NameNode Startup

On startup

- Load FSImage
- Check safe mode
- Start Daemons
    - HeartbeatMonitor
    - LeaseManager
    - ReplicationMonitor
    - DecommisionManager
- Start RPC services
- Start HTTP info server
- Start Trash Emptier

Load FSImage

Name Directory

- Dfs.name.dir: can be multiple dirs.
  ${dfs.name.dir}/current/VERSION
  /edits
  /fsimage
  /fstime
- Checks consistence of all name dirs.
- Load fsimage file
- Load edit logs
- Save namespace
    - Mainly setup dirs & files properly

Read File

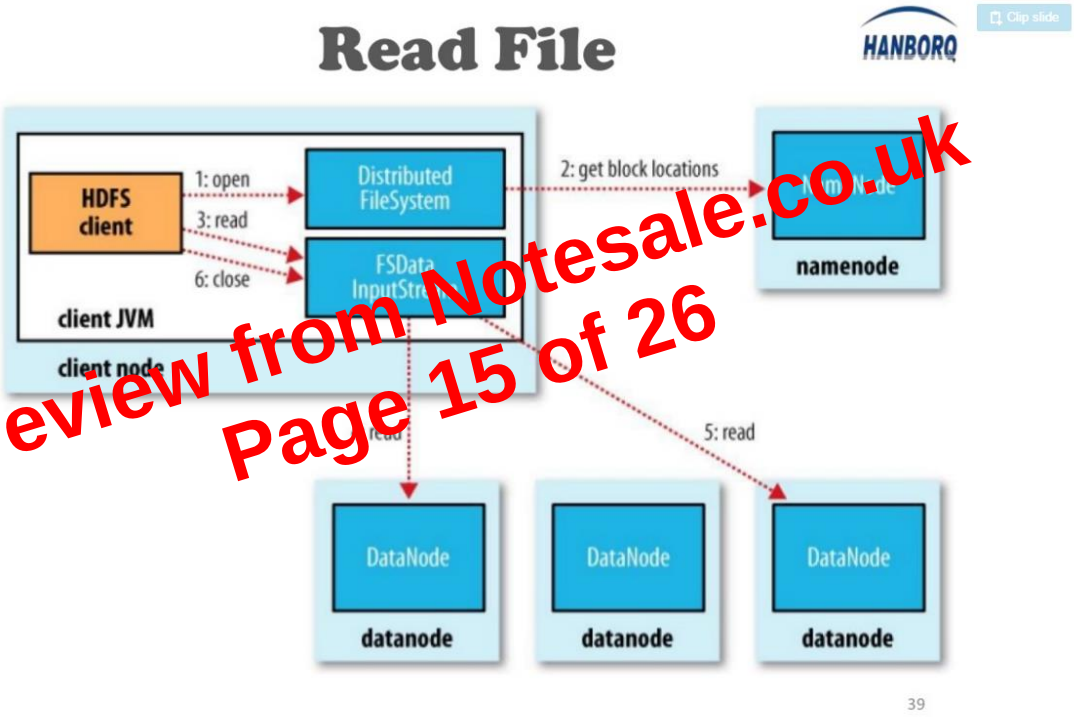
Sample Code:

DFSClient dfsclient = …

FSDataInputStream is = dfsclient.open(…)

is.read(…)

is.close()



- Change FSName replication factor
- Choose excess replicas
    - Rack number do not decrease
    - Get block from least available disk space node
- Add to invalidateSets(to- be-deleted clock set)
- ReplicationMonitor compute blocks to be deleted for eash DN
- DN delete specified blocks
- Update blocksMap when DN send blockReport

One DN Down

- DataNode stop sending hearbeat
- NameNode
    - HeartbeatMonitor find DN dead when doing heartbeat check
    - Remove all block belong to DN
    - Update neededReplication (block set need one or more replication)
    - ReplicationMonitor compute block to be replicated for each DN
    - On next DN's heartbeat, NameNode send replication block command
- DataNode
    - Replicate block

O&M, TOOLS

High Availability

NameNode SPOF

- NameNode hold all the meta
- If NN crash, the whole cluster unavailable

Through fsimage can recover from SNN

- Its not up-to-date fsimage

Need HA solutions