Spark Execution Model And Architecture - (Kafka).

2 hours session

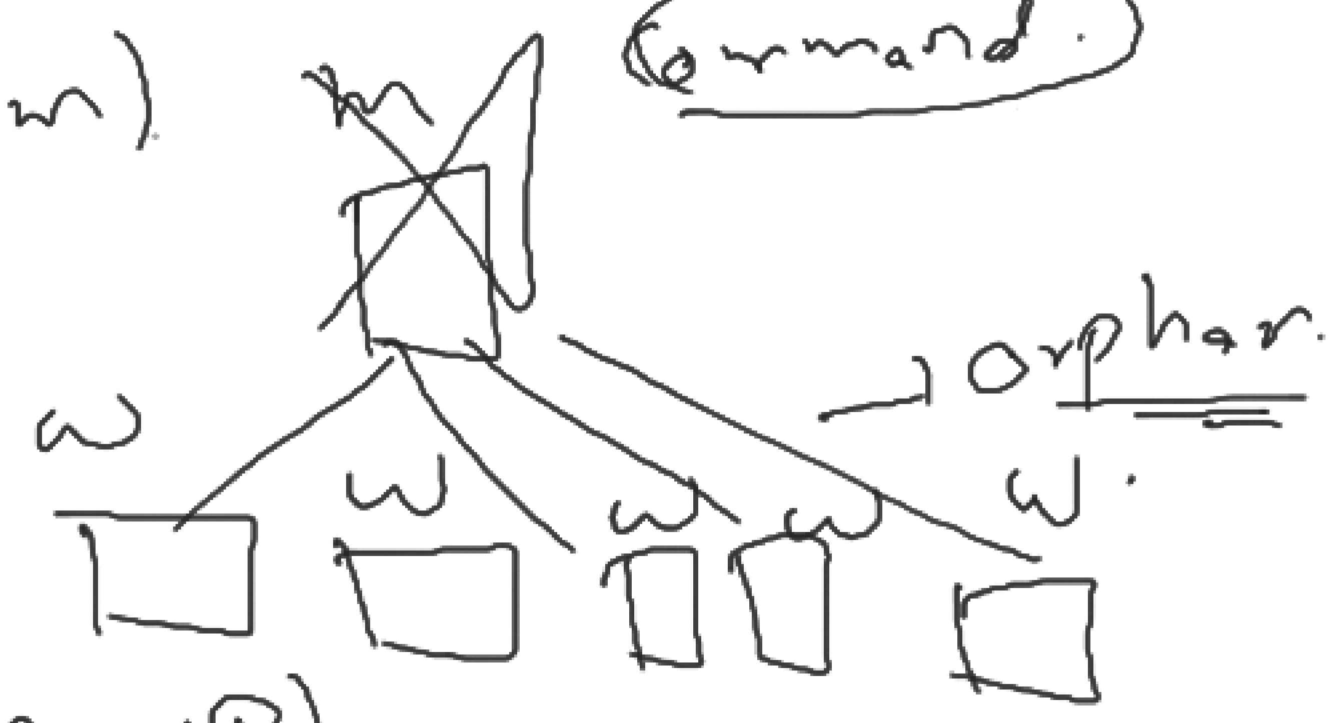
8:00

1. How can we create spark Programs?
2. How to execute spark Programs?

1. Interactive client (spark-shell, Notebooks) (learning phase).
2. Submit Job - spark-submit.

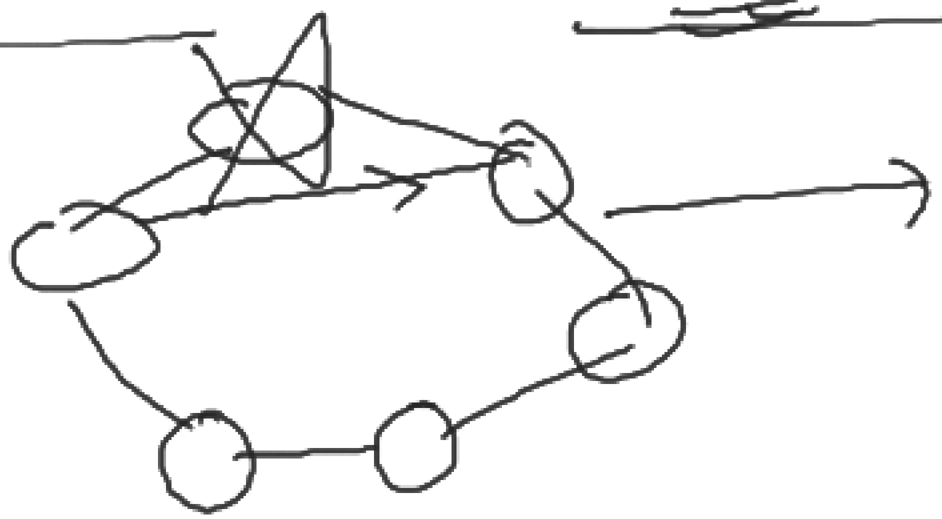
created wholeprogram → submitted to a cluster

