**ASSIGNMENT 3.1**

**Components of Hadoop 2.X:**

* HDFS (Hadoop Distributed File System) – Storage system
  + Name Node
  + Secondary Name Node
  + Data Node
* YARN (Yet Another Resource Negotiator) – Processing framework
  + Resource Manager
  + Node Manager
  + Application Master
  + Scheduler
  + Container

**HDFS Daemons:**

* **Name Node (the master):** Name Node handles the file system name space. It maintains the directory tree of all files in the file system and the metadata of those files in the tree. It determines the mapping of blocks to Data Nodes. It also performs typical file related operations such as opening, closing and renaming directories and files.
* **Data Node (workers):** Data Nodes perform the task of storing and reading data to and from blocks of disk whenever they are instructed to do so by clients or the Name Node. They report the list of blocks that they are storing to Name Node periodically. When a cluster get started, a Data Node connects to the Name Node, waits for a request and performs relevant operations upon getting a request.
* **Secondary Name Node:** The major role of secondary name node is to merge the namespace image with the edit log periodically which is used when the name node gets failed. It runs on a separate physical machine since it needs a lots of CPU and memory capacity to perform merge operations. It is not a backup for name node. Rather, name node takes help of secondary name node to make sure the file system image is intact with its current status periodically.

**YARN Daemons:**

* **Resource Manager:** This component is considered as the negotiator of all resources in the cluster. Every cluster has one instance of a resource manager. In Hadoop 2.0, MapReduce job is treated as an application. It gets information from node manager periodically (in the form of heartbeats) about containers which have resources, assigns resources among applications and makes sure that resources are utilized optimally.
* **Node Manager:** Node Manager runs on each node of the cluster and communicates with Resource Manager about usage of resources on the machine. It performs following tasks:
  + Manages Container lifecycle
  + Offers Containers to applications for lease
  + Monitors Node health
  + Sends heartbeat request to Resource Manager
* **Application Master:** This component is responsible for processing data in the MapReduce environment. It accepts job submissions, requests Resource Manager to get resources and communicates with Node Manager to get those resources to execute assigned tasks. Application Master could be MapReduce or any other processing framework.
* **Scheduler:** Scheduler is responsible for allocating resources to various applications running in the cluster. In this way it helps Resource Manager to perform its job. Each scheduler may use different algorithm to do this. Few commonly used schedulers are FairScheduler, CapacityScheduler, etc. CapacityScheduler is the default scheduler in Hadoop 2.x. Its purpose is to allow multi-tenancy and share resources between multiple applications on the same cluster.
* **Container:** A Container in YARN platform is an indication of successful resource allocation, which means Resource Manager has granted resources for lease to an application running on a specific node. To launch the container, the Application Master must provide a container launch context (CLC) that includes the following information:
  + Environment variables
  + Security tokens
  + Dependencies
  + Command to create a process to execute application tasks

In other words, it is a set of allocated system resources (CPU core and memory). Node Manager maintains Containers by offering them to tasks for their execution.