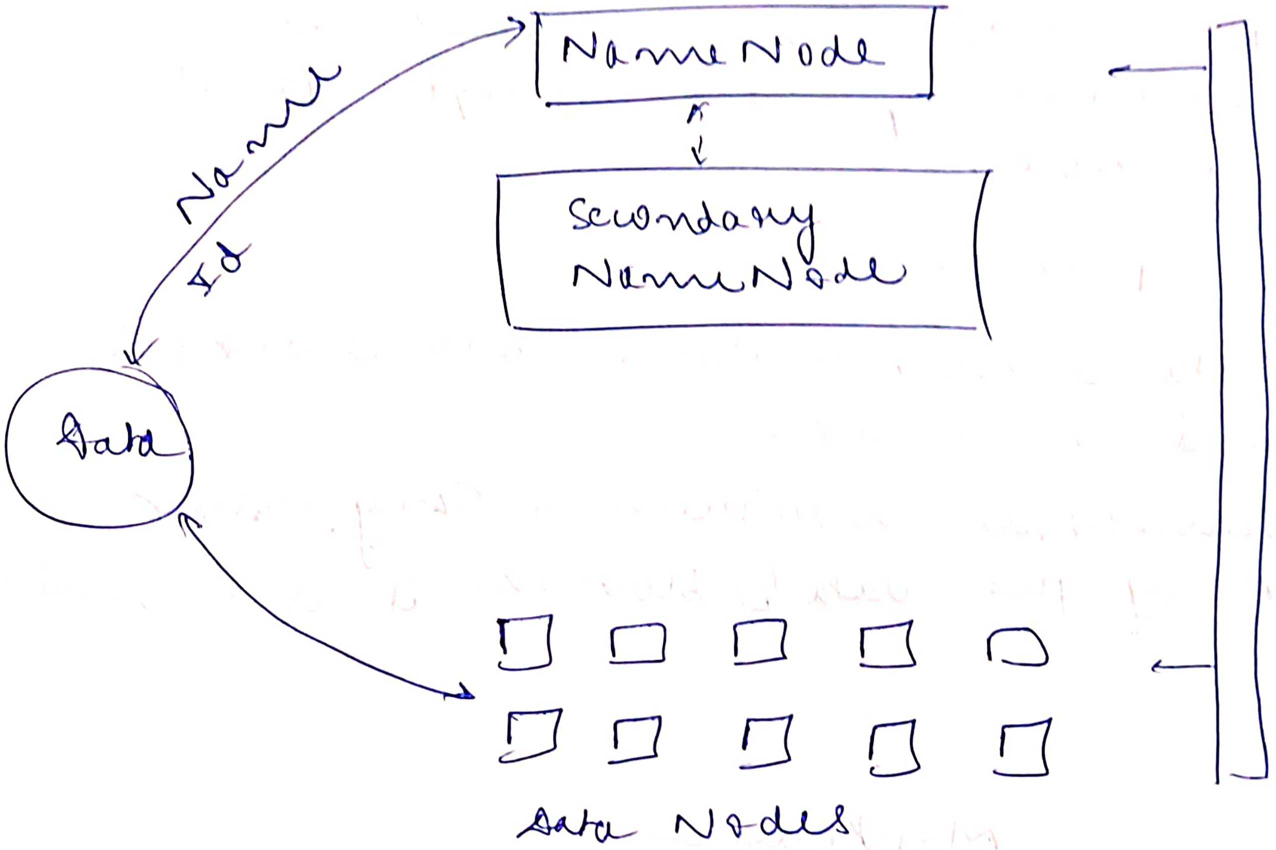
Ans1)

Om Shree
2006077
30-Jan-2023

~~Architecture of Hadoop :-~~

Architecture of HDFS :-



Name-node :-

- Maintains block IDs
- Maintains block health reports
- Runs on system memory
- single point of failure

Secondary NameNode :-

- NOT a back-up
- Performs memory intensive tasks.
- Runs on system memory
- Mainly a helper for NameNode

Data Node :-

- Stores broken down data.
- Slave in nature

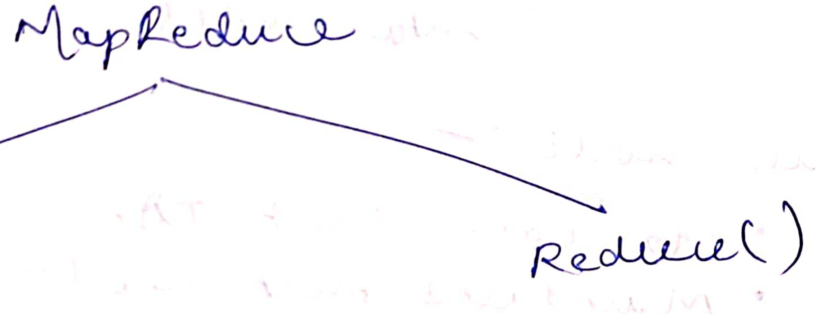
Read operation :-

- Namenode keeps track of block IDs & health reports.
- At the time of reading, pointer is shifted to the required block Id & data is read. Once a block is exhausted, pointer shifts to the next block.