cluster – (Hadoop). (Multiple system).



① Master-slave Architecture –

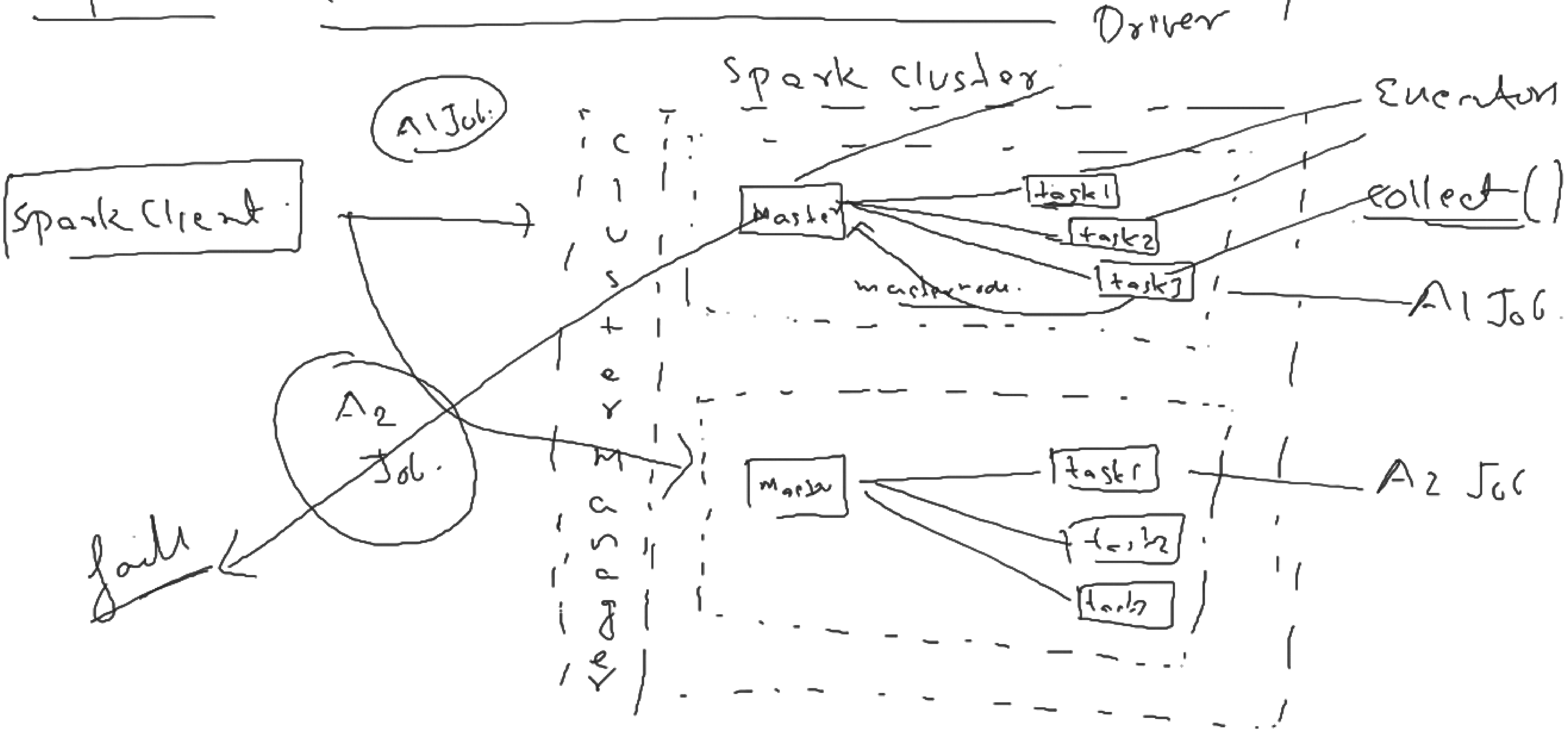
② Ring Architecture – ((Cassandra)/ MongoDB).

