Assign-2

Date:
Page:

1) List the use and Responsibility of following tools of Hadoop eco-System.

→ a) HDFS →

→ It stores the data across the multiple machines in a fault tolerant manner.

→ HDFS replicates data to ensure fault tolerance & reliability.

→ It is design to scale horizontally as more data is added to the System.

b) YARN

→ It allocates & Manages cluster Resources

→ It schedules Prothon task & jobs submitted to cluster

c) Flume:

→ Flume collects data from sources like logs, web servers, social media & transfers it to Hadoop store.

→ It can perform lightweight data transformation during data ingestion

d) Sqoop:

→ Sqoop is used to import data from relational databases into Hadoop & export data from Hadoop to Database

→ It can perform Hadoop to Database.

e) Zookeeper:
→ Help in coordinating Distributed Process ensuring the Large consistent view of Shared Resources

f) Hive:
→ Allows user to write SQL like queries to extract insights from data stored in Hadoop.

2) Explain Responsibility of Name & Data Nodes & Diff between Normal File System & HDFS

→ → NameSpace Manager Manages file System Namespace which means it keeps track of file & Directory Structure

→ Metadata Manager - Stores meta about each file such as file name etc.

→ Block Manager: keeps track of locations & replications

→ Heartbeat & Health Monitoring
It recieves heartbeat signals from DataNodes to ensure they are alive & monitors their health.

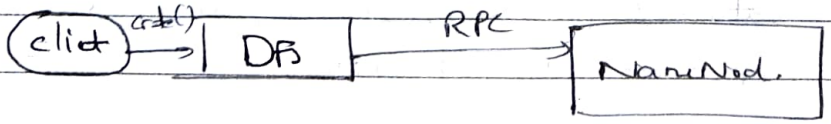
Data Nodes:
→ Data Storage: Data Node stores the actual Data Blocks
→ Block creation & Deletion - DataNodes create,

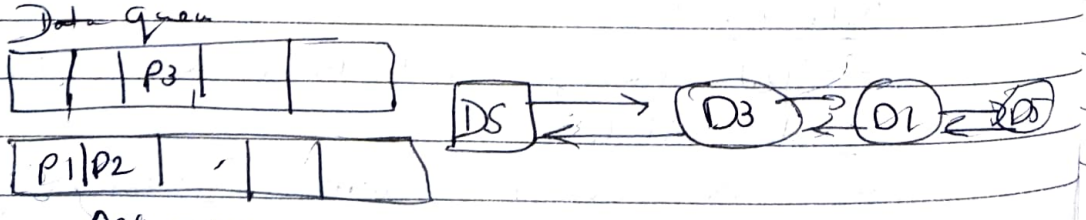
delete & replicate data blocks as instructed by the Name Node.

→ Heartbeat & Block Report - Periodically send heartbeat to the Name Node

HDFS	Normal File System
* Designed for Distributed Data Storage.	* Data is stored in Single Disk.
* HDFS separates Metadata from Data	* Meta Data & Data are correlated.
* HDFS replicates Data Blocks to ensure fault tolerance	* Replication & fault tolerance are typically not built in
* Follows a write-once & read-many model	* Read-Write can be done many times

3) With Diagram show two write operations performed in HDFS. List the steps to recover from failure during HDFS write operation.

