

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
Caserella

GFS → distributed storage → HDPS

MapReduce → distributed data processing → MR

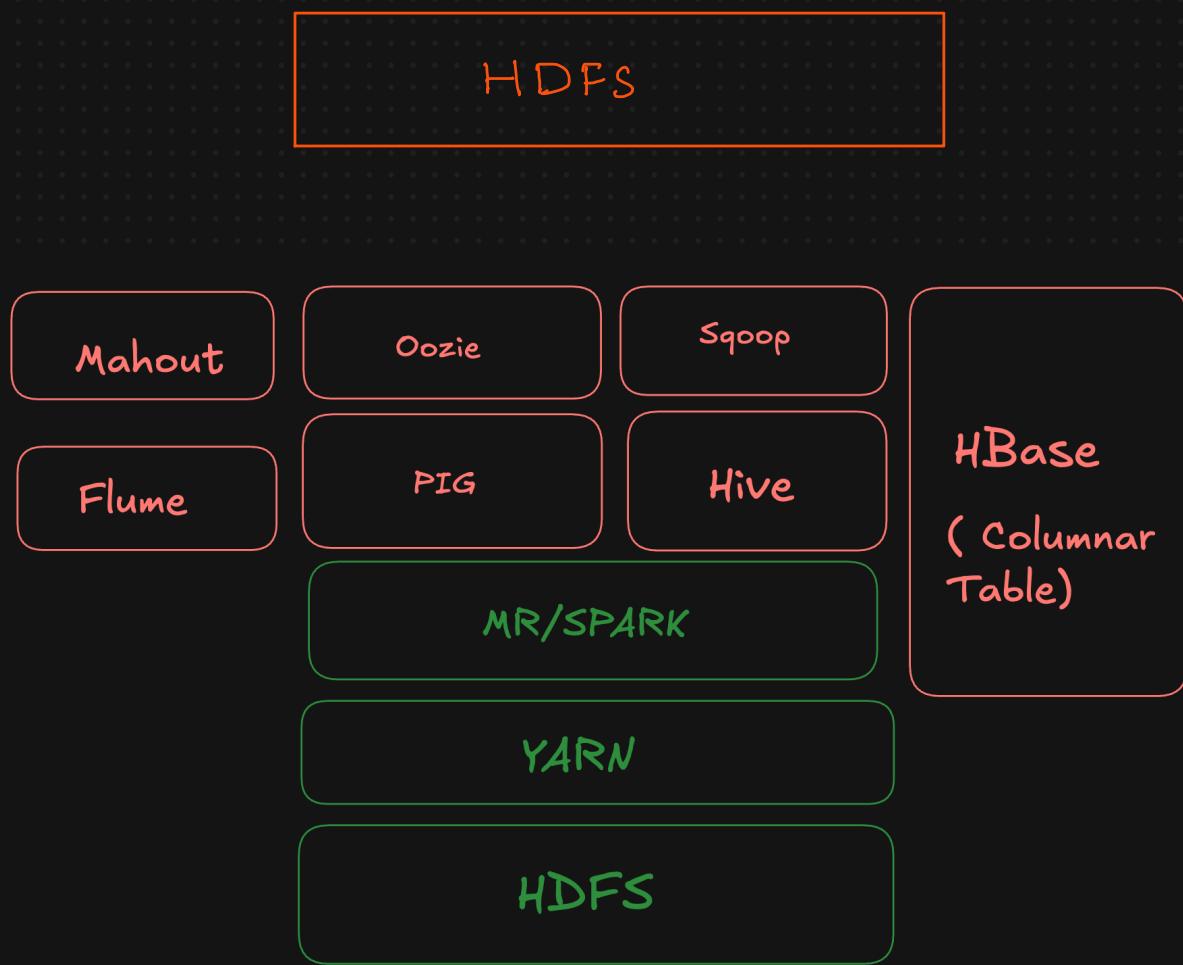
Hadoop can store and process data across many cheap, commodity hardware m/c working together

