15/3/23

Hadoop

- It is an open source software framework used for storing data and running applications on a group of commodity hardware.
- open source.
- Distributed data processing
- Map - Reduce programming.

Why hadoop

- handle any data type
  - structured / unstructured data.
- schema, schemaless
  - ↳ transaction → job.
- High volume / Low volume of data.
- All kind of analytic application
* proven with peta byte scale.
* Grows with busines
* capasity and performance grows.

featores

- No vendor block in
- Rich eco system and community developmen
- More affordable
- Scalable

- cost effective
- flexible
- fast
- Resilence to failure.

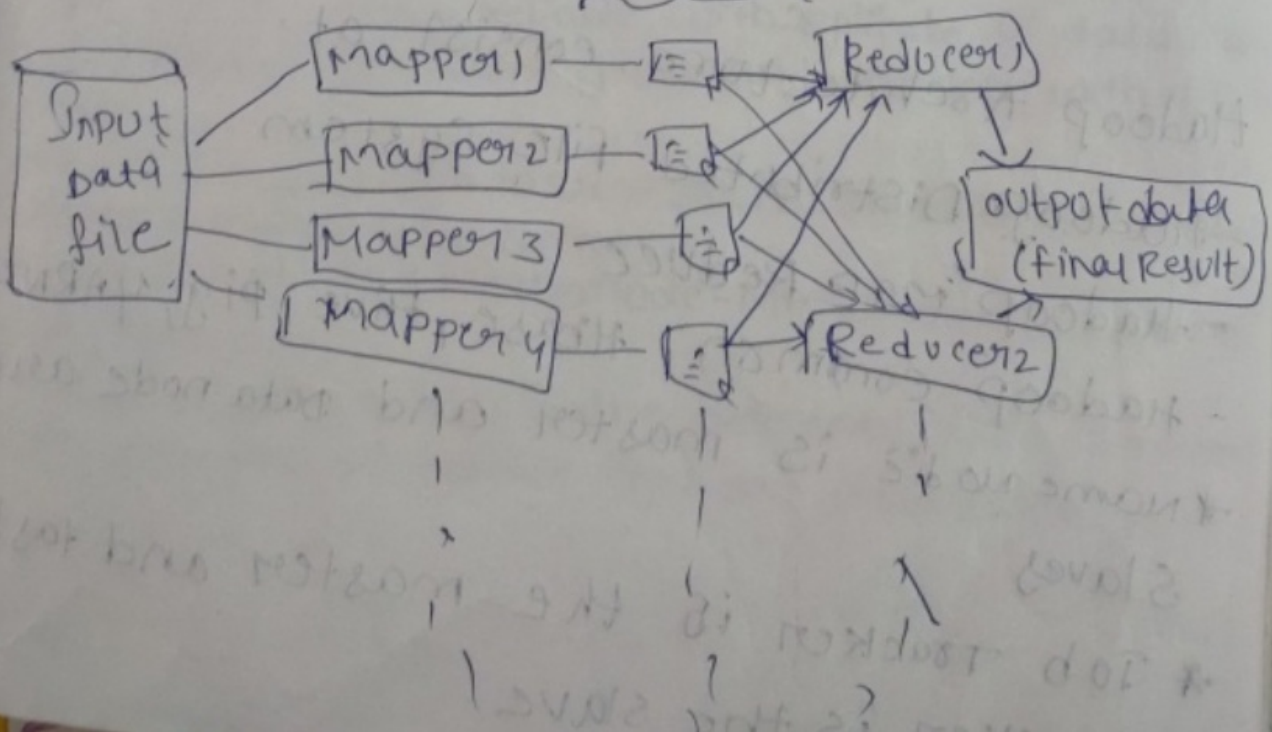
Hadoop Infrastructure.

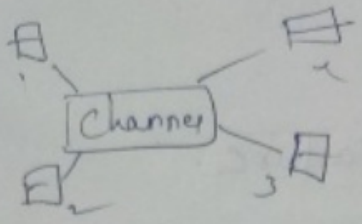
Two Infrastructure model of ~~Hadpoo~~ hadoop

1. Data model →
2. computing model →

Distributed data model	VS Hadoop Data model
- deals with tables and relation	- Deals with flat files in any format
- must have schema for data	- NO schema.
- Data fragmentation and Partitining	- Files are automatically divided into blocks.

