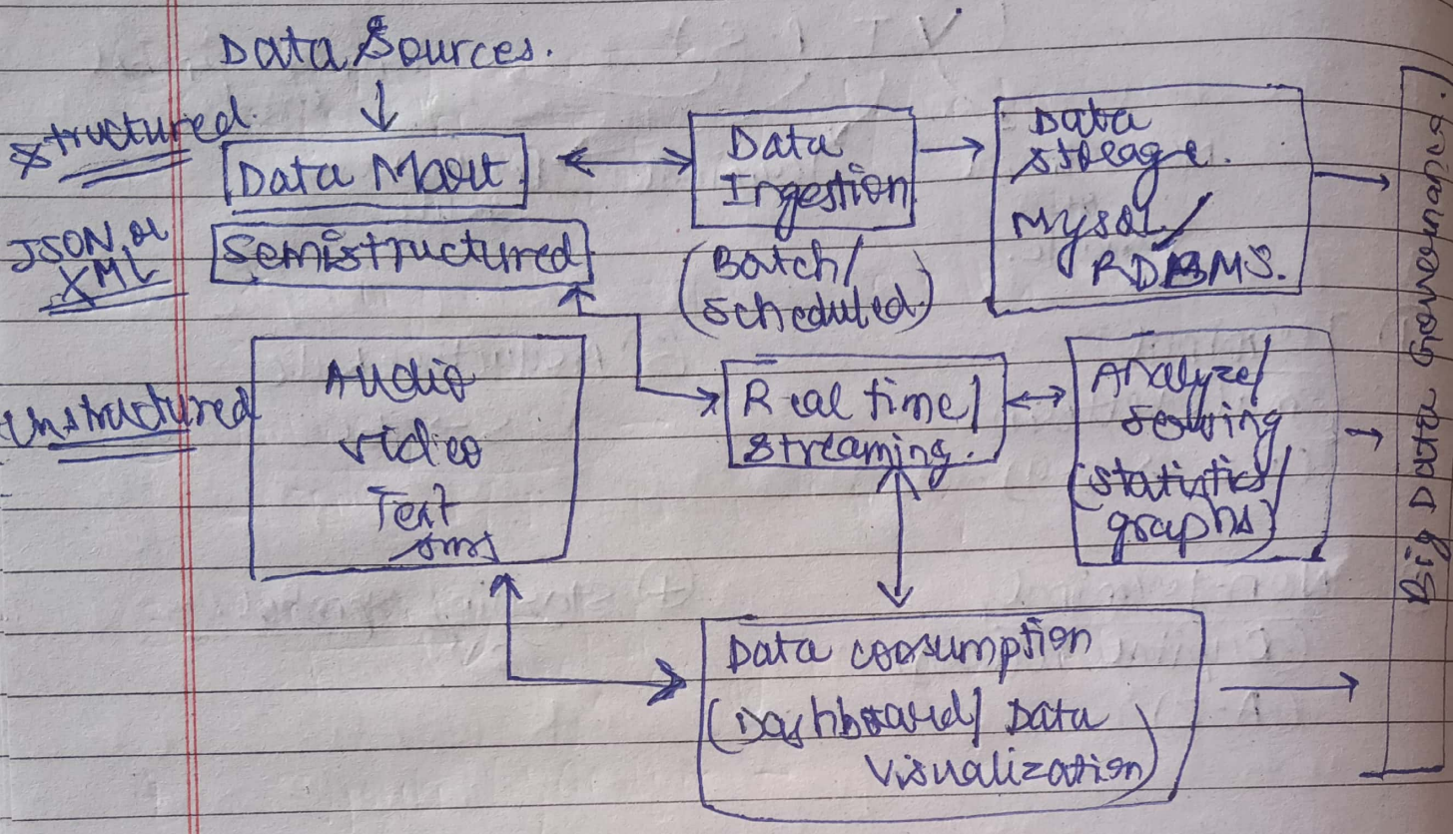
Date / / Big Data. Page No.:

Architecture of Big Data:



~~Importance of Big Data~~

Data Flow:

1. Capture Data.
2. Storing Data.
3. Processing Data / Data Ingestion
4. Data Visualization.

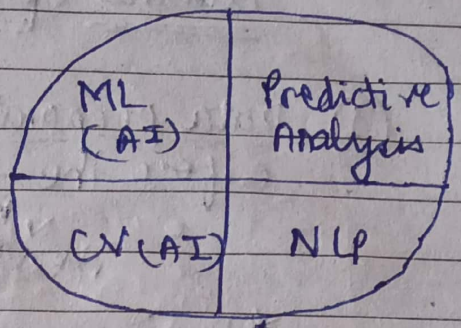
Fields in which Big Data is used :-

1. Predictive Analysis ~~(Machine Learning)~~ [Trends/ graphs]

"Each one prays to God according to his own light."

