BDA unit-1 - Hand written Notes

big data analytics (Jawaharlal Nehru Technological University, Anantapur)

UNIT - I

15/3/22

Introduction to Big Data

What is big data:

Data: Qualities, characters (or) symbols on which operations are performed by a computer which may be stored and transmitted in form of electrical signals & record on magnetic optical recording media like CD (or) pendrive is called data.

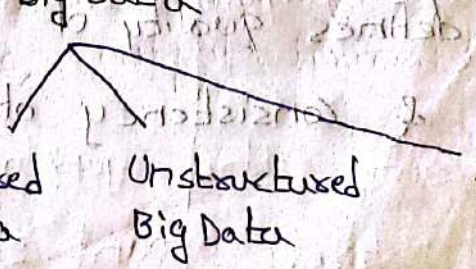
→ Big data is also data with huge size growing exponentially with time.

→ It is collection of large datasets that can't be processed using traditional & migrent tools efficiently.

Examples of Big Data:

* Social Media - Facebook generates 500 TB+ of data per day.
* NYSE (New York Stock Exchange) generates one TB of data per day.
* Google Maps
* Medical Industry
* Transportation

→ There are 3 types of bigdata as follows:



→ Structured Big Data : A data that can be stored & processed in fixed format.

Eg: Employee table in database

→ Unstructured Big Data: A data that can be stored without any fixed format.

Eg: Google search (80% is U.D)

→ Semistructured Bigdata: A data with combination of both structured & Unstructured Bigdata.

Eg: XML (extensible Markup Language)

Characteristics of Big Data:

These are 4 characteristics called as 4 V's

i) velocity
ii) volume
iii) Veracity
iv) Variety

Volume: Defines size (or) quantity of big data

Eg: bits < bytes < KB < MB < GigaBytes (GB) < TeraBytes (TB) < PetaBytes (PB) < ExaBytes (EB) < ZettaBytes (ZB) < YottaBytes (YB)

Velocity: It defines speed of big data.

Eg: KBPS, MBPS

Variety: It defines type of Big data.

Eg: structured BD, semistructured BD, unstructured BD

Veracity: It defines quality of data.

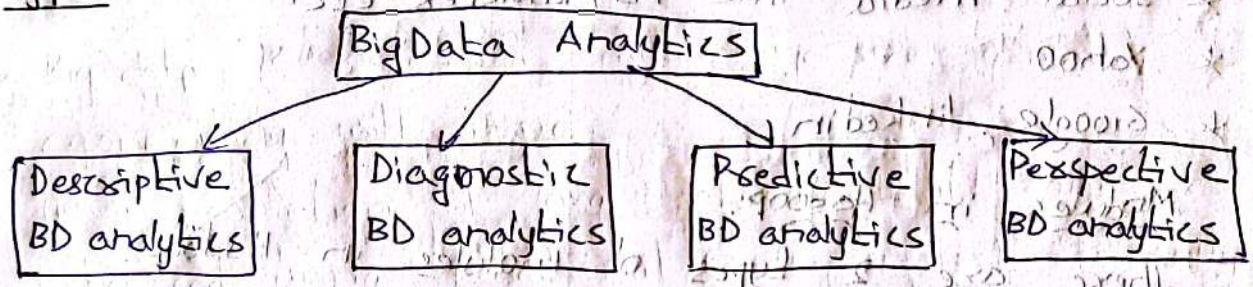
Eg: Accuracy & consistency of Big data

Designations for Big data specialisation:

1. Big data developer
2. Big data engineer
3. Hadoop developer
4. Big Data Analyst

Big Data Analytics: The process of analysis of large volumes of diverse data sets using advanced analytical techniques refer as Big Data analytics.

Types:



Descriptive analytics: It simplifies data & summarise past data into readable form.

Eg: Dow chemical company utilises its past data to increase facility utilization across offices & labs.

Diagnostic analytics: It gives a detail & indepth of data & finds route cause of problem.

Predictive analytics: It makes use of historical (past) & present data to predict future data.

Perspective analytics: It allows business to determine best possible solution to problem.

