Hadoop is also a frame work. It is open source, apache s/w found
It is free, Just download & install. It is basically for storing
& for processing huge data set not small data.

Hadoop is used for storing & processing huge data set with
cluster of comodity hardware.

cluster is set of machine in a single LAN. comodity H/W
cheap H/w like unwanted laptop, pc etc. because of using cheap H/w.

If I wanted to store 500TB; it needs 4-5 crore, but
if it is cheap H/w then it will be 1-1.25 crore.

How your are storing data? HDFS   } 2 core components.
"   " you are processing of data? Mapreduce.

It is specially design a file system for storing huge data set
with cluster of comodity H/w with streaming access pattern.

→ It is a sply designed

write one & runing any where (any platform) → Java.
write ones & read any no. of times but dont change content
of the files. this is called streamed.

block size
4kb
for
500GeB Harddisk
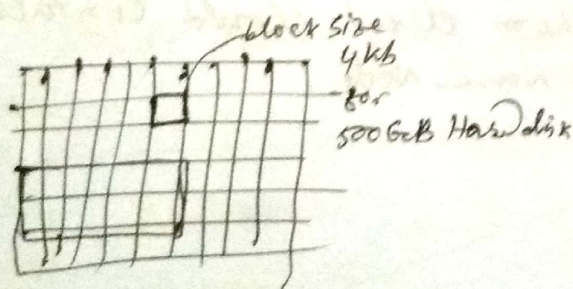
If I want to store 2kB of data
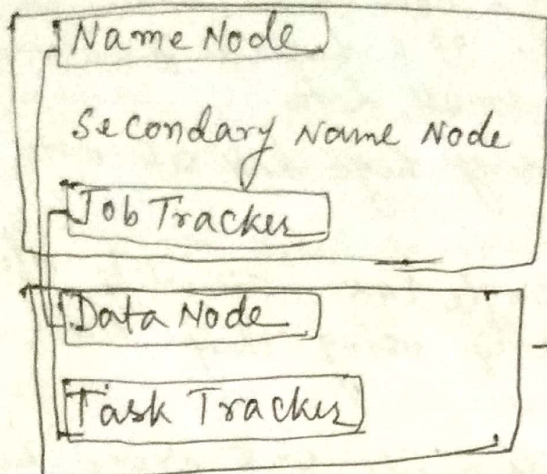into a 4kB space, so other 2kB
space is wasted.

In HDFS by default the block size is 64MB. It is zapped by
Hadoop install. After instaling of HDFS if I want to store
35MB, so rest space is 29MB, & it is possible to use
again. changing of setting can extened 64MB to 128MB.

HDFS has Five services.   → It is run on background.

```
┌─────────────────────────────┐
│ ┌─────────────┐             │
│ │ Name Node   │             │     ─ Master service
│ └─────────────┘             │       / Node.
│  Secondary Name Node        │
│ ┌─────────────┐             │
│ │ Job Tracker │             │
│ └─────────────┘             │
├─────────────────────────────┤
│ ┌─────────────┐             │
│ │ Data Node   │             │     ─ slave servic
│ └─────────────┘             │       / Node
│ ┌─────────────┐             │
│ │ Task Tracker│             │
│ └─────────────┘             │
└─────────────────────────────┘
```

Every Master service can talk each other
  "   slave   "   "   "   " "

Name Node can talk to ~~Job Tracker~~ Data Node.
Job tracker   "   "   " Task tracker.

file name → file.txt.
client foler   — splited
         200MB   into NO of 64MB
         like
         500GB both.

wanted to store
& processes some
data
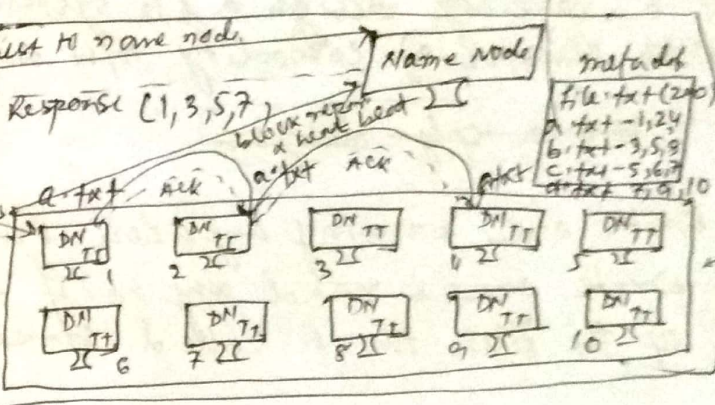
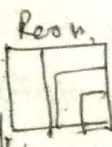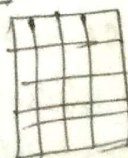How many of block for
  200 MB File.
  64, 64, 64, 8
like   a.txt 64
       b.txt 64
       c.txt 64
       d.txt 8 — 56 will be
                    used for
                    other.
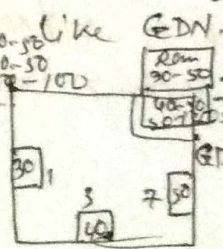
To whom client should Contact with,
   Any — Name Node.

for Eq.

A Farmer
having
one field
& small
room.

1st year
he generate.
10 rice packet.
20   "   "
50   "   "
200 "   "

Room.

30-50 like GDN.
40-50
50-100

Now it is not possible to store this room,
So through item no should keep in other ways

white paper & filled all details
He approaches a GDN with
the help of Loony.
Not possible to store directly,
He has to approach GDN Manager
The Manager knows how many
Spaces are there, where are they.

