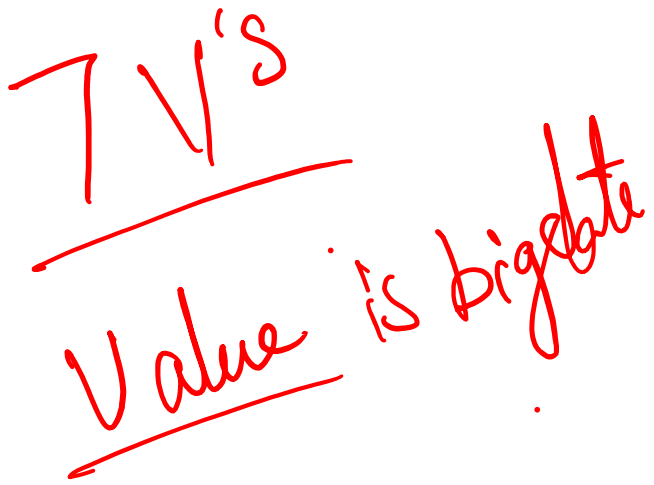
Why Big Data now?

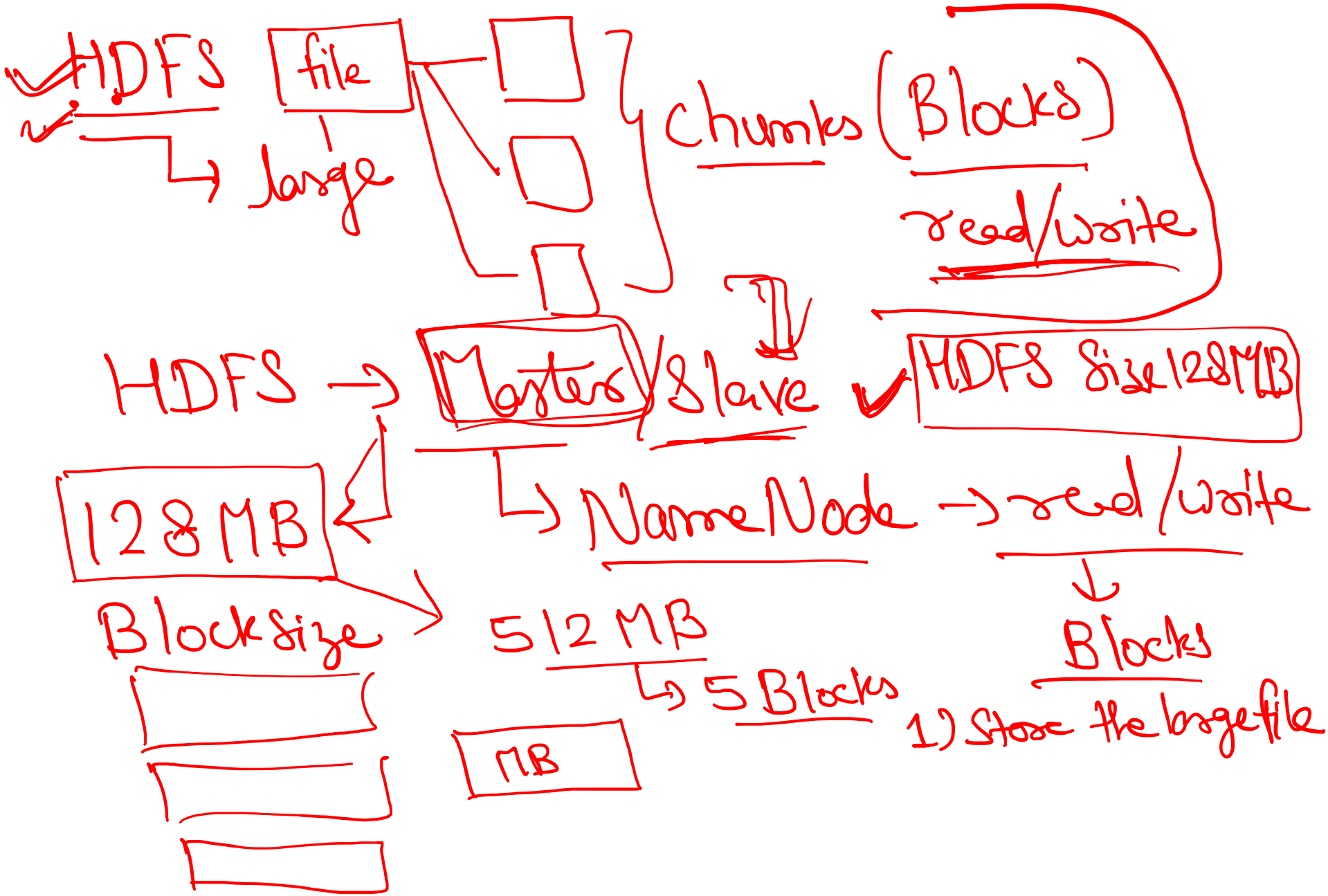
- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