Why Big Data important

* cost reduction
* Faster & better decision making
* Increased Productivity (Performance)
* Time saving
* Improved customer service
* Fraud detection
* Innovation

Meet Hadoop:

* It is an open source framework from apache.
* It is used to store, process & analyze the data which is in very large in volume.
* It is written in Java.
* It is used for offline processing / batch processing.

Applications:

* Social media like fb, twitter etc.
* Yahoo
* Google, linkedin

Modules in Hadoop:

There are 2 types of modules in Hadoop:

i) HDFS (Hadoop Distribution File system)

ii) Map-reduce

Advantages of Hadoop:

* Resilient to failure
* Fast
* Scalable
* cost effective

Data: Qualities, characters (or) symbols on which operations performed by computers which may be stored & transmitted in form of electric signals record on magnetic optical recording media like CD.

Examples:

Ancestry.com, the genology website (study of families) stores around 2.5 peta bytes per day.

→ large Hadron collider near Geneva, Switzerland will produce about 15 peta Bytes of data per year.

Data Storage & Analysis:

* In 1990's, one drive would store 1370 MB data & had transfer speed of 4.4 MB/sec, within 5 min, all data can be sent.
* Over 20 yrs later, 1 TB drives are the norms with a speed of 100 MB/sec and takes 2½ hours to read all data.
* Since, it takes long time to read all data on single drive and written is even slower.
* To reduce time, we use multiple disks at once but it seems diswasteful & cost effective.
* Hence, we can store 100 data sets, each of size 1 TB and provide shared access to them with shorter analysis types.

Problems in Data Analysis:

There are 2 problems:

i) H/w failure
sol: Replication — RAID: Redundant Copies of data, kept in case of failure
To combine data in disk with others.

ii) Combined data, solution is map reduced.
Hadoop provides - reliable data storage (HDFS)
Efficient analysis (Mapreduce).

Comparision with other systems:

i) RDBMS:

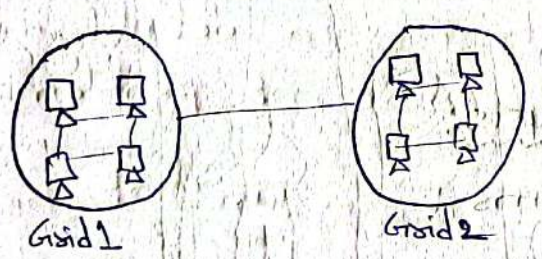
Type of DBMS that stores data in row based table structure.

Eg: MySQL

RDBMS	MapReduce/Hadoop
1) Data size in TB.	Data size in Petta Bytes.
2) It is an interactive(online) & offline processing(batch)	It is offline processing.
3) It reads the data & writes many times.	Data is write once, read many times.
4) Static structure schema	Dynamic structure schema
5) High integrity	less integrity
6) scaling is non linear	scaling is linear

8) Grid Computing:

It is a group of networked computers which works together to perform large tasks.



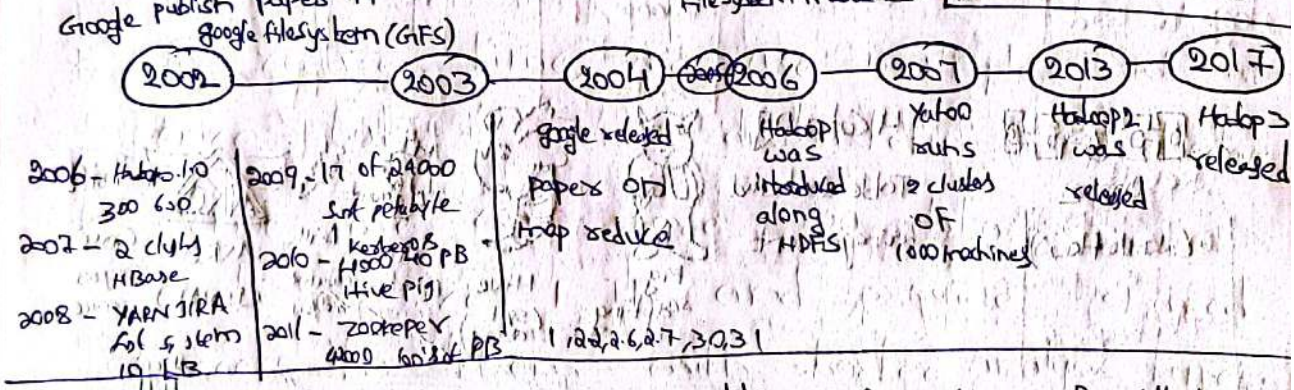
Grid Computing	Map Reduce
1) Data access is slow.	Data access is fast.
2) low level programmed.	High level programmed.
3) Highly expensive.	low expensive

iii) **Volunteer computing**

It is a type of distributed computing where people donate their unused computers to perform large tasks.