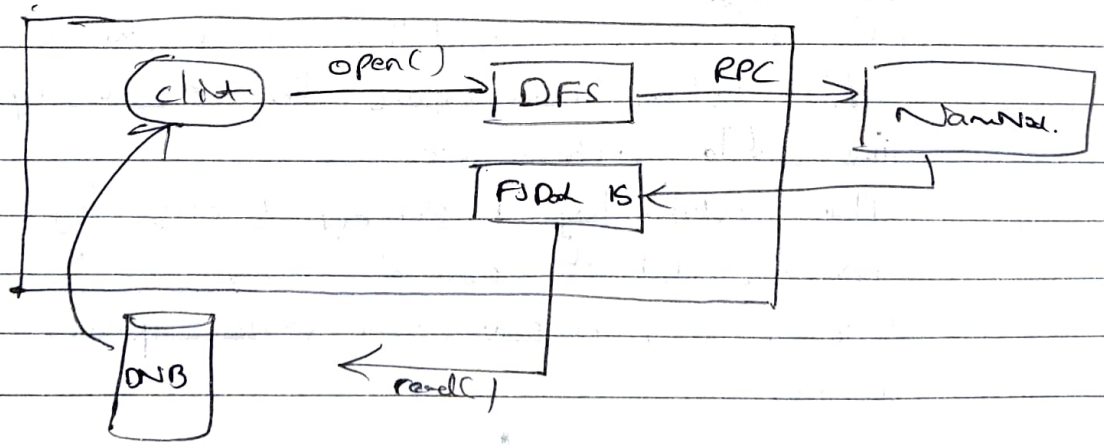




→ Let us assume DN1 failed while writing P3
→ DS puts packet P3 back to Data Queue & communicate failure of DN1 to NameNode.
→ NameNode creates & new id block -002 & add DN7.

4) Explain how HDFS ensures high Availability & Service of Data
→ Data Replication
→ Data Distribution
→ Data Block Placement.
→ Heart beats & Health Monitors
→ Secondary Name Node.
→ Load Balancing
→ Manual Failover.

5) With Diagram Explain, How read operation is performed in HDFS.



* client read a open()
* Through DFS object RPC is made.

* NameNode search in "metadata" data required file.

* It finds out the Datanodes which have blocks of data.

* User FSDD goes to data Node.

* If one of the Data node is down, it goes to next nearest Datanode.

* Reading operations happen Sequentially.

(a) Explain the terms heartbeat, checkpoint & edit logs.

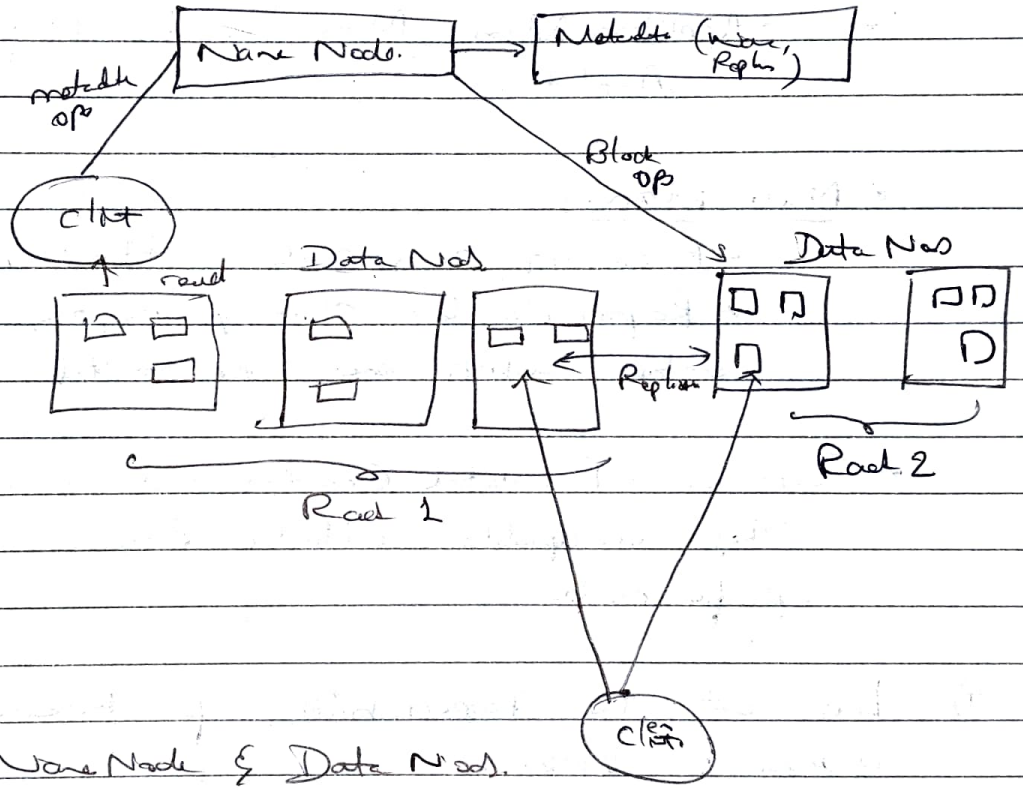
→ Heartbeat → It is a periodic call sent by a datanode to the master node in HDFS cluster known as the namenode. The signal serves as a way for data nodes to communicate their health.

Edit log: It is a log file that records all changes or modifications made to the file system namespace. It includes operations like file creation etc.
The edit log is an append only log means that it records changes as they happen and doesn't overwrite previous entries.

Checkpoint: This is a mechanism to periodically save a snapshot of the

File System metadata & Data structure

7) With a neat Diagram, Explain architecture of HDFS.



Name Node & Data Nodes.

→ HDFS is a master slave architecture.
→ An HDFS cluster consists of a single namenode, a master server that manages the file system namespace & regulates access to files by clients.
→ There are a no' of Datanodes usually one per node in the cluster.
→ HDFS exposes a file system namespace allows user data to be stored in files.
→ Data Nodes perform block creation, deletion & replication upon Instruction.

8) List the Major Components of YARN, & Responsibilities.

→ * Resource Manager
  It keeps track of available resources
  • Scheduler → allocates Resources
* * Application Master → Manages Lifecycle of Application

* Node Manager
  * Runs on every node in Hadoop.
  * Responsible for starting & Stopping Containers, which are the isolated execution.

* Containers :- Fundatal Execution Units in YARN
  * They encapsulate allocated Resources of a specific node.

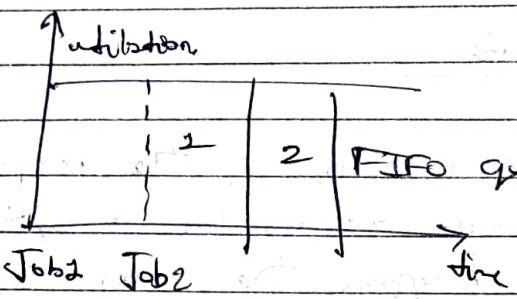
9) List out the Responsibilities of Resource Manager in YARN

→ * Resource Allocation
* Application life Cycle Management
* Scheduling
* Node manager Communication
* Fairness and Priority
* High Availability
* Handling multiple Application.

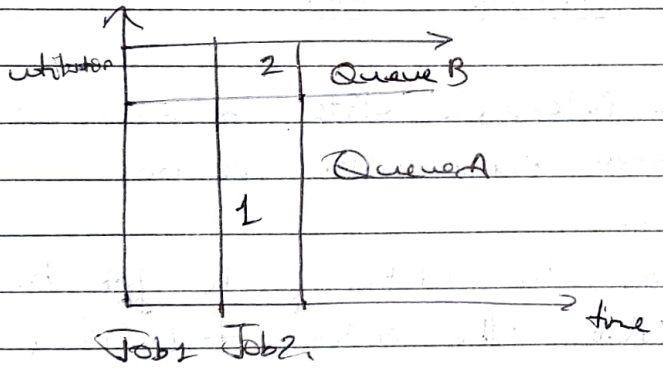
10) Explain Different Shedulers in YARN.

→ FIFO:-
→ Simple Scheduler.
→ Reqts for the first Application in the

queue as allocated first.



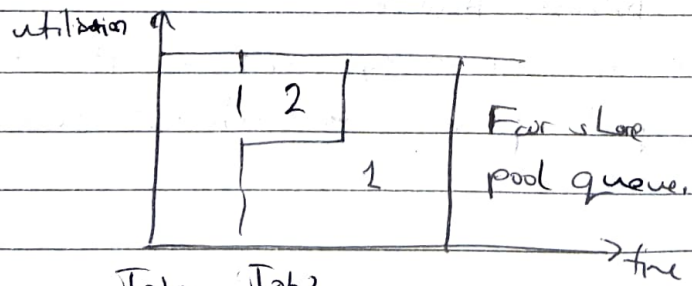
* Capacity Scheduler:-
  * A seperate dedicated queue allows the small job to start as soon as it is submitted.
  * It cost overall cluster utilisation since the queue capacity is served for jobs.



Fair Scheduler:-

→ No need of reserve a set amount of capacity

→ It will dynamically balance resources



11) Explain Different types of failures in Map reduce Job.

→ Task Failure

→ When user code in the map or reduce task throws as runtime exception
→ If this happens the task JVM reports the error back to its parent application master before it exits.
→ This error ultimately made it into the user logs.
→ The application master marks the task attempt as failed.

Application master failure:

* Resource Manager starts a new instance of master running in a new container.
* Recover the state of the job history so that the entire task need not be run again
* Map Reduce client polls the application master for progress. If it doesn't get response then asks resource Manager for the new instance address.

