BIG DATA (TFC303)

Pertemuan 7 – YARN

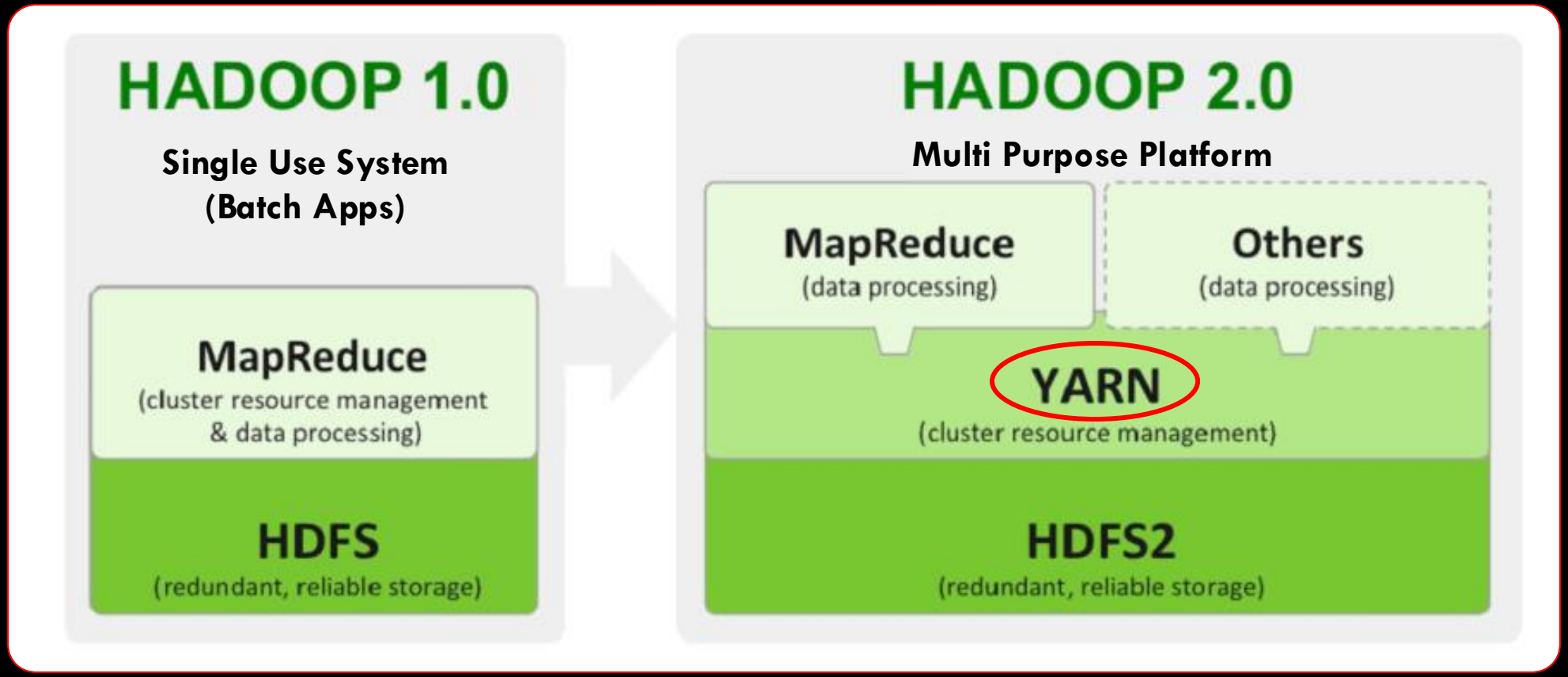
ALIFIA REVAN PRANANDA

Department of Information Technology
Faculty of Engineering
Universitas Tidar

WHAT IS HADOOP YARN ?