Hadoop computing model:-





16/3/23

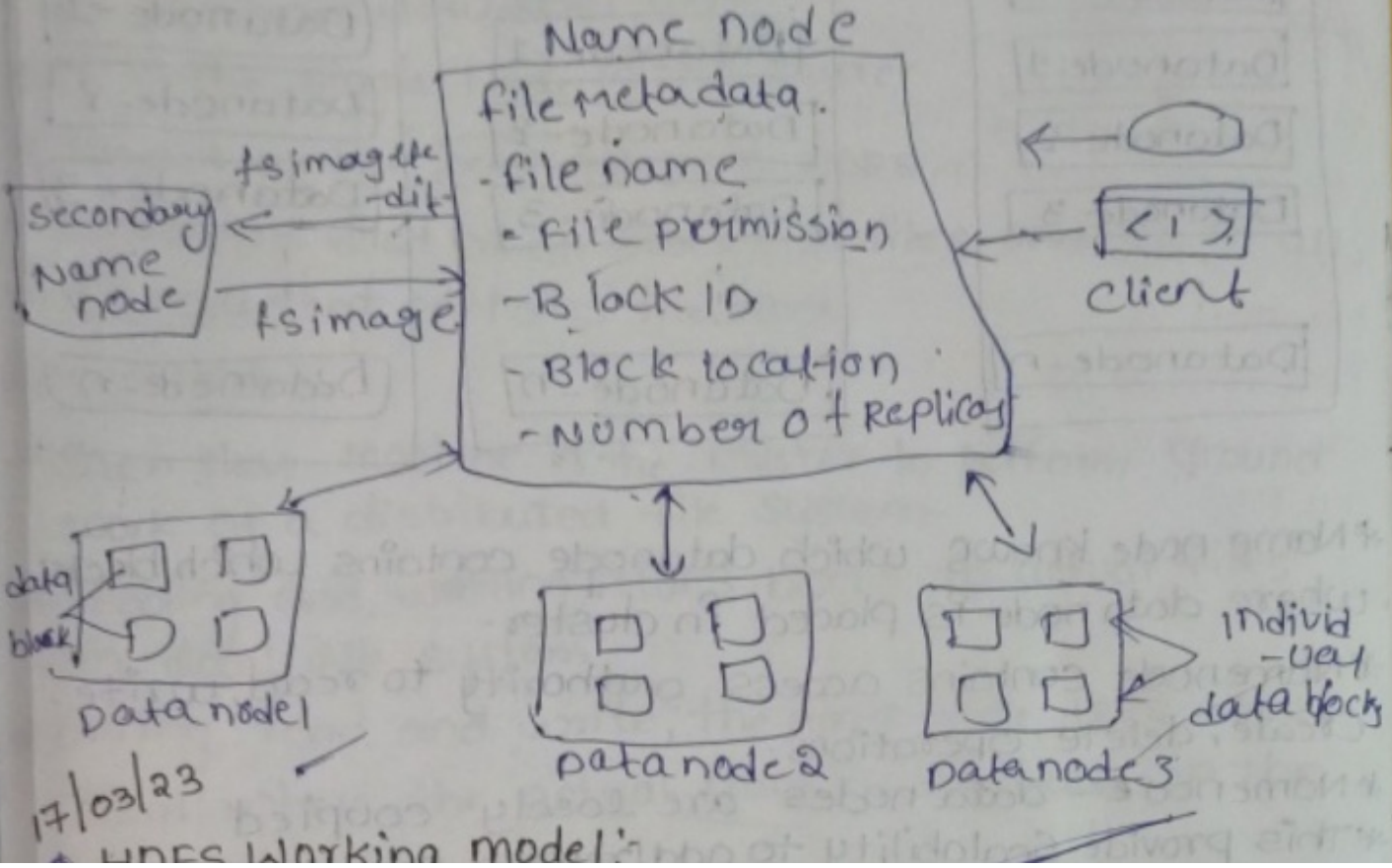
Hadoop Architecture:-

Hadoop Architecture consist of

- Hadoop Distributed file System
- Hadoop map Reduce
- Hadoop common, Hbase, Hive, Pig, YARN etc,

* NameNode is master and Datanode are slaves
* Job Tracker is the master and task tracker is the slave

* every slave node come with task track deamon and data node synchronize with the job tracker and namenode respectively.



17/03/23

* HDFS Working model: