Cauvery College for Women (Autonomous)

Nationally Accredited (III Cycle) with 'A' Grade by NAAC

Annamalai Nagar, Tiruchirappalli-18.

Name of the Faculty : Dr Sinthu Janita
Designation : Professor & Head
Department : Computer Science
Contact Number : 9894484436
Programme : MSc Computer Science
Batch : 2016-2017 Onwards
Semester : IV
Course : Big Data Analytics
Course Code :P16CSE5A
Unit : V
Topics Covered : Introduction to YARN
Components
Need and Challenges of YARN

Dissecting YARN

Hadoop MapReduce and YARN Framework

- Introduction to MapReduce
- Processing data with Hadoop using MapReduce
- Introduction to YARN
- Components
- Need and Challenges of YARN
- Dissecting YARN
- MapReduce application
- Data serialization and Working with common serialization formats
- Big data serialization formats

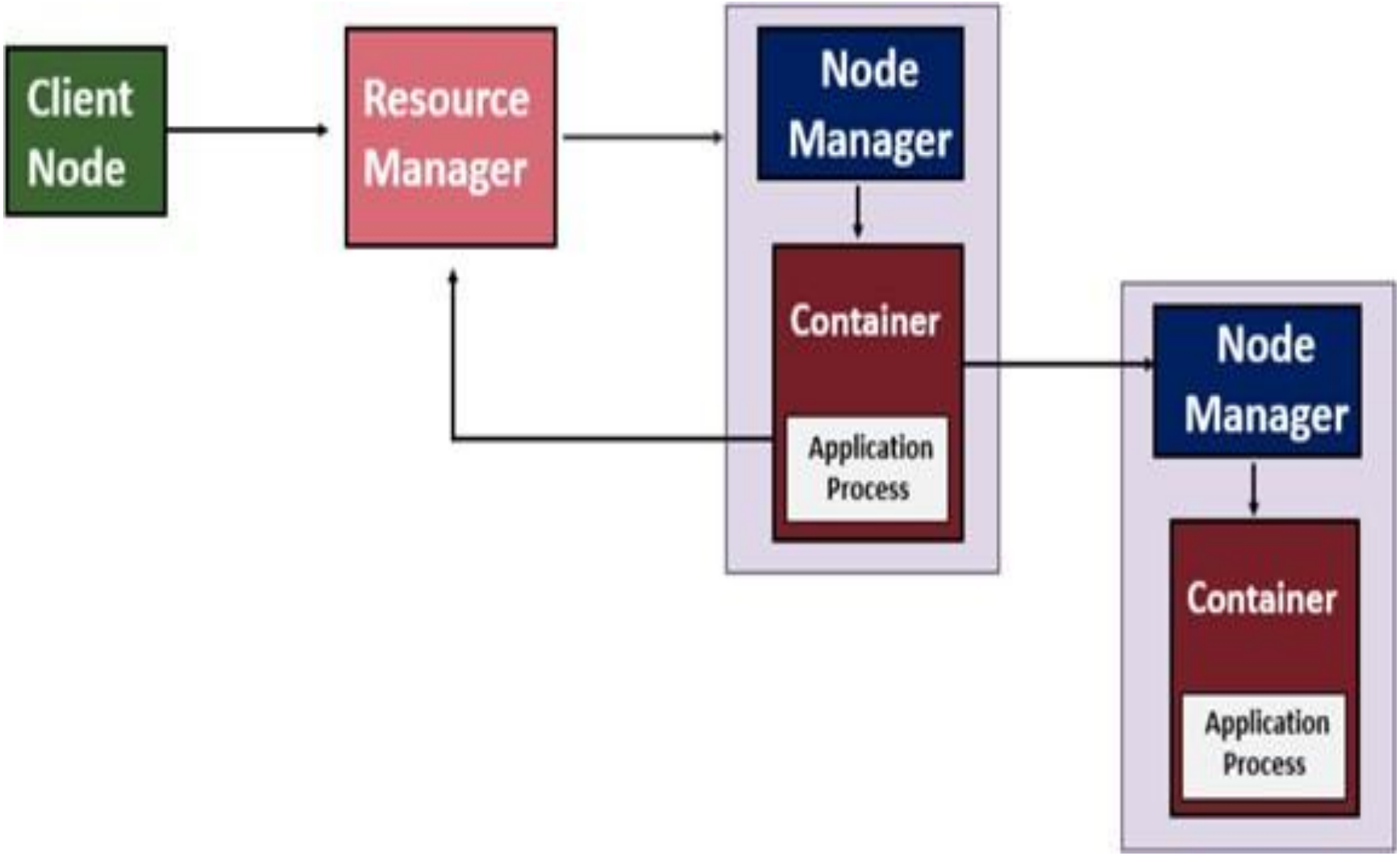