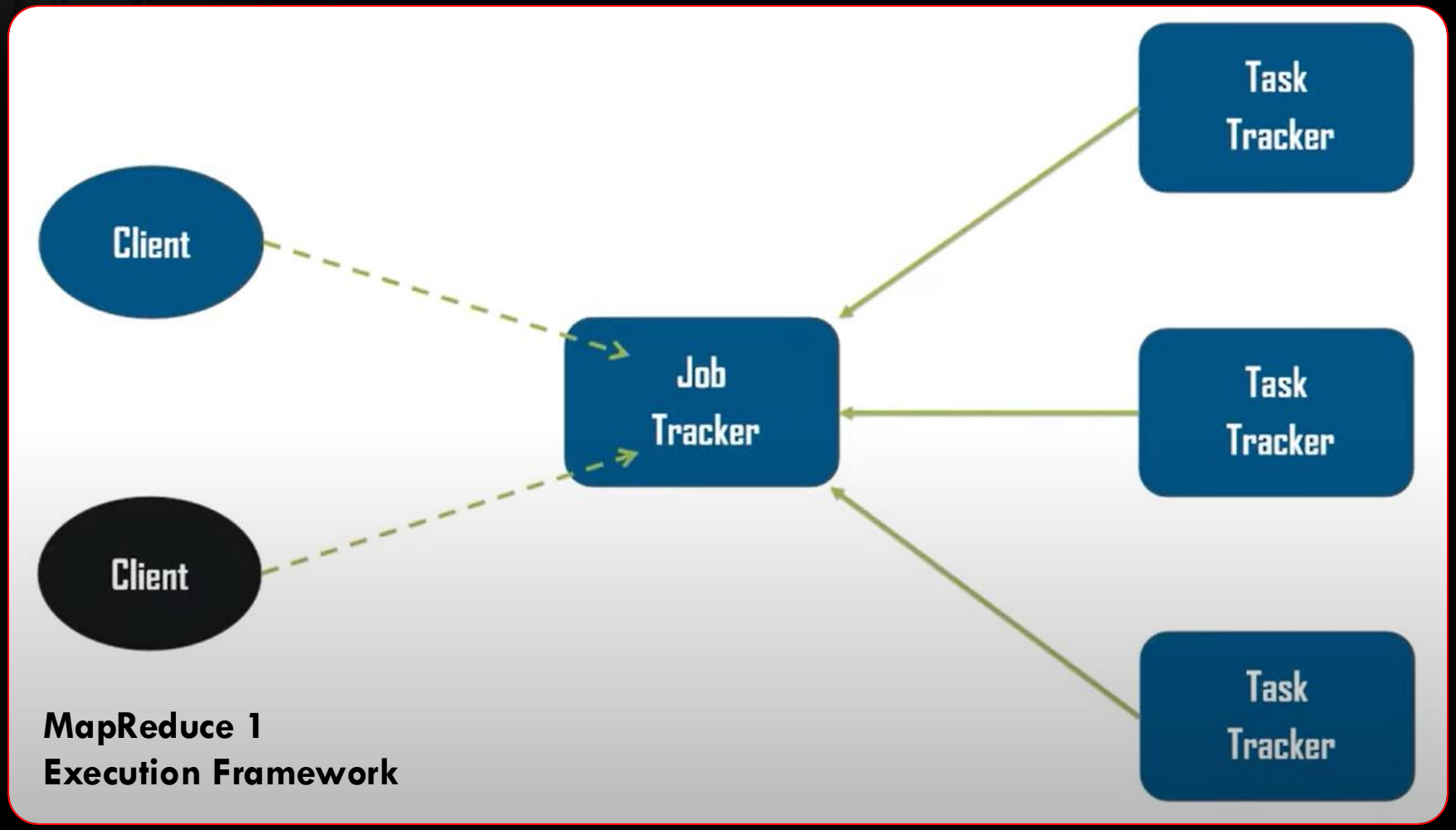
YARN

YARN is a core component of Hadoop 2.0 and is added to provide improved performance in the Hadoop. YARN is Hadoop computing platform which offer various advantages as compared to classic map reduce engine in the first version of Hadoop.



MAPREDUCE 1 VS YARN

Before learning about YARN, lets us see MapReduce 1 execution framework:

