Apache Oozie:-

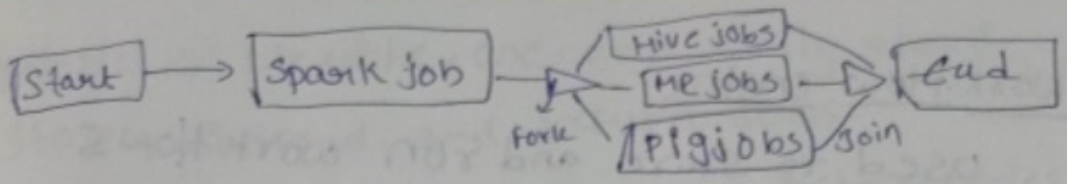
* Oozie is a job scheduling system integrated with Hadoop.
* used to execute multiple jobs in parallel.
* Integrated Spark, Hive, scala
* It is a workflow scheduler for Hadoop.
* It is a system which runs the workflow of dependent jobs
* user permitted to create DAG (Directed Acyclic Graph of workflow which can run in parallel and sequential

3 types of jobs

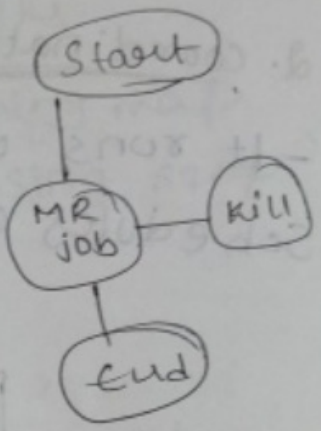
1. Oozie work flow jobs*
   - DAG is used to define jobs edges specifies action.
2. Oozie coordinator jobs
   - work flow jobs triggered by time and data availability.
3. Oozie bundle.
   - Multiple coordinator.

How Oozie works:-

(1) Oozie work flow jobs



- output of the previous action is the input of the present.
- output of the present is the input of the future.



31/3/23

features of oozie:-

* Oozie has client API and command line interface which can be used to travel, control and monitor jobs from Java application.
* using webservices API can control the jobs from anywhere.
* provision to execute jobs which are schedule to run periodically.
* provision to send email notification upon completion of jobs.

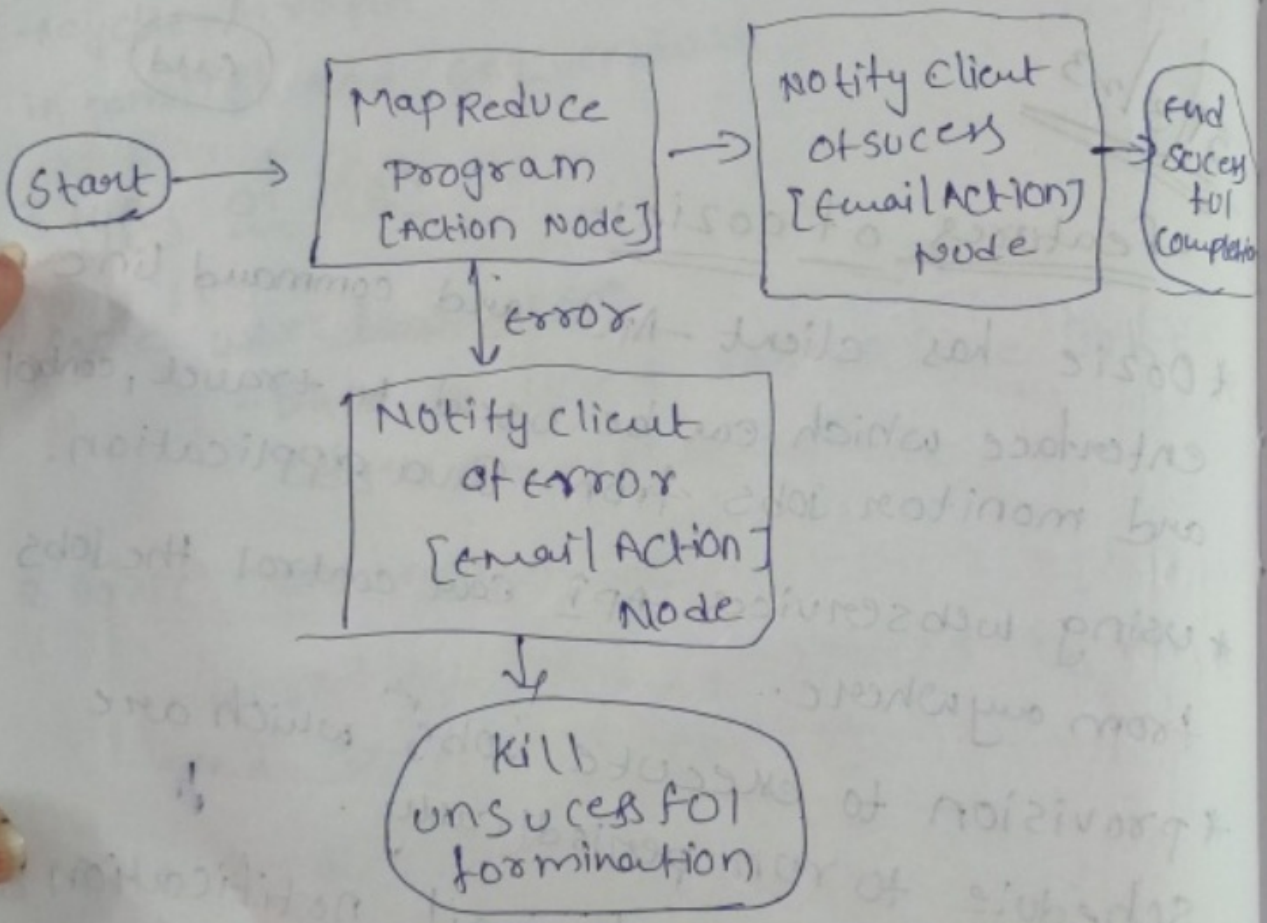
1. Workflow engine:

