SUNBEAM

Big Data Technologies

Trainer: Mr. Nilesh Ghule.

Big Data Technologies

Contents	Prerequisites
① Introduction	
② Hive (DWH)	– SQL
③ Spark	– Python
④ Kafka	– Python
⑤ Airflow	– Python
⑥ Hadoop	– Java → generics, array list & hashmap exception handling jdbc Stream programming oops basics.
⑦ HBase	–

Lecture:
8:00 am to 1:00 pm

Lab:
2:00 pm to 7:00pm

Introduction

Evolution - Data engg.

File based data handling

1970+ : RDBMS

1990 : internet + www

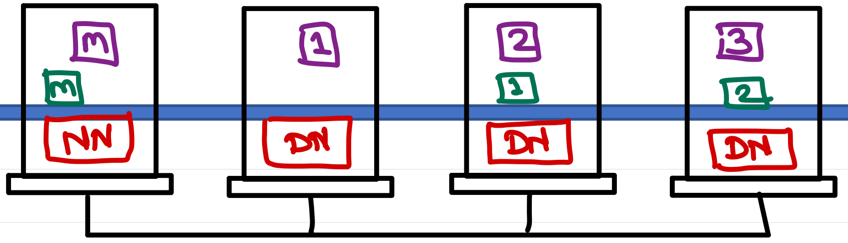
Java

1998+ : NoSQL

2000 : MPP

2001 - 2004+ : Google Big Data → 2003: GFS, 2004: MR

2006 : cloud computing



hadoop cluster

Distributed Systems

* cluster : set of computers in a close network doing dedicated task.
* Horizontal scalability
* Distributed Storage + Distributed Computing
* HA, Reliable, Fault tolerance,

*Distributed Storage challenges:
1. Block size
2. Data node failure
3. metadata node fail.

*Distributed Computing
1. synchronization
2. node failure

* Big Data Characteristics

1. Volume
2. Velocity
3. Variety
4. Veracity ← data quality
5. Value

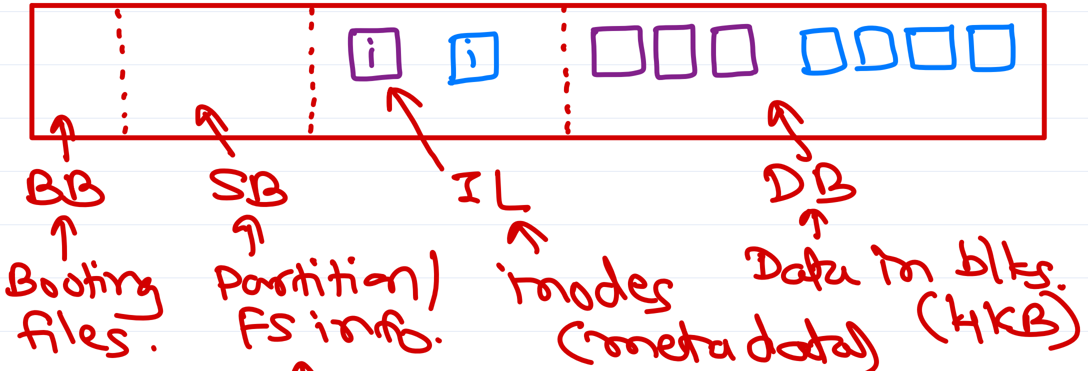
Variety →
- structured data - fixed schema → RDBMS
- semi-structured data - flexible schema → NoSQL
- unstructured data - no schema → BigData

Local storage vs Distributed storage (HDFS)

File = collection of data/info on a storage device.

File = Data + Info
(contents) (metadata)

File System = Organizing files on disk e.g. FAT, NTFS, EXT3/4,...



label, blk size, part size, free blocks,...

type, size, permissions, user, group, timestamps, name, ...

info about data blocks.

① Data divided into data blocks → stored on multiple nodes in a cluster.

