Architecture of Apache Hadoop YARN :-

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. There are mainly five building blocks inside this runtime envinroment. Those 5 blocks are :

1. The cluster is the set of host machines(nodes).Nodes may be partitioned in racks. This is the hardware part of the infrastructure.
2. The YARN Infrastructure (Yet Another Resource Negotiator) is the framework responsible for providing the computational resources (e.g., CPUs, memory, etc.) needed for application executions.

Two important elements of YARN are:

1)The Resource Manager (one per cluster) is the master. It knows where the slaves are located (Rack Awareness) and how many resources they have. It runs several services, the most important is the Resource Scheduler which decides how to assign the resources.

2)the Node Manager (many per cluster) is the slave of the infrastructure. When it starts, it announces himself to the Resource Manager. Periodically, it sends an heartbeat to the Resource Manager. Each Node Manager offers some resources to the cluster. Its resource capacity is the amount of memory and the number of vcores. At run-time, the Resource Scheduler will decide how to use this capacity: a Container is a fraction of the NM capacity and it is used by the client for running a program.

iii) The HDFS Federation is the framework responsible for providing permanent, reliable and distributed storage. This is typically used for storing inputs and output (but not intermediate ones).

iv) The other alternative storage solutions. For instance, Amazon uses the Simple Storage Service (S3).

v) The MapReduce Framework is the software layer implementing the MapReduce paradigm.

In YARN, there are at least three actors:

i) The Job Submitter (the client)

ii) The Resource Manager (the master)

iii)The Node Manager (the slave)

The application startup process is the following:

1) A client submits an application to the Resource Manager

2) The Resource Manager allocates a container

3) The Resource Manager contacts the related Node Manager

4) The Node Manager launches the container

5) The Container executes the Application Master The Application Master is responsible for the execution of a single application.

It asks for containers to the Resource Scheduler (Resource Manager) and executes specific programs (e.g., the main of a Java class) on the obtained containers. The Application Master knows the application logic and thus it is framework-specific. The MapReduce framework provides its own implementation of an Application Master.

The Resource Manager is a single point of failure in YARN. Using Application Masters, YARN is spreading over the cluster the metadata related to running applications. This reduces the load of the Resource Manager and makes it fast recoverable.