Job Post ✓

Brand Building ✓

Hadoop

2 things
1. lots of data
2. Cluster

Hadoop is an open source framework designed to handle massive amounts of data in a scalable and distributed way

⊖ with earlier soln
→ storage limits
→ Processing issue
→ expensive to add more memory

2000

Doug Cutting

GFS → distributed storage → HDFS

MapReduce → distributed data processing → MR

Hadoop can store and process data across many cheap, commodity hardware m/c working together

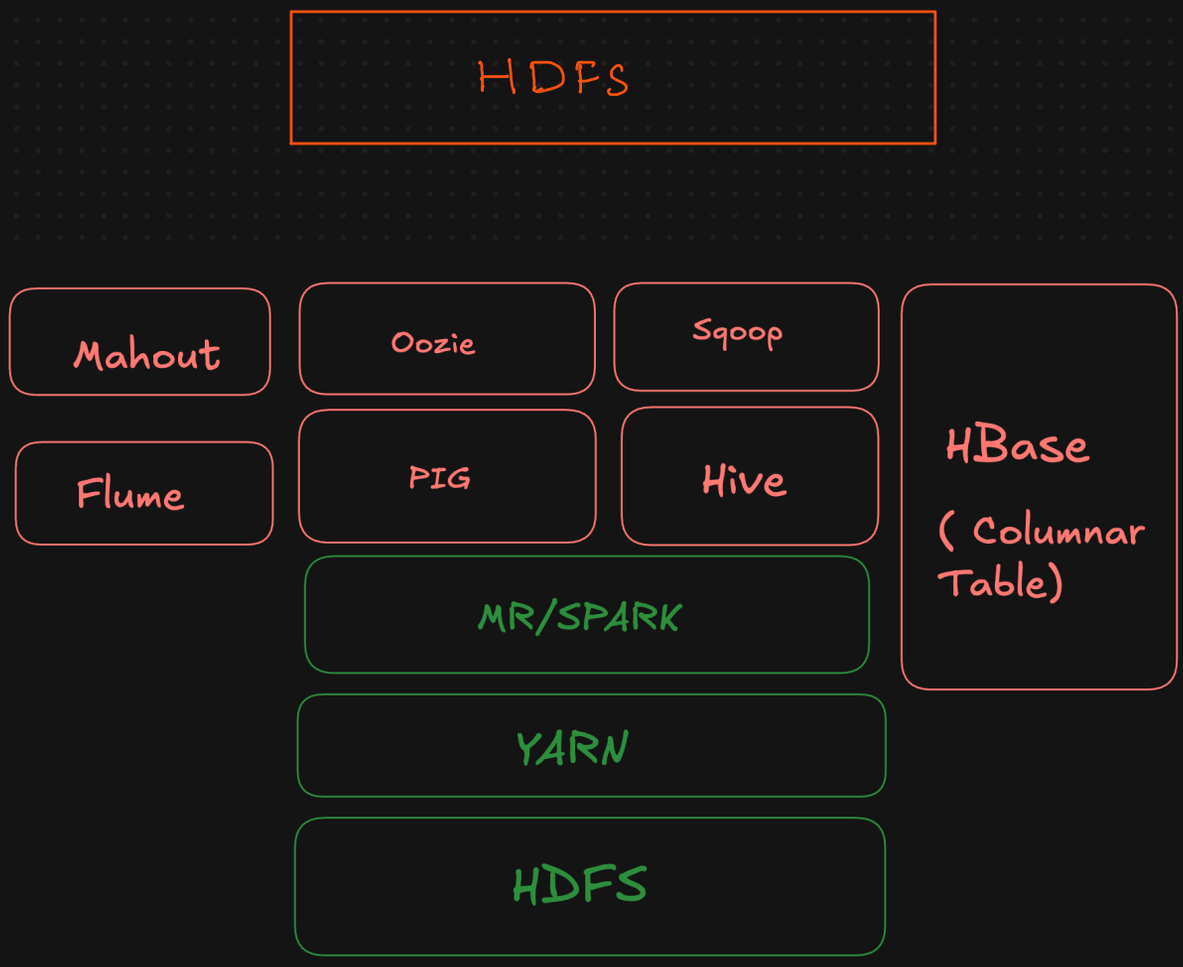
↙ Massive data storage

↘ faster parallel processing

Properties

1. Scalability
2. Fault Tolerance