Date ___/___/_____ Page No.: _____

(2) Machine Learning. (Pattern Matching / Recognition)
(part of AI)



(3) Natural Language Processing (NLP)
(Text as data)

(4) Computer Vision (CV).
(Images as data)

Technologies used for Big data:- (large dataset)

(1) Apache Hadoop (It is an open source framework)
(2) Spark. (by Apache)
↓ used in combination with Hadoop
→ open source processing engine.

↓ distributed computing.
→ Scalability
→ Availability
→ flexi-ble

(3) Apache Flink. (open source stream processing framework)
→ (User friendly api)
→ (live streaming of real-time data).

(4) Presto (Open source SQL engine support ~~all~~ interactive analysis on huge dataset)

(5) Druid (open source analytical data storage. designed for quering eventbased data)
→ (log files)

MapReduce Hortonworks
Cloudera Big data Appliance. (Oracle)
IBM BigInsights.

"Let us always meet each other with a smile, for a smile is the beginning of love." –Mother Teresa

Date __/__/____ Page No.: ____

Features of Big Data :-

(1) Data preparation → Before the ~~data~~ model is applied/ constructed.
iterative
*During model construction but, data iteration before.

(2) Data Exploration → (~~Insights~~ / Visuals ~~used to explore data~~)

(3) Scalability → ~~should~~ use less energy and less cost.

(4) Supports for Analytics → various types of analysis

(5) Version control → For every version, checks are made.

(6) Data Management →

(7) Data Integration → collection of different types of data from diff. sources.

(8) Data governance → ~~data should~~ accurate, usable, reliable, and easy to govern.

(9) IDS → Intrusion Detection System.

(10) Visualization → visual representation.

"Each one prays to God according to his own light." —Mahatma Gandhi

Date ___/___/___

Audit and Analysis → verification of strategies adopted by the organisation. and application of big data on it.

Audit and Analysis :-

1. Reduction in operational cost.
2. Improved decision making.
3. High customer retention.
4. High satisfaction rate.
5. Banking → Manufacturing.

* Prediction of Disease symptom
candidate selection algorithm.

BD + ① AI
② Automation
→ correct and accurate result.

* Knowledge base → dataset with output.

5. Fraudulent cases

Big Data

Unit-2 (Hadoop)

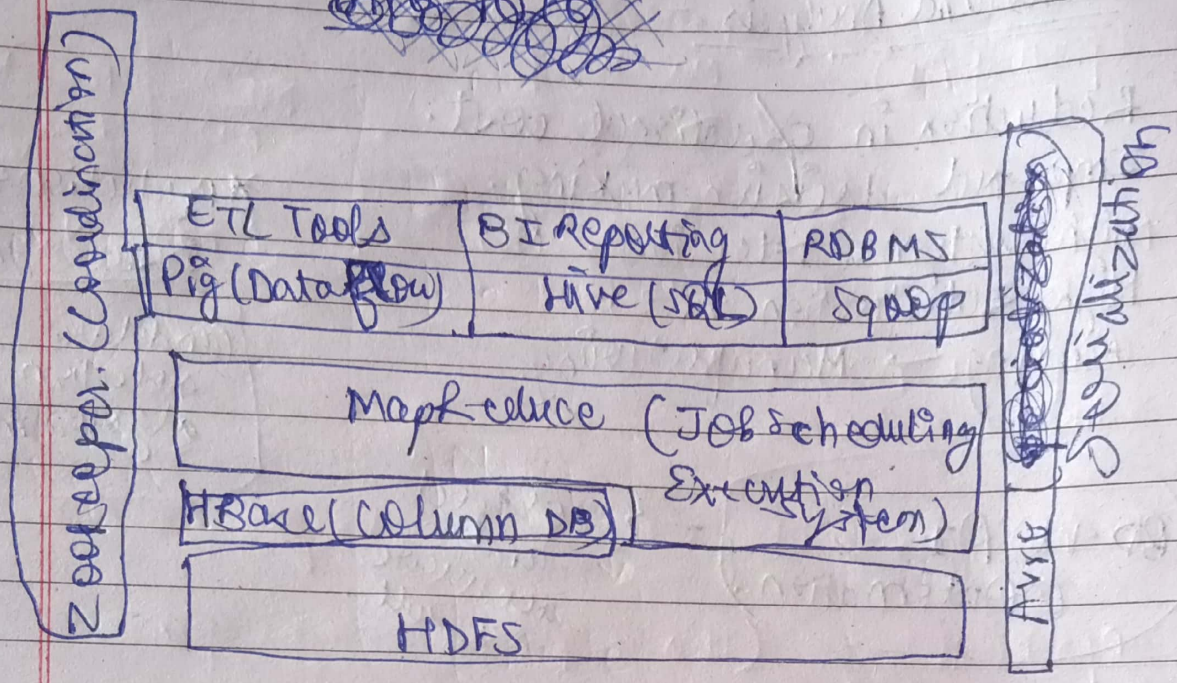
→ large scale batch data processing system.

