omsairam omsrivenkatesayanamaha ammahaganedhipatayanamaha ommamaheshwaryai

HADOOP

1) large storage & fast processing.
2) commodity Hardware
3) cluster

Hadoop → Storage : HDFS (Hadoop file System)
→ Processing : Map Reduce.

→ Streaming Access Pattern :-

HDFS: It is a specially designed file system for storing huge data set with cluster of commodity hardware and with streaming access pattern.

→ Streaming Access Pattern:
: Write Once read any number of times"

→

HDFS:

① normal n/w block size ; 4 KB.
② If 2 KB is used, then remaining 2 KB is wasted

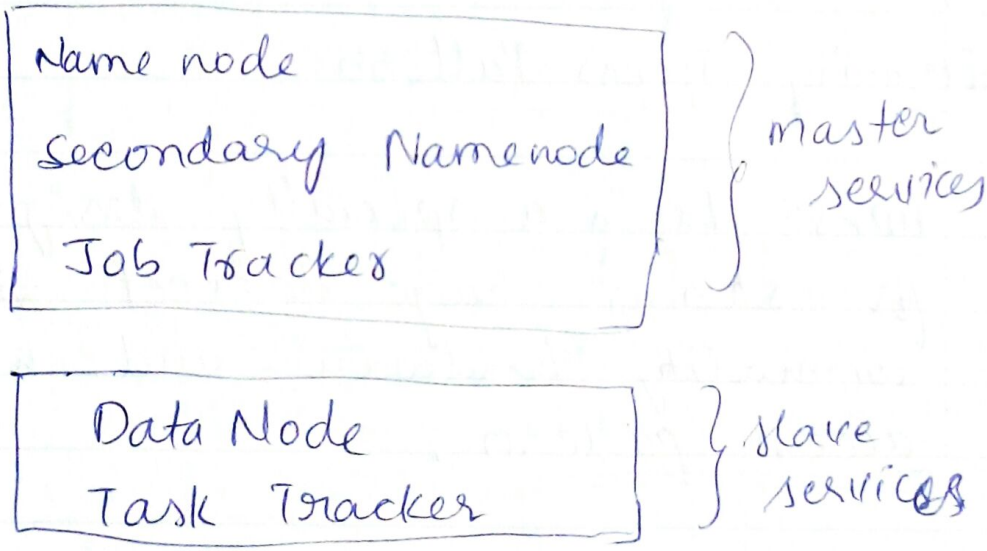
③ To store huge files, in 4 KB blocks & track the blocks, it requires to maintain huge meta-data

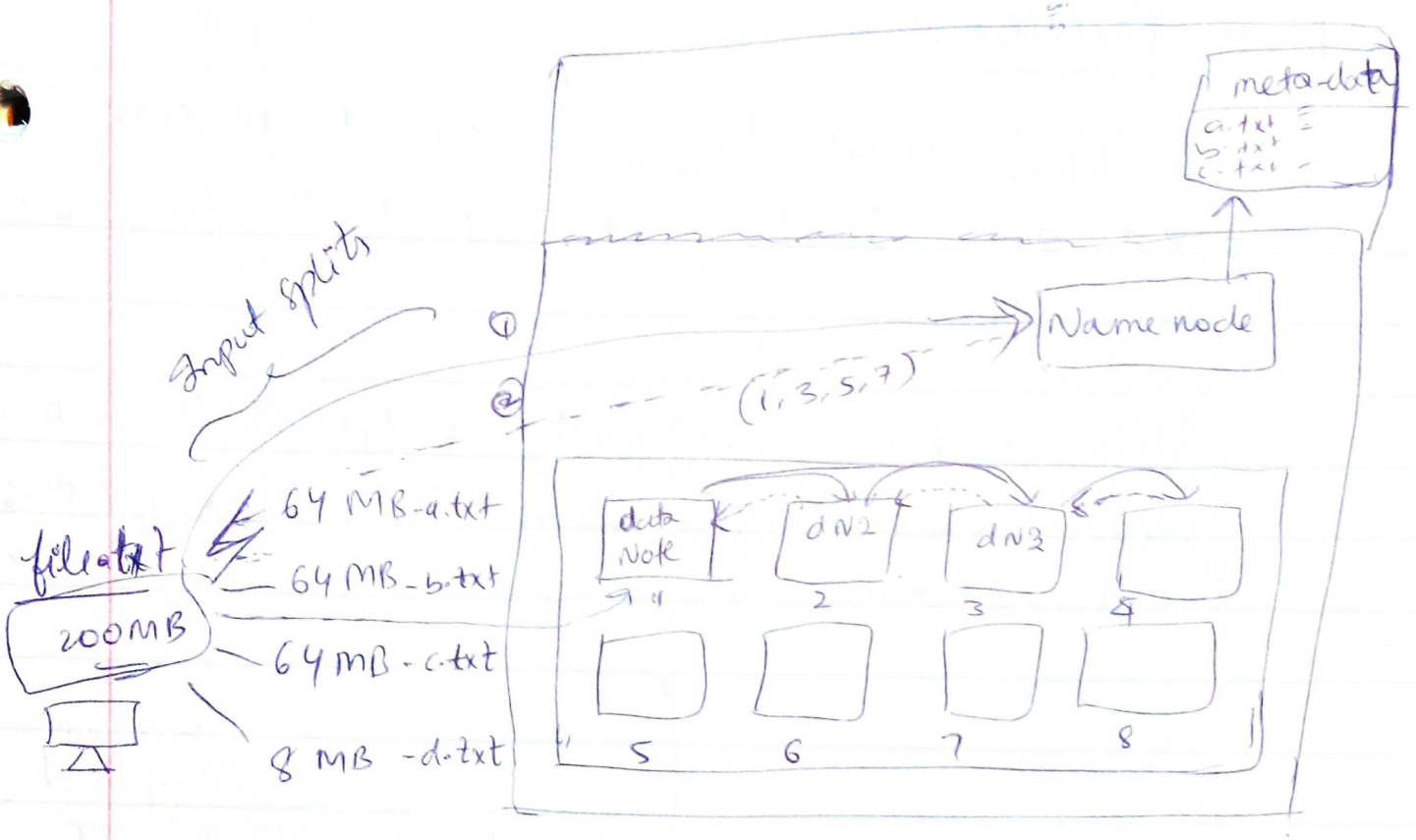
④ But in HDFS, default block size is 64 MB (and it is changeable), & if 10 MB is used in a block, remaining block can be reclaimed. & blocks are less, it creates less meta-data.

