Components of Hadoop 2.x

HDFS(Hadoop Distributed File System) is a storage system.  
YARN(Yet Another Resource Negotiator) is a processing system.

This consists of following daemons.

* **Name Node (HDFS):**

It stores metadata only. The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself.

* **Secondary Name Node (HDFS):**

It’s main function is to take checkpoints of the file system metadata present on namenode.

As the NameNode is the single point of failure in HDFS, if NameNode fails entire HDFS file system is lost. So in order to overcome this, Hadoop implemented Secondary NameNode whose main function is to store a copy of FsImage file and edits log file.

It is not a backup namenode. The Secondary NameNode is a helper to the primary NameNode but not replace for primary namenode.

* **Data Node (HDFS):**

Stores blocks from files. A functional file system typically have more than one DataNode in the cluster, with data replicated across them. On startup, a DataNode connects to the NameNode; spinning until that service comes up. It then responds to requests from the NameNode for file system operations.

Sends signals to NameNode periodically (called as Heartbeat) to verify it is active.

* **Resource Manager (YARN):**

The function of the Resource Manager is simple: Keeping track of available resources. One per cluster. It contains two main components: Scheduler and Applications Manager

* **Scheduler (YARN)**

It is plugged with Resource Manager to help in resource allocation.

Different schedulers allocate resources using different algorithms.

* **Application Master (YARN)**

It is the actual instance which does processing. It requests Resouce Manager for resources and works with NodeManager to get those resources for task execution. Application Master could be MapReduce or any other processing framework.

* **Node Manager (YARN):**

The NodeManager is the per-machine framework agent who creates container for each task. The containers can have variable resource sizes and the task can be any type of computations not just map/reduce tasks. It then monitors the resource usage (cpu, memory, disk, network) of the container and report them to the ResourceManager.

* **Container (YARN):**

It is a set of allocated system resources (CPU Core and Memory).

Containers are allocated and managed by NodeManager and are used by tasks.

Below is the high level diagram which depicts hadoop 2.x components