Hey Manager, have not enough space to
store my rice packet. so you give me some space for storing rice.
Manager ask how many packets are there and what quality of packets
Hey farmer go and store your packet in respective room.

3   Detail about data. Data about data is metadata.

Every data node is cheap H/W.

"   If the system is going to be down. To over come this problem, this HDFS given 3 replication. Actually it had 600 MB file. The client has to know the replication. with the help of ack. How this Name Node can know where a.txt is stored. In which data node.

For every short period of time all the slave node give report to Name Node. Block report a heart beat will give to Name node. 600 MB data has been occupied. (Still working & alive). Every data node have proper block Report a heart beat

If any of the DN not giving proper block report a heart beat then, it says that it is dead. It has been removed. After removing, 2 replica will be available, and they will provide report properly. Fraction of second 3rd replica will be create.

If the metadata is lost, simply say that, there is no use of Hadoop. DN are cheap H/W but NN should be high end system or high reliable H/W. If NN is not working then no metadata will work. Keeping of metadata req. good NN, but actual storage is DN. If we use 4kB of block, have to maintain 4kB of metadata, so more space will req. NN is calling as single point of Faliure.

So, If client want to store data, he has to Contact with NN.
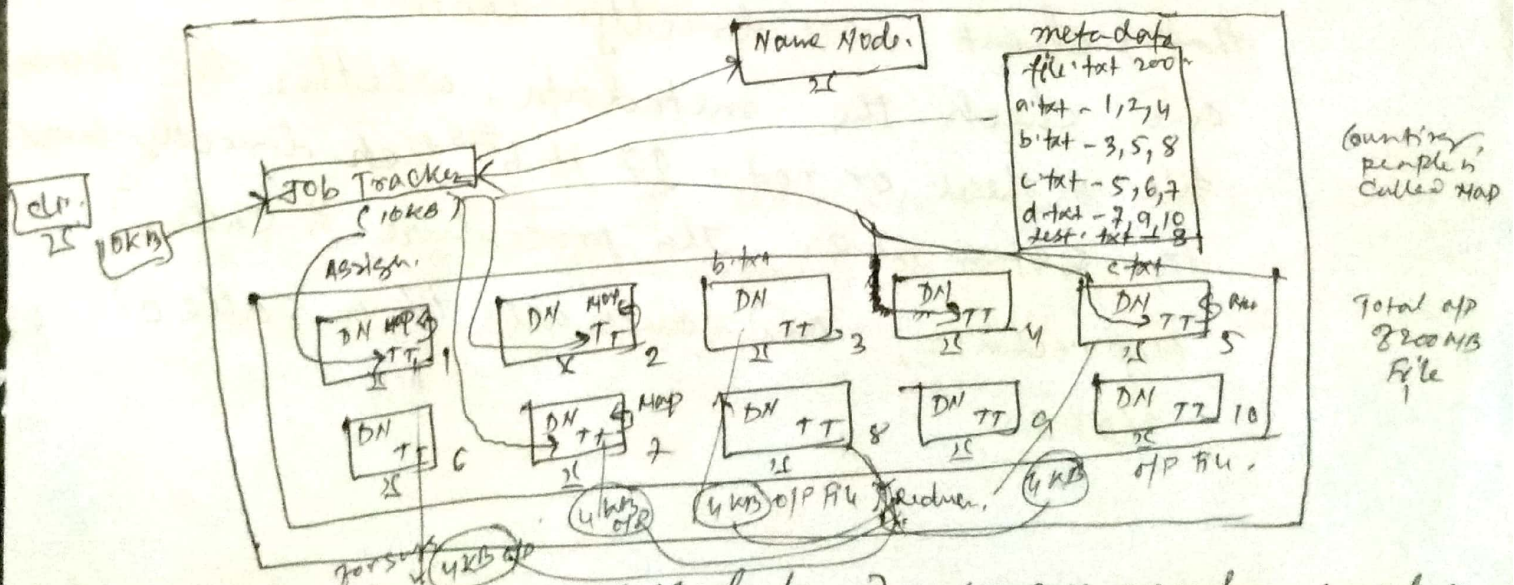
Store data & process data. For process need map reduce. 10kB of prog. to process of 200 MB data. Data should brough to your local machine. But In this case 10kB will send to Data Name. This will help by Job tracker. Job tracker send msg to Name Node.

Jobtrack is used to assigning job to tasktracker. In system
no 1, 2, 4 the artet is available, but nearest should be choosen by
Jobtraking. so time will take less if choose neares one. After taking
task by tasktracker, computation will start, so it is called Map.
four sub files is called input splits. Those many of no.
Mappers will be there, that many no. of input splits there.

It the 1st is going down, than data will be lost, tasktracker
will disturb. so this task will be reassign to another replica.
How the Jobtracker know wheter is dead or alive? or tasktracker
is working or not. All the tasktracker send heartbeat to Jobtracker.
Every 3sec tasktracker give heartbeat, and after 10 time waiting
if Namenode is not getting any report, that means its dead
or slow. some time max task may user for different blocks
so, tasktracker will work slowly. So, for that situation
Jobtracker assign task to farway system, because of less
tasktracking will working on, so it will take less time.

Reducer will combine all mapper o/p. And it the
Aetual o/p for 200 MB of data. Reducer will take care
of that. client is directly approach to system
no. 8 and he is going put back to o/p. Before
that, client will directly contact to Namenode
and check the metadata, whether the test file
file is there or not. If it is there then directly connect
to system no. 8. This procedure is called
Mapreduce, and over all it is called
HDFS.

4