→
- High-Volume
- High-Velocity
- High-Variety

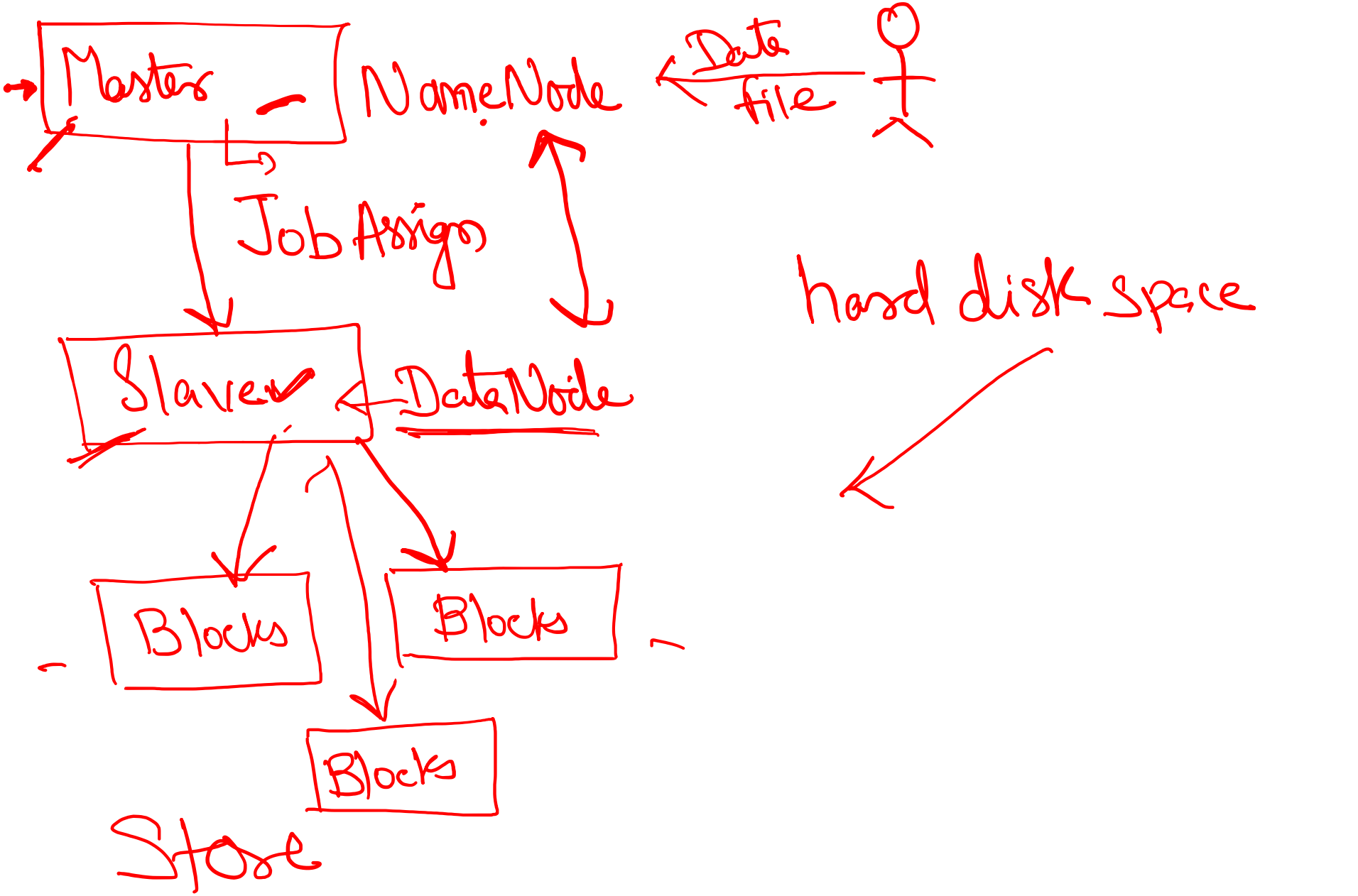


→ Artificial Intelligence

Hadoop Blocks – Namenode — Data Blocks ✓

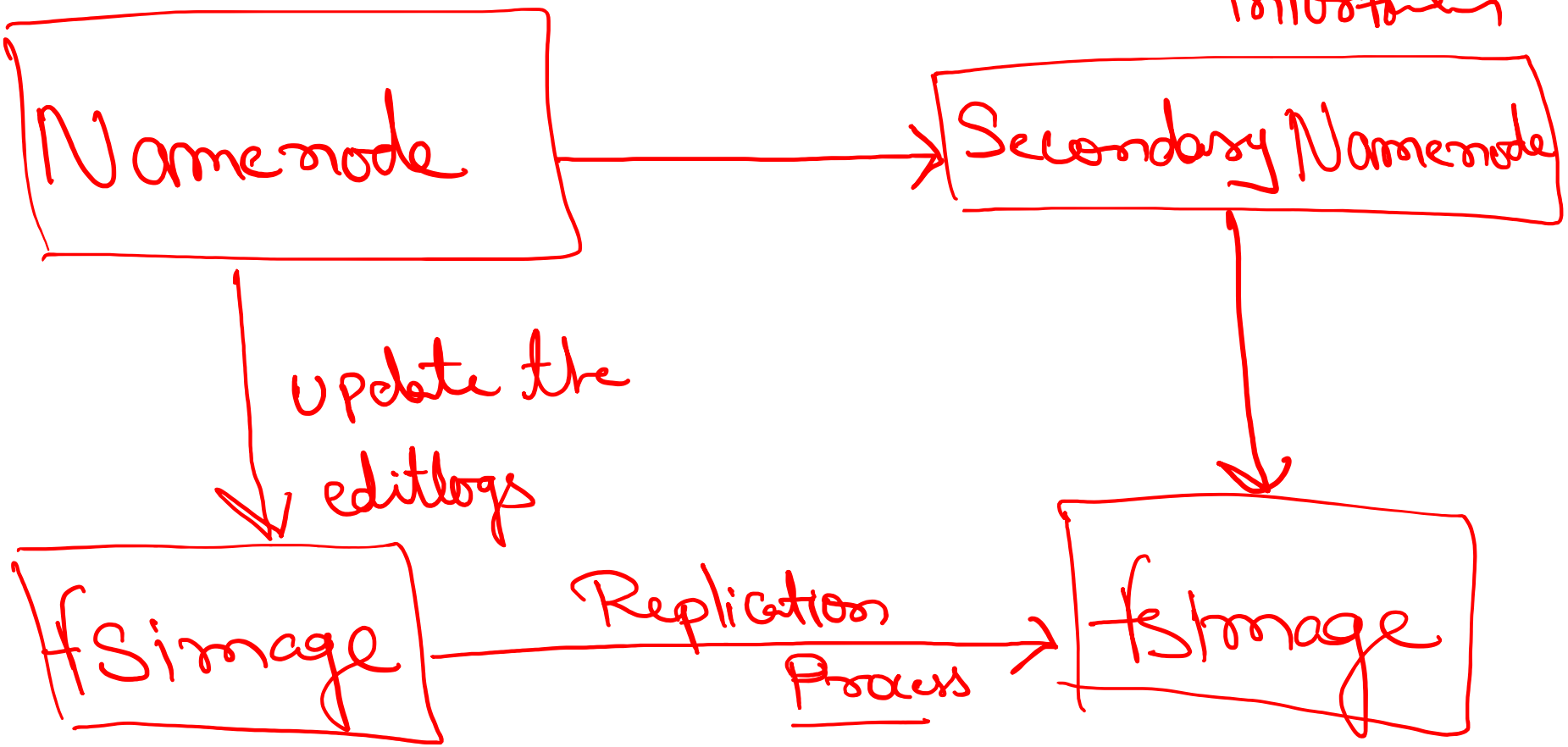


Hadoop Blocks – Datanode



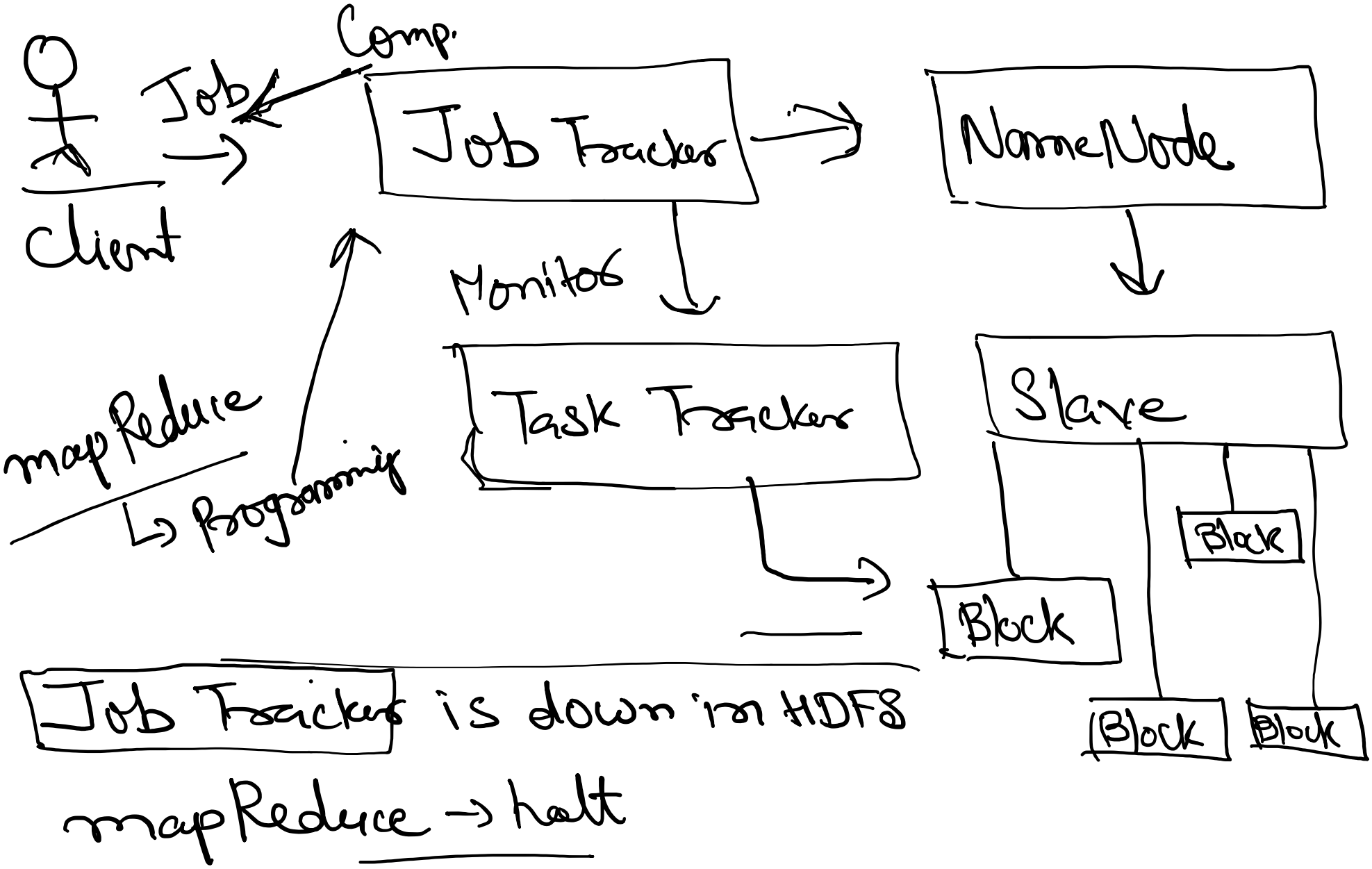
fsimage - Stores the snapshot of fs

Namenode - holds metadata for HDFS in blocks



Editlogs → keeps tracks of every change of HDFS.