Introduction to YARN

- YARN, which is known as Yet Another Resource Negotiator, is the **Cluster management** component of Hadoop 2.0.
- Hadoop YARN Architecture is the **reference architecture for resource management** for Hadoop framework components.
- It includes Resource Manager, Node Manager, Containers, and Application Master.
- The Resource Manager is the major component that manages application management and job scheduling for the batch process
- Node manager is the component that manages task distribution for each data node in the cluster.
- Containers are the hardware components such as CPU, RAM for the Node that is managed through YARN.
- Application Master is for monitoring and managing the application lifecycle in the Hadoop cluster.
-

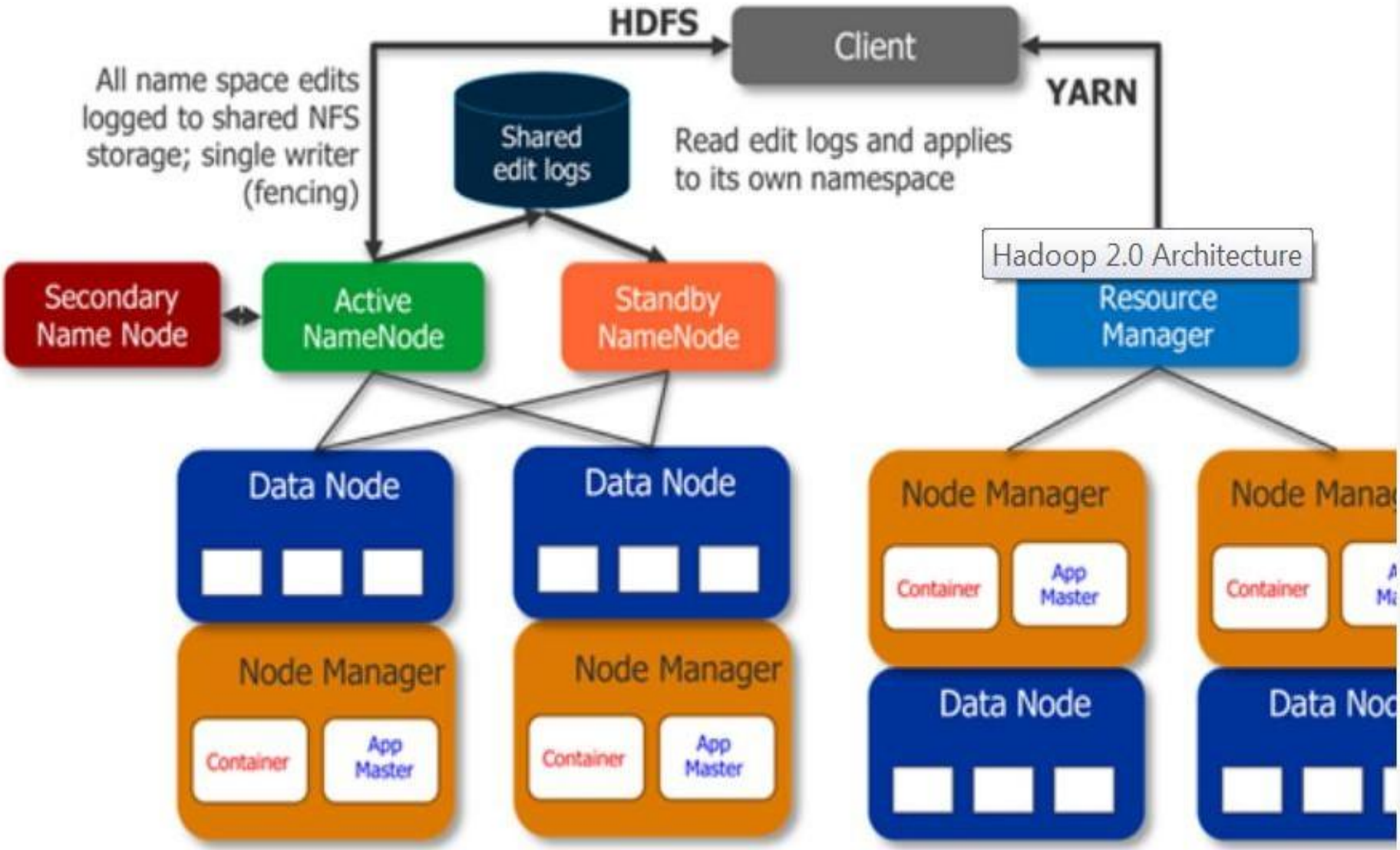
Limitations of Hadoop 1.0 (or) Need for YARN