→ Distributed cluster system.

→ parallel data processing

Features.

1. Access to file systems.
2. Hadoop common package
   - jar files
   - scripts.
   - source code.
   - contributor section.
   - docs



Hadoop Architecture.

Sqoop → CLI.

ETL → Extract, Transform, Load

BI → Business Intelligence.

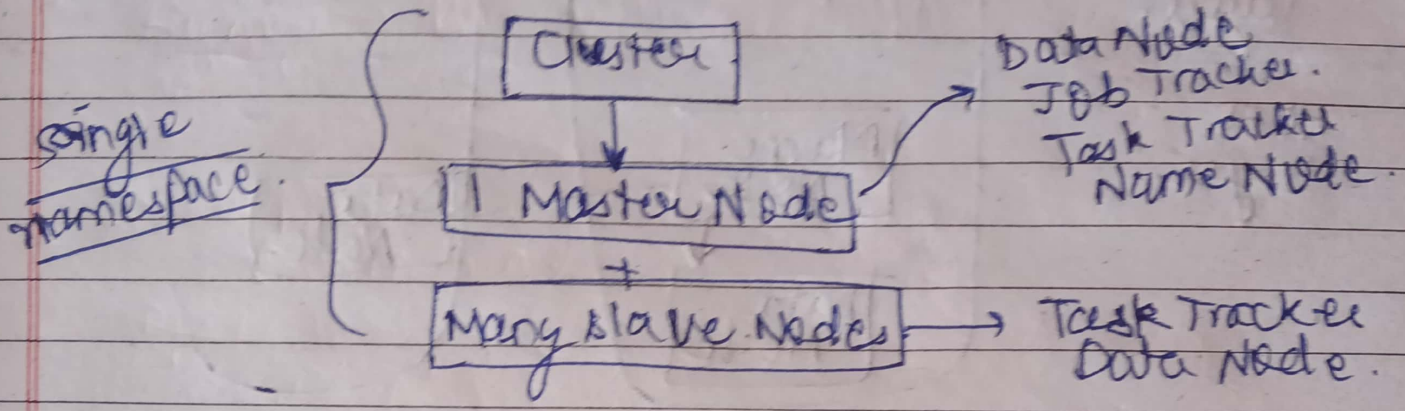
Core components of Hadoop.

- HDFS (Hadoop Distributed File System).
- MapReduce.