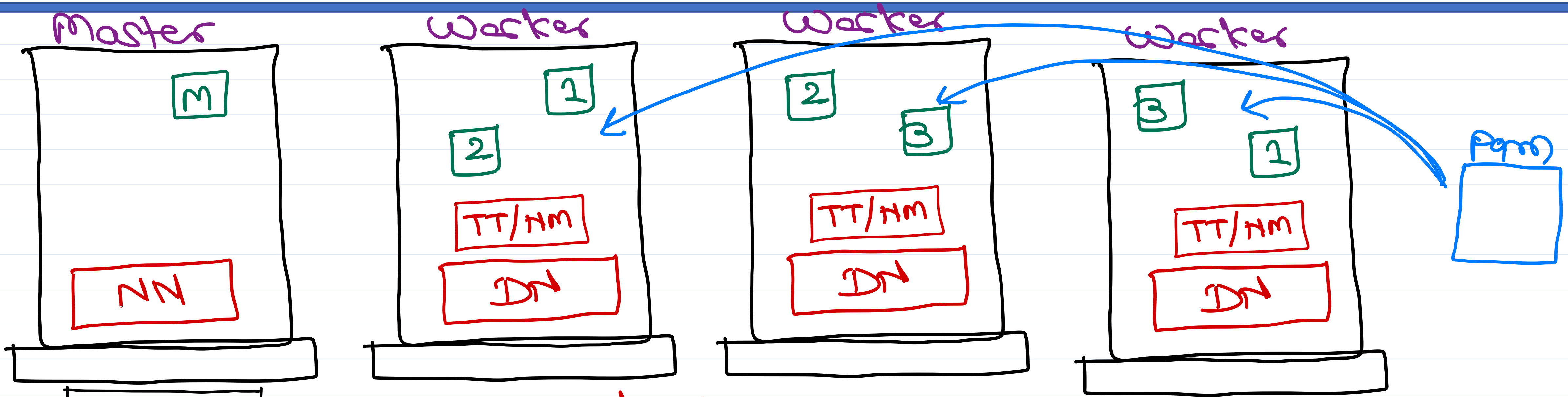
② Data Block size:

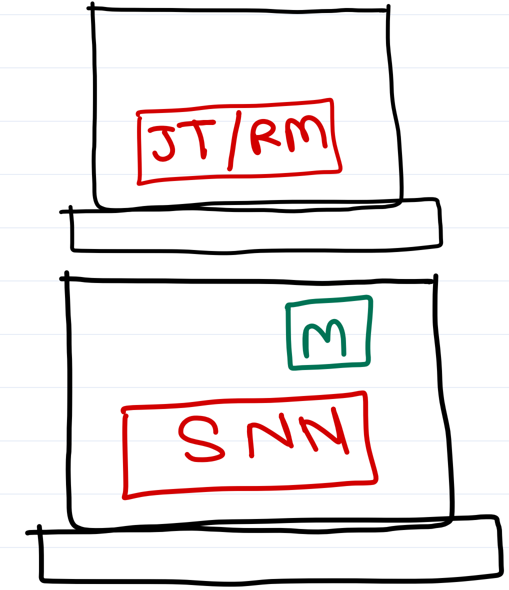
- Bigger block sizes to reduce overheads.
- HDFS block size
  - Hadoop 1.x → 64 MB
  - Hadoop 2.x+ → 128 MB
- Different/Custom block size can be given to each file.

③ Data Node failure

- Any node failure is handled by redundant storage.
- HDFS Replication: Each data block copied on multiple Data Nodes. Default: 3
- Hadoop 3.x: one more mechanism. i.e. Erasure Coding.
- Replication: Overheads 200%.
- Erasure coding: overheads 50%.
  + But need to re compute the lost block.

Distributed Storage (HDFS)

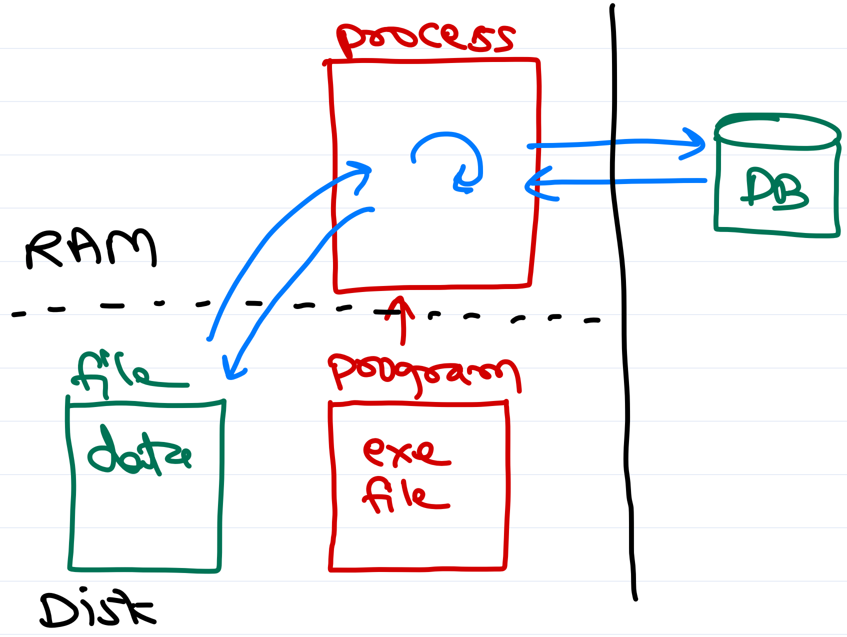




④ NameNode failure.

- If metadata is not avail, whole system is down
- keep metadata backup(s).
- In Hadoop 1.x, NN is SPOF.
  - SNN takes periodic backup
  - On failure, Admin switch SNN to NN (manual)
- In Hadoop 2.x, Standby NN takes active backup of meta.
  - on NN failure, Standby NN can auto become NN.
- Hadoop 3.x, can config multiple standby NNs.

Local computing vs Distributed computing (Map Reduc



Problems:
1. limited memory (RAM)
2. limited computing (CPU)
3. limited storage (Disk)
4. limited network (Band width)
5. Disk/Network speed.

✓ Data is distributed in multiple nodes (in blocks).

✓ Program (much smaller in size w.r.t. data size) It will be copied on all data nodes.

✓ Data processing will be planned/tracked/synchronized by special programs.

✓ Each node will have partial data processing and final result will be accumulated later

✓ Any node failure will lead to reassign that task on other node.

✓ Distributed Computing → Map Reduce.