⑦

⑧

HDFS





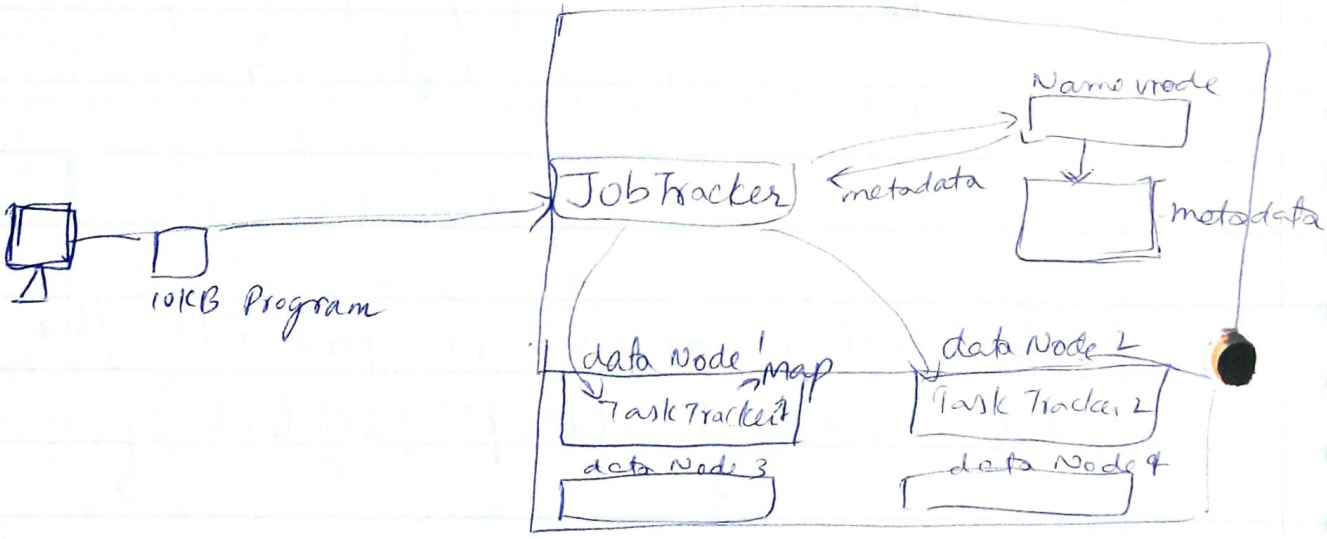
→ "Replication factor is provide the fault tolerance" hadoop system

→ To keep track of storage & running status for every '3' secs data node sends the block report & heart beat to Namenode

Map - Reduce :-

→ We can write a program to process data and give it to "Job Tracker" to process the data.

→ Job Tracker, with the help of Task Tracker runs (executes) program in data nodes and returns the output.



→ Map :- Task Tracker takes the program received from ~~Name node~~ Job Tracker and executes on local machine. It is called as "Map"

→ "Map" will be done on different blocks present at different data Nodes

∴ No of Mappers = # of input splits

REDUCE :-

→ Combining output of all mappers and returned by reducers

→ No of output files = No. of Reducers