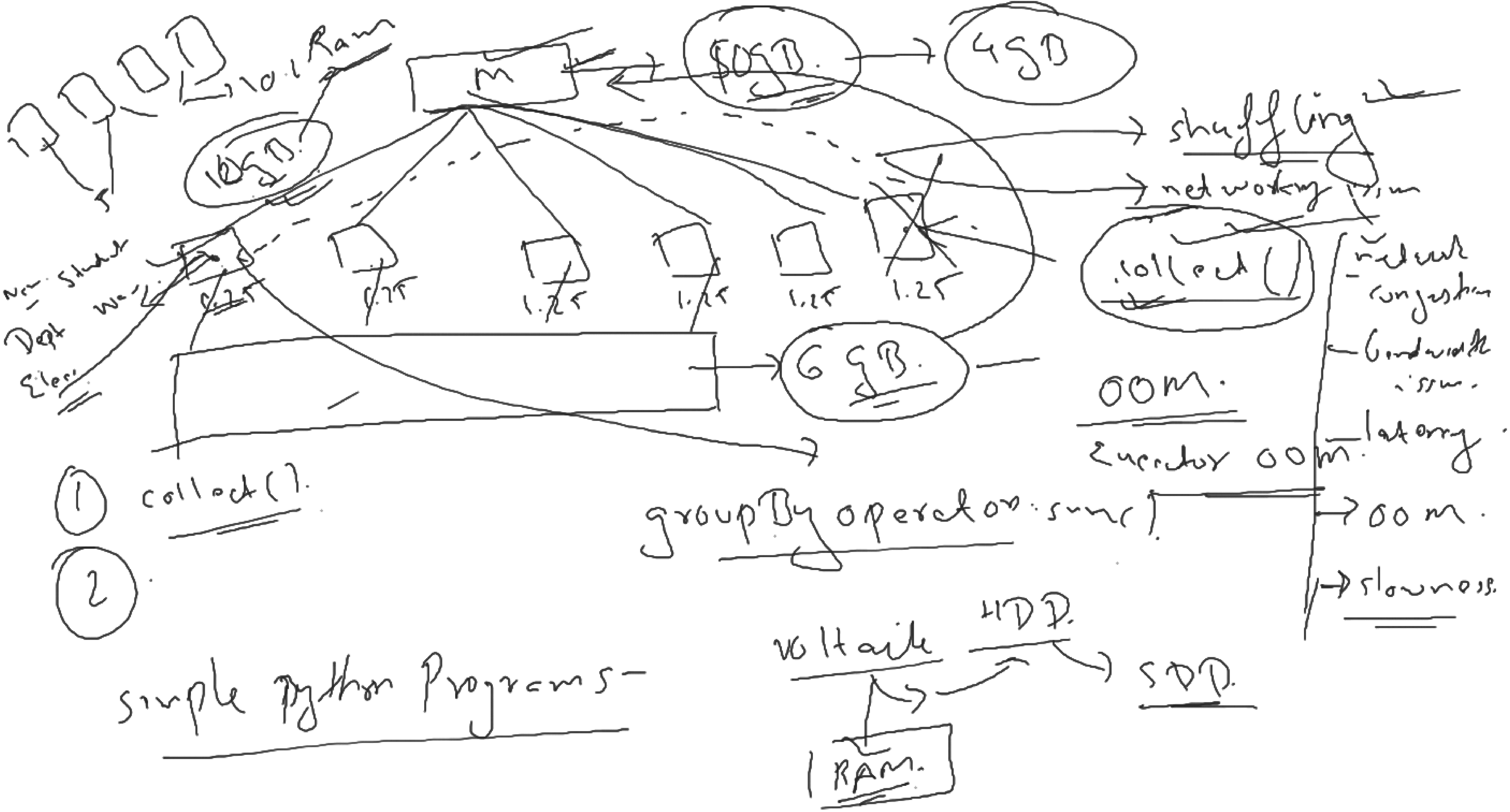


latency is very high.