3. Distributed Processing
4. Cost-effectiveness

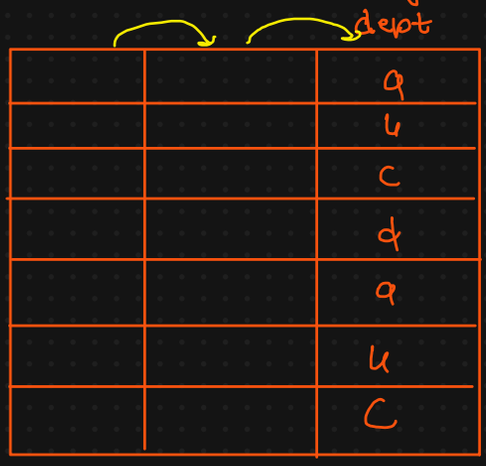
Hadoop Ecosystem / framework



2 imp things

1. Loosely coupled framework :
2. Integration

SQL (write heavy)



+

Database

OLTP

Columnar

Seek time

Data Warehouse

OLAP

Task for students ⇒ Master Basic Linux Command

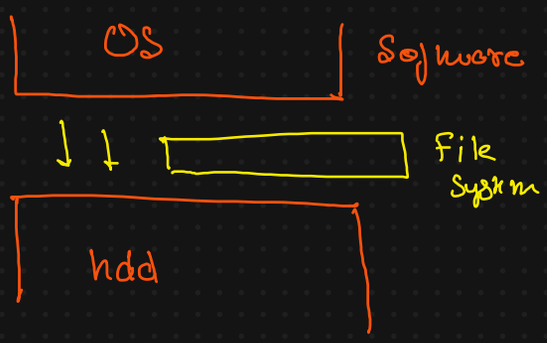
HDFS

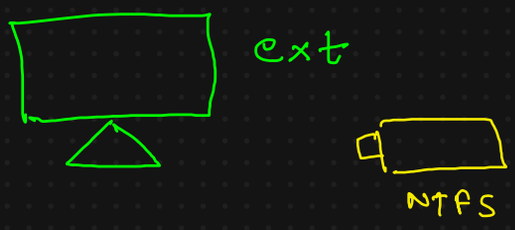
hadoop distributed file system

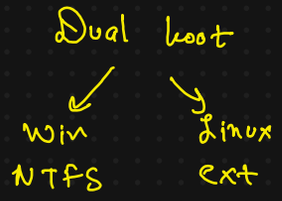
GFS

1. What is a file system?

Windows	NTFS / FAT32
Apple	APFS
Linux	ext







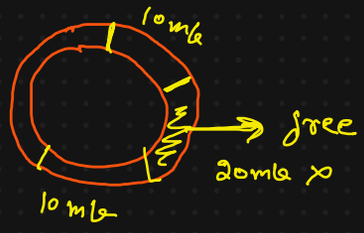
2. What is a block?