Write operation :-

- Data is broken down into chunks & stored on blocks.
- NameNode maintains a <key, value> pair of the data & block IDs it is associat to.

Ans2)



- Map → splitting

  Receives the data & breaks it into <key, value> pairs or tuples.

- Reduce → shuffling, reorganization

  Tuples are further broken down to reduce the size of data while maintaining integrity.

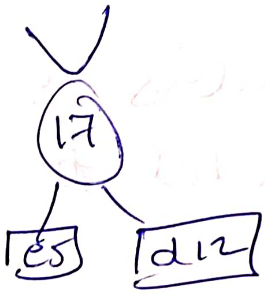
Ex – Huffman coding or compression algorithms.

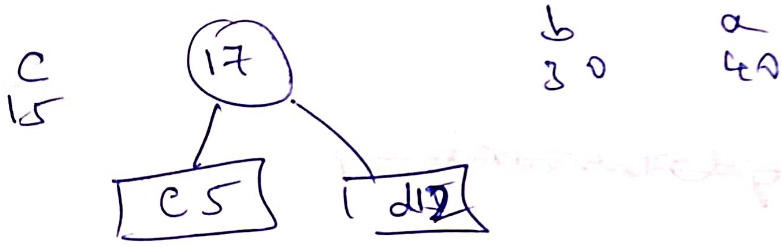
frequency of characters in a file is given:-