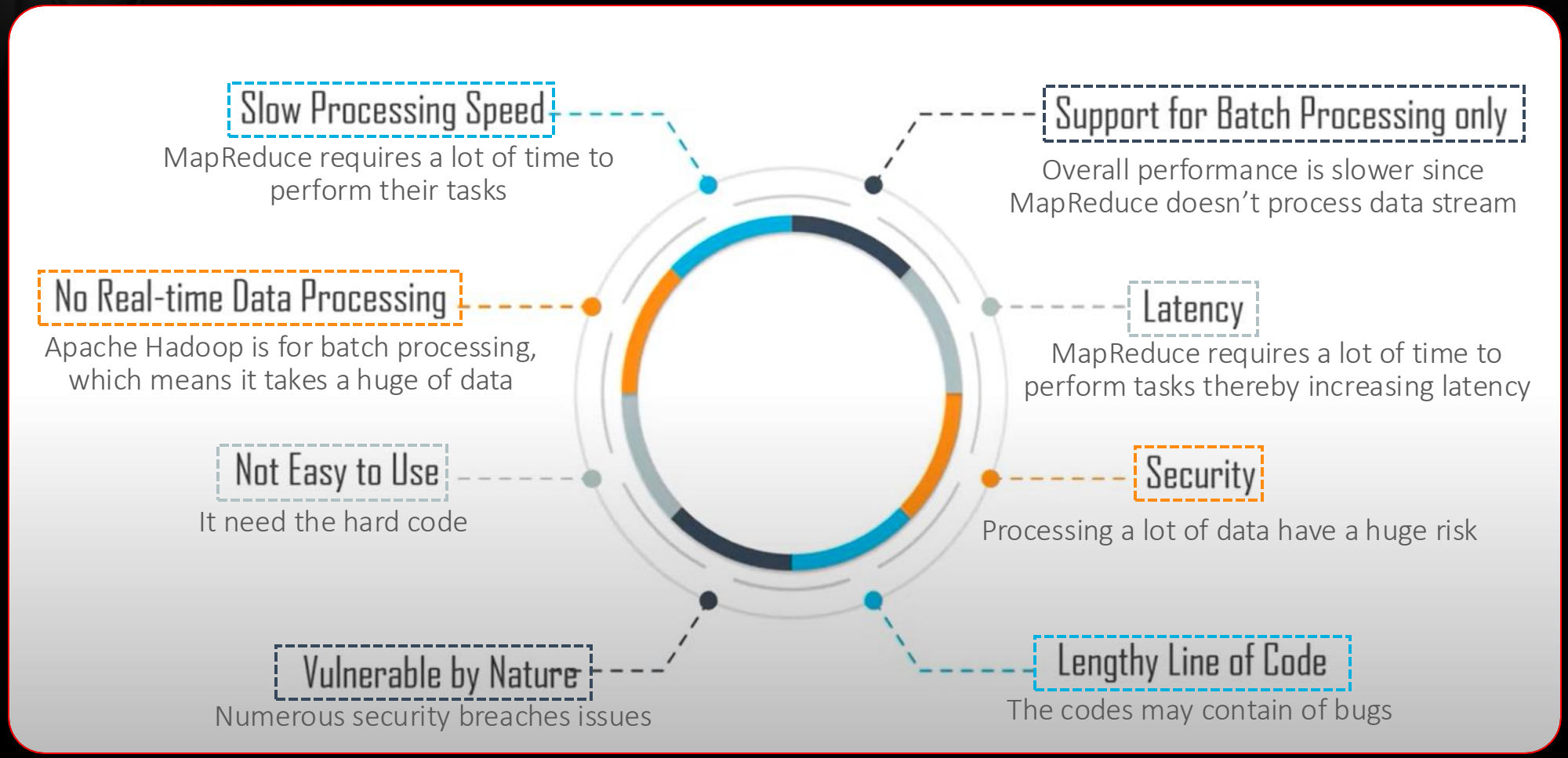


MapReduce 1 Execution Framework

- Job Tracker: a Master Daemon
- Responsible to assign and track task execution progress
- Task tracker are slave daemons
- They run on systems where data nodes reside
- Responsible to spawn a child jvm to execute Map Reduce and intermediate task.
- MapReduce 1 takes care of both job scheduling and task progress monitoring.

MAPREDUCE 1 VS YARN

Limitation of MapReduce 1



THE DEFINITION OF YARN

What is YARN ?

01 YARN stands for "Yet Another Resource Negotiator"

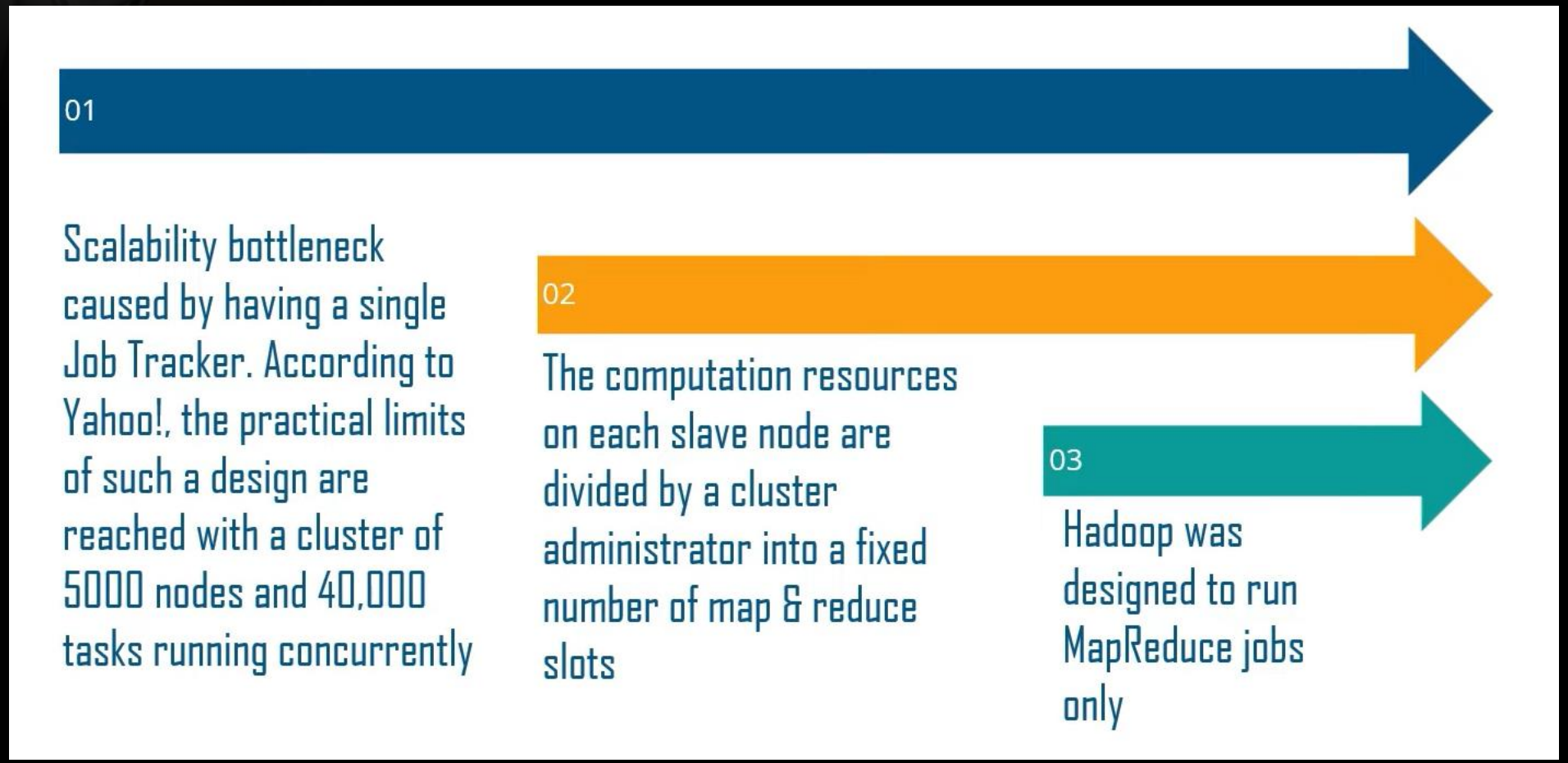
02 YARN/MapReduce2 has been introduced in Hadoop 2.0

03 It is a layer that separates the resource management layer and the processing components layer

04 MapReduce2 moves Resource management(like infrastructure to monitor nodes, allocate resource and schedule jobs) into YARN

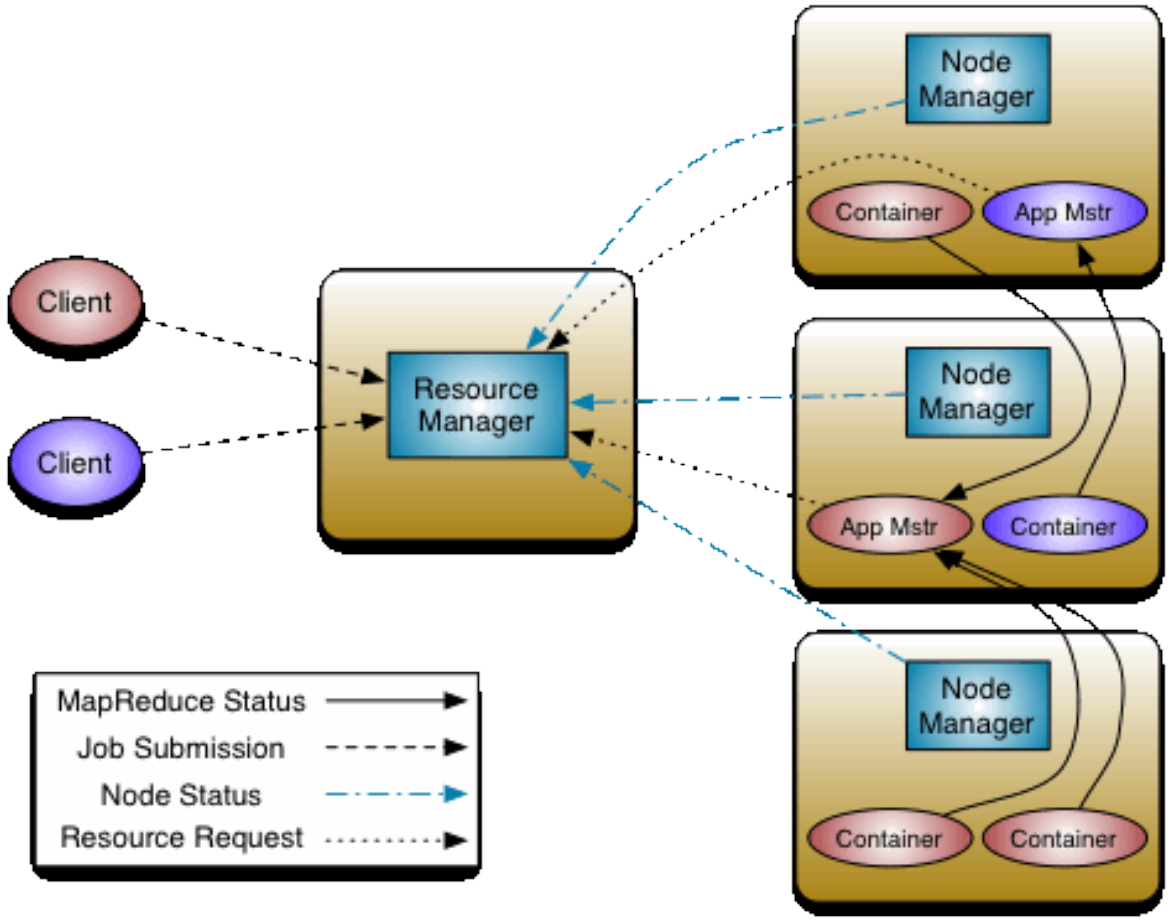
MOTIVATION OF YARN

Why does YARN is needed?



ARCHITECTURE OF YARN

How does it work?



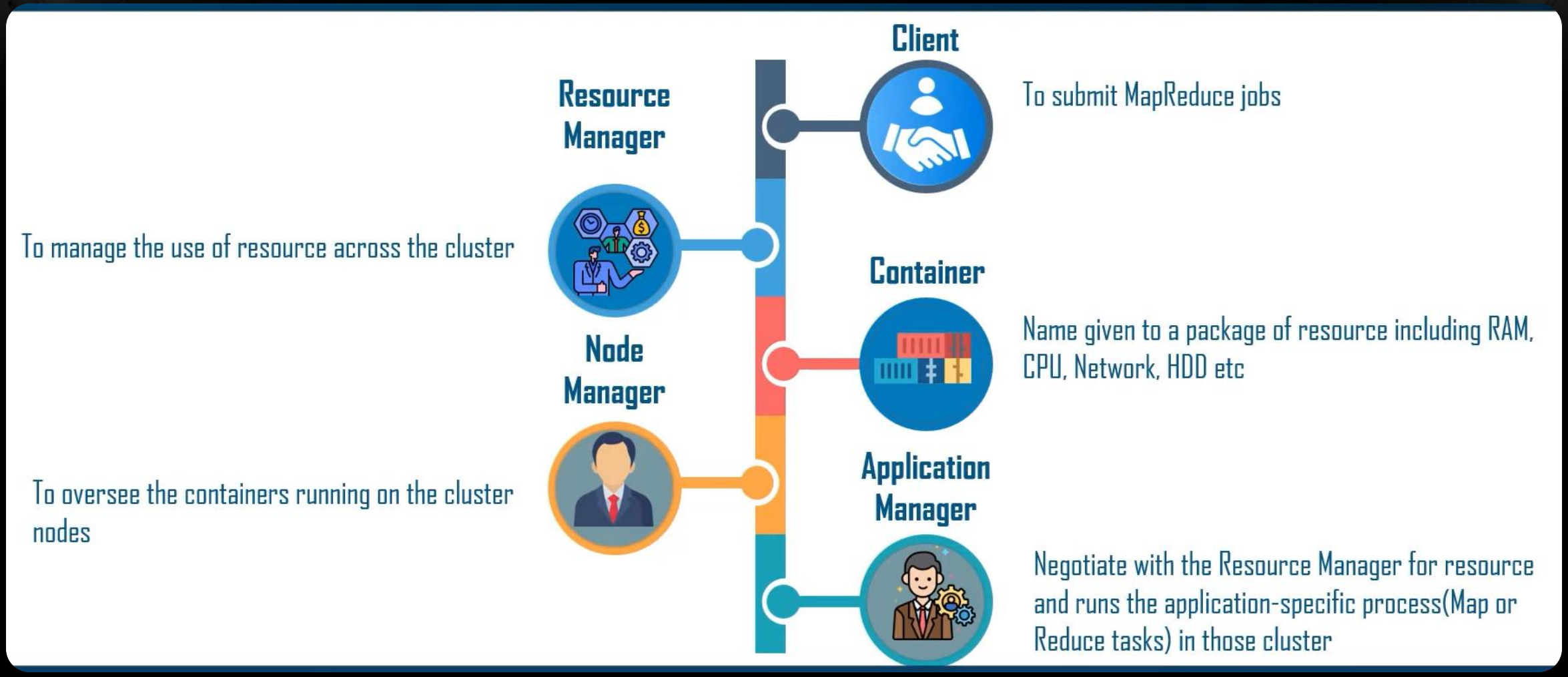
Resource manager manages the resource allocation in the cluster

Application master manages resource needs of individual applications

- Node manager is a generalized task tracker
- A container executes an application specific process

COMPONENT OF YARN

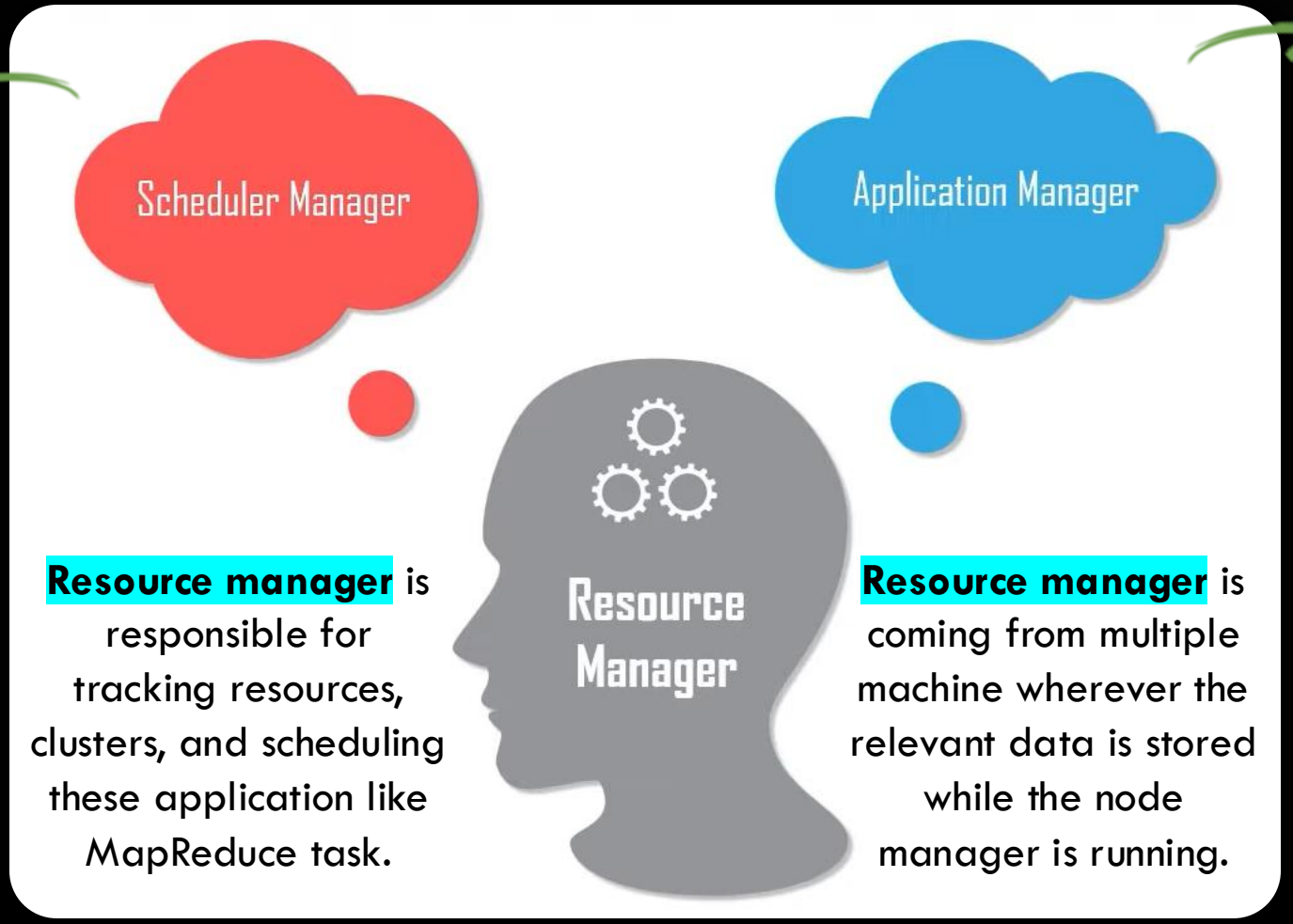
YARN Components



YARN DAEMON

What does the scheduler do ?

- Responsible for application
- Does not perform monitoring or tracking of status for the applications
- Offers no guarantee about restarting failed tasks due to hardware or application failures



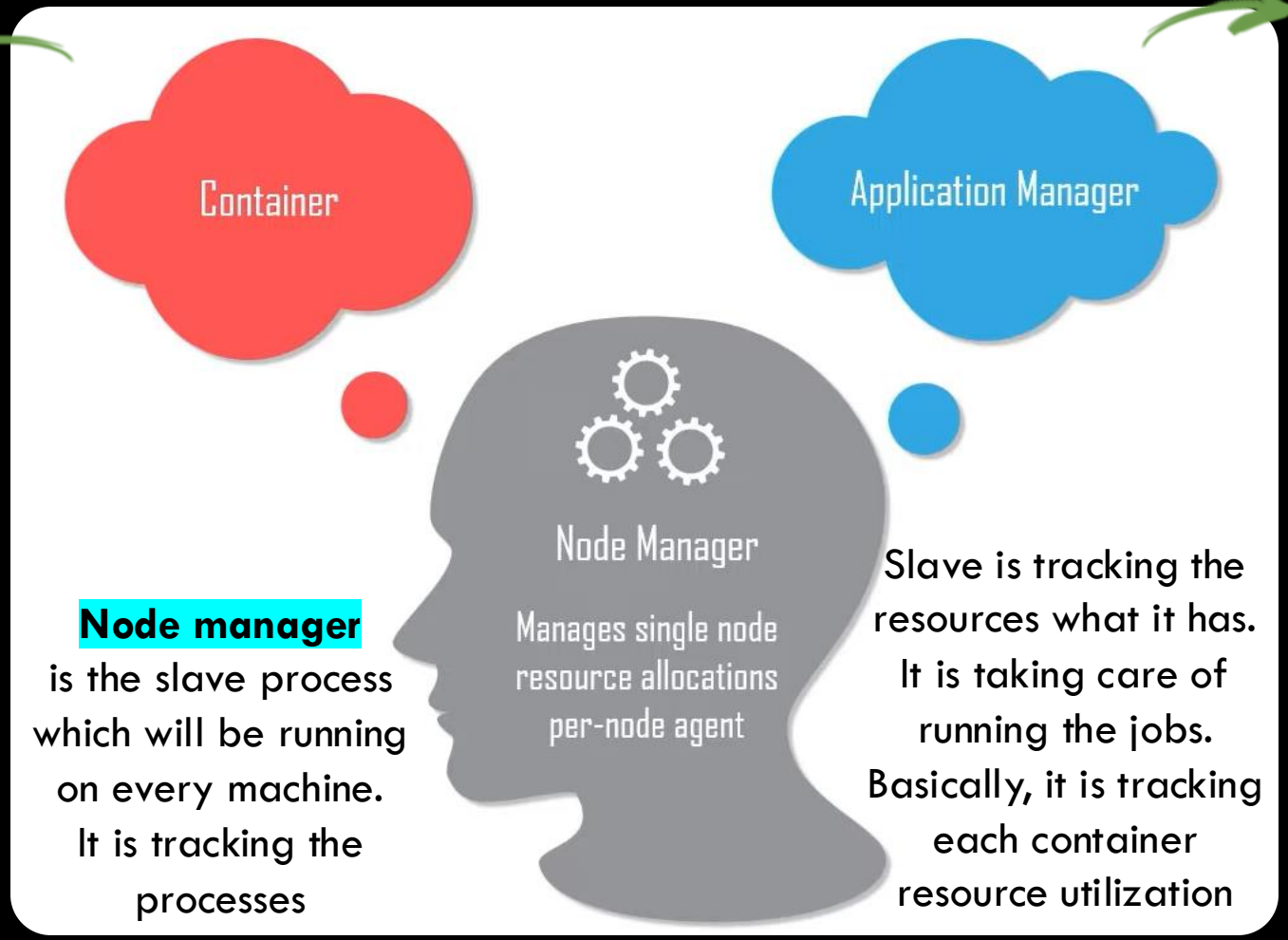
What does the application manager do ?

- Responsible for accepting job-submissions
- Negotiates the first container for executing the application specific application master
- Provides the service for restarting the Application Master container on failures.

YARN DAEMON

What is a container ?

- A container is a collection of resources like CPU, memory, disk, which could be used or which already has data and network, so the node manager is looking into the request from the application master.
- it authenticates and provides rights to an application to use specific amount of resource.



Node manager is the slave process which will be running on every machine. It is tracking the processes

Slave is tracking the resources what it has. It is taking care of running the jobs. Basically, it is tracking each container resource utilization

What is a Apps Master ?

- Application master manages resource needs of individual applications
- Interacts with scheduler to acquire required resources and node manager to execute and monitor tasks

INTRODUCTION TO MAPREDUCE IN HADOOP 2.0

MapReduce in Hadoop 2.0 (MR2)

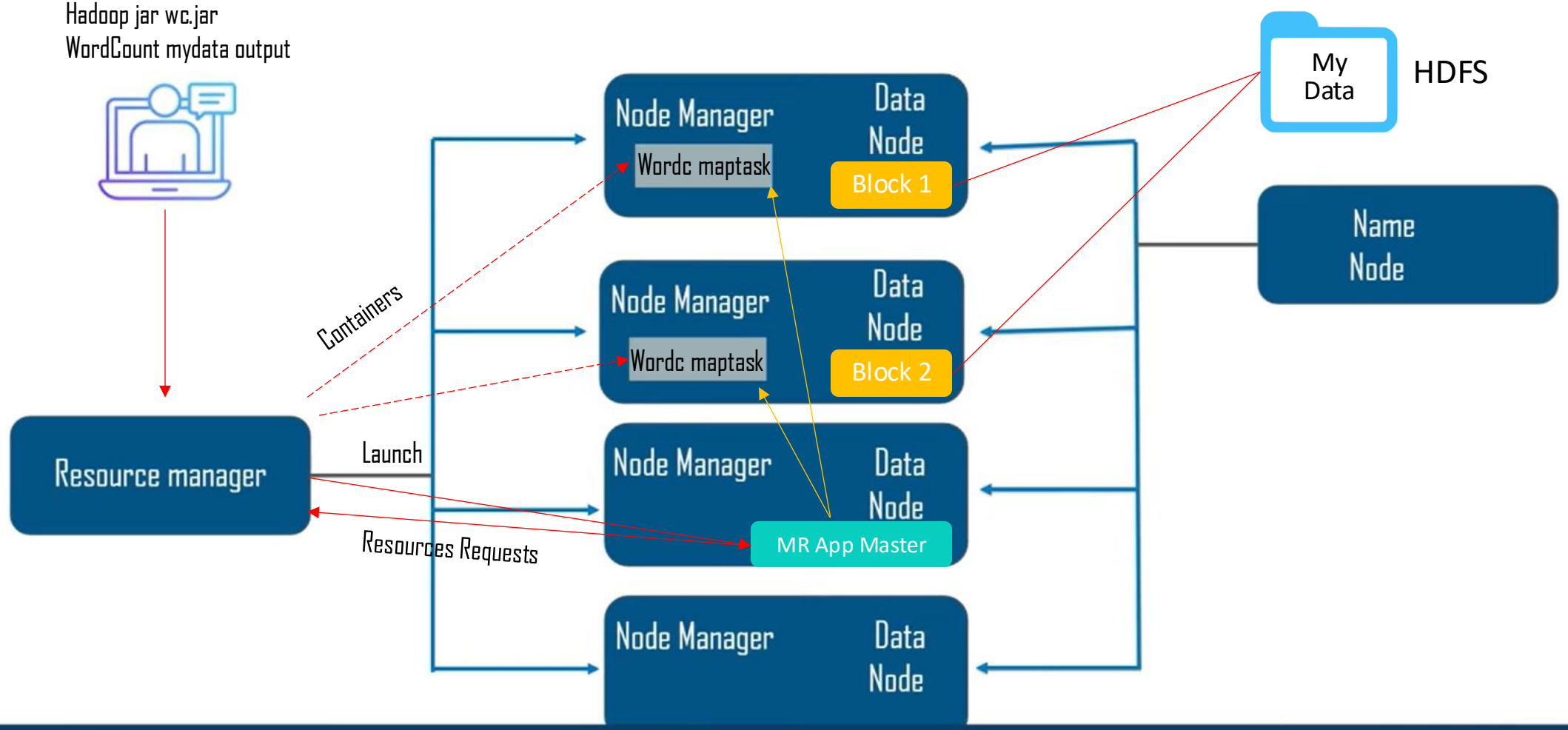
01 With YARN, there is no longer a single Job Tracker to run jobs and a Task Tracker to run tasks of the jobs

02 The old MRI framework rewritten to run with a submitted application on top of YARN. This application was christened MR2 or MapReduce version 2

03 It is the familiar MapReduce execution underneath except that each job now controls its own destiny via its own Application Master taking care of execution flow

04 It is more isolated & scalable model than the MRI system where a singular Job Tracker does all the resource management, scheduling & task monitoring work

HOW DOES MR2 WORK?



THE OVERALL MAP REDUCE WORD COUNT PROCESS