Smallest unit of data storage in a file system

Block size ⇒ depends on file system

NTFS → 4kb
Hadoop/HDFS → 128 mb

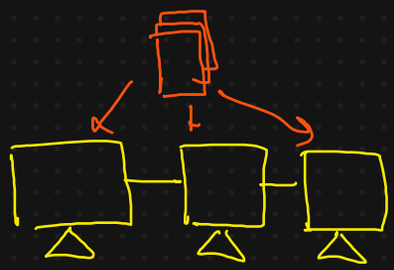
10 kb → 3 block



50 mb

3. Types of file system.

Standalone



Distributed

4. Cluster and Node ?

5. Process & Daemon Process ?

hadoop is a process

JP1
JP2 } HDFS
JP3
JP4
JP5 } MR

6. Metadata

data about data

Image

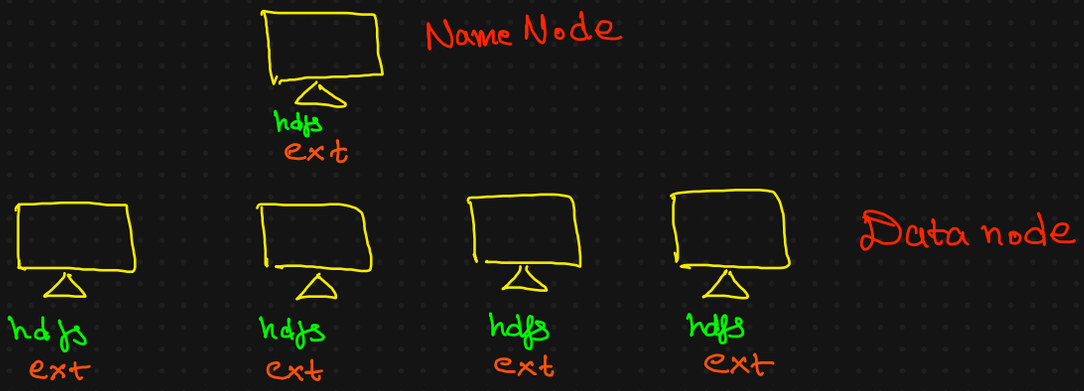
Size
Created at
Pixels
dimensions

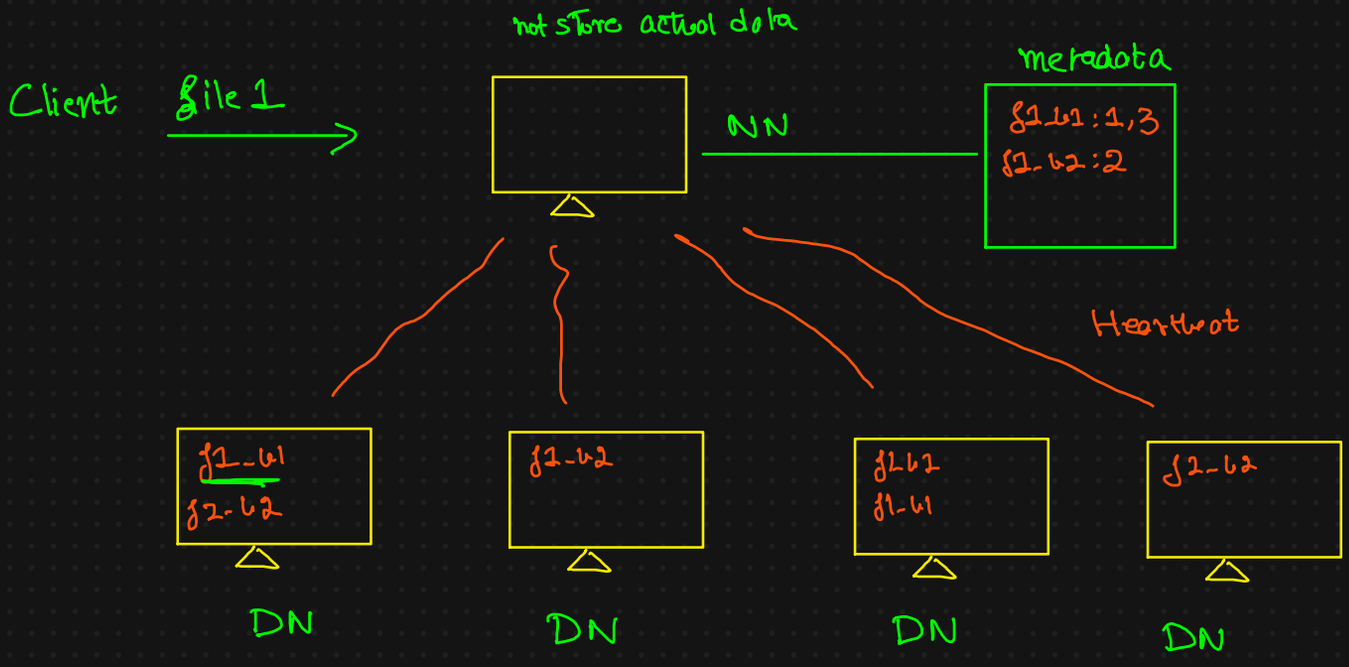
metadata

7. Replication

making copies of data

HDFS

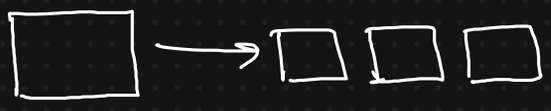




replication factor = 3

lgb

Blocks



128 mb by default size

256 mb → 2 blocks

Bazooka / Gun to kill an aunt

Key Points

1. efficient for large storage file
2. Distributed data

1Kb
64mb
64mb
127mb
1kb * 10 → 10 blocks

⇒ 1 block

Q: Can we inc or dec. block size?