System Design

Spark – (Master-Slove Architecture)




Ram
M
10 GB.
50 GB.
4 GB
shuffling
networking issue
collect()
Name Student
Dept
Elem
1.25
1.25
1.25
1.25
1.25
1.25
6 GB.
OOM.
Executor OOM.
network congestion
bandwidth issue.
latency.
OOM.
slowness.
groupBy operation sum()
1 collect().
2
volatile
HDD.
SDD.
RAM.
simple Python Programs -

spark programs - (Legacy software). Python - Java

Scala

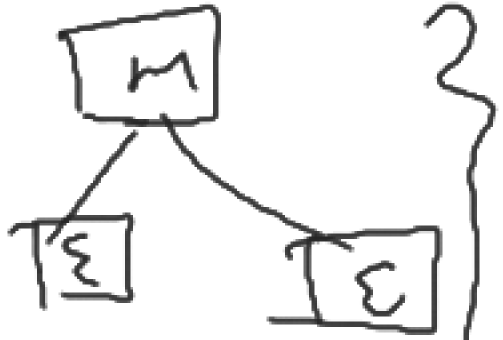
1 local[n] → local[2] →

local[*] → cores → 4

→ 4 thread

local[1] → one driver

one executor.



threading

threads

mimic the distributed system

single Python program

sending their distribution

2 YARN - (EMR cluster, ADB, dataproc).

Fargate)

3 Kubernetes - Flink spark. in kubernetes

→ very interesting

→ AWS

→ flink (Beam).

(orchestration tools).

4 Mesos - (ochrestation).

5 standalone - development -

① C, C++. (Evolution of technologies).

Advance Java (7,000/-)

→ J2EE.

→ Java

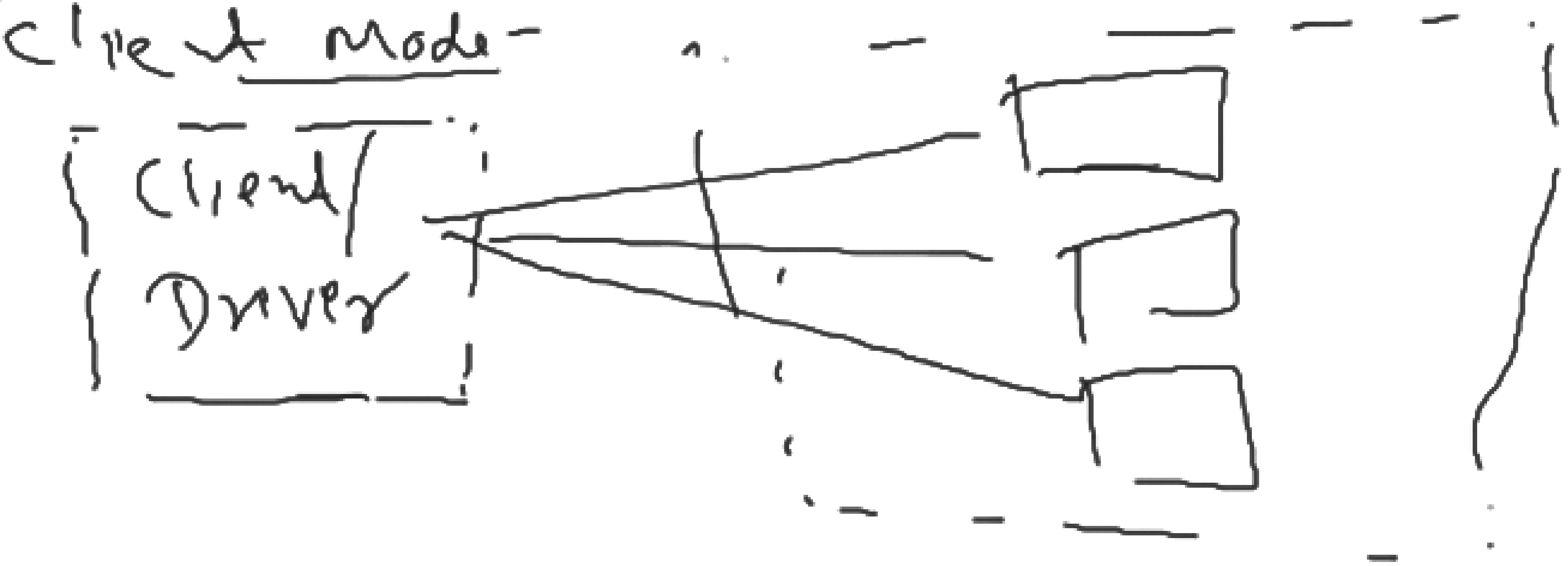
② Spark – (Entry level Big data engineer).

→ legacy. (system)

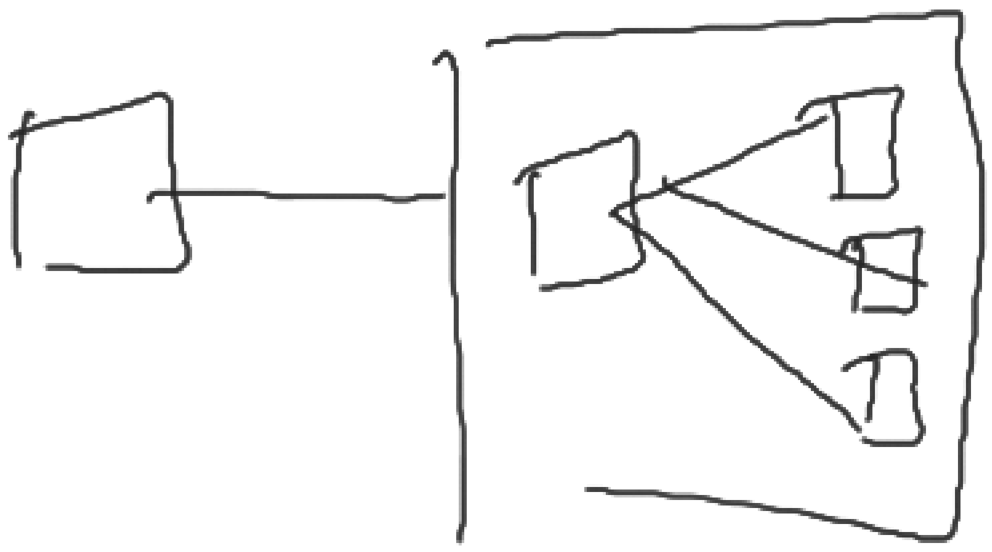
→ Flink/Beam

Spark with Interactive Client

client Mode-



Cluster



Cluster Manager	Execution Modes	Execution tasks.
1. local[n]	client	IDE, NoteBook.
2. YARN — on-prem. on-cloud.	client	Databricks Notebooks, Shell.
	Cluster	spark-submit

Debug mode.

In memory – RAM → (voltaile memory). → 10GB.

14,000 ₹ → 50 m

V
24,000

Hadoop Map-Reduce – Harddisk → knitting (slow process).

consumer ↔ DMS → computer

servics. cost of manufact

CA
Sales
CIO.
20.
50L/-

Technical.

1024 GB → 1 TB – 4000 → 7,000/-

8 GB – DDR3 → 6,000 –

(40-50%).

battery → 4000/-
→ 2650/- → 3,200

dell charger → 1200/-

2000/-

① Servers

350km. → ① ② ?

1,00,00/per month

Spark – R → flushed to datadisk.
(storage level).