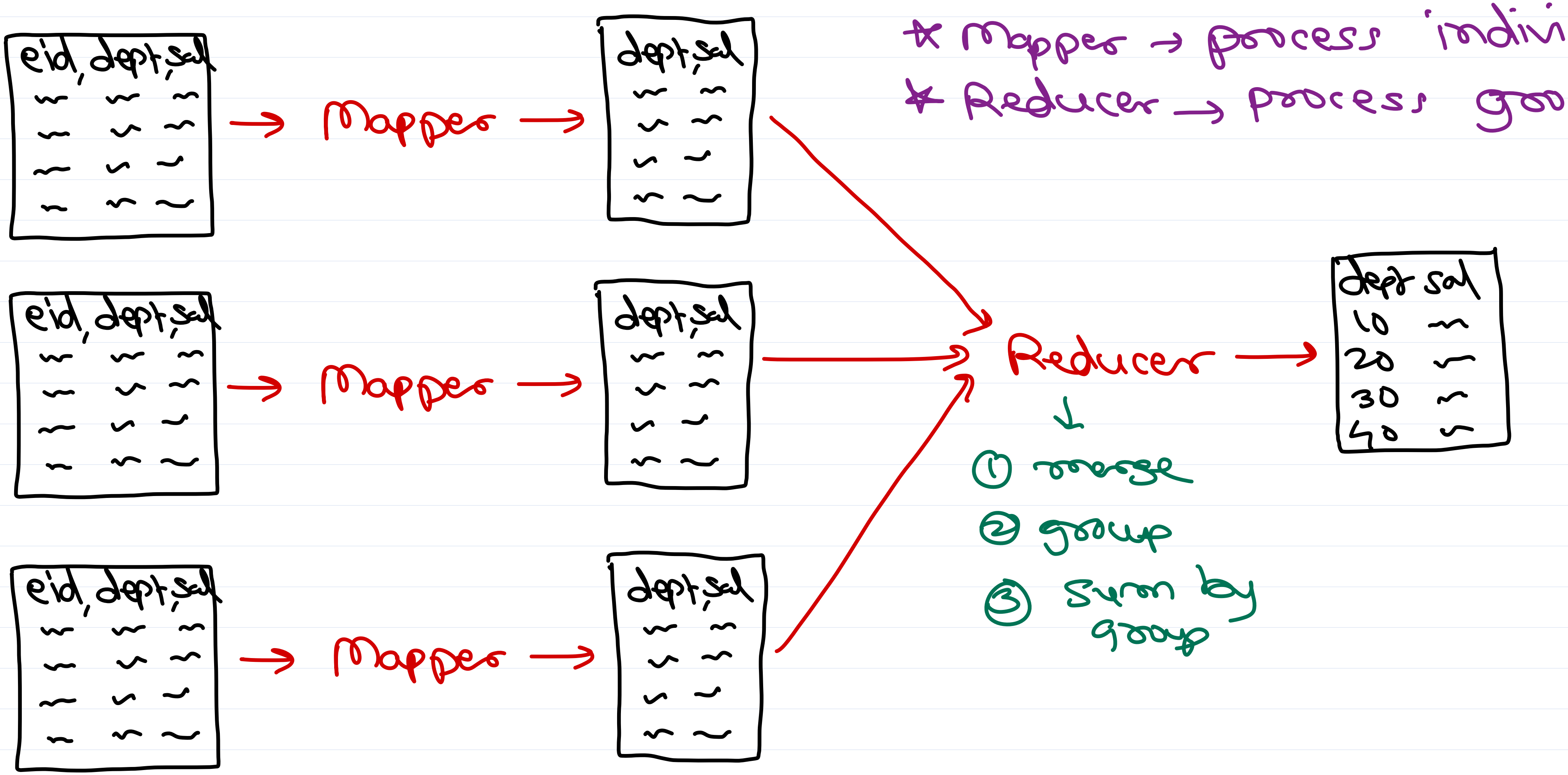
✓ Data to be processed is fetched in program memory and then processing is carried out.

✓ Suitable for smaller data volumes.

Map Reduce - Design pattern

* Mapper → process individual record

* Reducer → process group data.