↓
Massive data storage ↓
↓
Cluster 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)

			a
			b
			c
			d
			a
			b
			c

+

--

Database

OLTP

Columns

200

id

500

name

700

dept

--

--	--	--

		d
--	--	---

Seek time

Data warehouse

OLAP

Task for students

⇒ Master Basic

Linux Command

HDFS

hadoop distributed file system

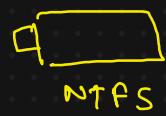
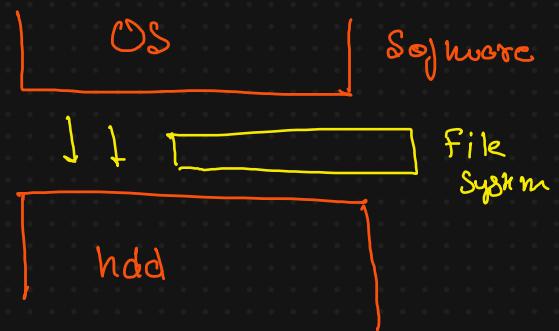
GFS

1. What is a file system?

Windows NTFS / FAT32

Apple APFS

Linux ext



Dual boot
win NTFS linux ext

2. What is a block?

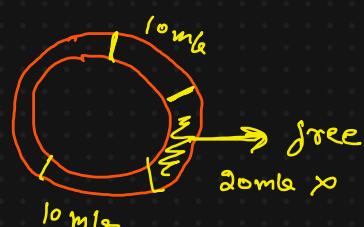
Smallest unit of data storage in a file system

block size \Rightarrow depends on file system

NTFS \rightarrow 4kb

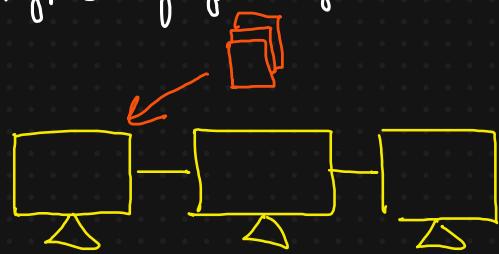
Hadoop / HDFS \rightarrow 128 mb

10 mb \rightarrow 3 blocks

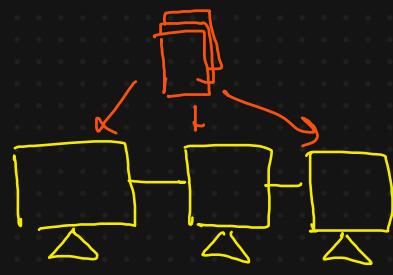


50 mb

3. Types of file system.



Standalone



Distributed

4. Cluster and Node?

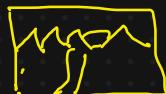
5. Process & Daemon Process?