Used to store and run workflows composed of hadoop jobs.

-Eg: Map reduce, Pig, Hive.

2. coordinator engines:-

- It runs workflow jobs based on predefined schedules and availability of data.



*Oozie is scalable and can manage the finally execution of thousands of workflow (each consist of dozens of jobs) in a hadoop cluster.

* Oozie is flexible, one can easily start, stop, suspend and return jobs.

* Oozie workflow consist of Action nodes and control flow nodes.

Action node:-

- Represents work flows tasks.

(Eg:) moving files to HDFS, running map reduce jobs import data using sqoop or running shell script.

* Control Node:-

control work flow execution b/w actions

→ allowing conditional logic where different branches may be followed depend upon the result of the earlier node.

→ Start node:- It used to start the work flow job.

→ end node: it signals the end of the job

→ error node:- it designated to the occurs-ce of the error and notify the error to notify print

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.

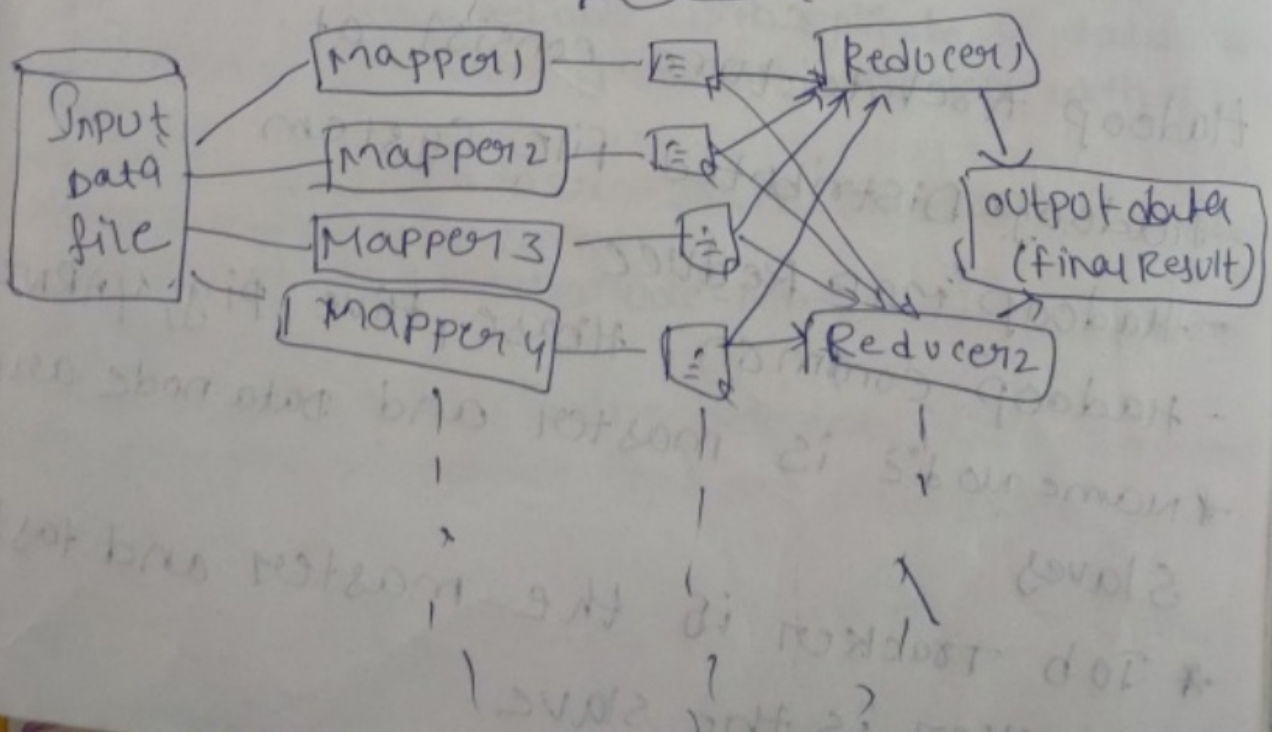
Hadoop Infrastructure.

