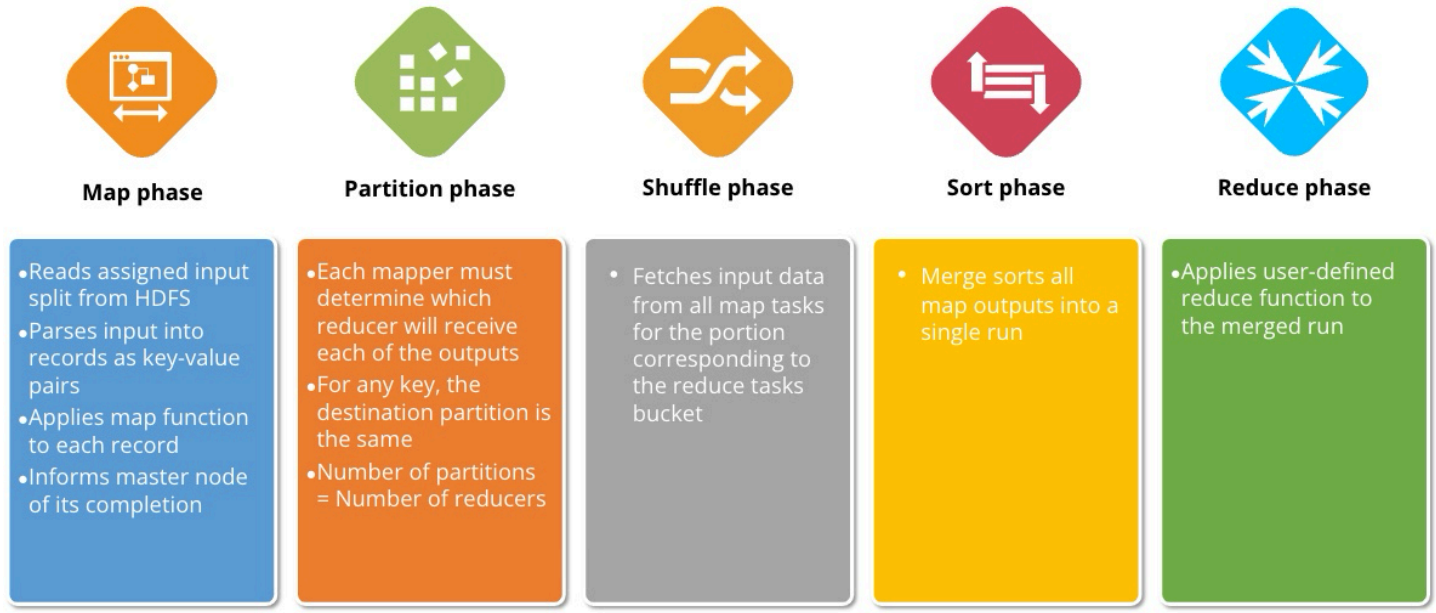
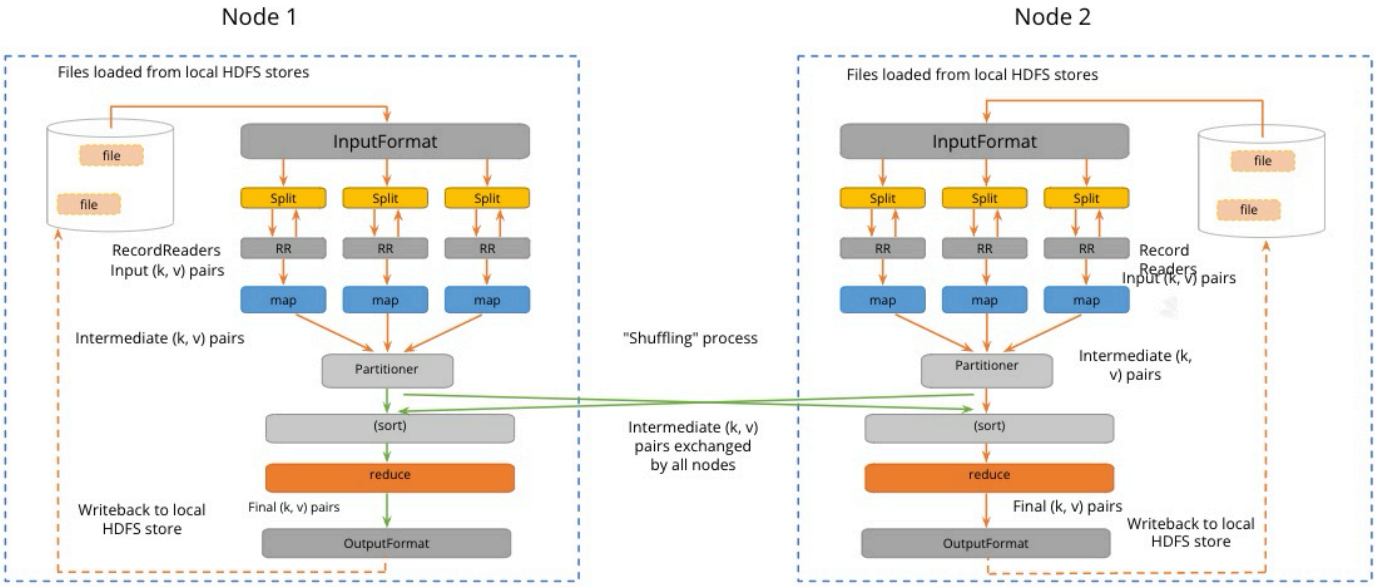
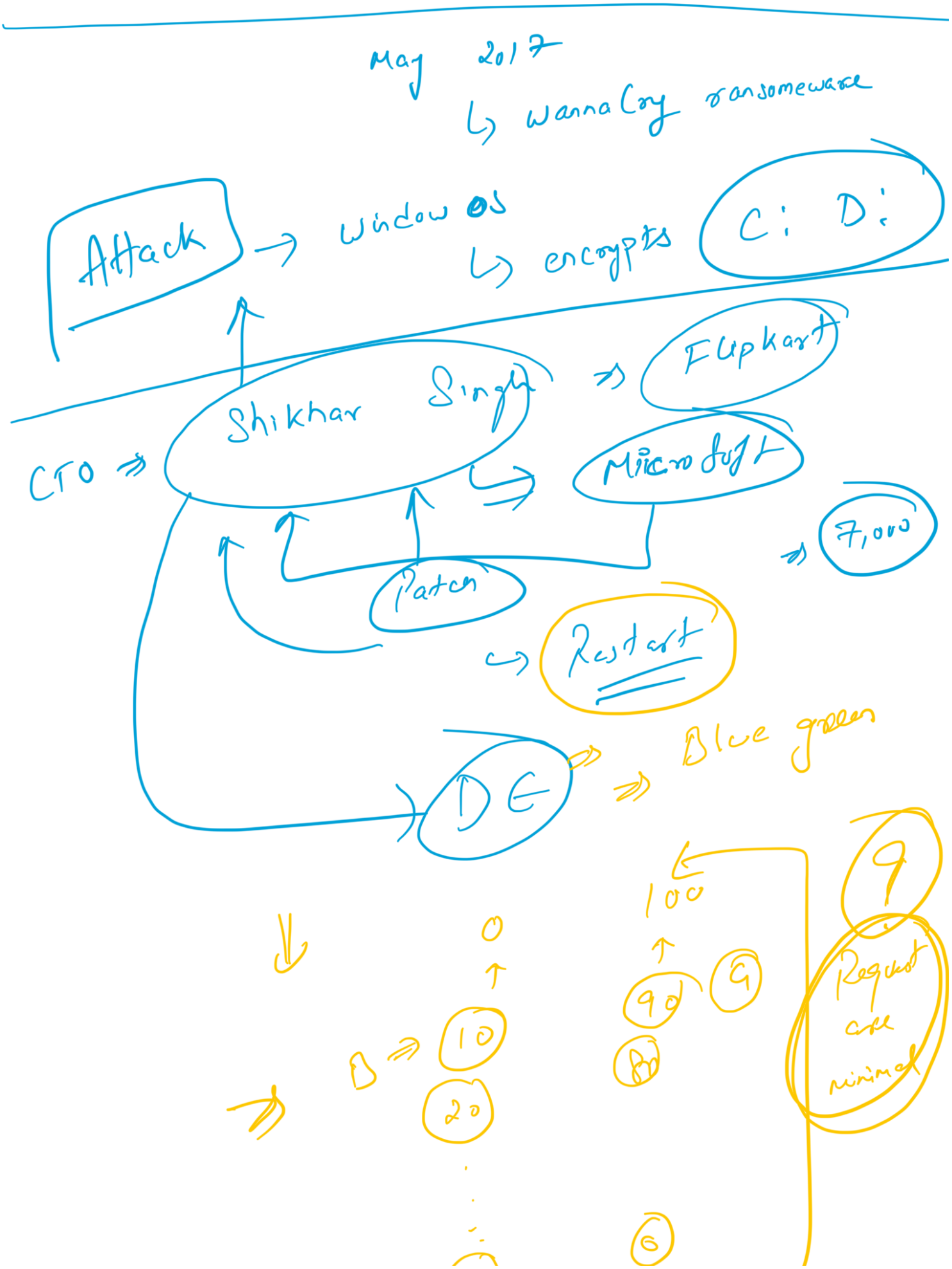
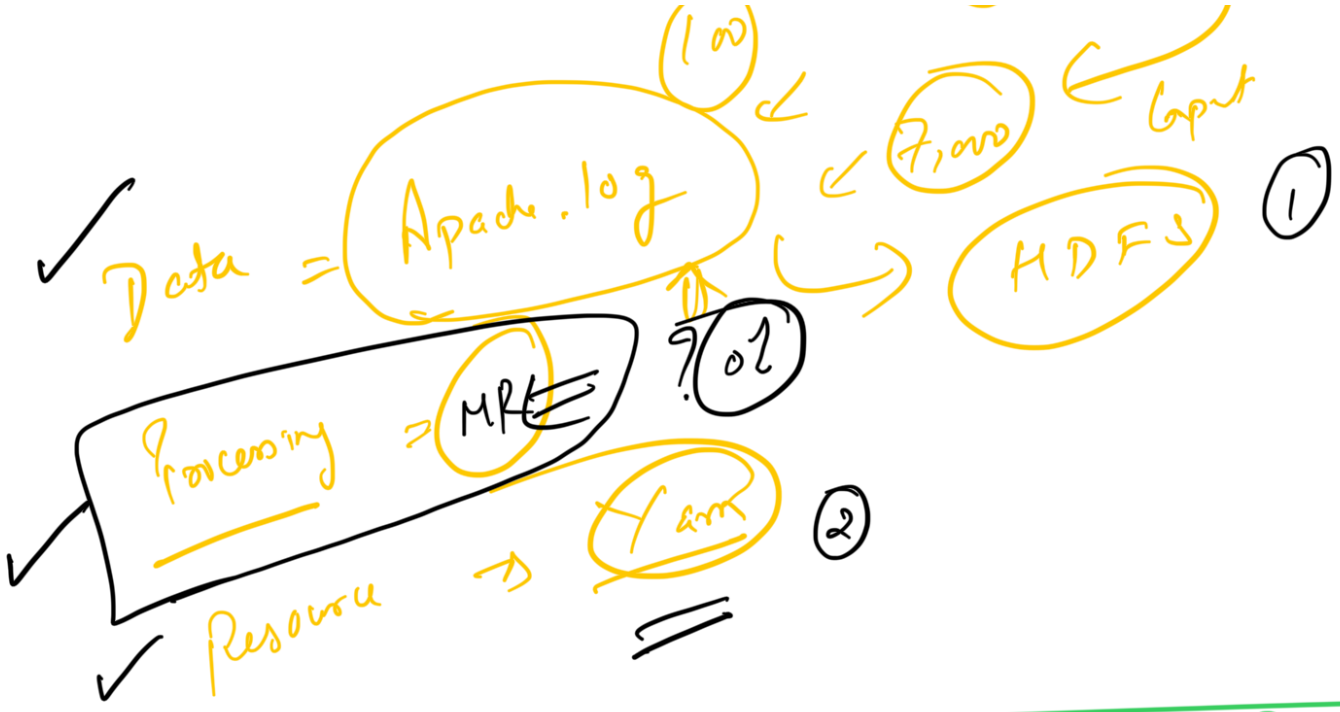
## Agenda

Learn the architecture of YARN and its key components.
Understand how MapReduce tasks are scheduled and executed using YARN.
Explore the communication between HDFS and YARN.
Work through examples to solidify these concepts.

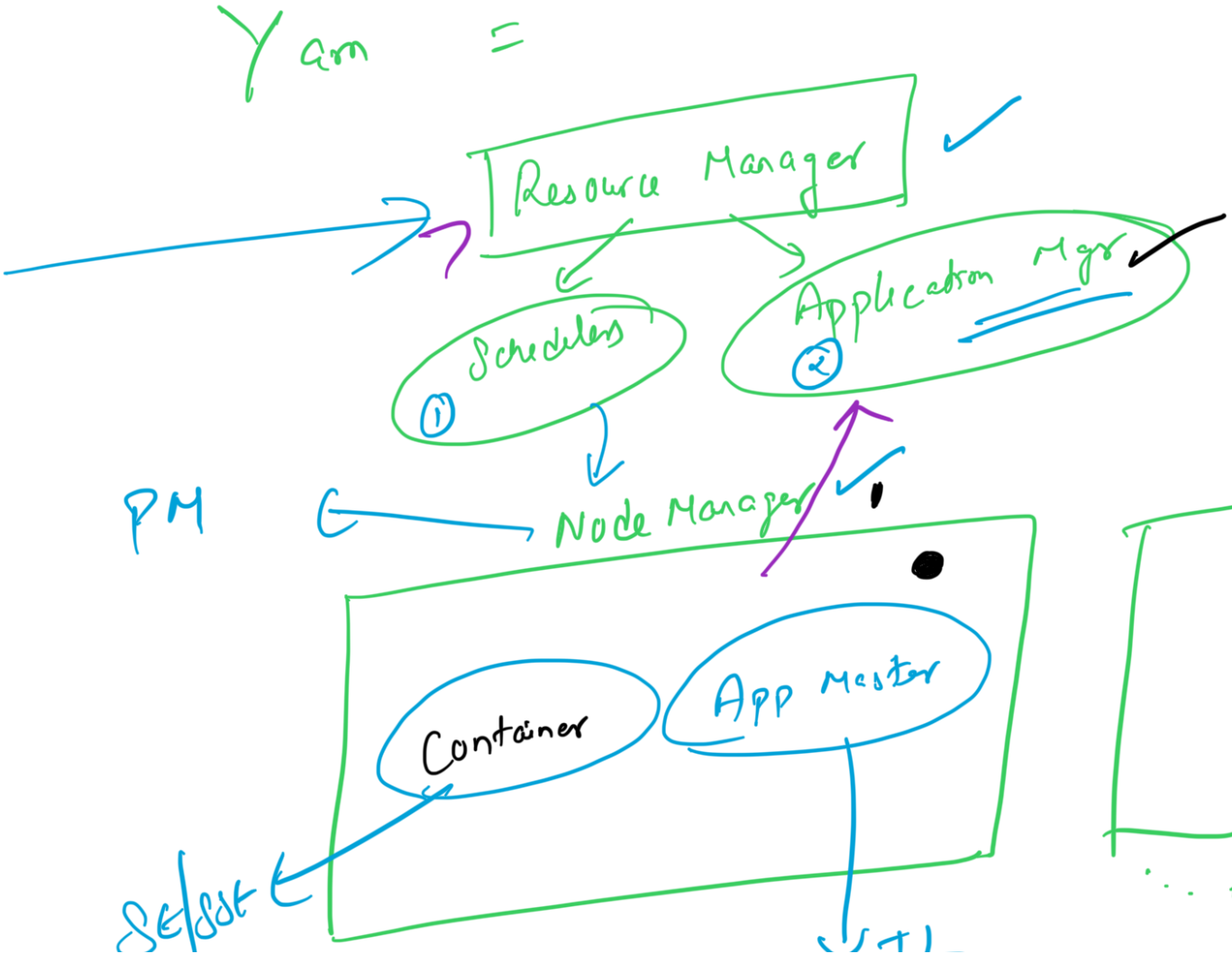




Node 1
Node 2
Files loaded from local HDFS stores
Files loaded from local HDFS stores
file
file
file
file
InputFormat
InputFormat
Split
Split
Split
Split
Split
Split
RR
RR
RR
RR
RR
RR
map
map
map
map
map
map
RecordReaders
Input (k, v) pairs
Record
Readers
Input (k, v) pairs
Intermediate (k, v) pairs
"Shuffling" process
Partitioner
Partitioner
Intermediate (k,
v) pairs
(sort)
(sort)
Intermediate (k, v)
pairs exchanged
by all nodes
reduce
reduce
Writeback to local
HDFS store
Final (k, v) pairs
Final (k, v) pairs
OutputFormat
OutputFormat
Writeback to local
HDFS store


May 2017
↳ WannaCry ransomeware
Attack
Window OS
↳ encrypts
C: D:
CTO
Shikhar Singh
Flipkart
Microsoft
7,000
Patch
Restart
DE
Blue green
100
90
9
10
20
8
6
Request can minimal

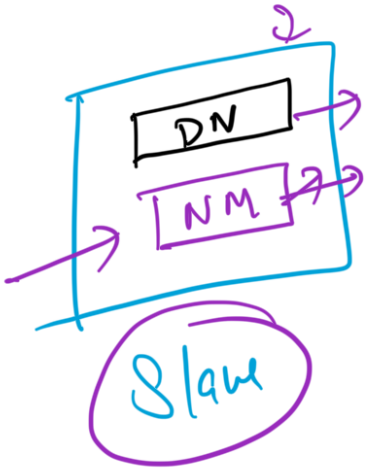

Data =
Apache.log
(100)
Flume
Input
HDFS
1
Processing =
MR
(01)
Yarn
2
Resource


HDFS = Apache.log
B1
B2
B3


Yarn =
Resource Manager
Scheduler
1
Application Mgr
2
PM
Node Manager
Container
App Master
Set/Get
. . n

① Node Manager (NM)

↳ runs on each node and manages ✓



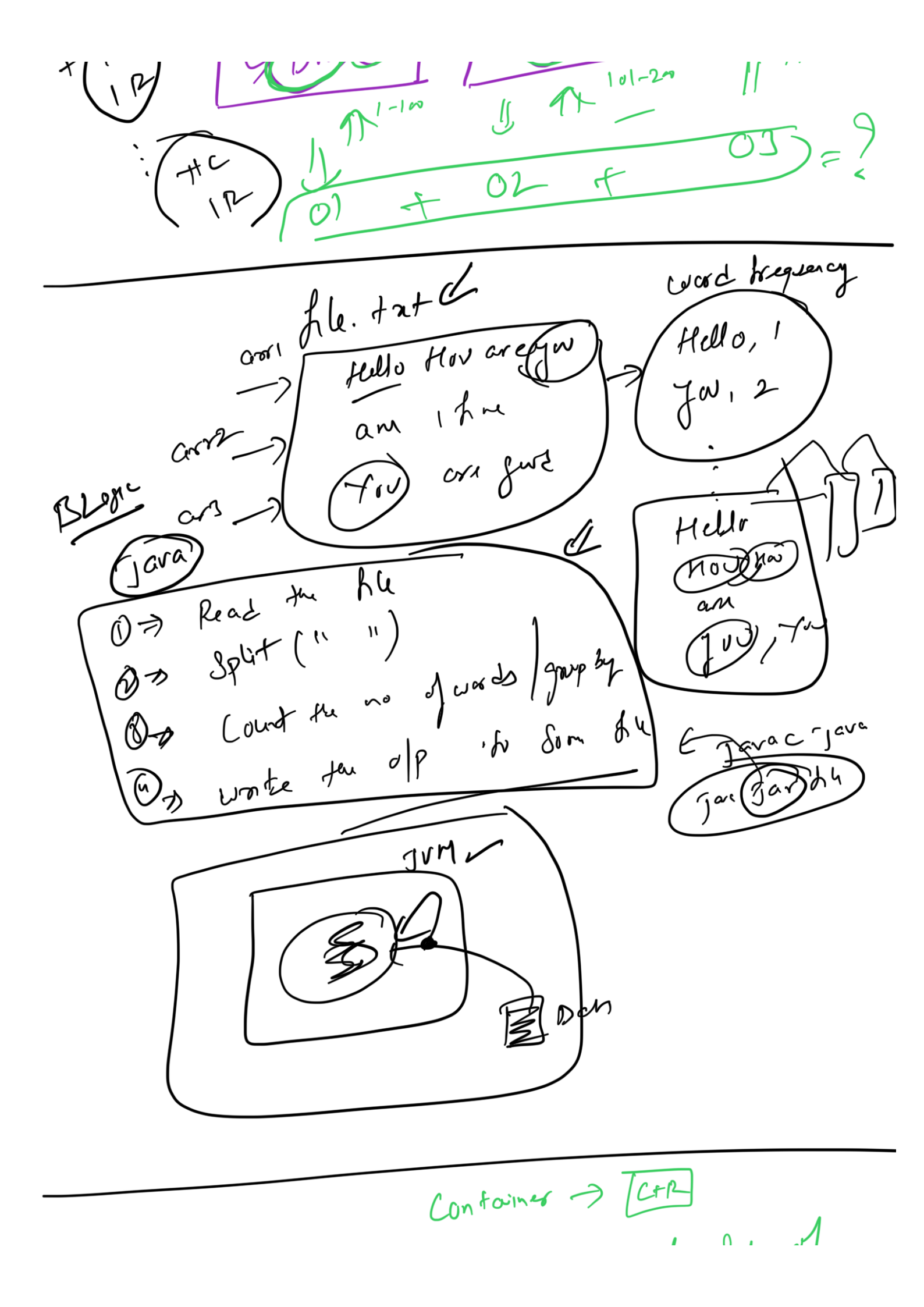

1. Container lifecycle
2. Container dependencies
3. Container leases
4. Node health | Log mgmt
5. Reporting for node status to RM.

R+C

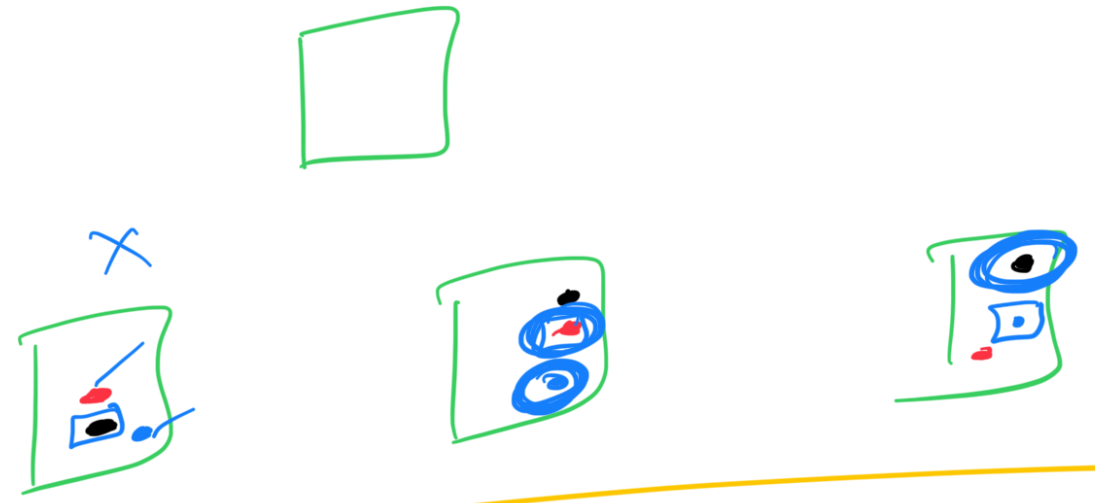
② Application Master :-

↳ this is your distributed job.

.jar / .py / .exe ⇒ is a framework specific library, which negotiates resources from RM and works with NM to execute it self with the help of Containers.


file.txt
word frequency
Hello, 1
Jw, 2
Hello How are you
am I fine
You are sure
arg1
arg2
arg3
BLogic
Java
Read the file
Split (" ")
Count the no of words / group by
Write the o/p in some file
Hello
How are you
am
Jw, You
Javac Java
Jar Jar file
JVM
Container → CLR

↳ a collection of set of resources (CPU + RAM) in certain numbers on a specific node.



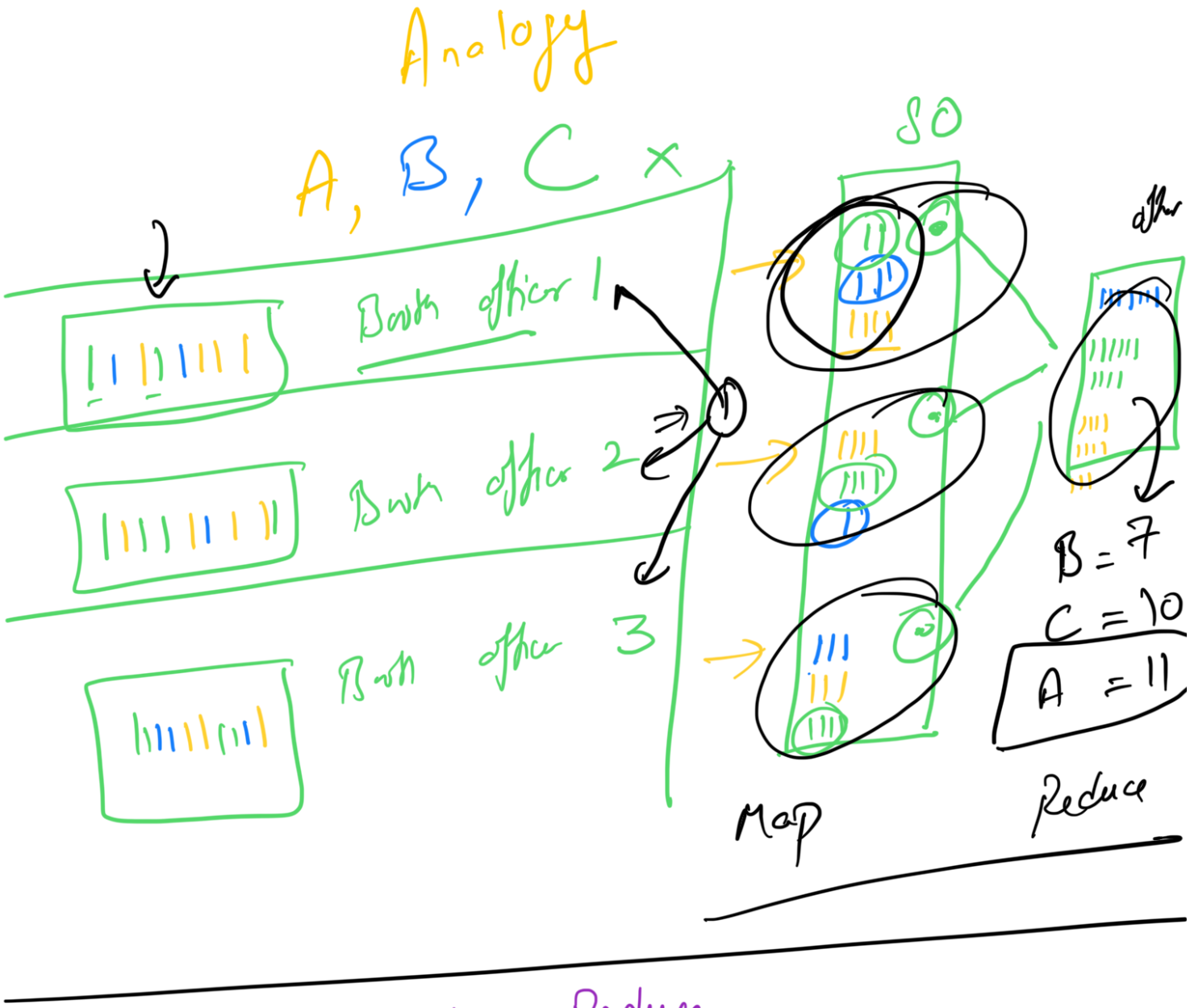
Job !

⇒ Distributed Processing

⇒ Map Reduce

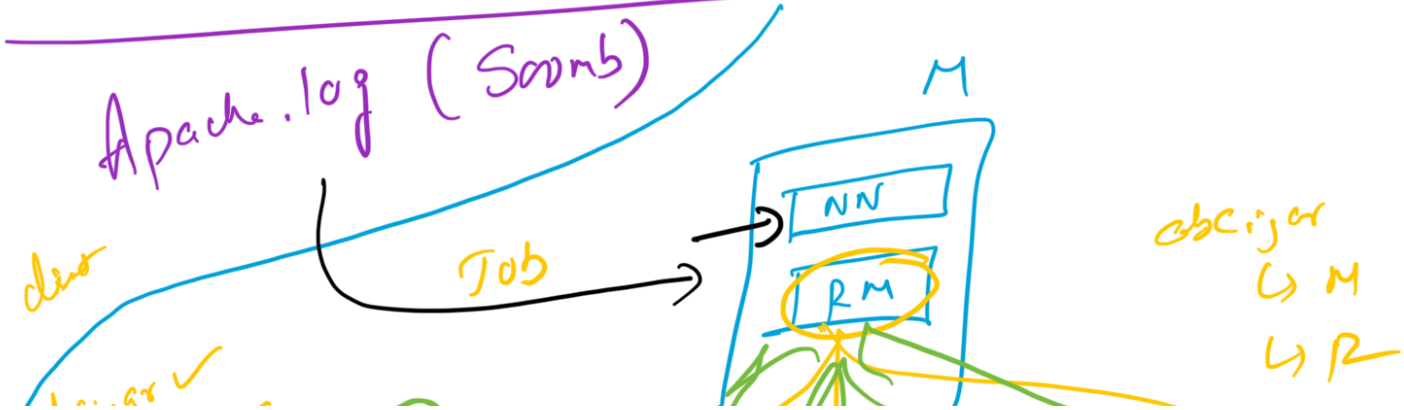
↳ is a programming model that simultaneously process and analyzes huge data sets logically into cluster.

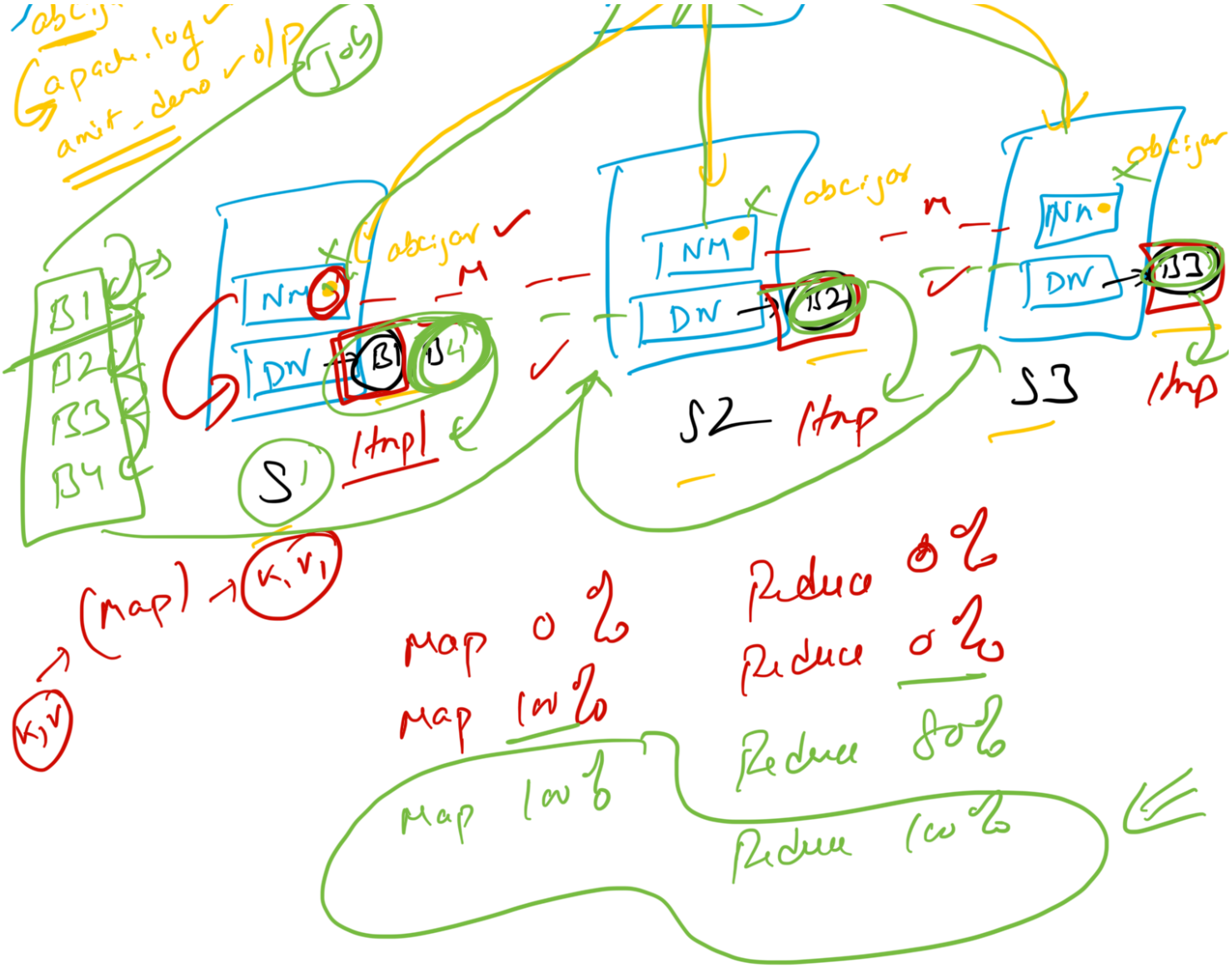
↳ while Map Reduce sorts the data, segregates the data




Map Reduce

? Phases




apache.log
amit_demo
o/p
job
B1
B2
B3
B4
NM
DN
B1
B4
S1
Impl
S2
Impl
B2
S3
Impl
B3
(map)
K,V
K,V
map 0%
map 10%
map 100%
Reduce 0%
Reduce 0%
Reduce 80%
Reduce 100%