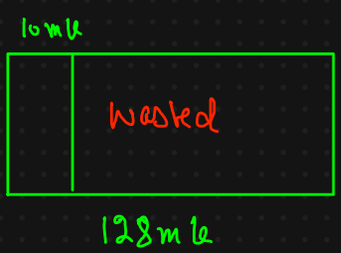
Yes, default is 128mb

1. decrease → 10mb

↑↑ parallelism but also ↑↑ metadata

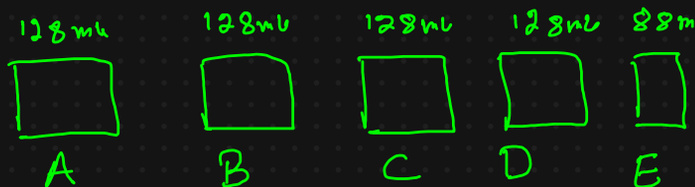
index metadata → Chapters 10 pages

metadata IT → 1 pages



2. increase : parallelism

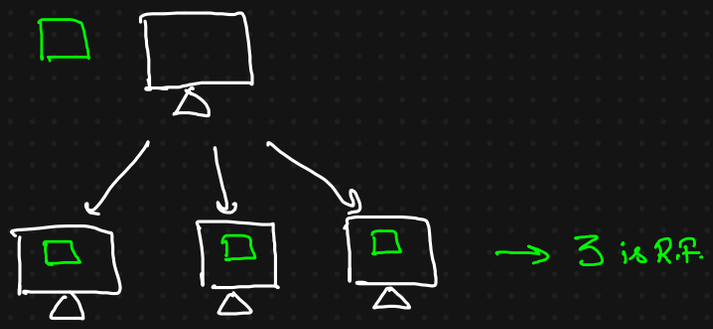
600mb file



900mb file

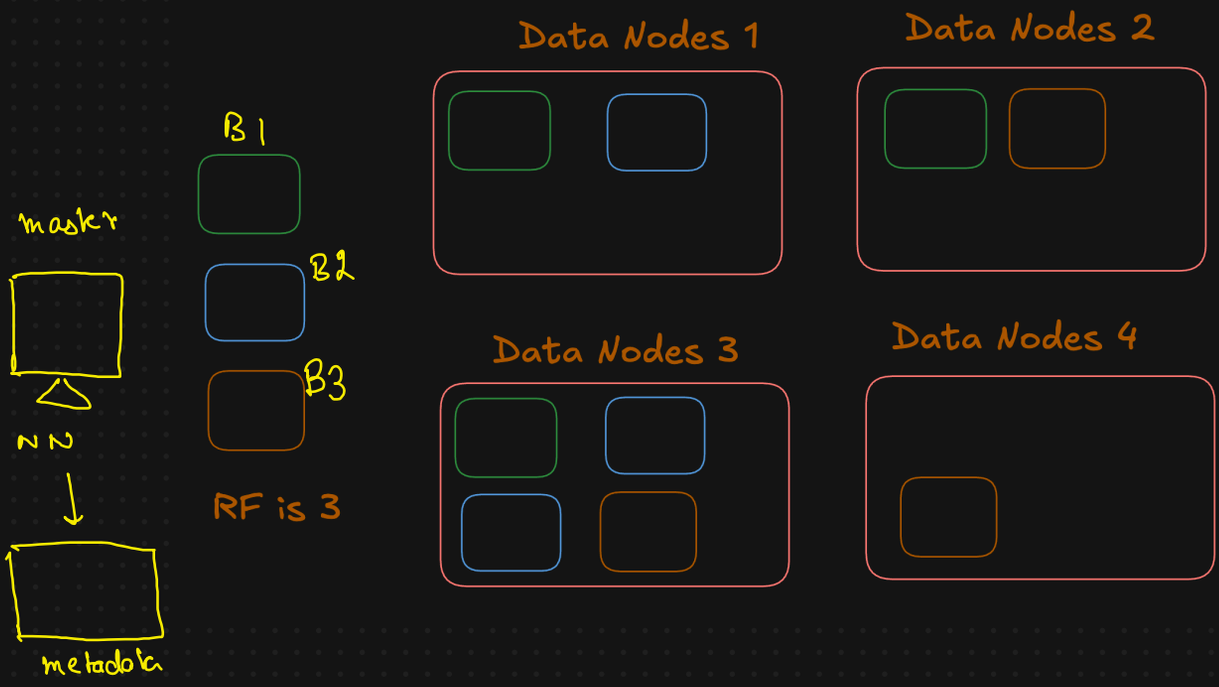
