Schedulers :-

→ In hadoop, we can receive multiple jobs from differed clients to perform.

→ The Map Reduce framework is used to perform multiple tasks in parallel in a typical Hadoop cluster to ~~provide~~ process large size datasets at a fast rate.

→ This Map-Reduce framework is responsible for scheduling and monitoring the tasks given by diffeent clients in a Hadoop cluster. But this method of scheduling jobs is used prior to Hadoop 2 .

→ Now in Hadoop 2, we have YARN (Yet Another Resource Negotiator)

→ In Yarn we have seperate Daemons for performing job scheduling, Monitoring and Resource Management as Application Master, Node Manager and Resource Manager respectively.

Resource Manager :- Resource Manager is the Master Daemon responsible for tracking or providing the resources required by any application within the cluster, a

Node Manager :- Node Manager is the Slave Daemon which monitors and keep track of the resources used by an application and sends the feedback to Resource Manager.

The Schedulor in YARN is totally dedicated to Scheduling the it cannot track the Status of the application.

→ On basis of required resources, the Schedular performs or we can say Schedule the Jobs.

Types of SCHEDULER

Fifo SCHEDULAR.

FAIR SCHEDULER

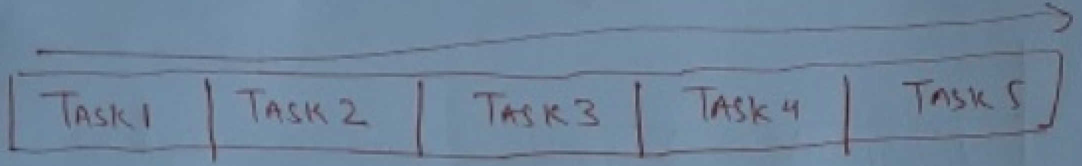
CAPACITY SCHEDULAR

These Schedulars are actually a kind of algorithms that we use to Schedule tasks in Hadoop Clusters when we receive requests from different - different clients.

→

A Job queue is nothing but the collection of various tasks that we have received from our various clients. The tasks are available in the queue and we need to schedule this task on the basis of our requirements.

JOB QUEUE



1 Fifo SCHEDULAR:

→ As the name suggests fifo i.e. first In first Out, so the tasks or applicaton that comes first will be served first.

→ This is the default Schedulor we use in Hadoop. The tasks are placed in a queue and the tasks are performed in their submission order.

→ In this method, once the job is scheduled, no intervention is allowed.

→ So sometimes the high-priority process has to wait for a long time since the priority of the task does not matter in this method.

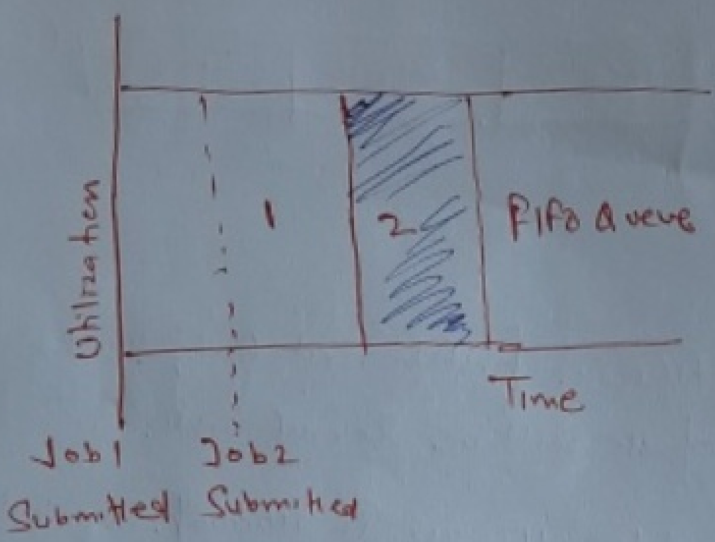
Advantage:

- No need for configuration.
- First Come First Serve
- Simple to execute.

Disadvantage:

- Priority of task doesn't matter, so high priority jobs need to wait.
- Not suitable for shared cluster.

Fifo SCHEDUER



2. <u>Capacity Scheduler</u> (default Scheduler with YARN) (client same cluster can be sorted out

1) In Capacity Schedular we have multiple job queues for Scheduling our tasks.

→ The Capacity Schedular allows multiple occupants to share a large size Hadoop cluster.

→ In Capacity Scheduler corresponding for each job queue, we provide some slots or cluster resources for performing job operation

→ Each job queue has its own slots to perform its task.

→ In case we have tasts to perform in only one queue then the tasts of that queue can access the slots of other queues also as they are free to use, and when the new tasks enters to some other queue then jobs in running in its own slots of the cluster are replaced with its own job.

Capacity Scheduler also provids a level of abstraction to know which occupant is utilizing the more cluster resource or slots, so that the single user or application doesn't take disapproprate on unnecessary slots in the cluster.

→ The Capacity Scheduler mainly contains 3 types of the queue that are root, parent and leaf which are used to represent cluster, organisation, or any subgroup, application submission respectively.

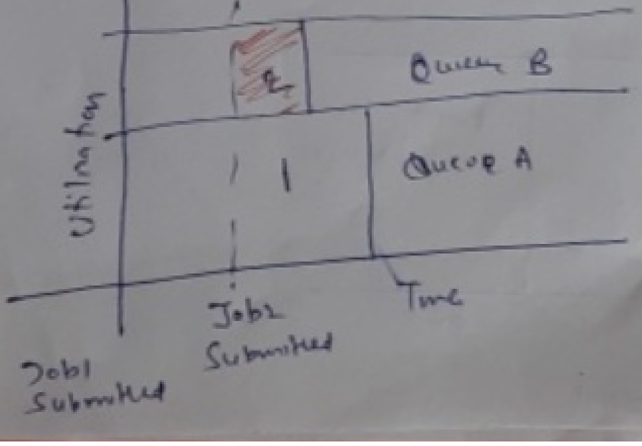
Advantage:

- Best for working with Multiple Clients or priority jobs in Hadoop Cluster.
- Maximizes throughput in Hadoop Cluster.

Disadvantage:

- More Complex
- Not easy to configure for everyone

CAPACITY SCHEDULER



3. Fair Schedular

→ The Fair Schedular is very much Similar to that of the Capacity Schedular.

→ The priority of the job is kept Consideration.

→ With the help of Fair Schedular, the YARN applications can Share the resources in the large Hadoop Cluster and these resources are maintained dynamically so no need for prior capacity.

→ The resources are distributed in such a manner that all applications within a Cluster get an equal amount of time.

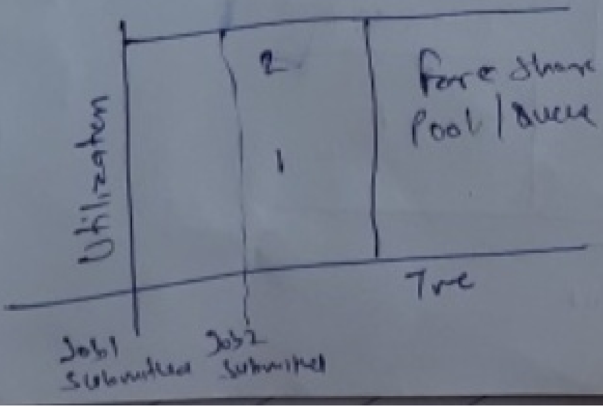
→ Fair Schedular takes Scheduling decisions on the basis of memory, we can configure it to work with CPU also.

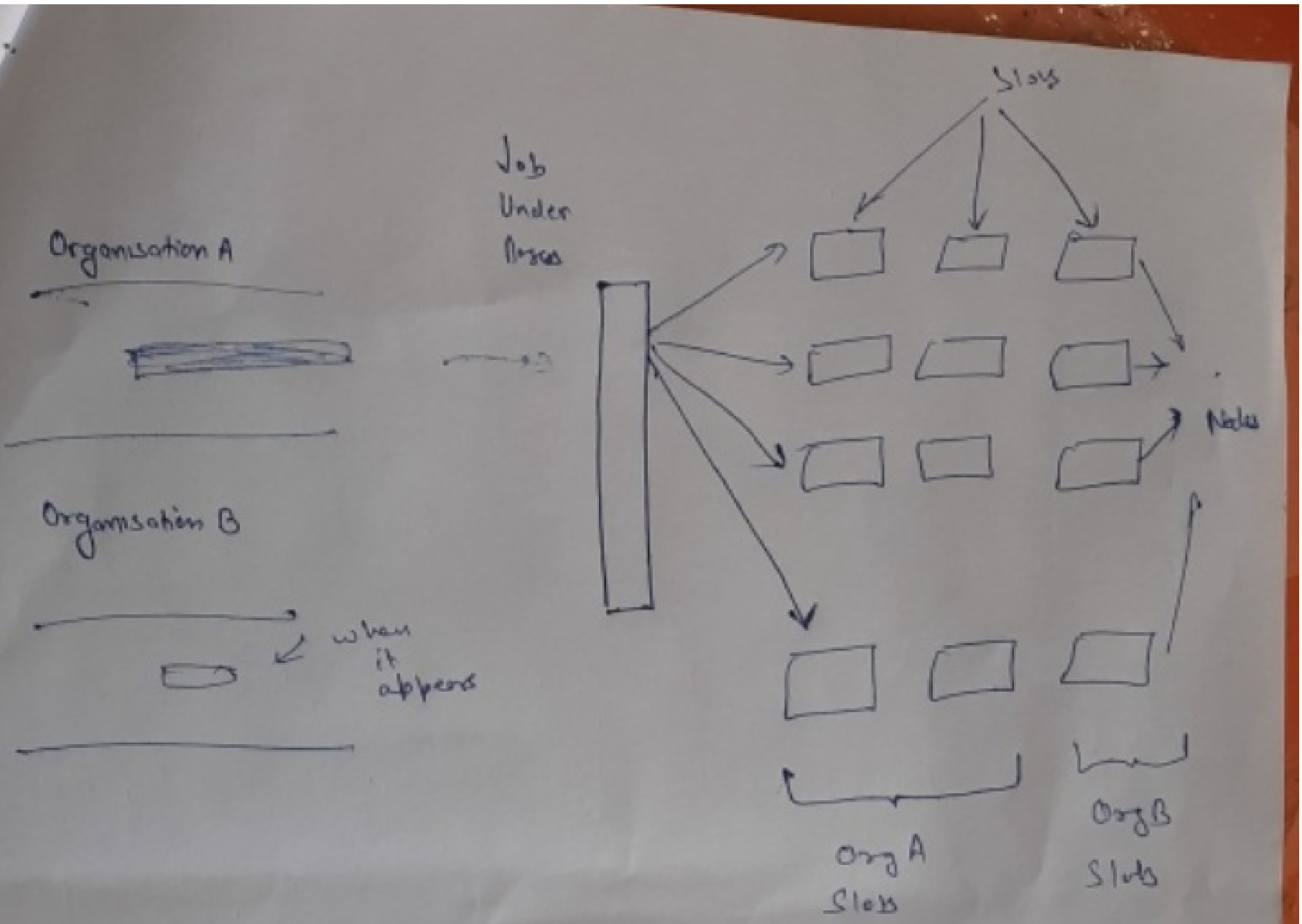
→ Fair Schedular whenever any high priority job arises in the same queue, the task is processed in parallel by replacing some portion from the already dedicated slots.

Advantages:
- Resources assigned to each application depend on its priority.
- it can limit the concurrent running task in a particular pool or queue

Disadvantages: The Configuration is required.

FAIR SCHEDULAR





CAPACITY SCHEDULAR

FAIR SCHEDULAR (Jobs are divided into pools)

- ~~[illegible]~~ In capacity Schedular a ~~has~~ small job having higher priority have to wait but in fair Schedular the job runs parallely and the resources are allocated accordingly.