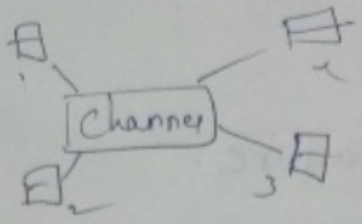
Two Infrastructure model of ~~Hadpoo~~ hadoop

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

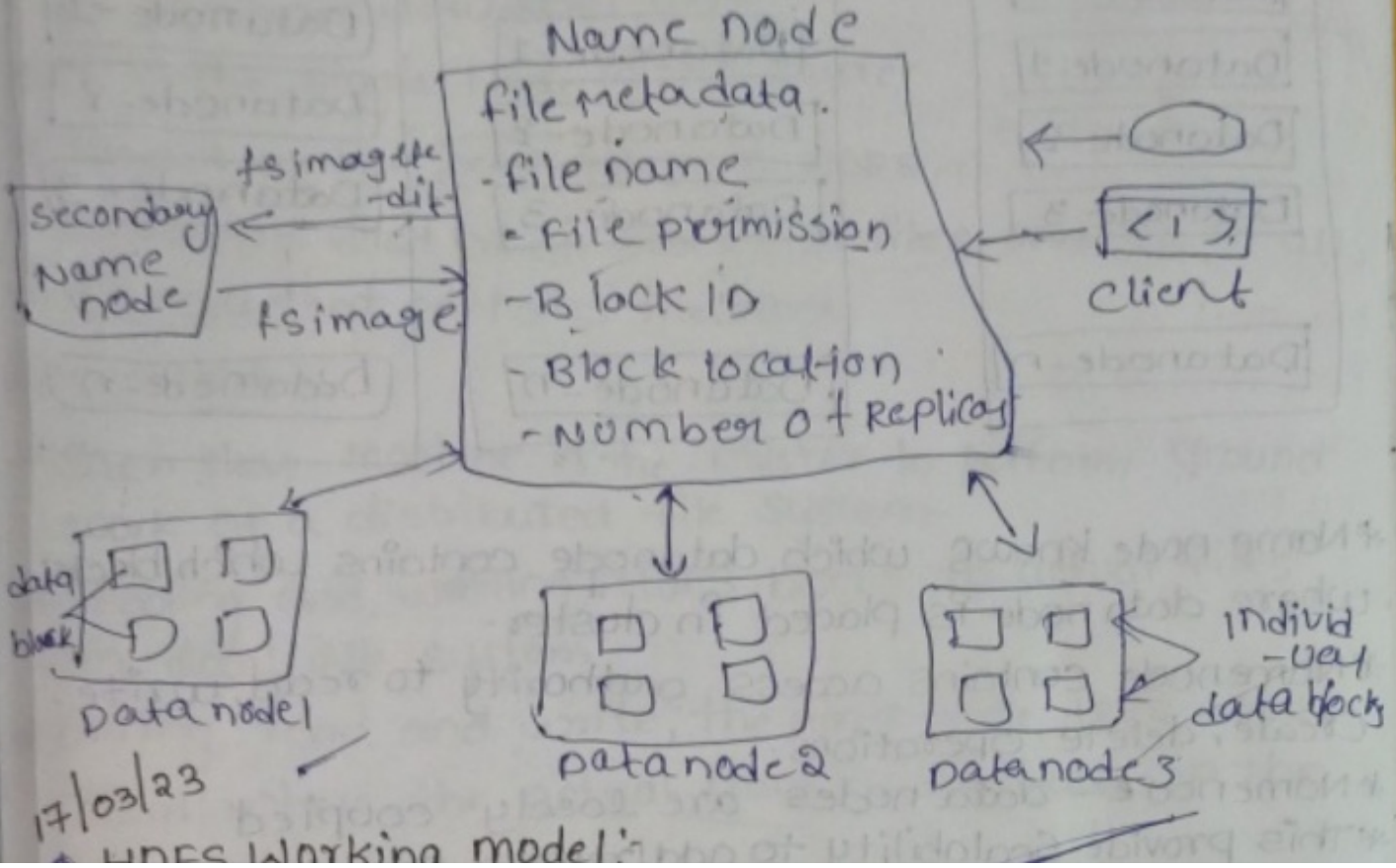
Hadoop Architecture:-

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: