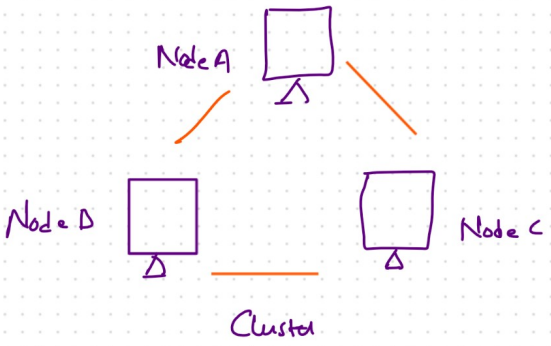
Key Terminologies — continuation

4. Cluster & Node

Cluster: Multiple machines connected together

Node: individual machine with in a cluster



5. Process & Daemon Process

Process: A program in execution (browser)

Daemon Process: background process that runs without user intervention

* Hadoop uses daemon process that run in the background to manage different roles

6. Meta data:

Data about data

eg: file properties (size, creation date, modification date, permissions)

7. Replication:

* process of making copies of data
* critical in HDFS to achieve fault tolerance
* If one machine fails, data can still be accessed from other nodes

Introduction to HDFS

* core component of Hadoop that provides a distributed storage solution for big data processing

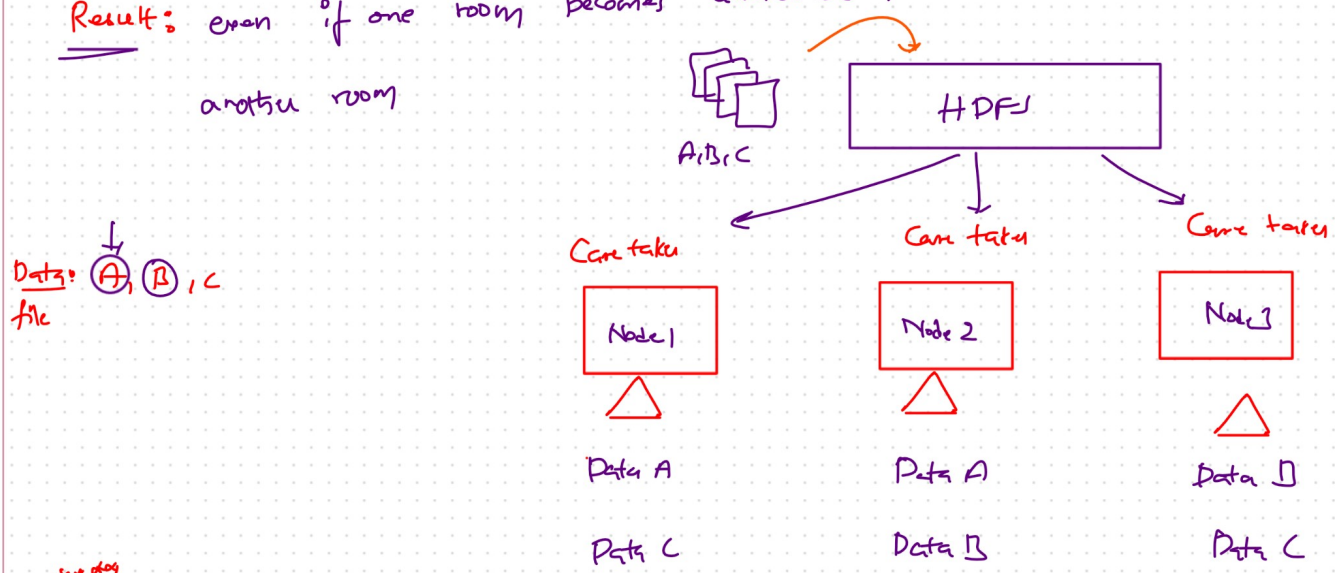
Library Analogy of HDFS:

Problem: millions of books in a single room ——— creates space & accessibility issues

✓ Sol: Distribute books across multiple rooms, each room with caretaker

✓ Added Benefit: storing copies of book across different rooms for redundancy

Result: even if one room becomes unavailable, we can get the book from another room



Core functions of HDFS:

1) Data Distribution:

→ distributes the data across multiple nodes / machines

→ prevents single point of storage bottlenecks

2) Data Replication:

→ creates multiple copies of data across different nodes

→ ensure fault tolerance

→ prevents data loss if a node fails

3) Efficient Access:

- provided optimized methods to access data
- Enables efficient bid data processing