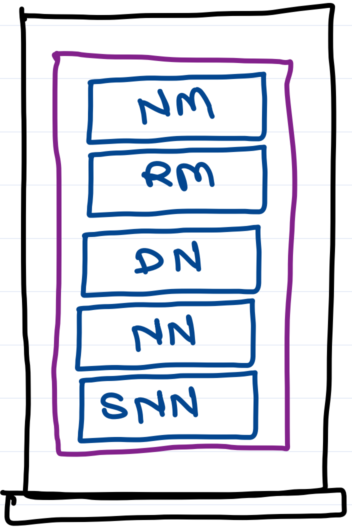


① merge

② group

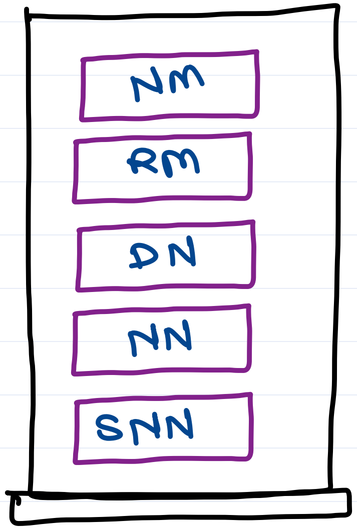
③ sum by group

Hadoop Installation Modes



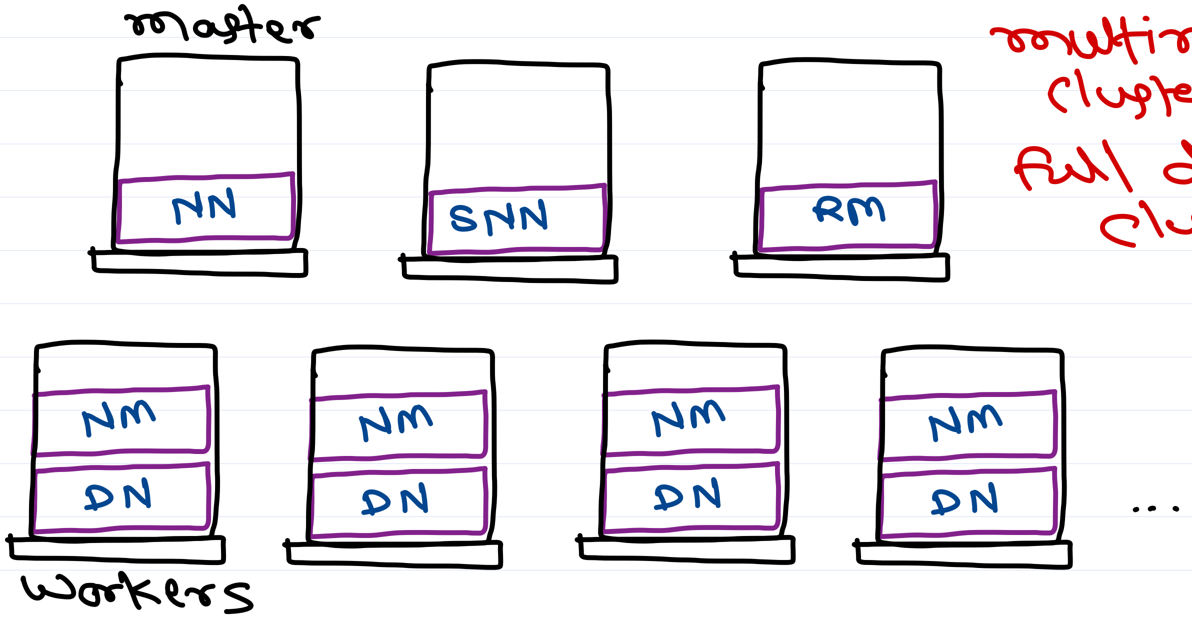
- ✓ fast (no ipc)
- × small data process only
- ✓ only for testing

local mode



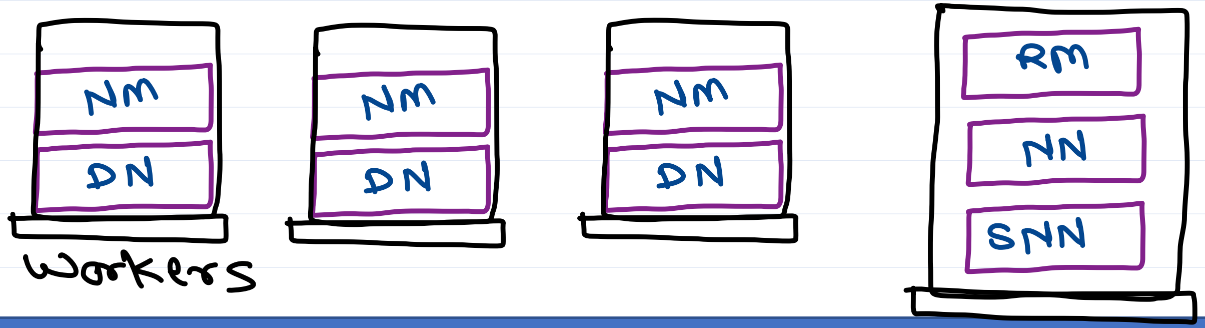
- ✓ Dev machine setup
- × slower (ipc)
- ✓ only for dev & testing.
- × small data process.

single node cluster or psuedo dist mode,



multinode cluster or full dist cluster.

lab assign cluster



Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>