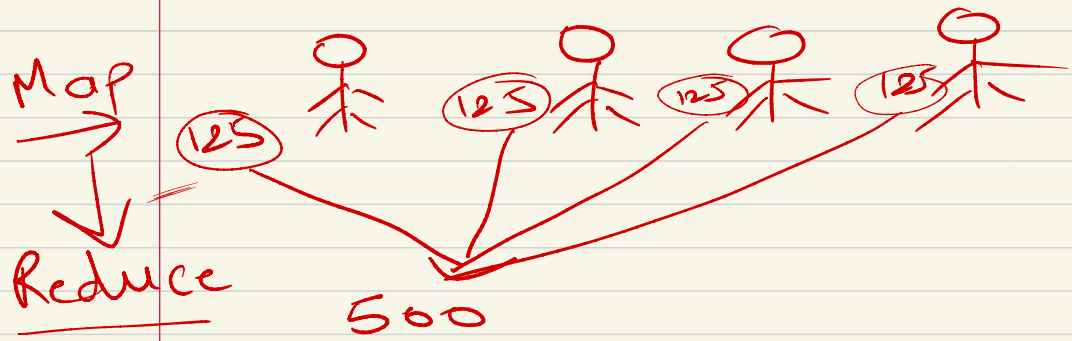
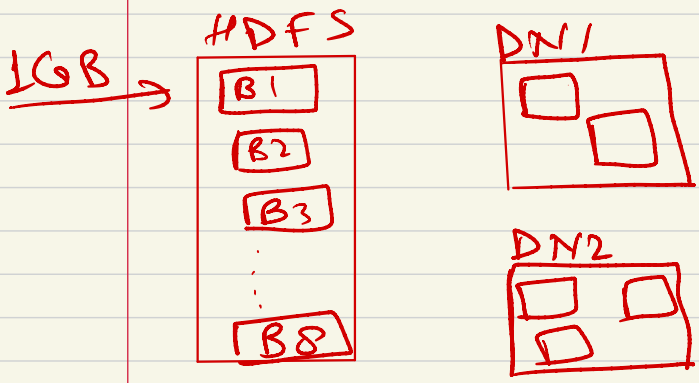
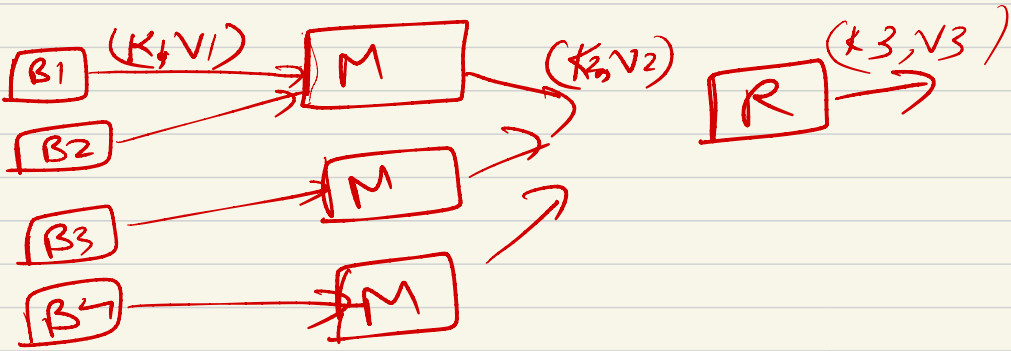
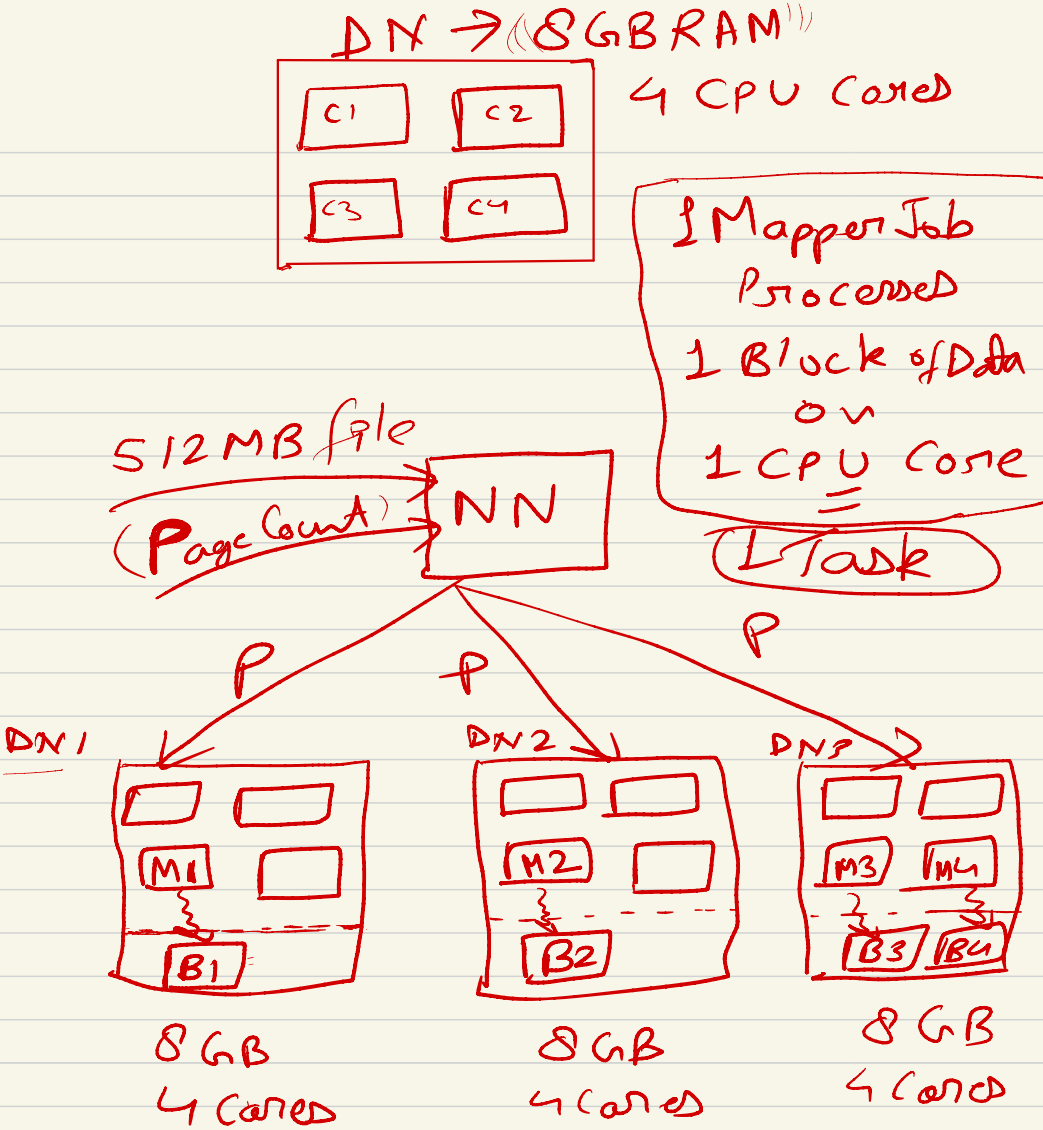
500 pages


Map
Reduce
125
125
125
125
500


1GB
HDFS
B1
B2
B3
B8
DN1
DN2

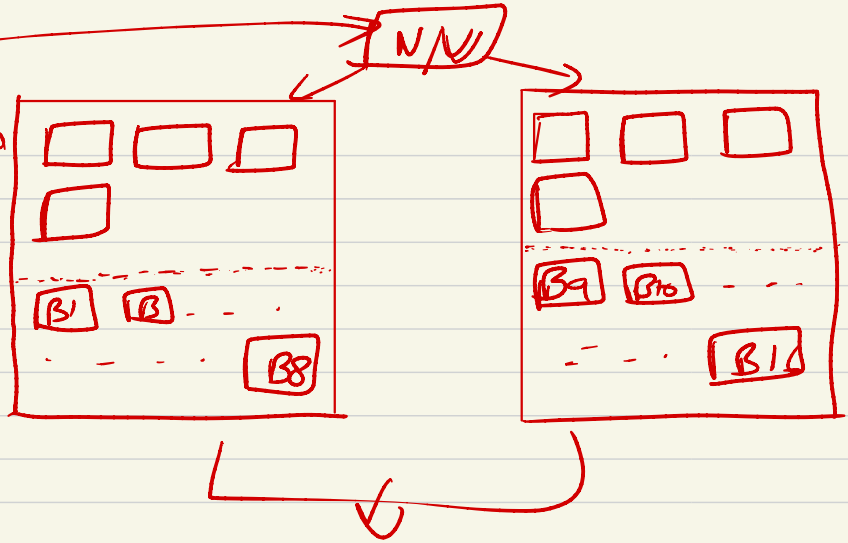

B1
B2
B3
B4
(K1,V1)
M
M
M
(K2,V2)
R
(K3,V3)


DN → (8GB RAM)
C1
C2
C3
C4
4 CPU Cores
1 Mapper Job
Processes
1 Block of Data
on
1 CPU Core
1 Task
512MB file
(Page Count)
NN
P
P
P
DN1
DN2
DN3
M1
M2
M3
M4
B1
B2
B3
B4
8GB
4 Cores
8GB
4 Cores
8GB
4 Cores

2 GB

16 Block

128 MB



1 Batch ⇒ 8 Mapper Job
8 Data Block

2 Batch ⇒ 8 Mapper Job
8 Data Block

32 MB

128M

512 MB

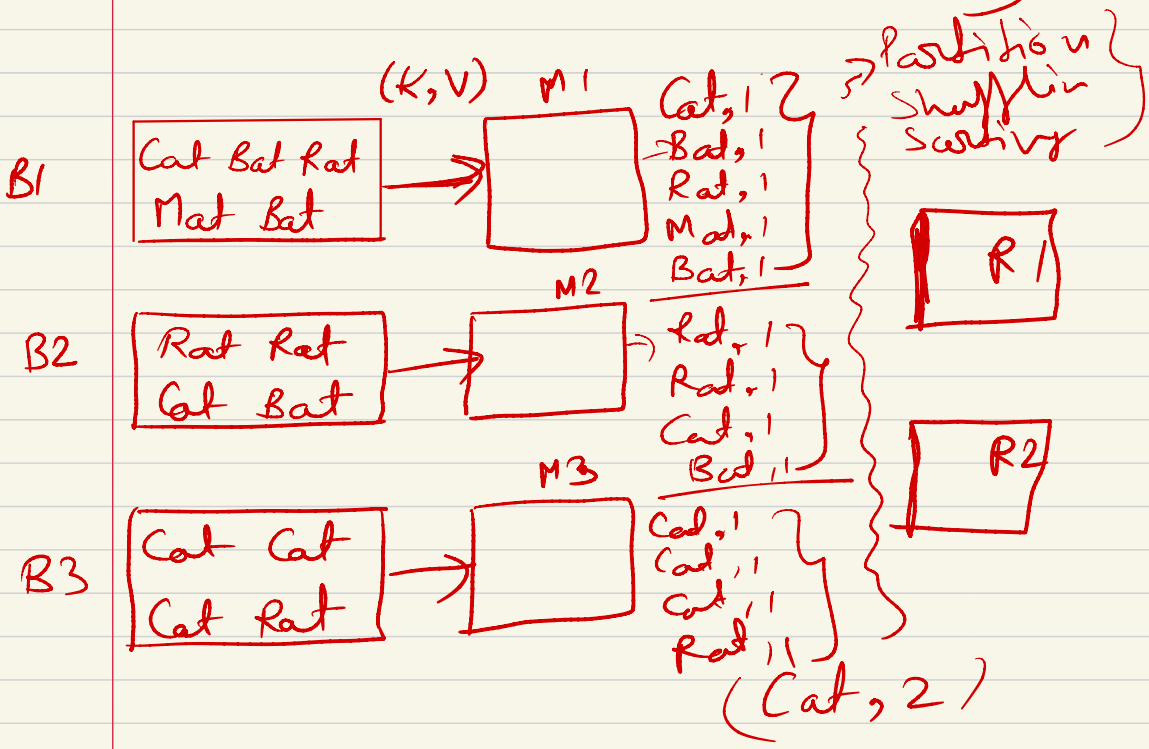
What if we decrease?

What if we increase it?

Increasing number

(Cat, 5)
(Bat, 3)

Word Count → (Mat, 1) (Rat, 4)



line1 → Cat_Rat_Rat → Split(_)

↓

([Cat] [Rat] [Rat])

[Cat, 1] [Rat, 1] [Rat, 1]

line2 → Cat, Bat → Split()

↓

([Cat] [Bat])

[Cat, 1] [Bat, 1]

Map()

Intermediate

(Cat, 1) (Rat, 1) (Rat, 1)

(Cat, 1) (Bat, 1)

~~something~~

(Cat, [1, 1]) → (Cat, 2)

(Rat, [1, 1]) → (Rat, 2)

(Bat, [1]) → (Bat, 1)

Reducer

(Cat Rat Bat)

MM

→ Value

→ Memory Address with Key

Combine

DN1

M1

Cat, 1
Cat, 1
Rat, 1
Bat, 1

→ Cat, [1, 1]
Rat, [1]
Bat, [1]

DN4

R1

Bat, [1, 1]
Bat [1]
Bat, [1, 1]

DN2

M2

Rat, 1
Cat, 1

Rat [1]
Cat [1]

DN3

M3

Bat, 1
Cat, 1

Bat, [1]
Cat, [1]

DN5

R2

Cat, [1, 1]
Rat, [1]
Rat, [1]
Cat, [1]
Cat [1]

Shuffle Sort

Cat, Rat, Bat

Cat → [1, 1, 1, 1]
Rat → [1, 1]

Partitioner

hash_function = len(Key) + Rando(1,2) % no. of Reducer

Cat $\Rightarrow f(x) = (\text{len}(\text{Cat}) + 2) \% 2$

$= (3+2) \% 2$

$= 5 \% 2$

$= 1$

Bat $\Rightarrow f(x) = (3+1) \% 2$

$= 4 \% 2$

$= 0$

Rat $= f(x) = (3+2) \% 2$

$= 5 \% 2$

$= 1$

write

$\boxed{R_1}$ (Bat, 2)

write

R2 → (Cat, 4)
(Rat, 2)

(Bat, 2)
(Cat, 4)
(Rat, 2)

/output/file1
file2

B1 (128 MB) B2 (128 MB)

(Cat, Bat, Rat) (Cat, Rat) (Bat, Rat)

Line1 Line2 Line3

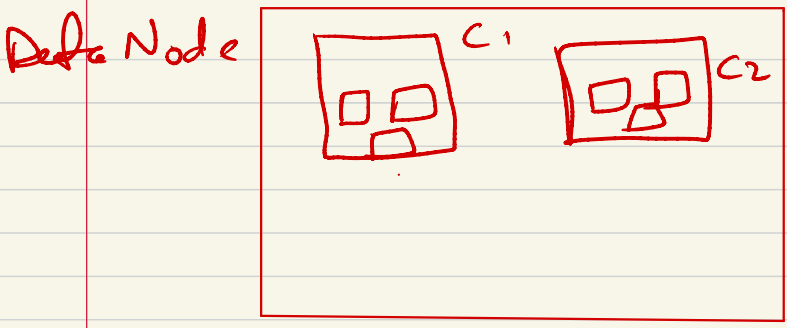
M1

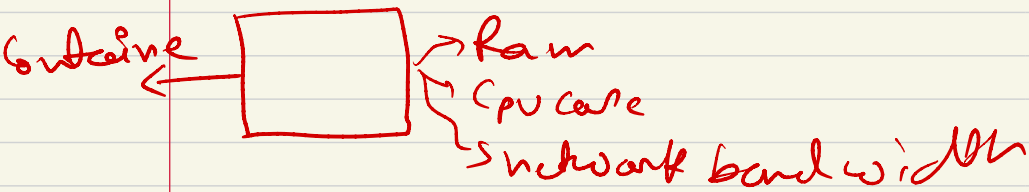
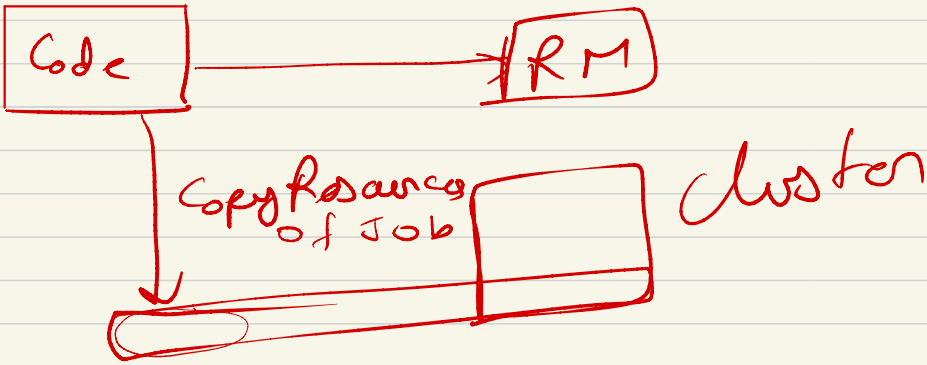
Slave

DataNode Daemon

Node Manager

64GB RAM, 16 CPU Cores





6 CPU Cores for the computer