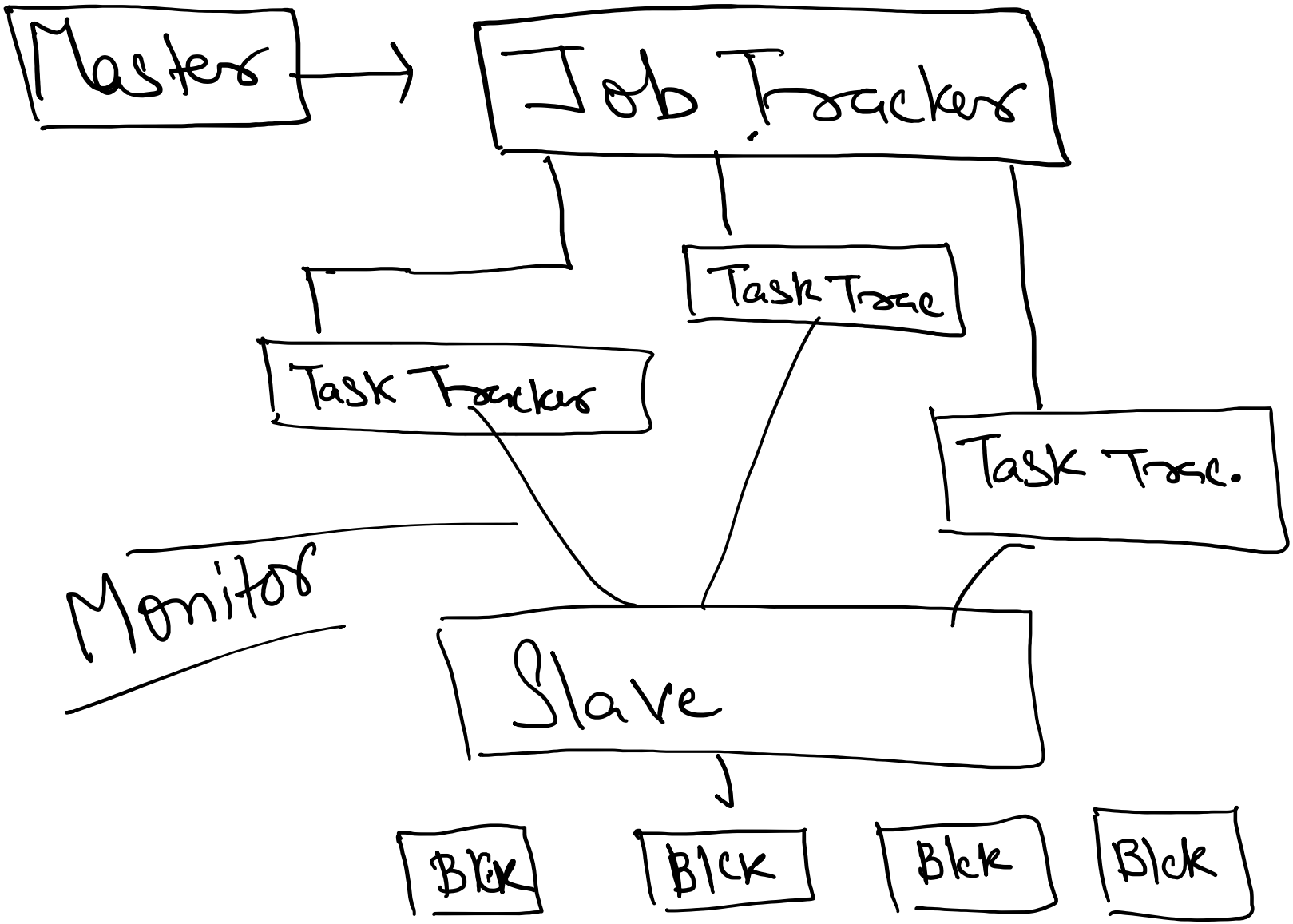
Hadoop Blocks – Job Tracker ✔



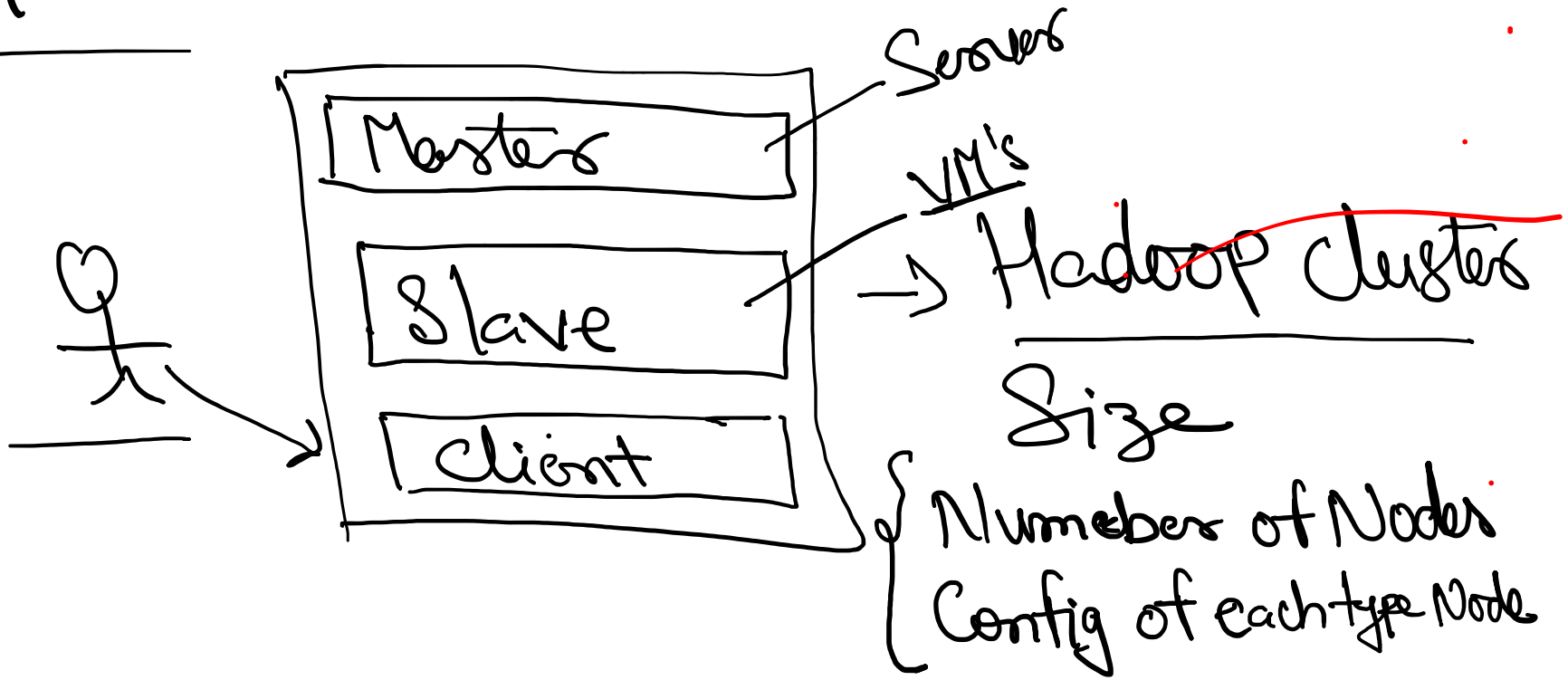
Job Tracker is down in HDFS

mapReduce → halt

Hadoop Blocks – Task Tracker ✓



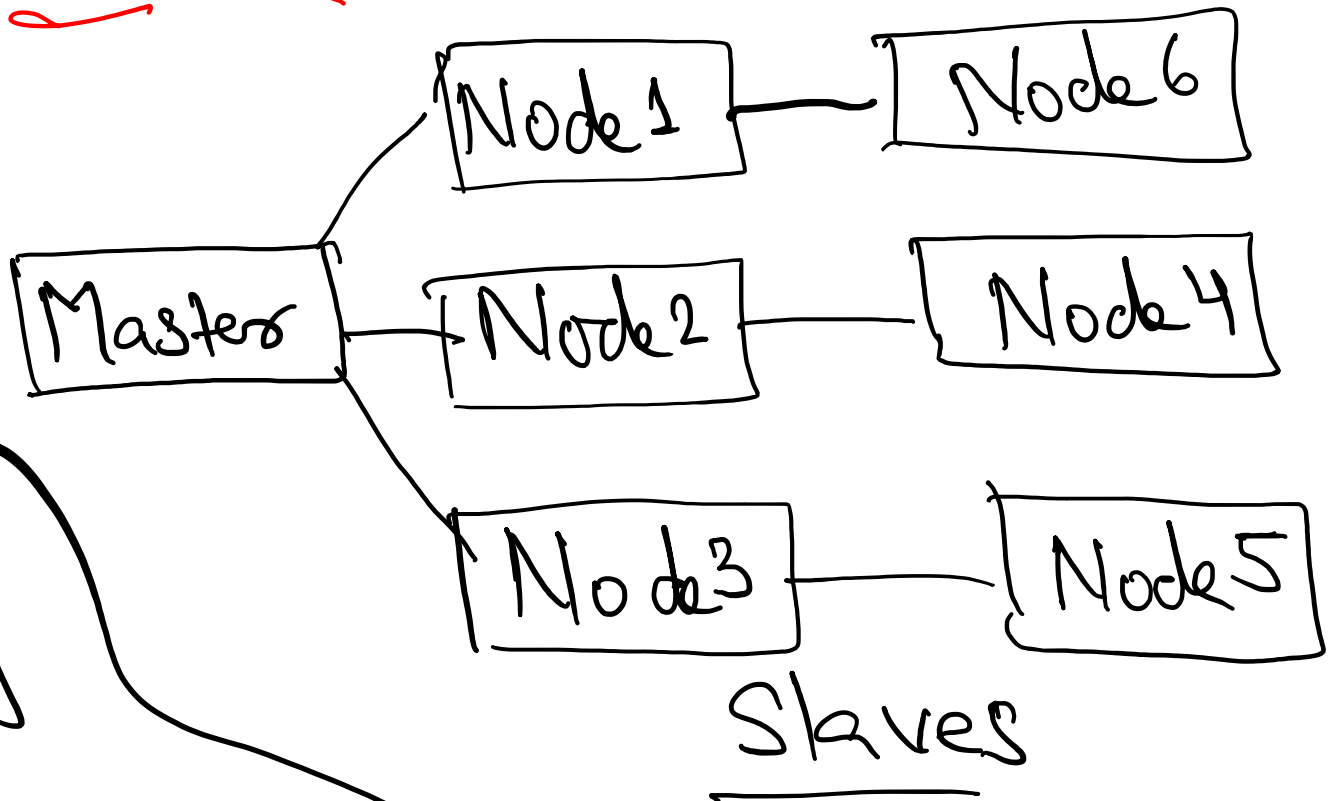
Apache ⇒ big data analytics, process tasks to be broken down into smaller tasks that can be performed into parallel by using any algo like mapReduce.



Web Portal - Sqoop