hadoop is a process

{ JP1
JP2
SP3
JP4
JP5 } HDFS
3 MR

6. Metadata

data about data



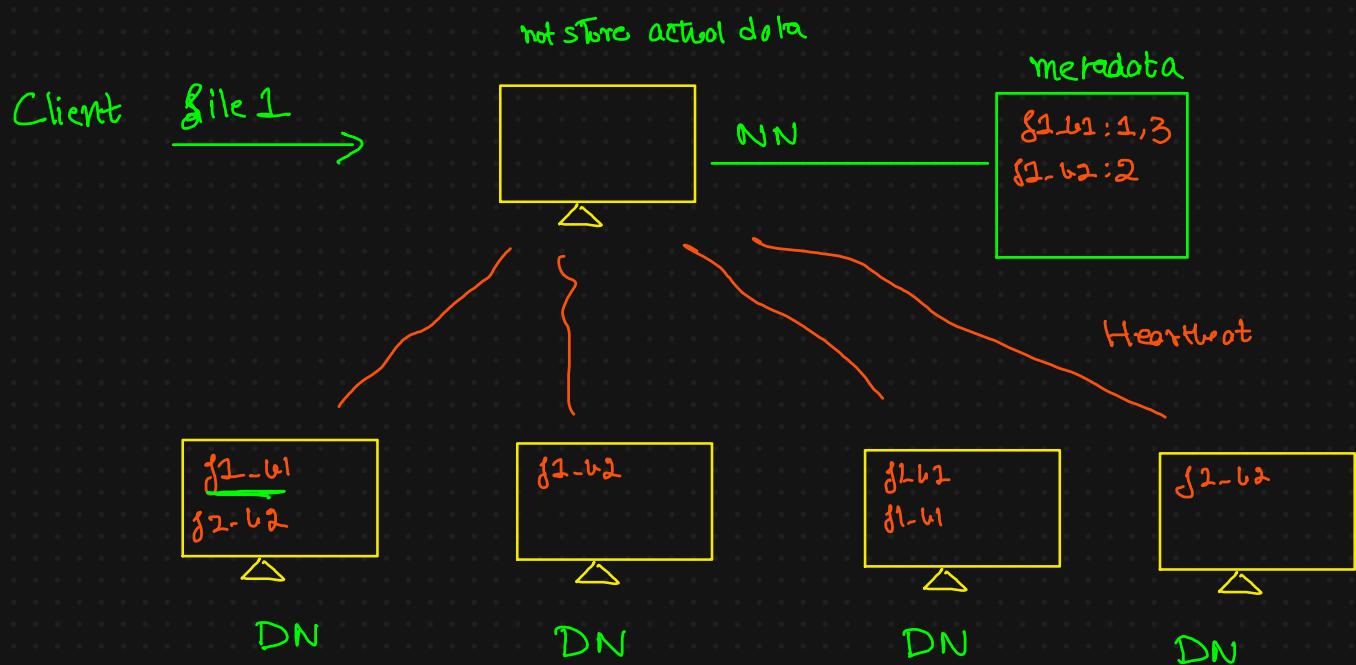
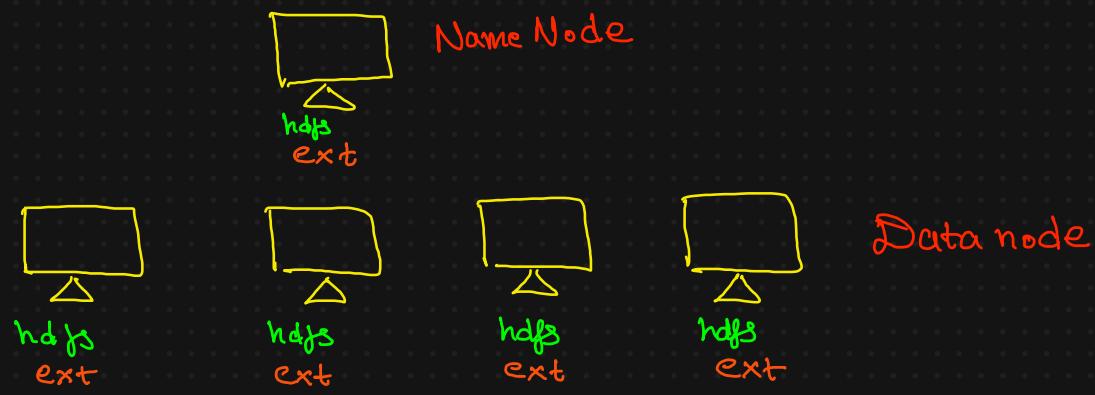
image

size
located at
pixels
dimensions
metadata

7. Replication

making copies of data

HDFS



replication factor = 3
cycle

Blocks



128 mb by default size

256 mb → 2 blocks

Bazaar / Gov

no will on count

1Rb

64mb

→ 1 block

64mb

127mb

1kb = 10 → 10 blocks

Key Points

1. efficient for large storage file
2. Distributed data

Q: Can we inc or dec. block size?

Yes, default is 128 mb



1. decrease → 10 mb

↑↑ parallelism but also ↑↑ metadata



128 mb

index	Chapters
metadata	10 pages
metadata ↑↑	1 pages

2. increase: parallelism



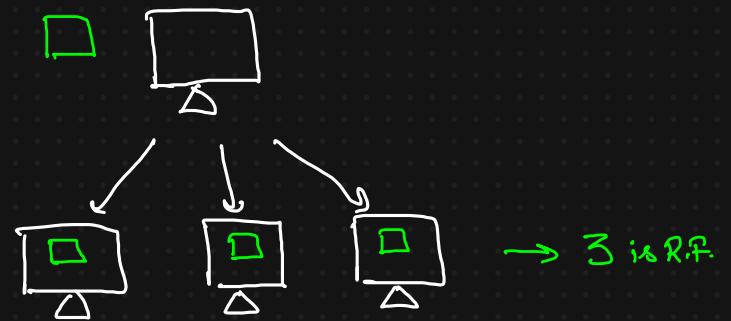
900mb file

129mb file

and 64mb block size

→ 3 blocks

Data Node Failures



1. Replication

R.P. = 3 \Rightarrow 1 original
2 copies