129mb file and 64mb block size → 3 blocks

Data Node failures

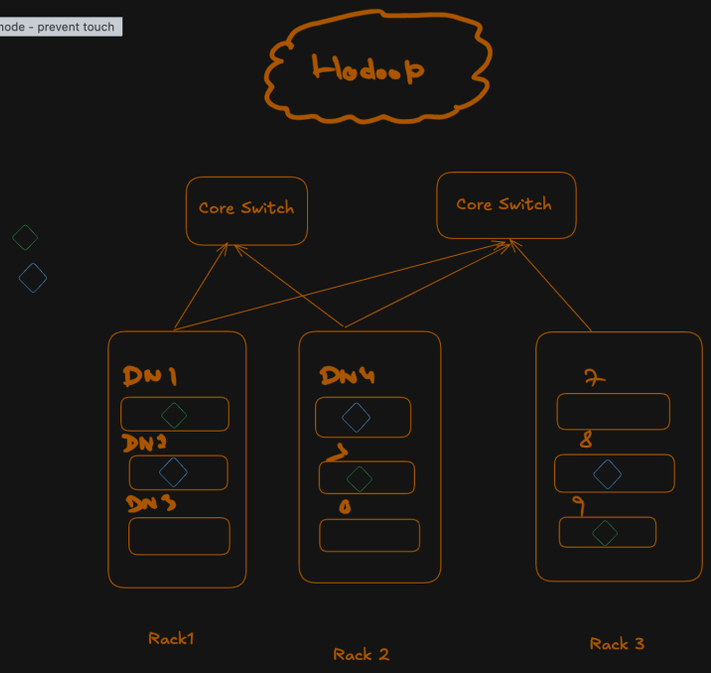


1. Replication

R.F. = 3 ⇒ 1 original
2 copies



Pen mode - prevent touch



1. Fault tolerance
2. Optimized Network Traffic

Q:

1. Temporary failure
2. Permanent failure

   Name node failure

   +

   hdfs Practical
3. YARN + MR

Linux + hdfs) Quick Guide