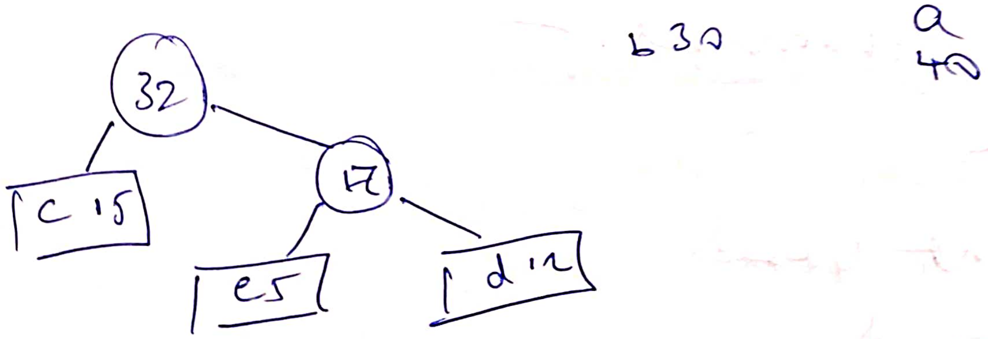
a	b	c	d	e
40	30	15	12	5

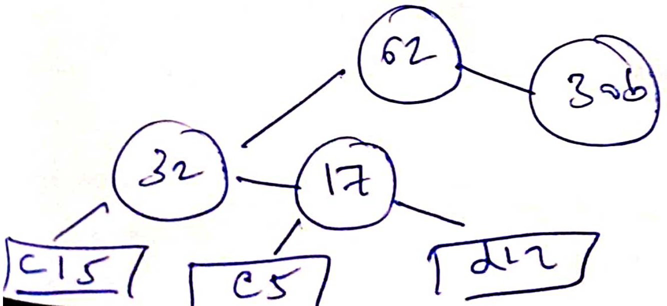
On reducing :-

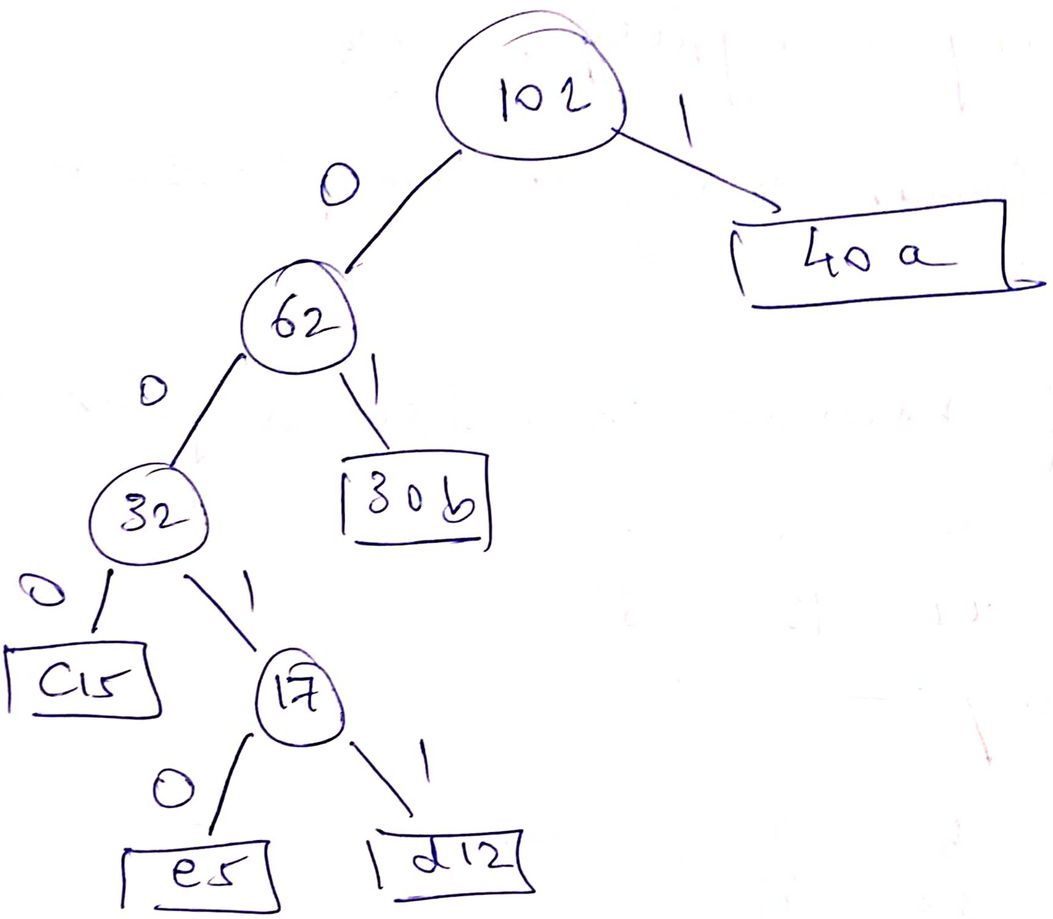
e	d	c	b	a
5	12	15	30	40











Now the frequency of characters is same but the memory needed to to store them has been reduced.

Ans3) ~~NoSQL replacement~~

NOSQL schemas:—

1. Key-value
2. Document based
3. Column based
4. Graph based

Key-value :-

	ID	Name	dept
SQL	1	Om	IT

Key value : 1 : { Name : Om
dept : IT }

- Key-value dictionaries are used to store data.
- Best for shopping carts.

Document based :-

	ID	Name	dept
SQL :	1	Om	IT

document : JSON {
ID : 1
Name : Om
dept : IT
}

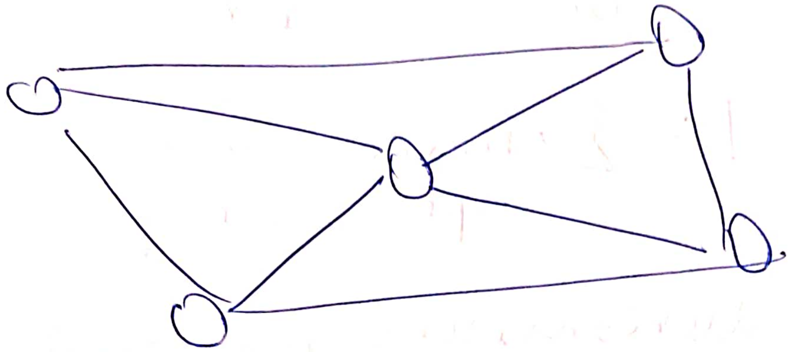
- Best for content management service

column based :-

- Best for data warehousing
- Reduces the size of data using vectorization.

Graph based :—

- Best for social media sites



- stores entity - relationships

Ans4) Replication management policy :—

↳ When a rack goes offline, ~~the~~ all the nodes on that rack will be unavailable.

↳ Not more than 1 replica of block is stored on the same node, not more than 2 replicas of a node are stored on the same rack.

↳ Default replication is 3-fold.

Rack separation

↑

Node separation

↑

Block separation

Ans5) 4Vs :-

(1) Velocity :-

(a) On-going transactions

(b) DB side balance management

(c) Hashing the user data to maintain data integrity.

(2) Volume :-

(a) Loan book

(b) Customer details

(c) Transaction details

(3) ~~Visibility~~ Visibility :-

(a) Bank's loan book should have limited visibility

(b) Account details should be visible to account owner only.

(c) Internal-employee data must not be visible to anyone who is not in a managerial position.

(4) Variety :-

(a) Customer details are stored as strings.

(b) Verification records are stored as PDF.

(c) Transaction details are stored as data dictionary.

Om shree
2006077
IT-02