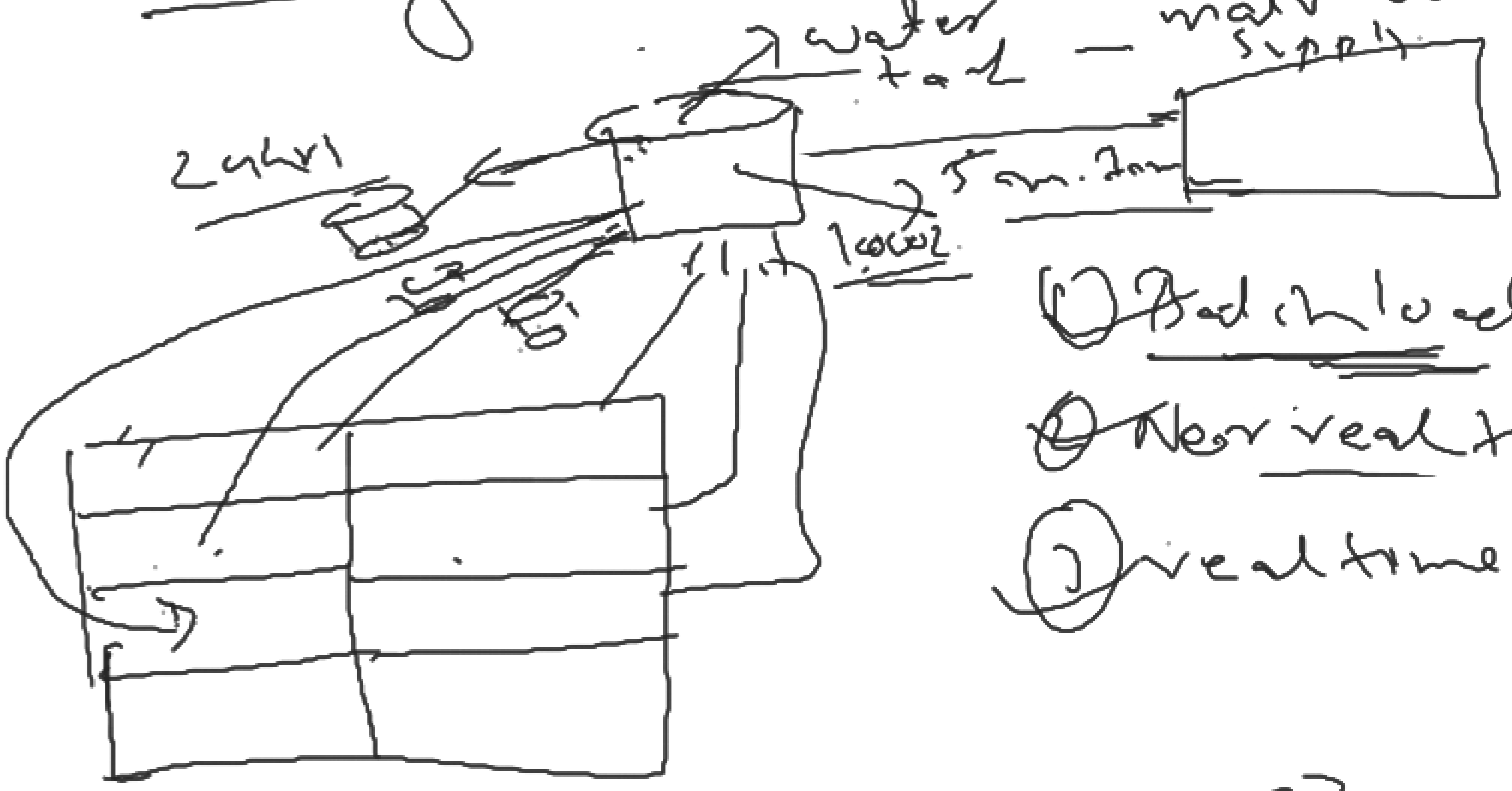
→ RAM.

① Batch form (Real time data).

① Data is coming in files, (historical data)

② Streaming - (Real time

5am-7am.



① Batch loading

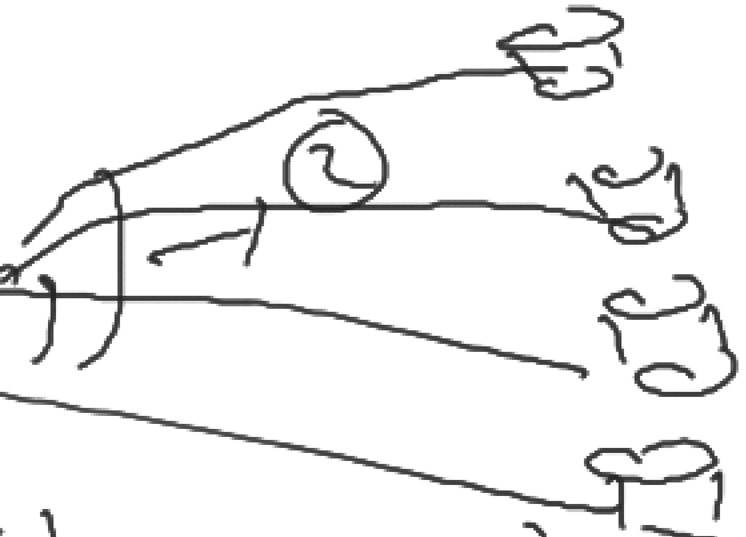
② Near real time

③ real time.

spark supports.

