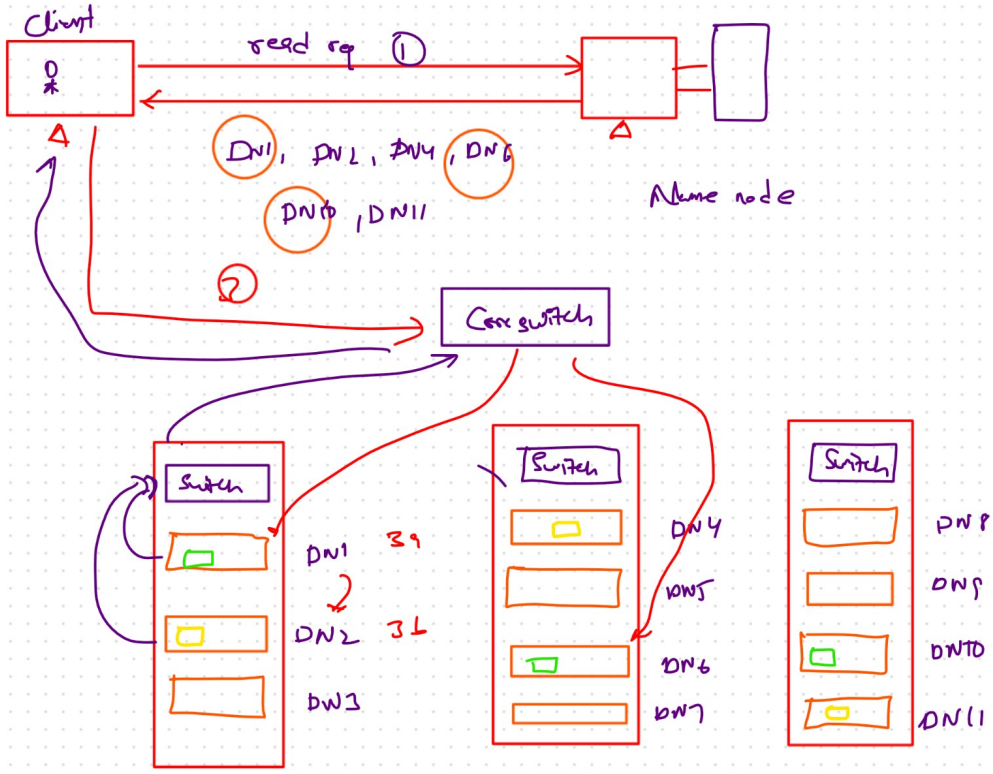
- It help to continue cluster operations

Data Node Failure (Serious)

- Blocks on failed node marked as under-replication
- Name node initiates replication to maintain its factor
- metadata is updated accordingly

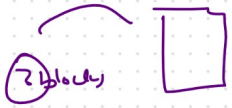
HDFS Read Operation

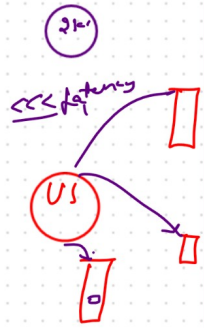
Read Operation Process



Step1: Client Request to Name node:

- Client sends a request to the Name node
- NN checks permissions & access metadata
- Name Node returns IP addresses of data nodes containing the required blocks





Step2: Data Node Selection

- Client doesn't need to read from all replicas
- Client selects the optimal data node based on
  - geographic location / rack awareness
  - N/w performance
  - proximity to Client

Step3 : Data Retrieval

- Client connects to the selected data node through the network (switches)
- Data node streams the requested block to the client
- Data flows back to the client through switches

Note Handling multiple Blocks

- for multiple blocks, the process is repeated for each block
- each block can be read from a different data node

Fault Tolerance in Read Operation

- If a data node fails during read, client automatically switches to another replica
- No metadata updates needed for read operations
- System provide data redundancy through replicas