Node Manager Failure:

* Resource Manager recovers it from its pool of nodes to schedule containers

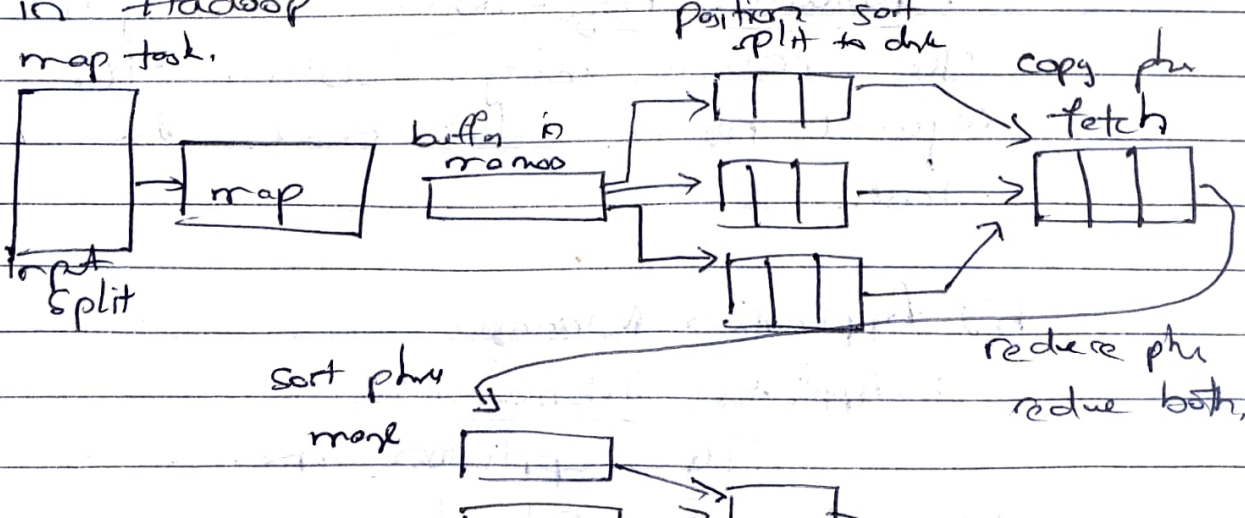
12) Give Different Reason for task failure, Give steps for recovery from application master failure

→ → User code in the map or reduce task throws a runtime exception
→ In this case task JVM reports the error back to its parent application master before it exits.
→ The application master marks it as failed and frees up the container.

Application master failure Recovery:-

An application master sends periodic heartbeats to the resource manager. In the event of application master failure the resource manager will detect the failure, the resource manager will detect the failure and start new instance of the master running in a new cluster.

13) Show how map reduce jobs are executed in Hadoop



Reducer

The map output file is sitting on the local disk of the machine that ran the map task but now it is needed by the machine. That is about to run the reduce task for the partition. The reduce task need the map output for its particular partition from several map task across the cluster.

14) Give the Responsibilities of

a) Application Master:

→ Application master is responsible for managing the lifecycle of a specific application or job running on the cluster. It is created when an application is submitted and monitors its progress until completion.

→ Resource negotiation: It negotiates the resources manager obtain the required resources for executing the tasks.

b) Application Manager

→ Application Submission:

The application manager is responsible for submitting new applications to the YARN cluster.

c) Node Manager:

→ Node Manager:

Node Manager is responsible for monitoring the utilization of Resources

d) Resource Manager:

Hadoop map-reduce is a programming framework for handling large-scale data processing tasks in a distributed and parallel manner.

~~It has map phase & Reduce phase.~~

~~Example:- word count.~~

15) With example, explain Hadoop map reduce process

→ Hadoop map-reduce is a programming framework for handling large-scale data processing tasks in a distributed and parallel manner.

It has map phase & Reduce Phase:

ex:- word count

Map-Reduce:-

In the Map phase the input data is divided into chunks and each chunk is processed

Input data:
→ Hello world, how are you? Hello world?

→ Mapper task tokenize the input into words & emit pairs for each word.

(Hello, 1)
(world, 1)
(how, 1)
(are, 1)
(you, 1)
(Hello, 1)
(World, 1)

Each word is a key and the value associated with it is 1.

Shuffle Sort:
Hadoop framework sorts & Shuffles intermediate key value pairs group them by key ensures that all key value pairs with same keys are grouped

Reduce Phase:
Reduce task process the sorted & shuffled key value pairs

output of the Reducer will be the count of each words.
→ (Hello, 2)
(world, 2)
(how, 1)
(are, 1) (you, 1)