ASSIGNMENT 1

ANALYTICS FOUNDATION (IB211)

Kartikey Garg
18010 9015
CSE-BDA

Q-1 → Explain the characteristics of Big Data

Answer → A. Variety → Big Data is collected & created in various formats & sources. It includes structured as well as unstructured data like text, multimedia, social media, business reports etc. Structured data such as bank records, demographic data, inventory database, business data, have a defined structure and can be stored and analysed using traditional data management and analysis methods. 1 of the 8 main objective of Big Data is to collect all the unstructured data and analyse it using the appropriate technology. Data Crawling also known as web Crawling is a scraper technology used for systematically browsing the webpages. There are algorithm designed to reach the maximum depth of a page and extract useful data worth analysing.

Example → 1. Health Care → 150 Exabytes

As of 2011 the global size of data in health care was estimated to be 150 Exabytes (Billion GB).

2. Social Data i). 30 Billion pieces of content are shared on facebook every month.

ii). 4+ Billion are hr/min of videos are watched on Youtube.

iii). 400 M tweets/day by about 200 M monthly active users.

B. Volume → The main characteristic of Big Data is its huge volume, collected through various sources. The data used to measure is in GB or TB. However big Data volume created is in Zetta Bytes (ZB) which is equal to a trillion GB {1 ZB ≃ 3 Million Gaalaxies equivalent to appeon. Of Stores.

Example → i). 40 ZB of data will be created by 2020, an increase of 300 times from 2005.

ii). 100 TB → Most companies in US have atleast 100 TB of data stored

iii). 6 Billion Cell Phones are used to Generate data and world population is 7B.

iv). 2.5 B GB of data created each day
↳ Billion Giga Bytes

...ntify the Synonyms of the given word.

d) Vitalize

Directions : Identify the Antonyms of the given word.

1.Provident

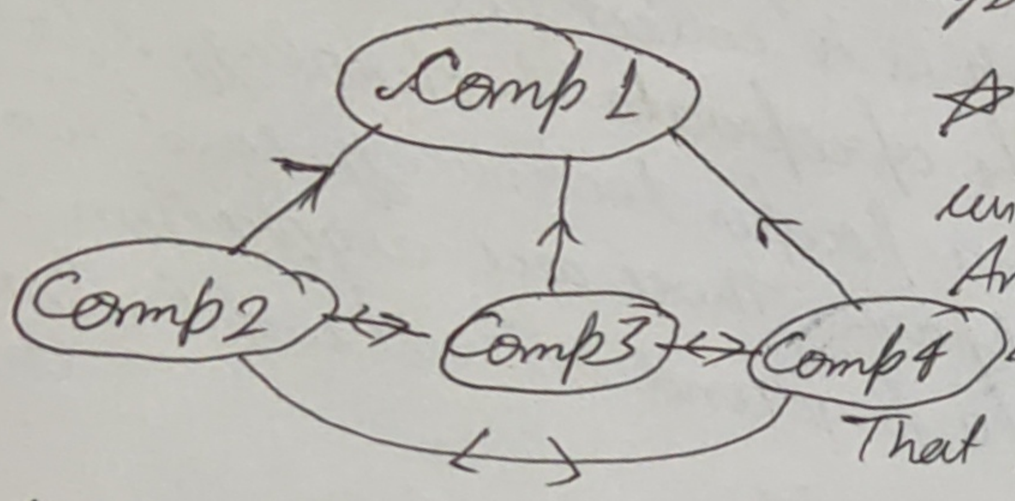
...areless✓

The 4 industries that categorise the data :-
Industry 1 - Water
Industry 2 - Electric Power (Generated by Water)
Industry 3 - Computing (Using Social Appl)
Industry 4 - Sensors (IOT Things)

To overcome the problem of excess of unstructured data we use Hadoop. Its an Open Source Framework uses HDFS

→ It is a master detailed File System.

Hadoop Distributed File System.



☆ All these PC's are interconnected with each other (or known as "slaves"). And There 3 have direct access to the main computer (master slave (comp 4)). That tracks each of them performances.

3. Velocity -> It refers to the speed at which vast amount of data being generated, collected & analysed. Everyday the no. of emails, twitter, messages, photos etc., increases at lightning speed around the world. Big Data Technology allows us to analyse the data while it is being generated without putting it into databases.

