Assignment 3.3

**Components of hadoop2.X**

* **YARN**
* **HDFS**
* **Map Reduce**

Under this three component the Resource Manager, Node Manager, History Server is present.

New Features added to hadoop 2.X

* Hadoop 2.x has some common Hadoop API which can easily be integrated with any third party applications to work with Hadoop
* It has some new Java APIs and features in HDFS and MapReduce which are known as HDFS2 and MR2 respectively
* New architecture has added the architectural features like HDFS High Availability and HDFS Federation
* Hadoop 2.x not using Job Tracker and Task Tracker daemons for resource management now on-wards, it is using YARN (Yet Another Resource Negotiator) for Resource Management.

HDFS High Availability:

Problem in 1.X: IF the name node fails then there is no access to the hadoop cluster.

Solution:

* Hadoop 2.x supports two Name Nodes at a time one node is active and another is standby node
* Standby Name Node manages metadata same as Secondary Name Node in Hadoop 1.x

Hadoop 2.x Maintains the resource manager and node manager instead of job tracker and task tracker.

**Resource Manager:**

The resource manager manages the resources among all the applications and it act as a central control for all the node managers.

Two Components:

1. Scheduler

The Scheduler is responsible for allocating resources to the various running applications. The Scheduler performs its scheduling function based the resource requirements of the applications. It is responsible for partitioning the cluster resources among the various queues, applications etc.

There are many schedulers present in the resource manager:

* + Capacity Scheduler
  + Fair Scheduler
  + FIFO Scheduler etc.

1. Application Manager

The Application Manager is responsible for accepting job-submissions and executing the applications specific Application Master and provides the service for restarting the Application Master container on failure. The Application master is responsible for resource container negotiate from the scheduler and tracking the status and monitoring for progress.

**Node Manager:**

* Node Manager is a per-node level component.
* This process runs on slave nodes (normally on HDFS Data node machines)
* It is responsible for coordinating with Resource Manager for task scheduling and tracking the resource utilization on the slave node
* It also reports the resource utilization back to the Resource Manager
* It uses other process like Application Master and Container for MapReduce task scheduling and execution on the slave node

The Node manager manages the following:

* + Container lifecycle management
  + Container dependencies
  + Node and container resource usage
  + Log management
  + Reporting node and container status to the ResourceManager.

**History Server:**

A YARN component called the History Server archives job metrics and metadata. Status of the completed applications are available in the history server.