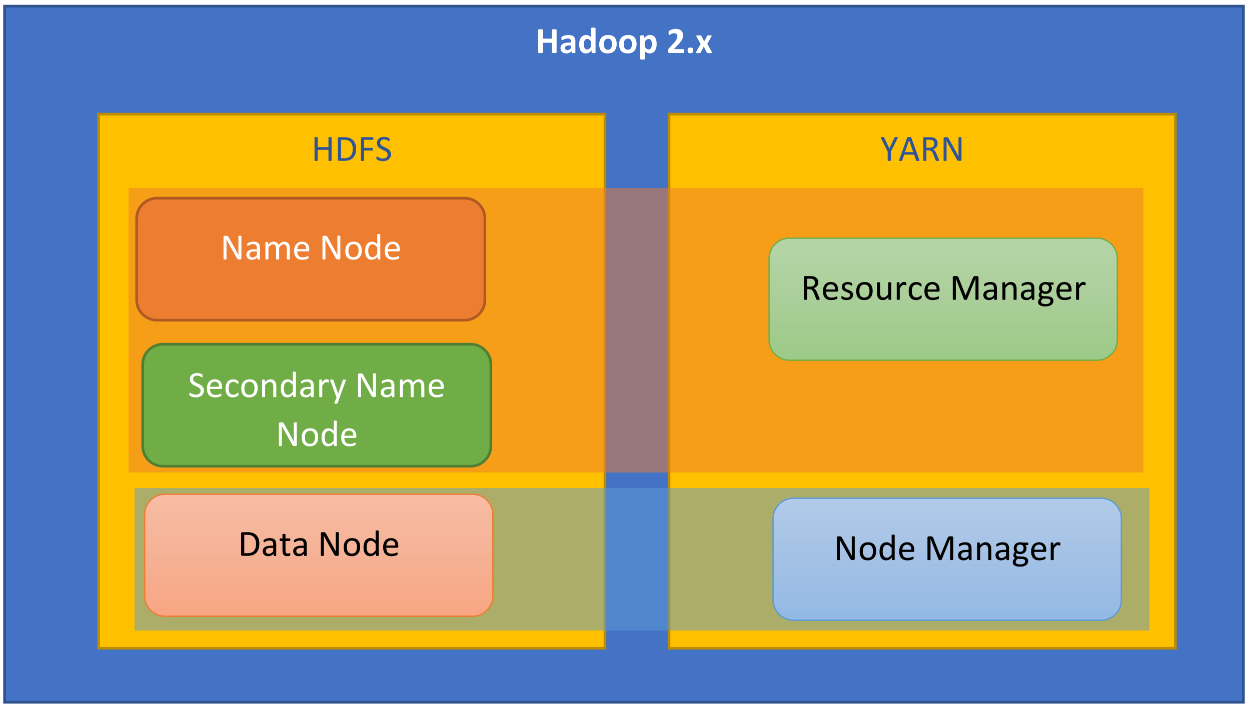
Assignment 1.5

Components of Hadoop 2.x

Hadoop comprises of 2 main core components:

1. **HDFS (Hadoop Disributed File System)** – Storage Unit
2. **YARN (Yet Another Resource Negotiator)** – Processing Unit



HDFS

HDFS comprises of 3 components:

1. Name Node

Name Node is the 'master' component of HDFS system. It stores the metadata of HDFS, which is essentially the directory tree of the files in HDFS and a track of all the files across the cluster.

Name node does not store the actual data. It knows the list of blocks and locations for any given file.

When name node is down, the cluster itself is considered down as the cluster becomes inaccessible.

It stores fsimage(snapshot of filesystem at strt-up) and editlogs(sequence of changes made to filesystem after start-up).

2. Secondary Name Node

Secondary Name node is NOT a backup name node as the name might suggest. It is more of assistive node.

It is responsible to taking checkpoints of the file system metadata present in namenode. It does this by checkpointing fsimage. It gets the edit logs from name node in regular intervals and creates the checkpoint. Then it copies this fsimage back to name node.

Name node uses this fsimage in next start-up, so that it reduces start-up time.

3. Data Node

This is where the actual data is stored. Name node keeps track of each block that is stored in these data nodes. It is also known as the 'slave'.

When it starts-up, it announces it's presence to name node and also the data blocks it is responsible for. This is the workhorse of Hadoop HDFS.

YARN

YARN comprises of 2 components:

1. Resource Manager

Resource Manager is responsible for taking inventory of available resources and run critical services, most critical of which is the Scheduler.

The scheduler allocates resources. It negotiates the available resources in the cluster and anages the distributed processing. It works along with the Node Manager to attain this.

2. Node Manager

Node manager acts as the 'slave' to the Resource Manager.

It keeps track of the tasks and jobs that are being deployed to the data nodes. It helps Resource manager in keeping track of available space, processing power. Memory, bandwidth etc. so that tasks can be distributed to the data nodes.