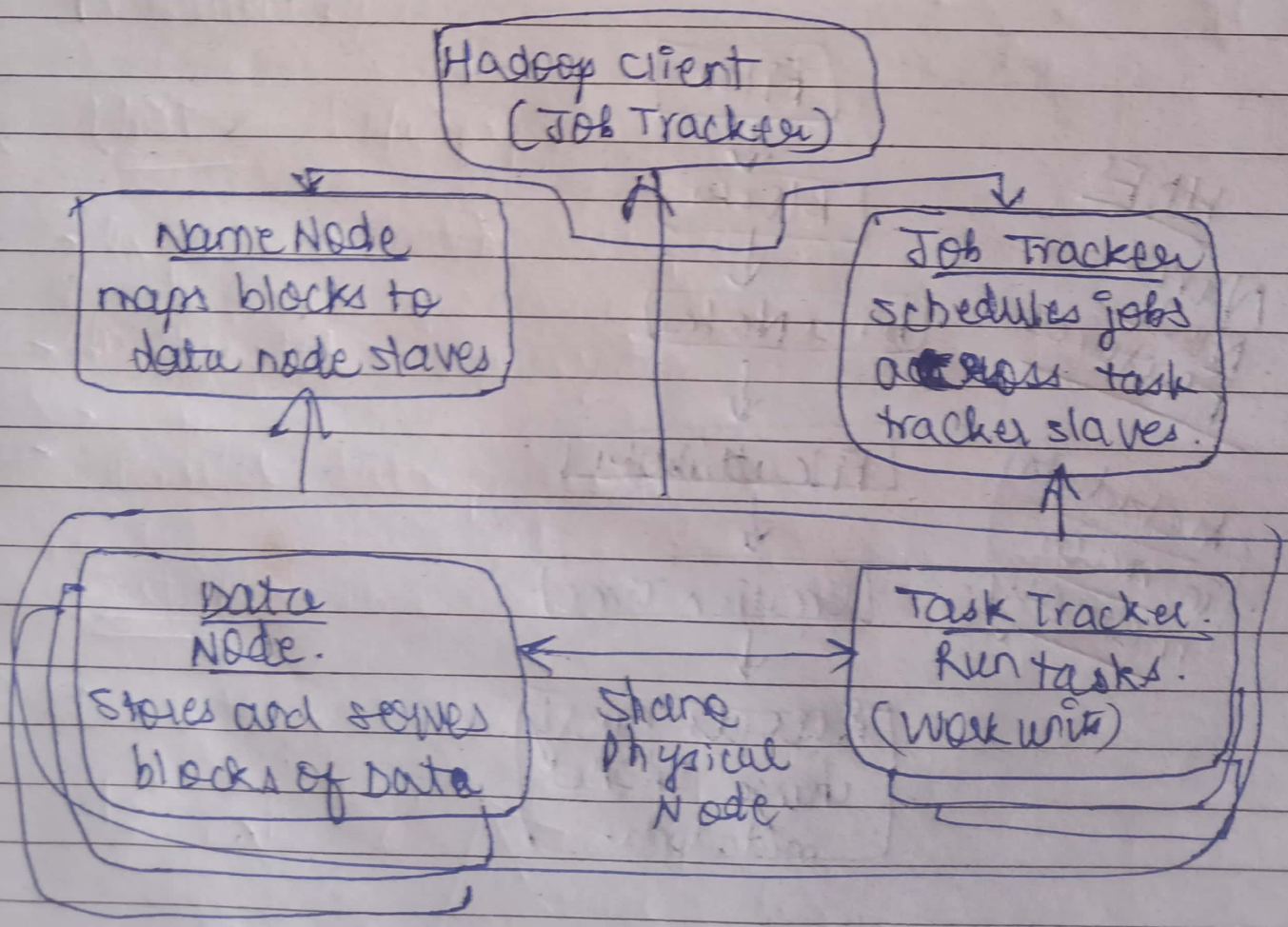
Date __/__/____ Page No.: ____

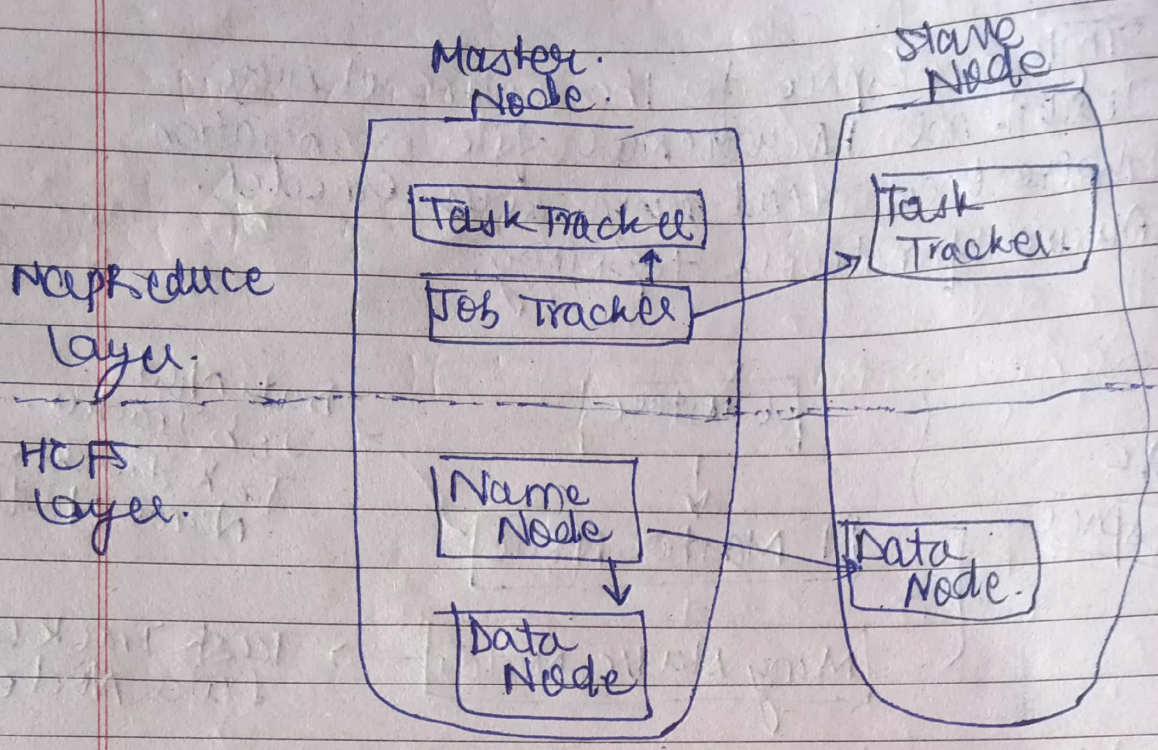
HDFS

→ Single namespace for the entire cluster.
→ Traditional Hierarchical file organization
→ write-once read-many access model.
→ Aware of network topology.



High Level Architecture.





Hadoop Cluster

HDFS

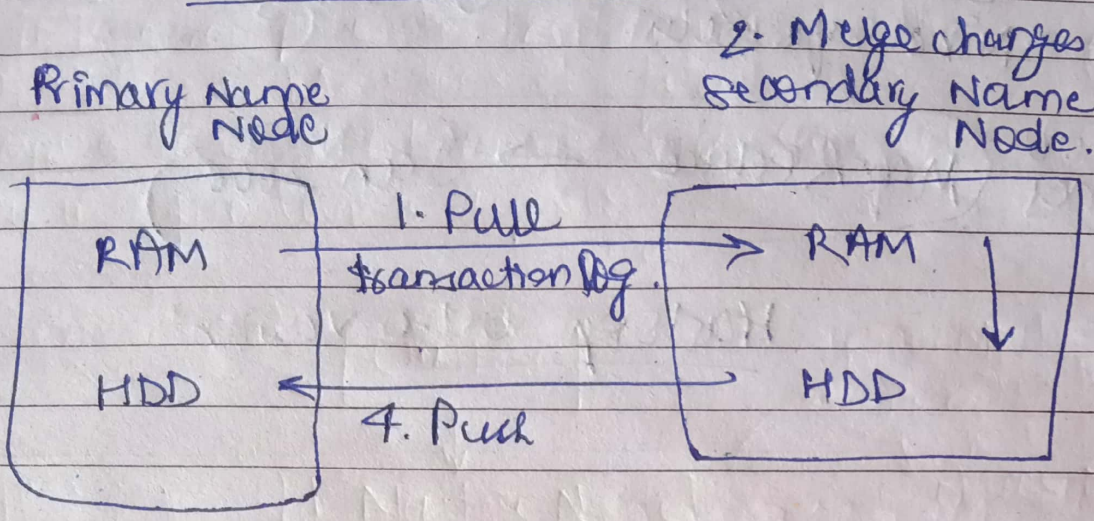
Name Node Features

Files
Blocks
Data Nodes
File attributes
Creation Time
Records every change in the metadata.

Metadata in main memory

Date __/__/____ Page No.: ____

HDFS – Name Node Architecture.



Data Nodes.

→ Block stores data
→ Periodic validation of checksums.
→ Periodically sends a report of all existing blocks to the name nodes.

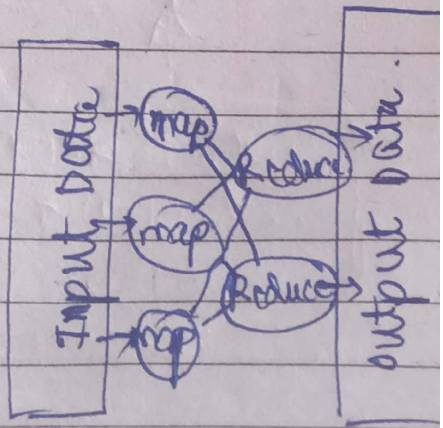
* MapReduce.

Job Tracker. Splitting into map and reduce tasks. Scheduling tasks on a cluster node.

based on weights → split [K1, y1]

sort by K1

Merge. [K1, y1, y2, y3]



Task Tracker. Runs MapReduce task periodically.

5000 lines

HDFS ★ Apache Software Foundation.

→ Doug cutting

→ Mike cafarella. Oct. 2003

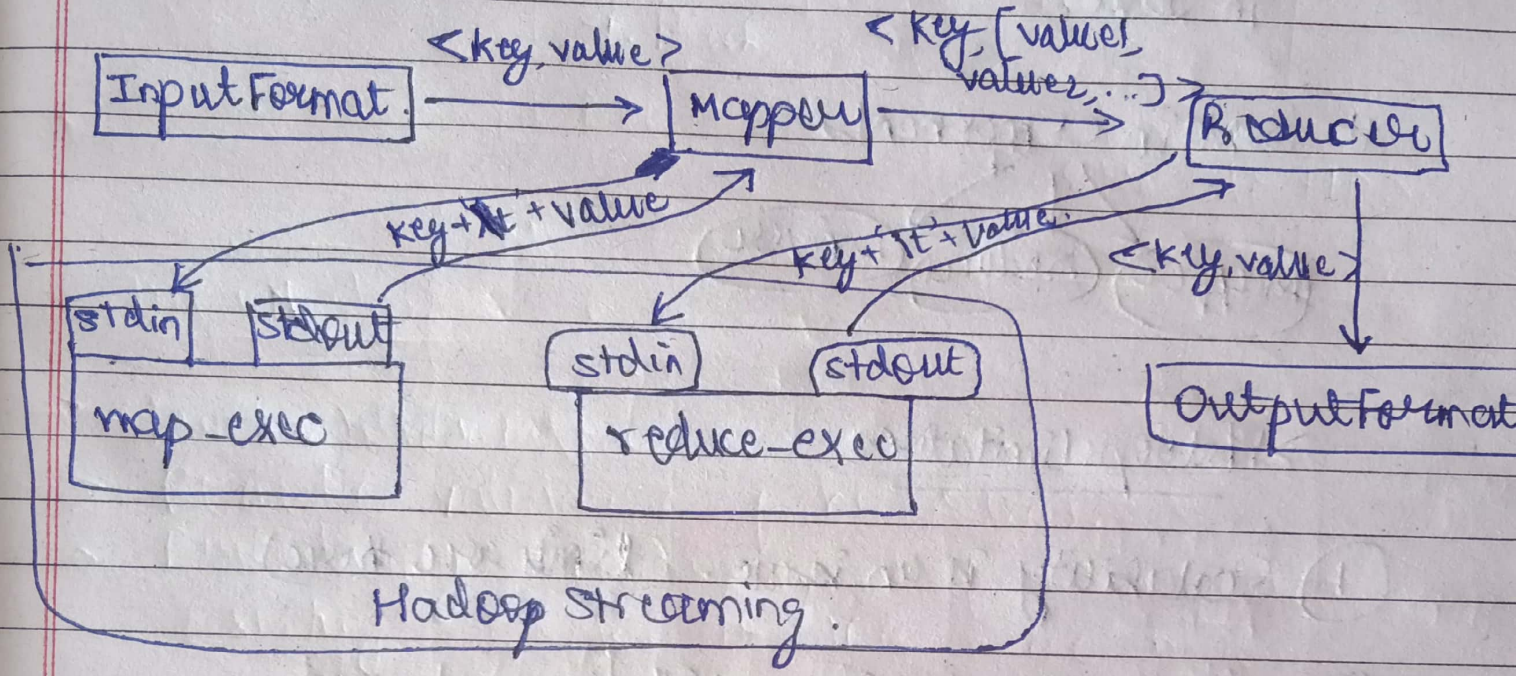
6000 lines (MapReduce → Jan 2006.)

Hadoop 0.1.0 version.

Q5) What is string? Explain any 5 string library functions used in Python with examples?

Submit → 6th September (Wednesday).

Big Data. (Hadoop Pipes).



Second Diagram.

Mapper → PipeMapRunner + PipeMapper.

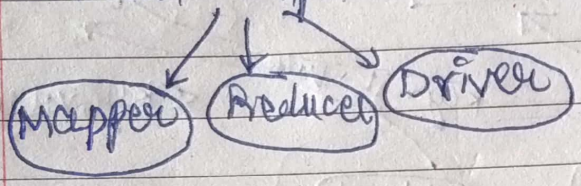
Reducer → PipeReducer.

Date I/O Format.

- inputFormat <javaClassName>
  - JobConf: setInputFormat()
- outputFormat <javaClassName>

written in Java but not necessary to code in Java.

3 Main components



→ Hadoop limitations

(1) Scalability is an issue. (limits are there)

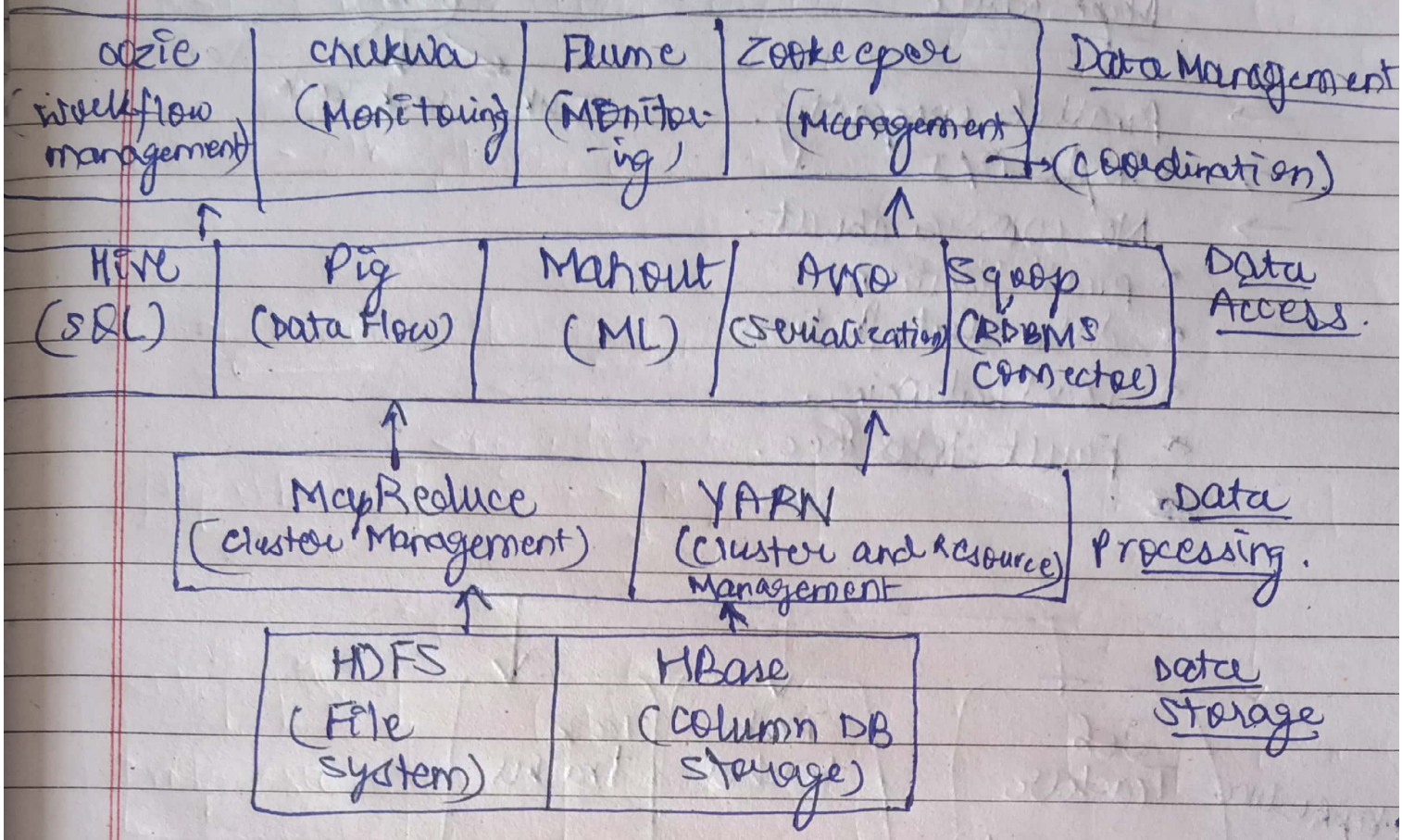
(2) Availability (Failure kills queued and running jobs)

(3) Hard partition of resources into map and reduce slots.

(4) lacks support for alternate paradigms and services.

Date __/__/____ Page No.: ____

Hadoop Ecosystem.



Pig → Pig runtime.

YARN → Resource manager
→ Node manager
→ Application manager.

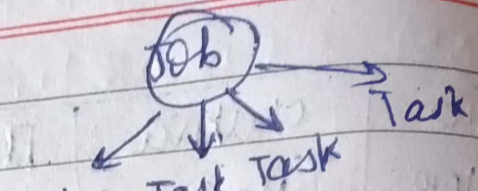
(Yet Another Resource Negotiator.)

"Let your life lightly dance on the edges of time, like dew on the tip of a leaf." –Rabindranath Tagore

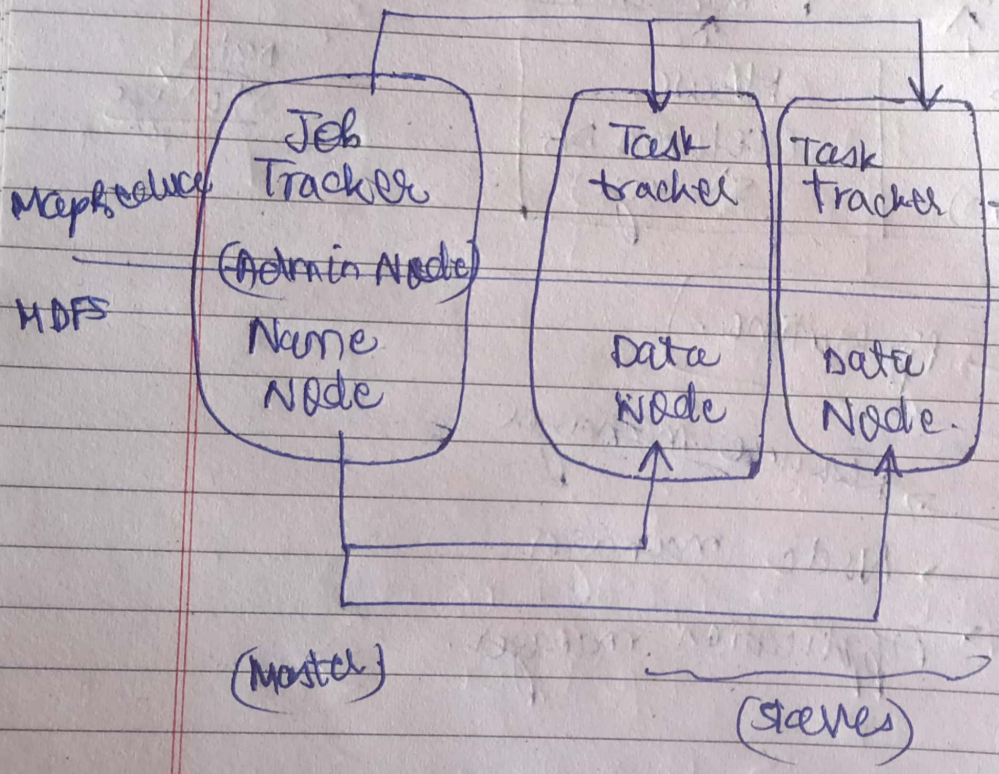
Page No.:

Date __/__/____

Mapreduce.



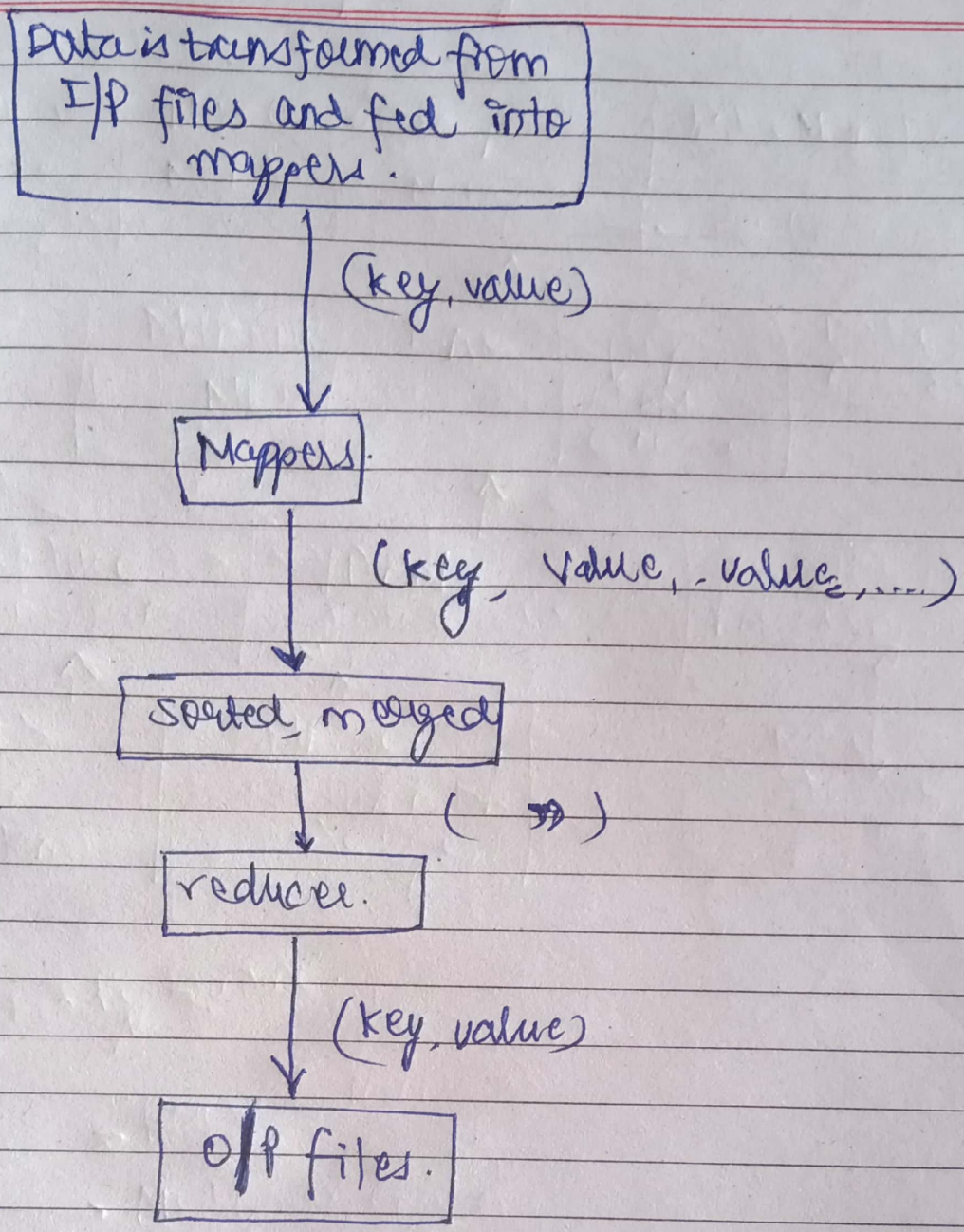
→ Parallel computing.
→ Simple API
→ No worries about:
- parallelization
- data distribution
- Load balancing.
- Fault tolerance.



(*) Heartbeat ⇒ Acknowledge.

[from task tracker to job tracker.]

Date ___/___/_____ Page No.: _____



"Let your life lightly dance on the edges of time, like dew on the tip of a leaf." –Rabindranath Tagore