eg: Analytical operations → filtering, grouping

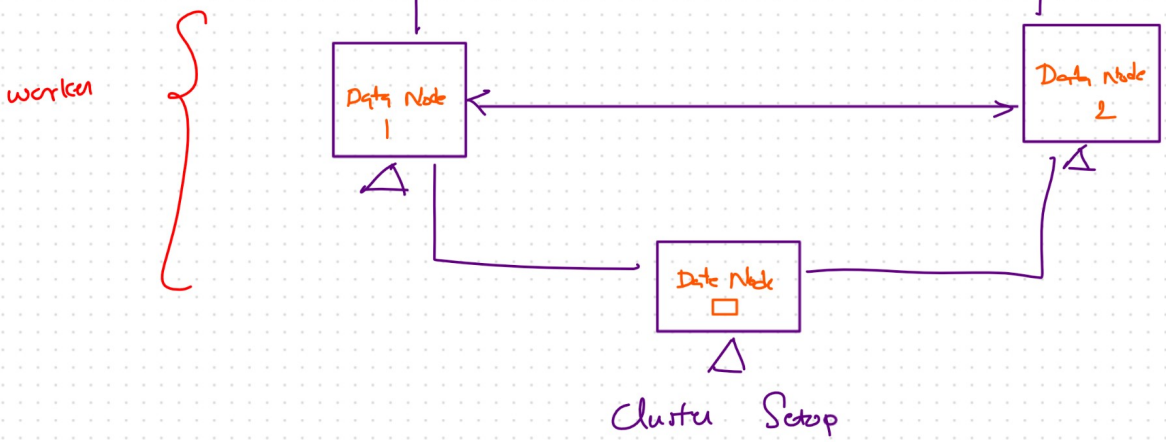
HDFS concept
- Amazon S3
- Azure blob

HDFS Architecture

Hadoop

master node — name node

worker node — data node



Note: HDFS is installed on top of Linux

Heart beats (every 3s)

Cluster Setup

Cluster Setup:

1) cluster consists of multiple machines running on Linux
2) HDFS is installed on top of Linux file System
3) When setting up hadoop, roles must be specified for each machine
   - which machine is master
   - which machines are workers

2 file System in Cluster:

- Linux commands works only on the local machine
- HDFS / Hadoop commands work across the entire cluster

Architecture Components of Hadoop:

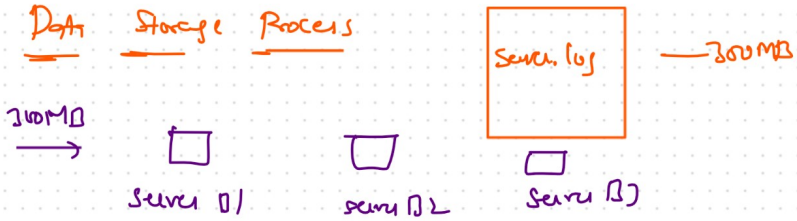
Name Node (Master)

* central control system of HDFS
* manages meta data but doesn't store actual data i.e creates & maintains metadata about file locations
* Co-ordinates how files are stored & accessed

Data Node (Worker)

* store the actual data in blocks
* sends heartbeat signals every 3 seconds to the name node
* perform read/write operations as instructed by the name node

Data Storage Process



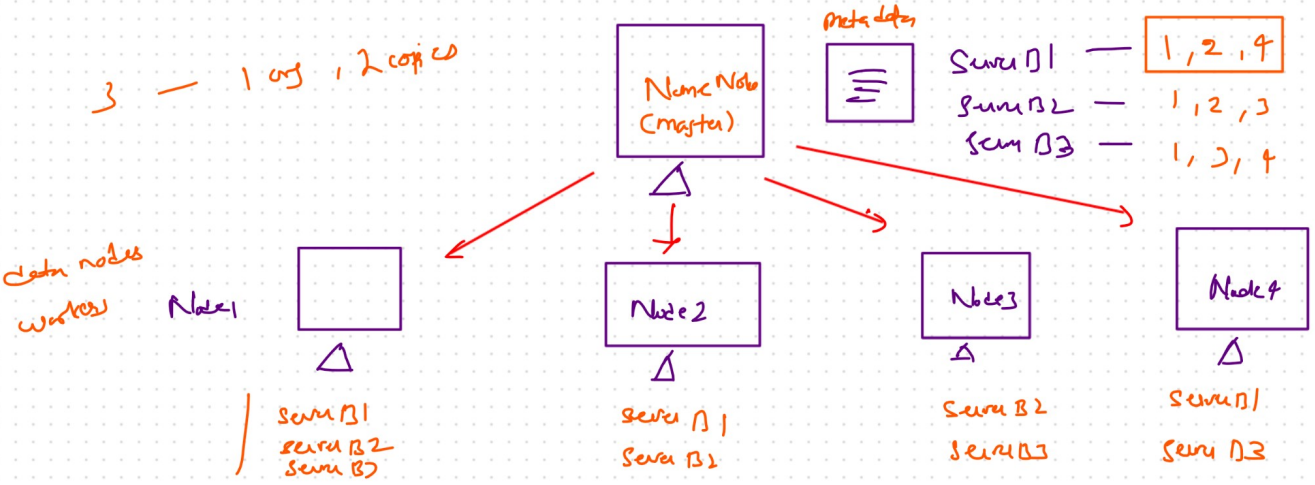
default size of block → 128 MB in HDFS

replication factor: 3 — 1 original, 2 copies
2 — 1 original, 1 copy
4 — 1 original, 3 copies

} Copies of blocks are stored across different data nodes

Server.log → 300MB
- Server B1 128
- Server B2 128
- Server B3 128

3 — 1 org, 2 copies



* If a client requests to store a file (server.log) 300 MB
* Name node divides file into blocks (default block size is 128 MB)
* Blocks are distributed across data nodes (eg server1, server2, server3)
* Based on replication factor (eg: 3) copies of block are stored on different node i.e
  1 — original & 2 — copies
* name node maintains meta data about where each block is stored