Data Store – Struct, Semi-Struct & Unstruct Data

Schema ⇒

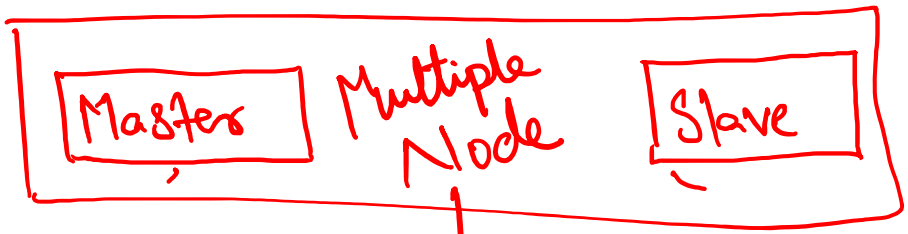


Properties

- Scalability
- flexibility
- Speed
- No data loss
- Economical

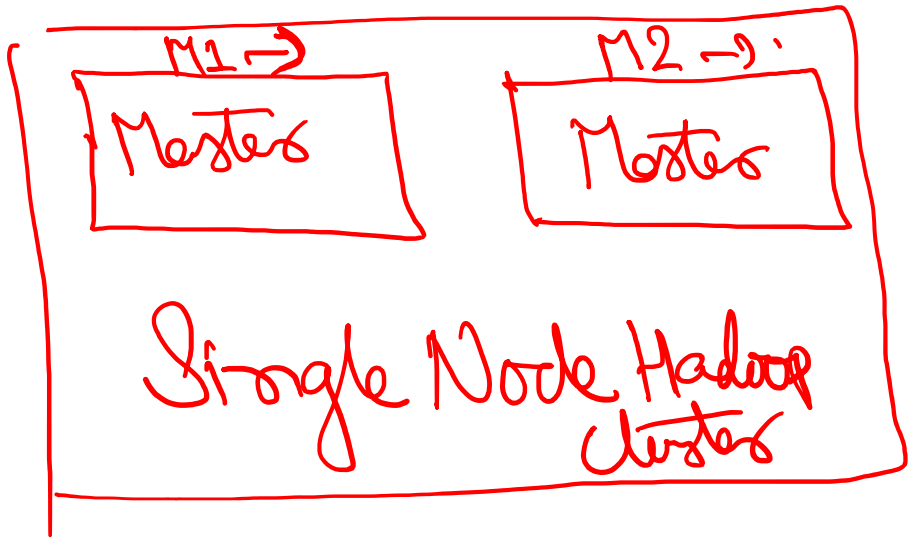
Type of Hadoop Cluster

1) Single Node Hadoop | 2) Multiple Node



Why Bigdata

Quick Proxy -> H/W



Types of Mode of Operation

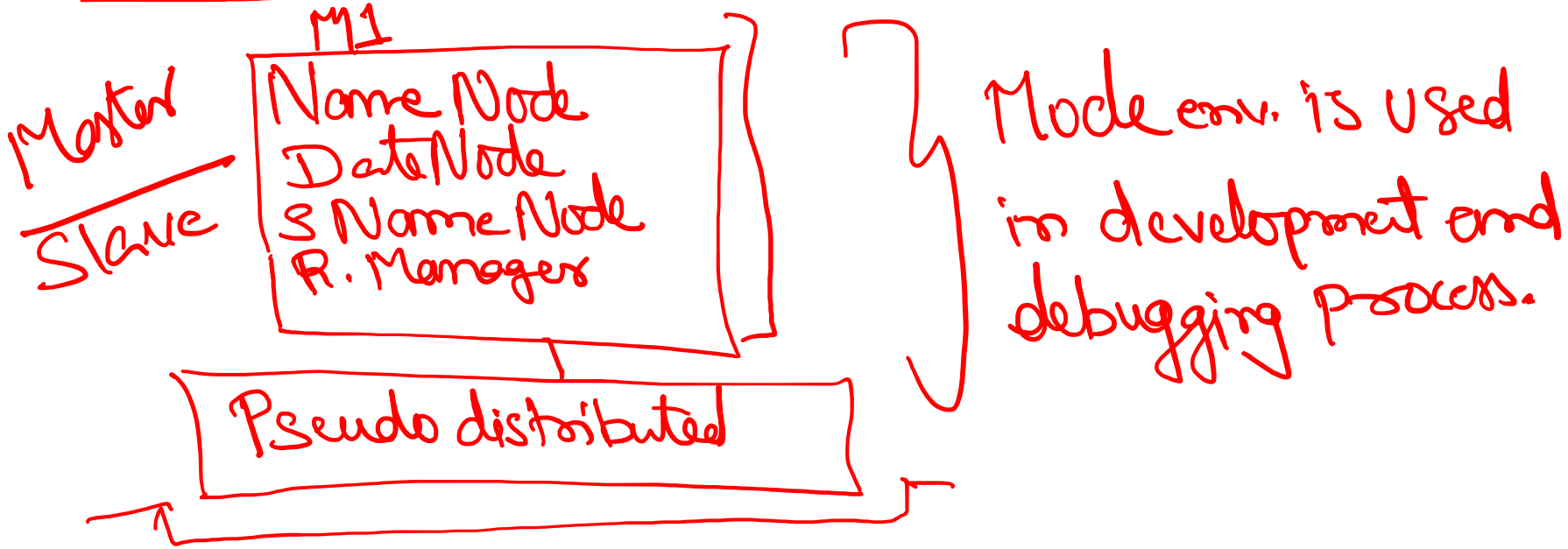
1). Standalone Mode -> M1 (All nodes execute)