Volunteer computing	Map Reduce
1) No data locality	with data locality
2) Multiple data centers	Single data center
3) low bandwidth	High bandwidth

History of Apache Hadoop:



→ Invented by Doug cutting & Mike Cafarella in 2002

→ while doing a project on Apache nutch and publish a papers in GFS

→ while working on apache nutch project they may dealing with big data to store that data they have to spent lot of cost. These pob becomes one of imp reason for emergence of Hadoop

→ In 2004, Google released papers on map reduced

→ In 2005, NDFS was introduced

→ In 2006, Doug cutting quit Google and joined Yahoo and on basis of NDFS, he introduced HDFS & released Hadoop 1.0.

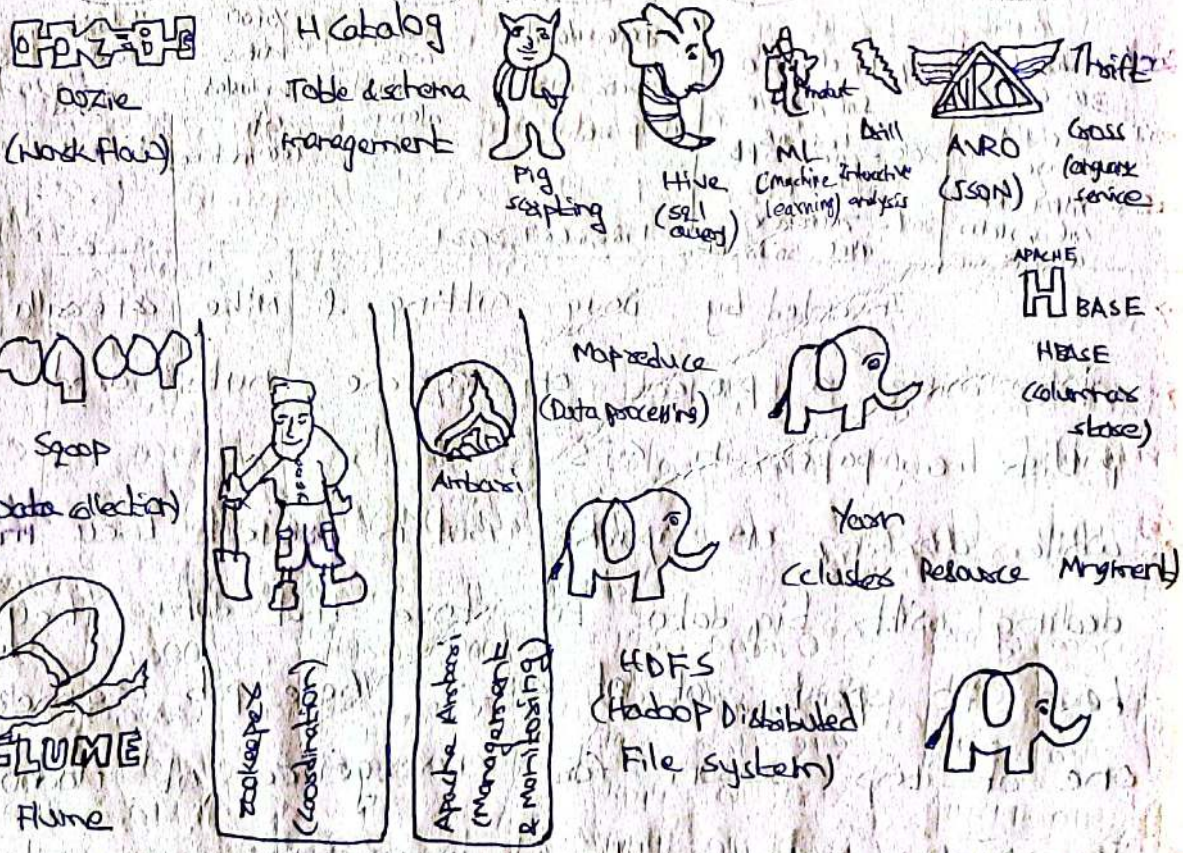
→ He gave named his project Hadoop after his son's toy elephant.

→ In 2008, Hadoop becomes fastest system to sort 1 TB of data on 900 load clusters within 201 seconds.!

→ In 2018, 3.1 released.

Hadoop ecosystem / Hadoop stack / Hadoop structure

Hadoop Ecosystem



Yet Another Resource Negotiator

HDFS (Hadoop Distributed File system) which is used to store large data sets.

MapReduce: It is a programming based data processor.

YARN: Yet Another Resource Negotiator used to manage the resources.

SPARK: Used for in memory data processing.

PIG, HIVE: Used for query based processing of data services.

Eg: SQL

HBASE: Used for NO SQL database

Mahout, Spark MLlib: Used for ML algorithm libraries

Zookeeper: Used to manage the clusters.

Oozie: Used for job scheduling.

Solar, Lucene: Used for searching an index.

Analyzing the data with Hadoop:

Map reducing is a programming model to analyse an process data with Hadoop

Hadoop can run mapreduce written in various language like Python, Java, C++.

Eg:

Map reduce program works in 2 phases.

1. Map phase
2. reduce phase.

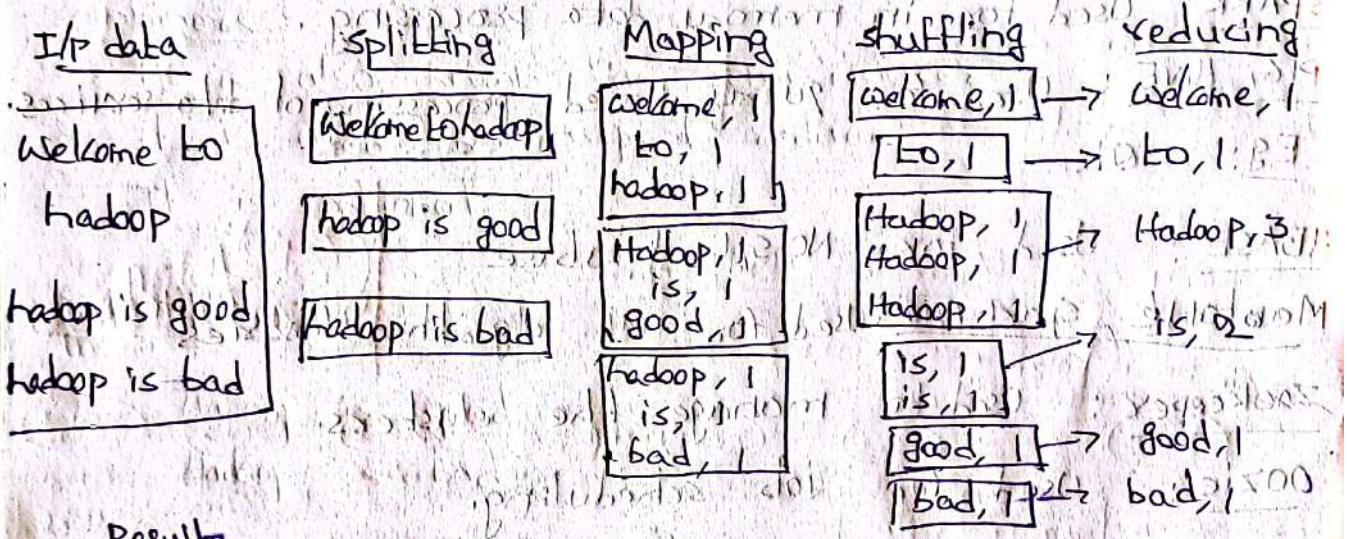
In Map phase it has 2 sub phases: 1. I/P splitting
2. Mapping

→ I/P to map reduce job is divided into fixed size pieces called i/p splits.

→ Mapping is a phase where each i/p split is passed to mapping function.

Eg: <word, frequency>

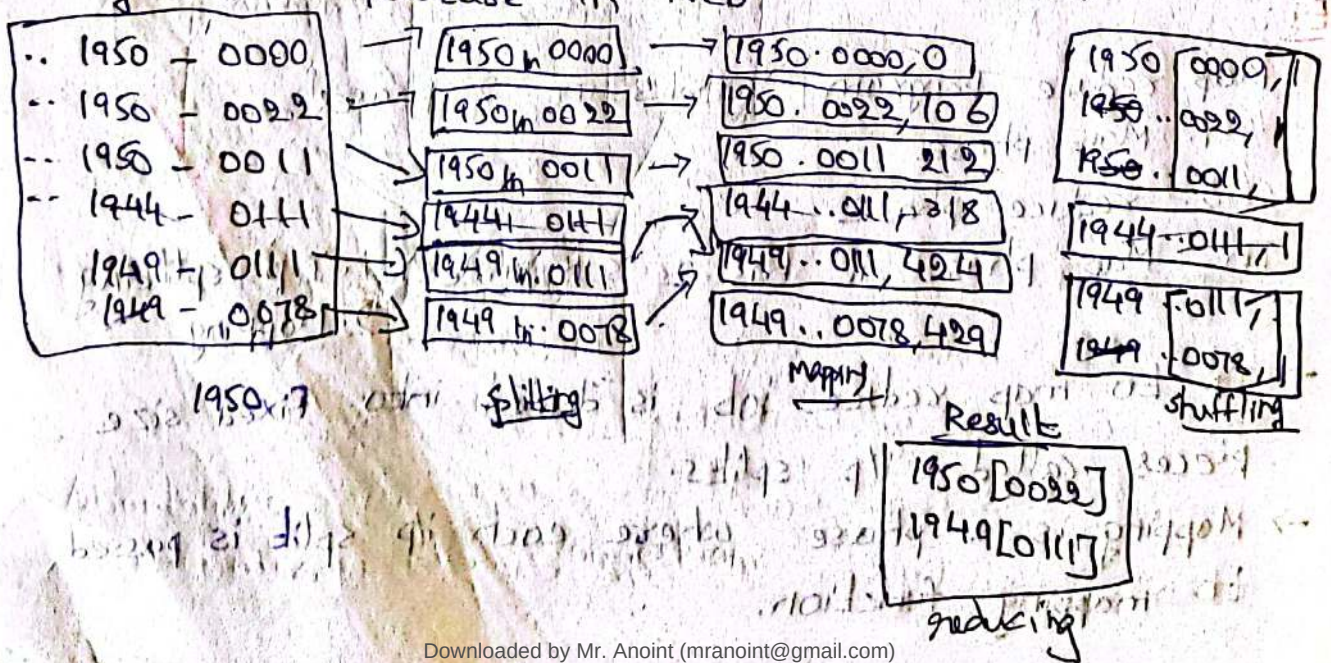
- Reduce phase is again subdivided into 2 phases.
  I - shuffling II - Reducing
- shuffling phase is used in to console data relevant parts.
- Reducing phase is used to aggregate data.



Result

welcome, 1
to, 1
Hadoop, 3
is, 2
good, 1
bad, 1

Ex:2:- weather climate set below data format, weather data set from NCDC (National Climate Dataset). Find the highest temperature in NCD. center



Scaling out:

- To scaling out the Hadoop we need to store the data in distributed File System Typically HDFS.

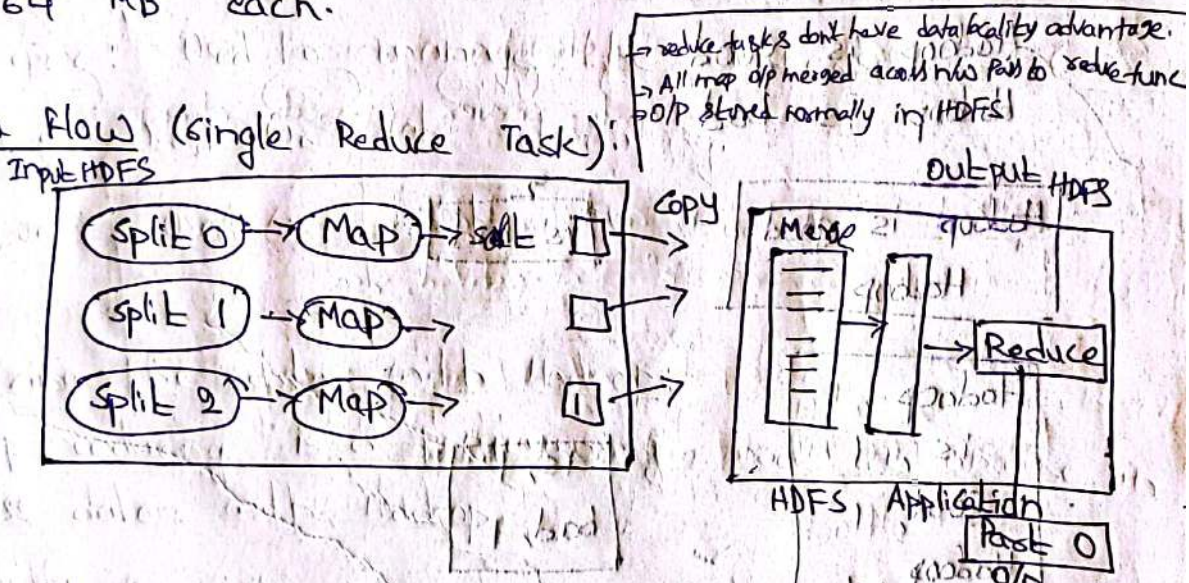
Data Flow in HDFS:

A map reduce job consists of i/p data, map reduce program & configuration info.

- Hadoop runs the job by 2 type of tasks:
  - map task
  - reduce task
- Hadoop contains 2 types of nodes
  - job tracker (single)
  - multiple task trackers (several)
- Hadoop divides data into fixed pices with size of 64 MB each.

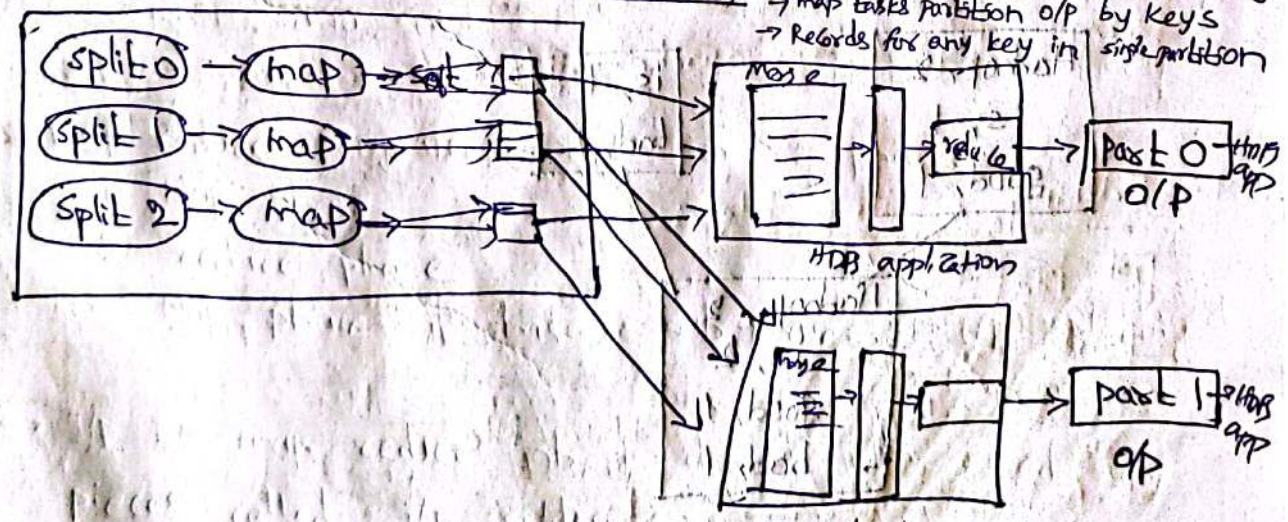
Data Flow (single Reduce Task):

→ reduce tasks don't have data locality advantage.
→ All map o/p merged across nw pass to reduce func
→ O/P stored normally in HDFS

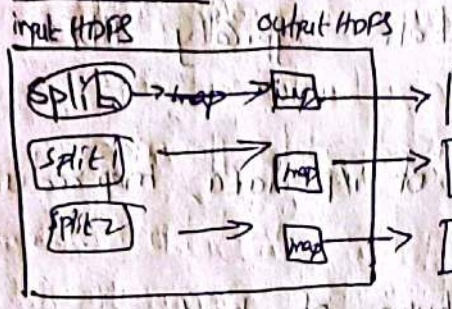


Data Flow (Multiple Reduce Task)

→ no. of reduce tasks specified independently
→ map tasks partition o/p by keys
→ Records for any key in single partition



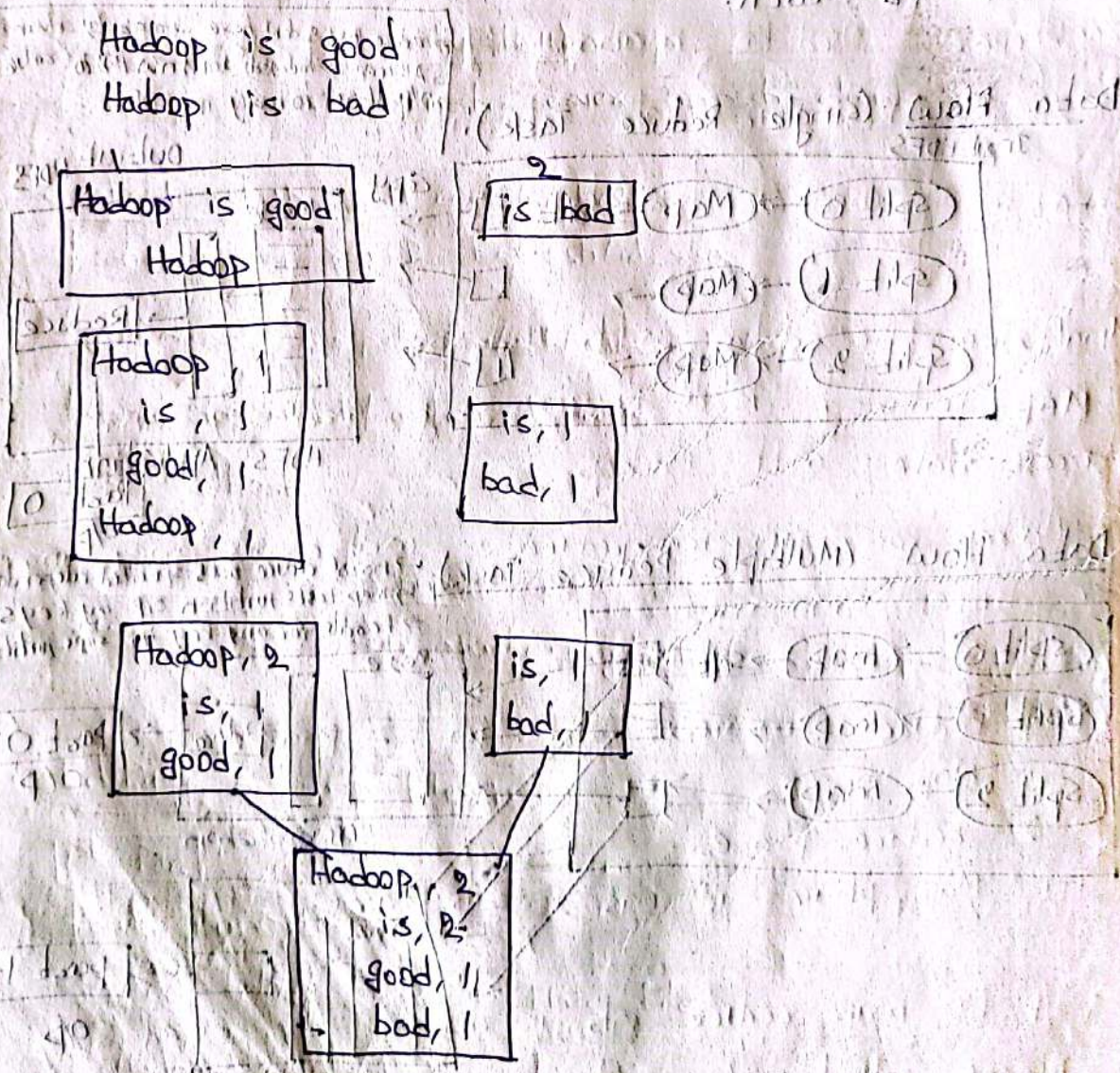
Data Flow — No Reduce Task



Combiners:

- It is also known as semi-reducer
- It is used b/w map & reduce
- It is optional

Data Flow



Advantages:

Minimize N/w Congestion

Experiment 1: Install Apache Hadoop

VMware installation for Hadoop:

Aim: To install single node Hadoop cluster backed by the Hadoop Distributed File system on Ubuntu using VMware

Different types of Linux OS:

parrot, cent OS, kali, Ubuntu

Installing VMWare:

i) Double click to launch VMWare-workstation-full-15 application.
ii) Security warning panel & click on Run to continue.
iii) Initial screen will appear, wait for process to complete.
iv) VMware Workstation setup wizard open, click next.
v) select I accept terms within license agreement, click on next.
vi) select directory during which you'd wish to install the appliance also select enhanced keyboard drivers checkbox
vii) leave it to defaults setting & click next.
viii) select both options desktop & start Menu programs folders & click next.
ix) click install to start installation process.
x) Installation in progress, wait for this to complete.

Install ubuntu in VMware:

i) open VMware workstation & click on create new VM.
ii) choose Installer disc image file (iso): to make ws to detect that iso file is appropriate or not.
iii) Fill info about full name, user name & PW.
iv) Then click next & give your VM relevant name.
v) Allocate size of Hard disk.
vi) Run VM
vii) Install Ubuntu 20.04 LTS Desktop.
viii) To begin installation, click install Ubuntu
ix) choose keyboard layout
x) This will take while to complete

3. Run Ubuntu in VMware

Open VMware run ubuntu 64 bit virtual machine

login to ubuntu OS

open terminal or use "ctrl + alt + t"

4 - check ubuntu is updated

$ sudo apt update

5. Installing Java:

$ sudo apt install openjdk-8-jdk -y

checking Java version

$ java -version

$ javac -version

6. Installing SSH (security shell)

$ sudo apt install openssh-server openssh-client -y

7. Create & setup SSH certificates:

$ ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa

$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

$ chmod 0600 ~/.ssh/authorized_keys

$ ssh localhost

8. Downloading Hadoop:

$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz

9. Editing 6 imp files:

1st file:

$ sudo nano .bashrc

Add below lines in this file

Hadoop related operations

export HADOOP_HOME = /home/bigdata/hadoop-3.3.1
export HADOOP_INSTALL = $HADOOP_HOME
export HADOOP_MAPRED_HOME = $HADOOP_HOME
export HADOOP_COMMON_NAME = $HADOOP_HOME
export HADOOP_HDFS_HOME = $HADOOP_HOME
export YARN_HOME = $HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR = $HADOOP_HOME/lib/native
export PATH = $PATH: $HADOOP_HOME/sbin: $HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.librarry.path = $HADOOP_HOME/lib/native"

Now type:
$ source ~/.bashrc

2nd file:
$ sudo nano $HADOOP_HOME/etc/bigdata/hadoop-env.sh
# Add below like in this file in end
$ sudo nano $HADOOP_HOME/etc/bigdata/hadoop-env.sh
$ export JAVA_HOME = /usr/lib/jvm/java-8-openjdk-amd64

3rd file:
$ sudo nano

10 Marks

1. What is big data? Explain it detail
2. Why big data is important? Discuss in detail
3. Explain (i) data (ii) Data storage & analysis
4. Distinguish hadoop with other systems
5. Elaborate history of apache hadoop
6. Explain hadoop ecosystem
7. Discuss about VMware
8. List steps in hadoop installation
9. How can you analyze data with hadoop
10. Discuss scaling out in hadoop