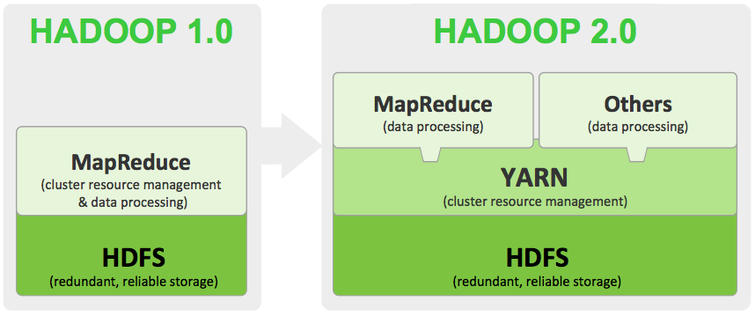
**Assignment 8.2**

**Question:** Explain the core changes made in Hadoop 2.x.

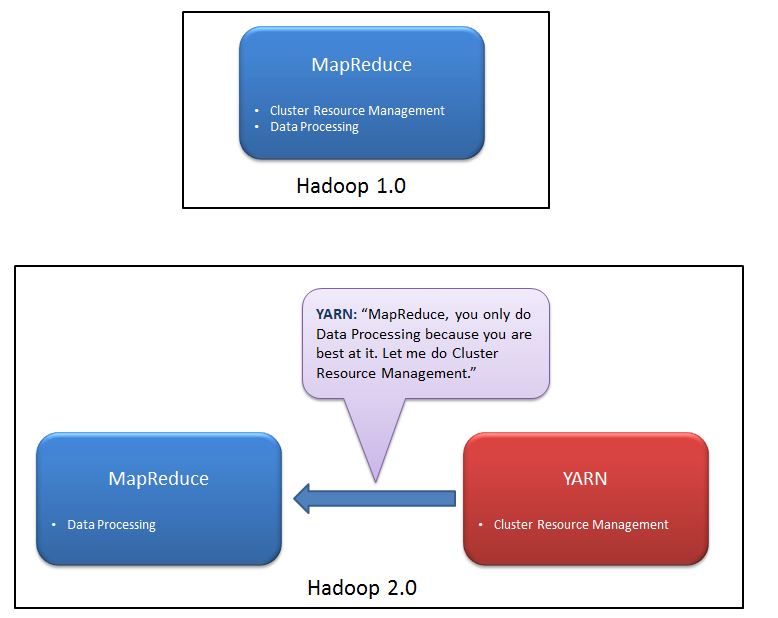
* Earlier every non-Map Reduce applications were forced to be modeled as MapReduce as compute resources in Hadoop1.X were only available to MapReduce programs. In Hadoop 2.X, YARN component generalizes the compute layer to execute not just MapReduce style but other new breed of applications, such as stream processing. The new architecture is more decentralized and allows Hadoop clusters to be scaled significantly to more cores and servers.
* In hadoop 1.X we have only JobTracker to manage both the compute resources and the jobs that use the resources. In hadoop 2.X, YARN is a resource manager that splits function into two. First is a Resource Manager (RM) which focuses on managing the cluster resources and second is an Application Master (AM), which manages each running application such as a MapReduce job. The AM requests resources from the RM, based on the needs and characteristics of the application being run. YARN is designed to allow multiple, diverse user applications to run on a multitenant platform. In addition to MapReduce YARN supports multiple processing models. YARN is also called as next generation execution layer of Hadoop. YARN eliminates bottleneck problem of job execution in MapReduce.
* In Hadoop, the cluster’s storage resources are available only to HDFS. But now in Hadoop 2.X the new storage architecture generalizes the block storage layer so that it can be used not only by HDFS but also other storage services which is as similar to that of YARN. Hadoop 2.X also support for heterogeneous storage. Hadoop 1.X treated all storage devices such as spinning disks on a DataNode as a single uniform pool. Hadoop 2.X will differentiate between storage types along with making the storage type information available to frameworks and applications by which they can take advantage of storage properties.
* Hadoop 1.X has a single master server called NameNode where all the metadata is stored. When the NameNode is brought down by any software or hardware failure, the cluster would be unavailable until it is restarted. Hadoop 2.X handles this situation by triggering automatic failover by which the standby NameNode becomes active. Here ZKFC (Zookeeper-based Failover Controller) manages failover of NameNodes. On each of the NameNodes this daemon runs and a session is maintained with the Zookeeper. An active local NameNode is elected by one of the ZKFC with the coordination of Zookeeper. Periodically NameNode health check is done by ZKFC. The local ZKFC resigns as the leader when the active NameNode fails health check. Similarly, when failure occur in the active NameNode machine, Zookeeper detects the loss and removes the ZKFC from the failed node as the leader and the ZKFC running on standby becomes the leader by makes the local standby NameNode active. This results in automatic failover.
* Hadoop 1.X was developed to support only the UNIX family of operating systems. But With Hadoop 2.X, the Windows operating system is indigenously supported because of the fact that Hadoop was written in Java. The compute and storage resource which were dependent on UNIX have been generalized to support Windows. This broadens Hadoop to reach Windows Server market.



**Question:** Explain the difference between MapReduce 1 and MapReduce 2 / Yarn.

**MapReduce 1:** which is also called as Hadoop 1 where the **HDFS** (Resource management and scheduling) and MapReduce(Programming Framework) are tightly coupled. Because of this non-batch applications can not be run on the hadoop 1. It has single namenode so, it doesn't provides high system availability and scalability.

**Map Reduce 2:** which is also called as **YARN** in this version of hadoop the resource management and scheduling tasks are separated from MapReduce which is separated by **YARN(Yet Another Resource Negotiator).** The resource management and scheduling layer lies beneath the MapReduce layer. It also provides high system availability and scalability as we can create redundant NameNodes. The new feature of snapshot through which we can take backup of filesystems which helps disaster recovery.



* No more JobTracker and TaskTracker needed in Hadoop 2. With the introduction of YARN in Hadoop2, the term JobTracker and TaskTracker disappeared. MapReduce is now streamlined to perform processing data.
* The new model is more isolated and scalable as compared to the earlier MR1 system. MR2 is one kind of distributed application that run MapReduce framework on top of YARN. MapReduce perform data processing via YARN. Other tools can also perform data processing via YARN. Hence Yarn execution model is more generic than earlier MapReduce model.
* MR1 was not able to do so. It would only run MapReduce applications.