- In Hadoop 1.0 , **HDFS and MapReduce are core components** , while other components are built around the core
- **Single NameNode** is responsible for managing entire namespace for Hadoop clusters .
- It has restricted processing model which is suitable for **batch oriented** MapReduce jobs
- Hadoop MapReduce is **not suitable for interactive analysis** .
- Hadoop 1.0 is **not suitable for Machine Learning algorithms , graphs and other memory intensive algorithms** .
- MapReduce is **responsible for cluster resource management and data processing .**
- In this architecture , mapslots might be “full” which reduced lots are “empty” and vice versa. This causes **resource utilisation issues** .
- This needs to be improved for proper resource processing .

YARN Architecture(Dissecting YARN)



YARN Architecture



YARN Architecture

A client program submits the application which includes the necessary specifications to launch the application specific **ApplicationMaster** itself

The ResourceManager launches the ApplicationMaster by assigning some containers

The ApplicationMaster on boot-up registers with the ResourceManager. This helps the client program to query the ResourceManager directly for the details

During the normal course, ApplicationMaster negotiates appropriate resource containers via the resource- request protocol.

YARN Architecture ctd..

- On successful container allocations, the ApplicationMaster launches the container by providing the container launch application to the NodeManager

- The NodeManager executes the application code and provides necessary information such as progress status etc. to its ApplicationMaster via an application specific protocol. During the application execution, the client that submitted the job directly communicates with the ApplicationMaster to get status, progress updates etc. via an application specific protocol.

- Once the application has been processed completely, ApplicationMaster deregisters with the ResourceManager and shuts down, allowing its own container to be repurposed

Components of YARN

YARN introduces the concept of a Resource Manager and an Application Master in Hadoop 2.0.

The Resource Manager sees the usage of the resources across the Hadoop cluster

The life cycle of the applications that are running on a particular cluster is supervised by the Application Master.

For cluster resources, the Application Master negotiates with the Resource Manager. This task is carried out by the containers which hold definite memory restrictions. Then these containers are used to run the application-specific processes.

These containers are supervised by the Node Managers which are running on nodes in the cluster. This will confirm that no more than the allocated resources are used by the application.

I) Resource Manager

- The Resource Manager sees the usage of the resources across the Hadoop
- YARN works through a Resource Manager which is one per node
- Node Manager runs on all the nodes
- The Resource Manager manages the resources used across the cluster the Node Manager lunches and monitors the containers
- Scheduler and Application Manager are two components of the Resource Manager.

I) Resource Manager ctd...

i)Scheduler:

Scheduling is performed based on the requirement of resources by the applications. YARN provides few schedulers to choose from and they are Fair and Capacity Scheduler.

In case of any hardware or application failure, the Scheduler does not ensure to restart the failed tasks.

Scheduler allocates resources to the running applications based on the capacity and queue.

ii) **Application Manager**:

The life cycle of the applications that are running on a particular cluster is supervised by the Application Master.

It manages the running of Application Master in a cluster and on the failure of the Application Master Container, it helps in restarting it. Also, it bears the responsibility of accepting the submission of the jobs.

2) Node Manager

- Node Manager is responsible for the execution of the task in each data node.
- The Node Manager in YARN by default sends a heartbeat to the Resource Manager which carries the information of the running containers and regarding the availability of resources for the new containers.
- It is responsible for seeing to the nodes on the cluster individually and manages the workflow and user jobs on a specific node.
- Chiefly it manages the application containers which are assigned by the Resource Manager.
- The Node Manager starts the containers by creating the container processes which are requested and it also kills the containers as asked by the Resource Manager.

3) Containers

- The Containers are set of resources like RAM, CPU, and Memory etc on a single node and they are scheduled by Resource Manager and monitored by Node Manager.
- The Container Life Cycle manages the YARN containers by using container launch context and provides access to the application for the specific usage of resources in a particular host.

4) Application Master

- For cluster resources, the Application Master negotiates with the Resource Manager. This task is carried out by the containers which hold definite memory restrictions. Then these containers are used to run the application-specific processes.
- It monitors the execution of tasks and also manages the lifecycle of applications running on the cluster. An individual Application Master gets associated with a job when it is submitted to the framework. Its chief responsibility is to negotiate the resources from the Resource Manager. It works with the Node Manager to monitor and execute the tasks.