For Example i). 1 TB of Trade Information -> The New York Stock Exchange captures 1 TB of Trade Information during each trading session.

ii). 100 Sensors -> Modern cars have close to 100 sensors that monitor items such as fuel level & tyre pressure.

iii). 18.9 Byte Network Connections -> By 2016 it is predicted, there will be almost 2.5 connections per person on earth.

iv). IOT (Internet of Things).

4. Vacanity -> It is the quantity i.e., quality or trustworthiness of the data just how accurate is all the data.

i.e., Gathered & Analysed.

It is defined as the uncertainty of data.

Eg) i) 1 in 3 business leaders → Don't trust the information they use to make decisions.

ii). 27% of respondents → In 1 survey it is unsure that how much of data was inaccurate because not everyone responded to that survey.

iii). $3.1 trillion a year → Poor data quality costs the US economy around $3.1 Trillion a year.

Q-2. What is the Difference between Hadoop and Traditional Relational Database Management System?

Answer → i) Data Volume → RDBMS works better when the volume of data is low (in GB). Otherwise, it fails to give the desired Results.
Hadoop work better when better when data size is big. It can easily process & store large amount of data efficiently as compared.

ii) Architecture → Hadoop → HDFS (Hadoop Destributed File System),
→ Hadoop MapReduce (a programming model to process large data set),
→ Hadoop YARN (used to manage computing resources in clusters).

Traditional RDBMS → ACID properties => Atomicity, Consistency, Isolation & Durability.

iii) Throughput → RDBMS fails to achieve a higher throughput as compared to Apache Hadoop Framework.

iv). Data Variety → Hadoop has the ability to process and store all variety of data whether it is structured, semi-structured, unstructured. Although it is mostly used in all amount of unstructured data.
Traditional RDBMS is used only to manage structured and semi-structured data. So Hadoop is better than Traditional RDBMS.

v). Latency / Response Time → Hadoop has higher throughput, you can quickly access batches of large datasets very quickly. Then Hadoop is said to have low latency. But RDBMS is comparatively faster in retrieving the information from the datasets. It takes a very little time to perform the function provided that there is a small amount of data.

vi). Scalibility → Hadoop is fault tolerant but RDBMS not.

vii). Data Processing → RDBMS is comparatively faster than Hadoop (OLAP) (OLTP)

Hadoop is a free open source software framework.
RDBMS is licensed software.

Q-3> Explain the concept of NameNode, BackupNode & CheckPoint Node.

Answer -> Name Node -> The Name Node stores the metadata of the HDFS. The state of HDFS is stored in a file called FsImage and is the base of the metadata. During the runtime modifications are just written to a log file called edits. On the next start up of the Name Node the state is read from FsImage the changes from edits are applied to that and the new state is written back to FsImage. After this edits is cleared and is now ready for new log entries.

CheckpointNode -> It fetches periodically FsImage and edits from the NameNode and merges them. The resulting state is called checkpoint. After this it uploades the result to Name Node.

Backup Node -> It provides the same functionality as that of Checkpoint Node, but is synchronized with the Name Node. It doesn't need to fetch the changes periodically because it recieves a stream of file system edits from Name Node. It holds the current state in-memory and just need to save this to an image file to create a new checkpoint.

Q-4> Explain the different between parallel & distributed system.

Answer-

Parallel System	Distributed System
1. Memory :- Tightly coupled system Distributed/Shared Memory	Loosely coupled system distributed Memory
2. Control :- Global clock control	No Global clock control.
3. Many operation are performed Simultaneously	System components are located at different locations.
4. Processor communicate with each other through bus	Computers communicate with each other through message passing.
5. Multiple processor perform multiple operation.	Multiple computers perform multiple operation.
6. Single Computer is required	User multiple Computers.

Q-5 ⇒ Illustrate the major blocks in HDFS Architecture

Answer ⇒ Hadoop HDFS Architecture consist of a Master/Slave architecture in which Master is NameNode that stores meta-data & Slave is Data Node that stores the actual data. HDFS Architecture consist of single NameNode & all the other nodes are DataNodes.

i). Master slave Architecture

ii). Master - Namespace

a). It manages the file system, namespace & meta data. The entire file system namespace including the mapping of blocks, files & file system properties is stored in a file called Fs Image

b). The NameNode uses a transaction Log called the "Edit log" to persistently records every single change that occurs to file-system metadata & synchronizes with meta-data in ram after each write.

c). It regulate client accessfiles.

iii). Slave - DataNode a). Many Per cluster

b). It manages storage attached to the nodes

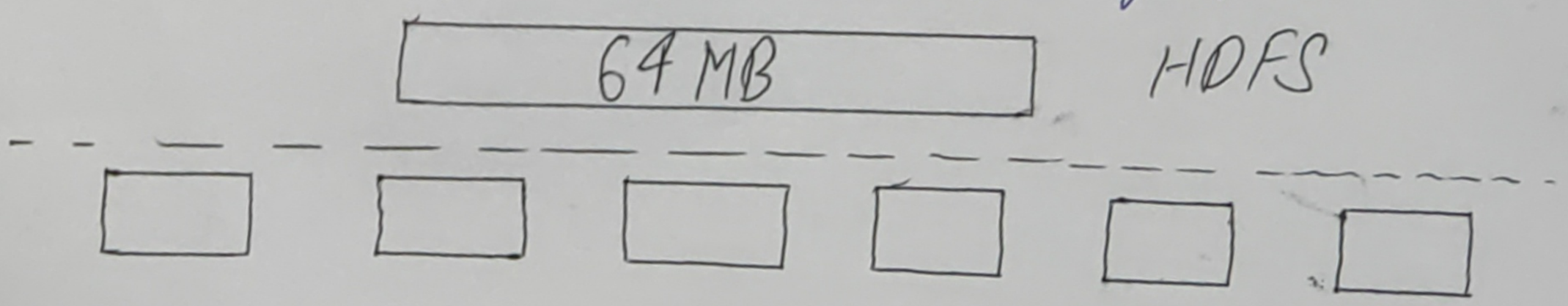
c). It periodically reports states to the NameNode.

Functioning

Data in Hadoop cluster is broken down into smaller pieces & distributed through the cluster, in this way Map & Reduce function can be executed on smaller subset of our larger data set and this provides the scalability that is needed for Big data processing.

i). HDFS is designed to support very large files; each file is split into blocks.. Hadoop default is 128 MB.

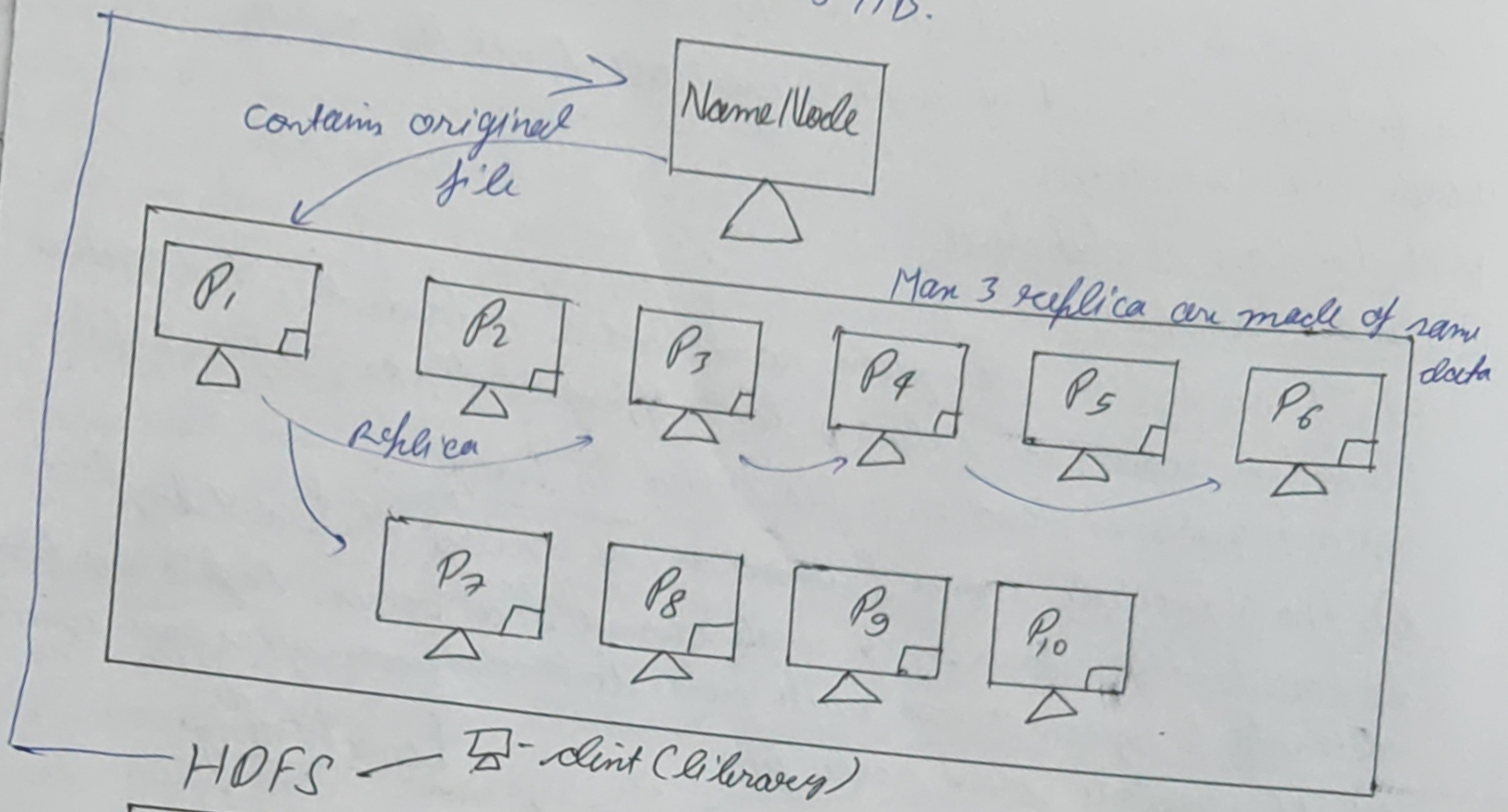
ii). Blocks reside on different physical dataNodes behind the scene HDFS block is supported by multiple operating system blocks.

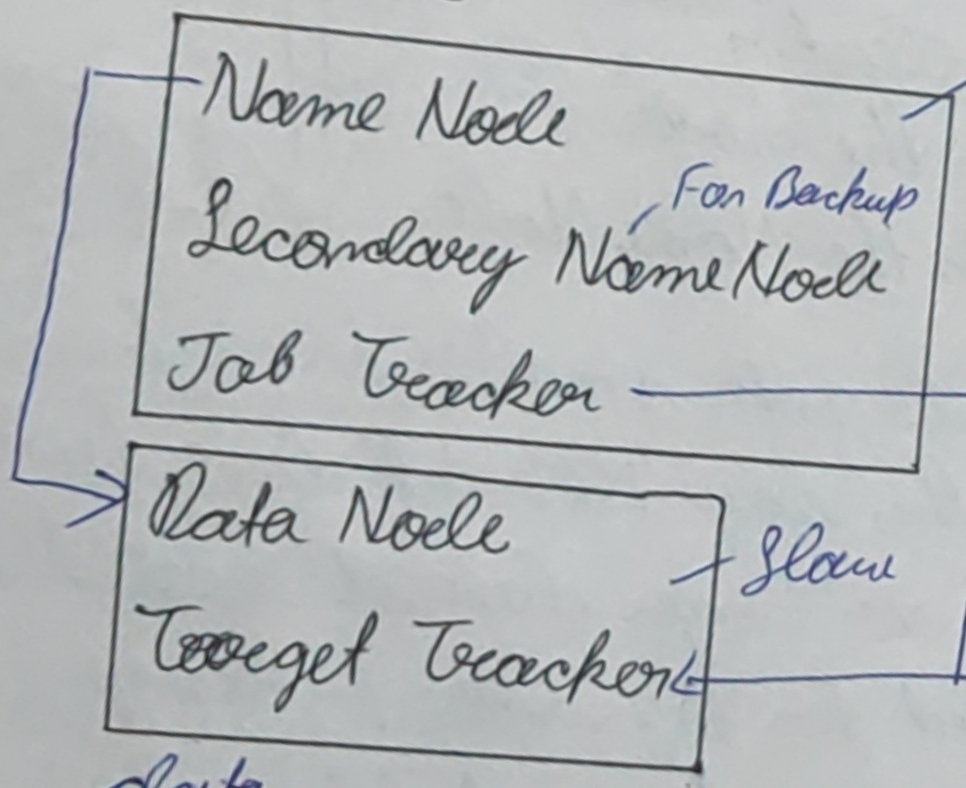


iii). If a file or a chunk of the file is smaller than the block size. Only the needed space is used. Eg → A 210 MB file is split as

64 MB	64 MB	64 MB	64 MB

For a size of 500 MB the split will be done in 128 MB.





Master Name Node tells the name to data Node. Data Node checks the various type of data in it like videos, photos etc. Job Tracker determines the way in which the data can be transferred Eg copy, delete, cut-paste. This all is done for client (Library)

The ~~process~~ given to Name Node create Replica of it & is stored in its slave so that if any 1 of data got corrupted then the other copied data may come in use.