① Batch loads → 10,000

② Near real time streaming (micro batching)



up-stream. (Flink). (Real time streaming).

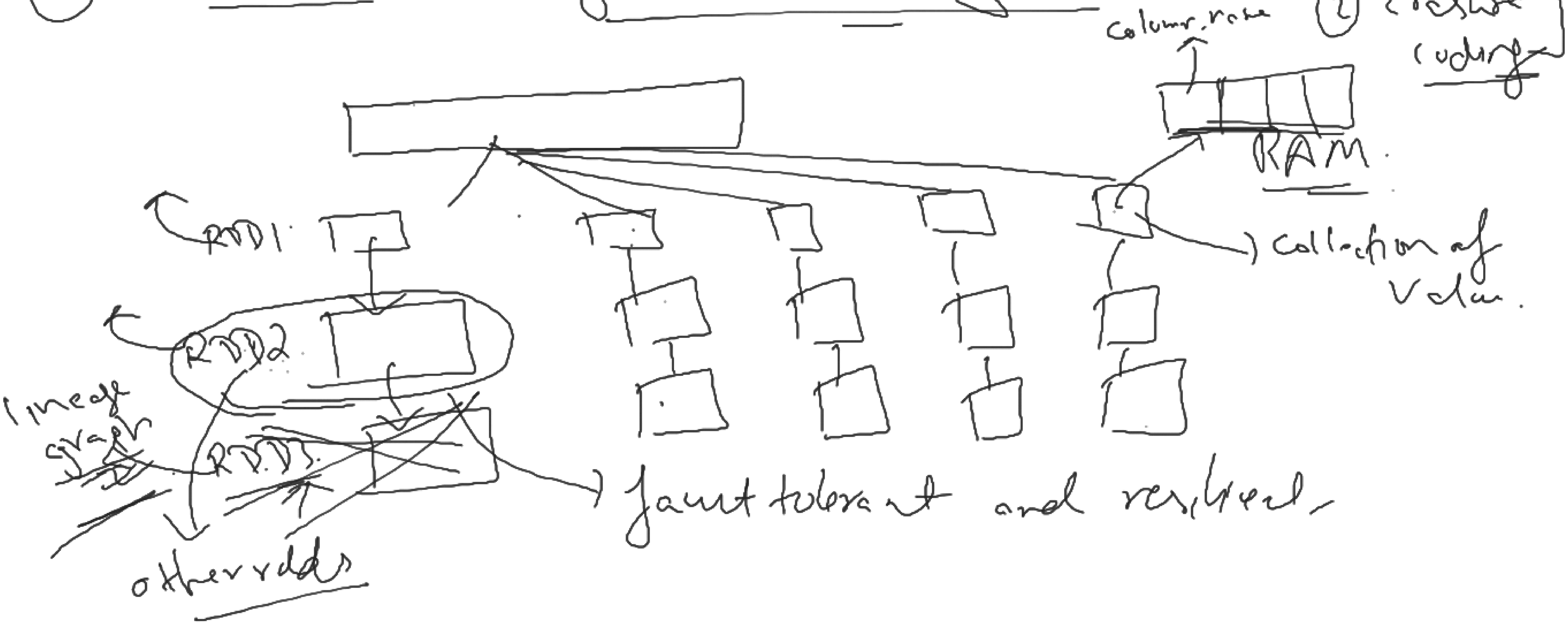
RDD — Resilient Distributed Dataset

Hadoop

1) Replication

2) Erasure coding

1) Immutable - change the contents of an RDD.



① Transformations —

↓

input RDD — output RDD

① wide → shuffling.

② narrow → no shuffling

② Action.

↓

input RDD → non-RDD.

until and unless an action is applied, transformation, won't be executed. (spark is lazy evaluated)