M2 (Manage & Monitoring purpose)

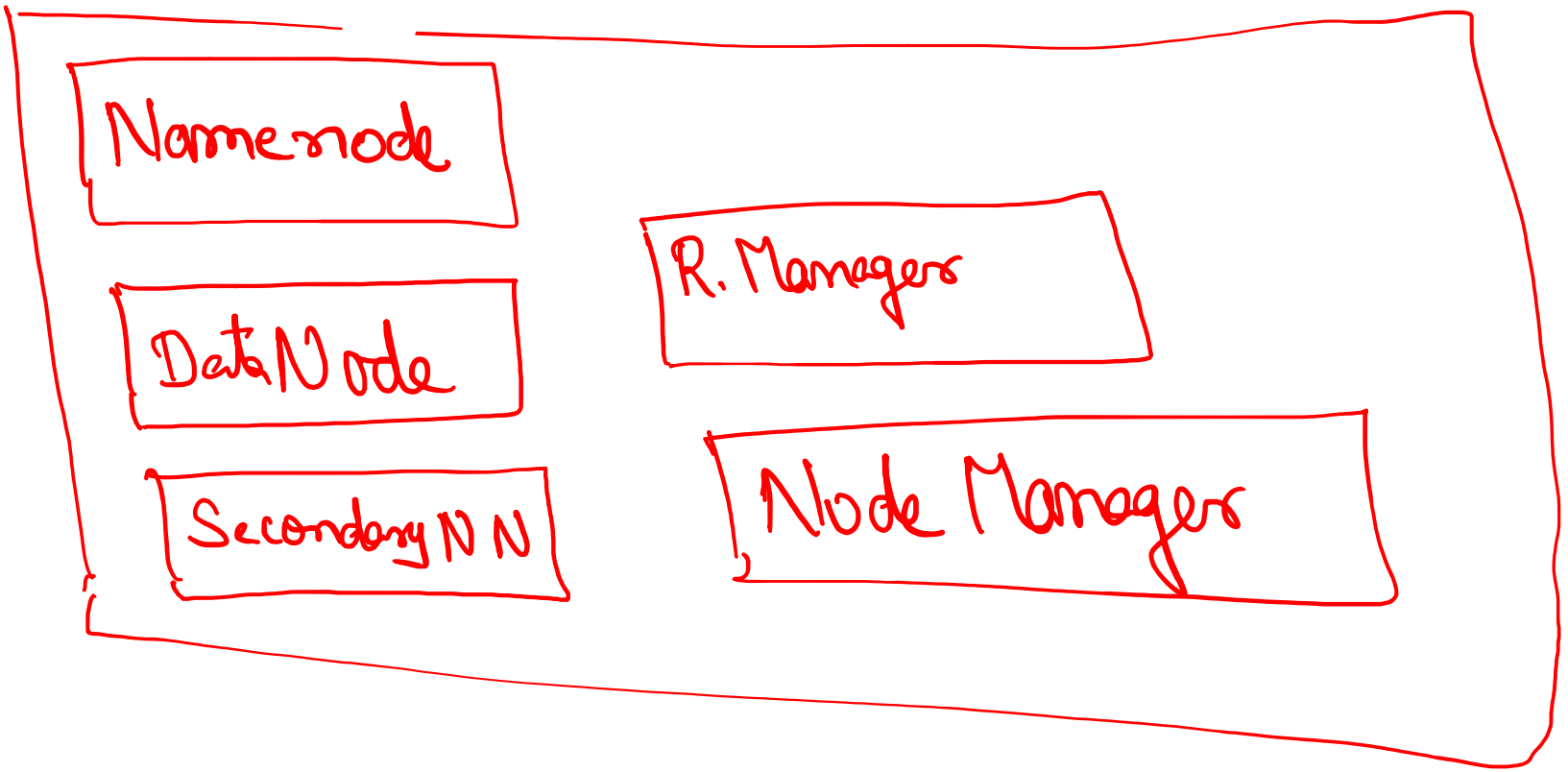
↓

Fastest

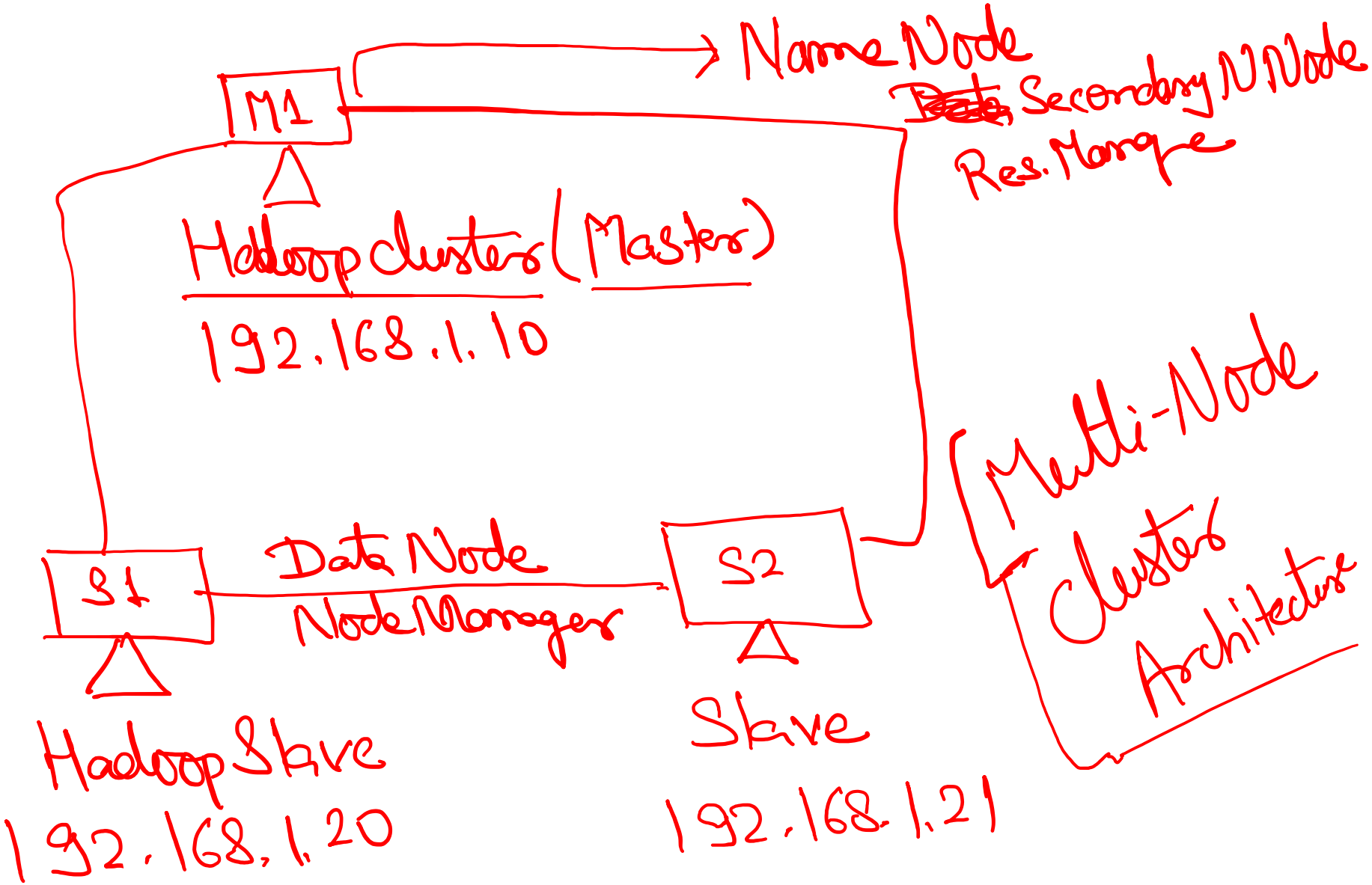
2) Pseudo Distributed Mode -> ✔



3). Fully Distributed — Multi Node cluster(✓)



fully Distributed Mode



Config file — Imp → Apache Hadoop docs.

$URBY

1). HADOOP-ENV.sh — JDK

$JAVA_Home

2) CORE-SITE.xML → Cluster Info

↳ NameNode - IP, Port#.

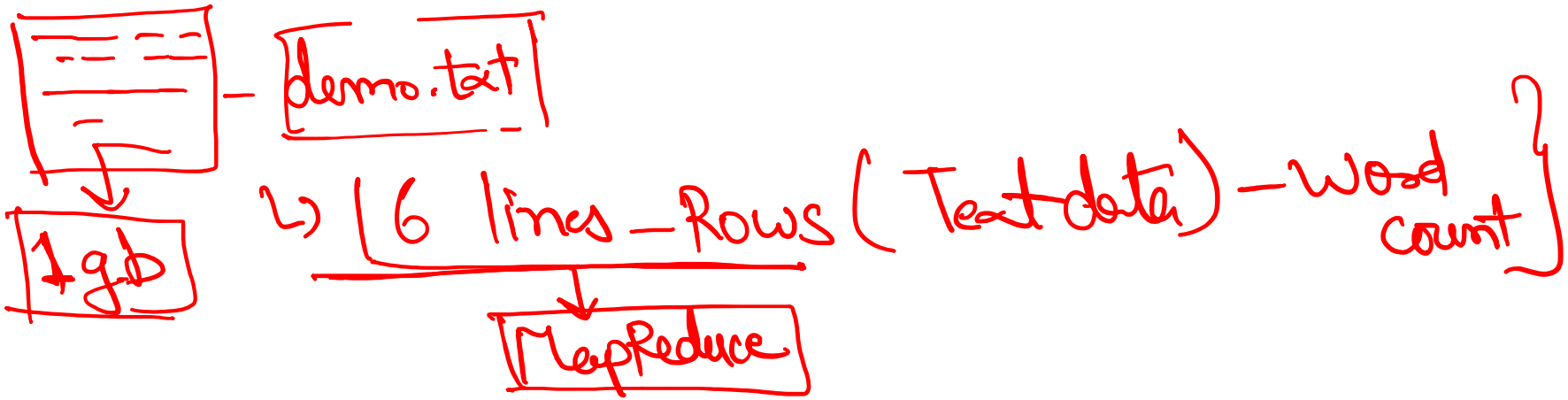
3) HDFS_SITE.xML → Replication of Block

4). MAPREDSITE.XML → MapReduce Program setting

5) Masters -

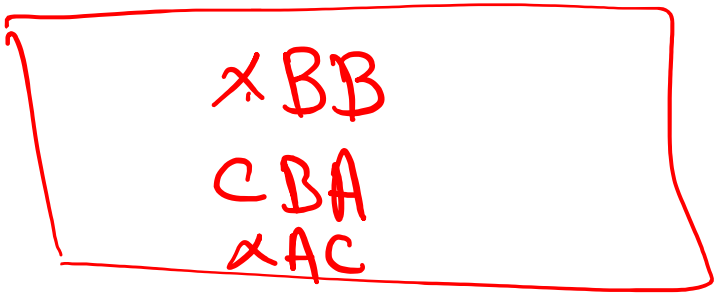
6) Slave - Master Node info a list of hosts, one per line IP address of Slave nodes.

MapReduce => it is responsible for processing the data file in hadoop ecosystem



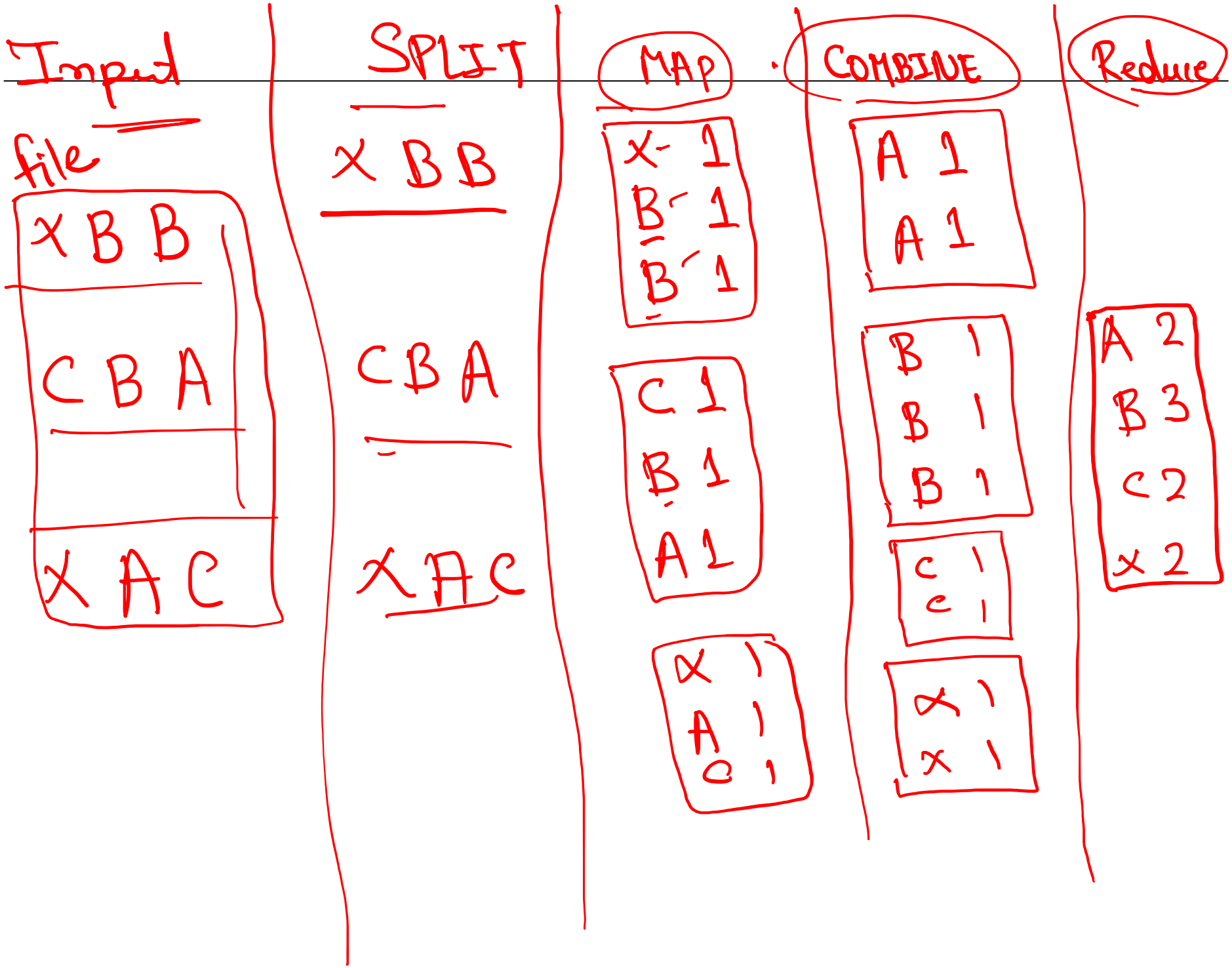
MapReduce

- Map Task
- Reduce Task

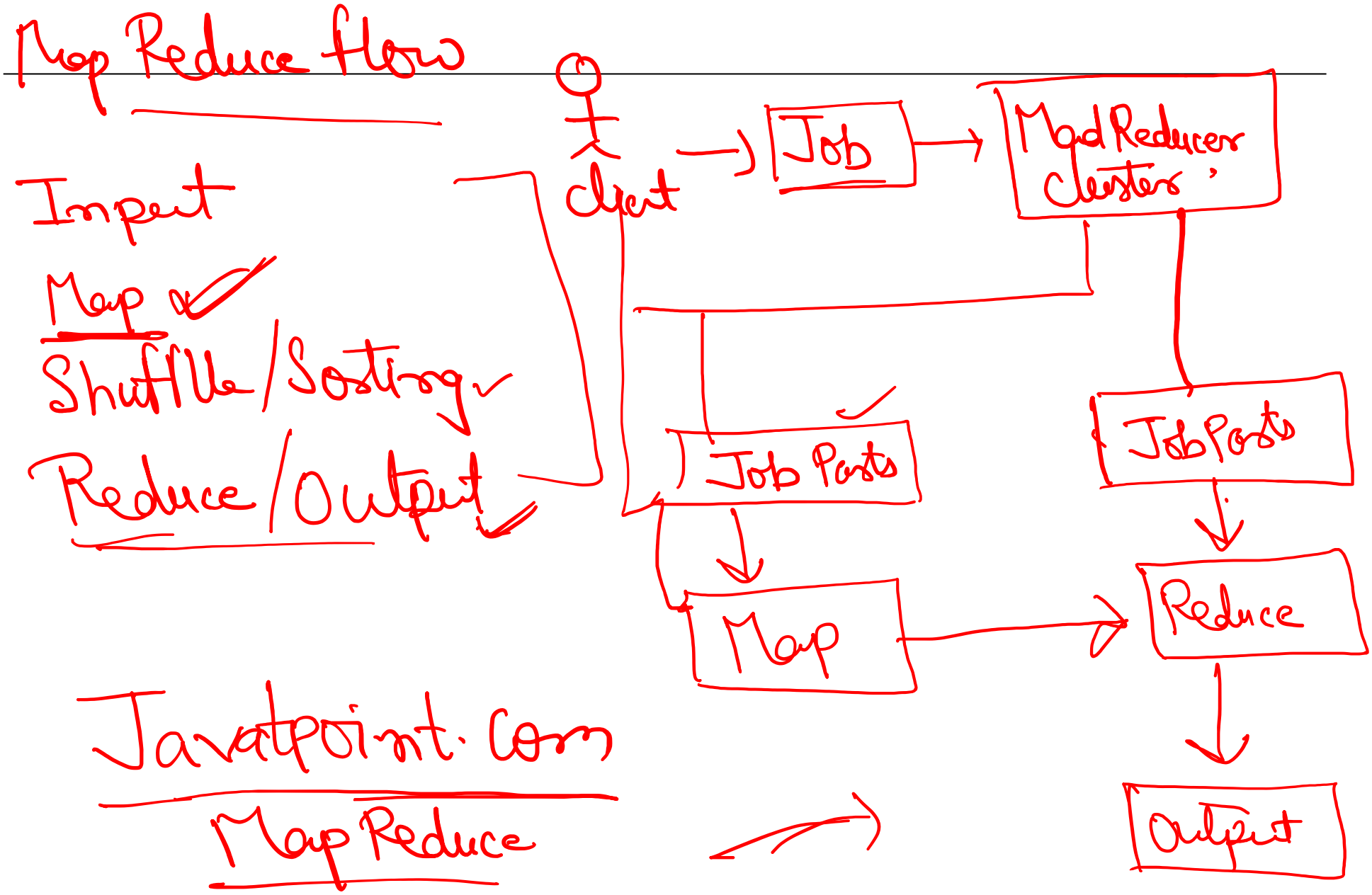


XBB

Example of MapReduce Program →


Input file
SPLIT
MAP
COMBINE
Reduce
X B B
C B A
X A C
X B B
C B A
X A C
X 1
B 1
B 1
A 1
C 1
B 1
A 1
X 1
A 1
C 1
A 1
B 1
B 1
B 1
C 1
C 1
X 1
X 1
A 2
B 3
C 2
X 2

Map Reduce Flow



Javatpoint.com

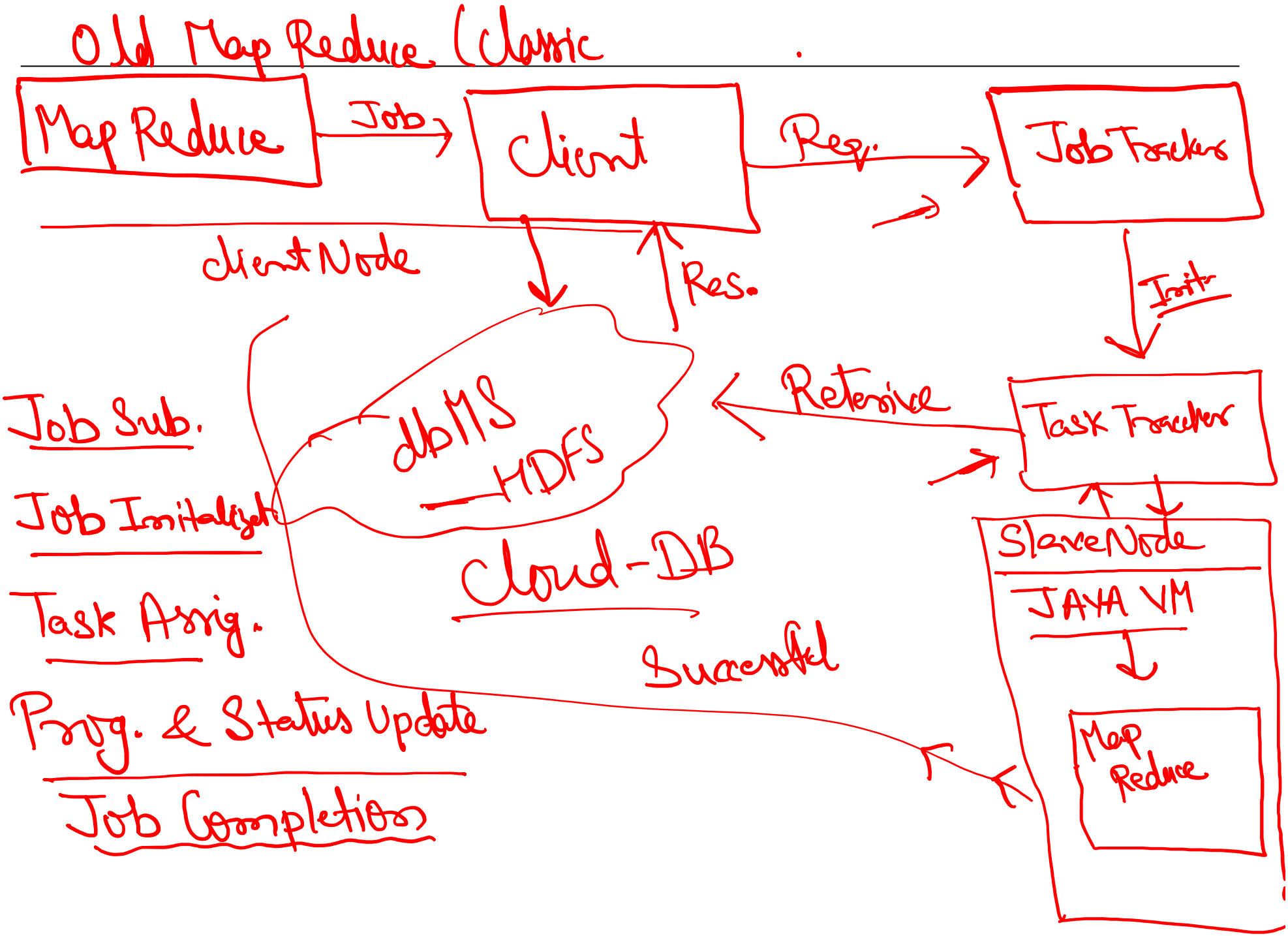
Map Reduce

Weather Dataset

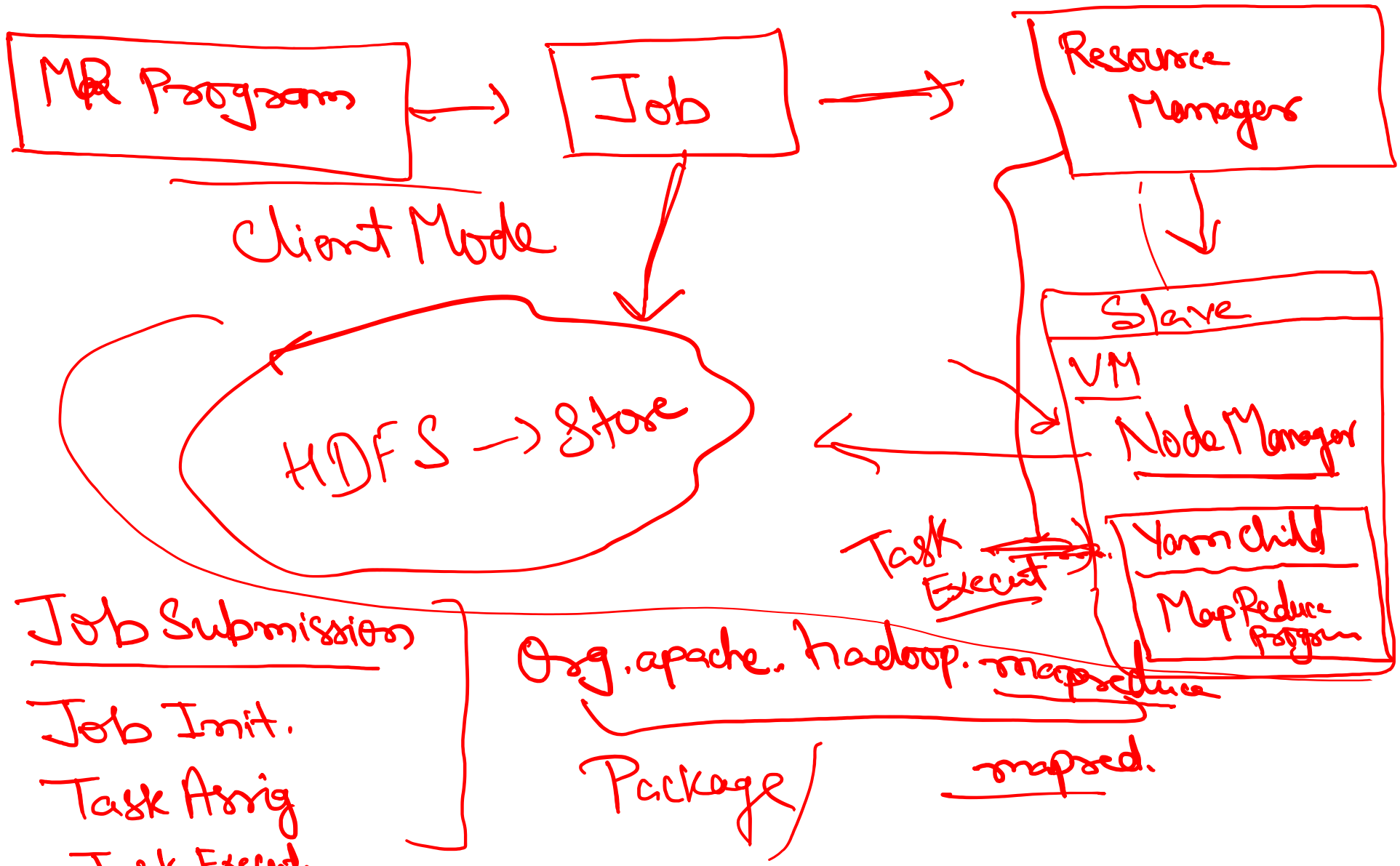
NCDC

Old & New - MapReduce framework

Old Map Reduce (Classic)



YARN -> (New MapReduce



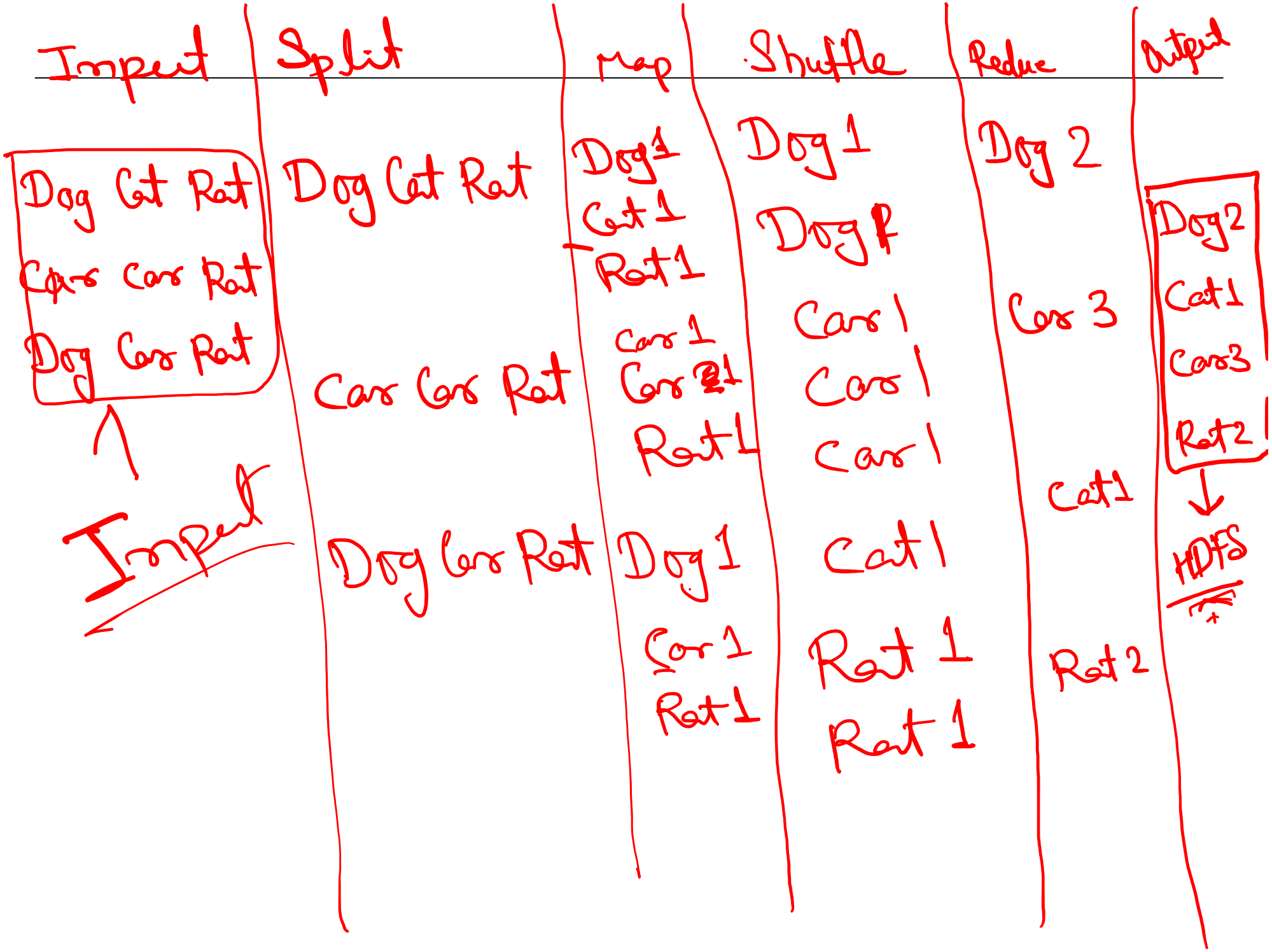
Drivers Code

Mappers Code

Reducer Code

Recorder Code

Combiner Code


Input
Split
map
Shuffle
Reduce
Output
Dog Cat Rat
Car Car Rat
Dog Car Rat
Input
Dog Cat Rat
Car Car Rat
Dog Car Rat
Dog 1
Cat 1
Rat 1
Car 1
Car 1
Rat 1
Dog 1
Car 1
Rat 1
Dog 1
Dog 1
Car 1
Car 1
Car 1
Cat 1
Rat 1
Rat 1
Dog 2
Car 3
Cat 1
Rat 2
Dog 2
Cat 1
Car 3
Rat 2
HDFS