**Assignment 3.1**

**Components of Hadoop 2.x**

Hadoop 2.x or later versions are using the following Hadoop architecture. When compared to Hadoop 1.x, Hadoop 2.x has added one new component **YARN** and also updated HDFS and MapReduce components’s Responsibility.

**Hadoop Common Module**

**HDFS V.2**

**YARN (MR V2)**

**Hadoop Ecosystem**

**MapReduce**

**(MR V1)**

* Hadoop Common Module is a Hadoop Base API (A JAR file) for all Hadoop Components. All other components work on top of this module.
* HDFS stands for **Hadoop Data File System**, also known as HDFS v2. It is used as Distributed Storage System in Hadoop Architecture.
* YARN stands for **YET ANOTHER RESOURCE NEGOTIATOR**. It is a new component in Hadoop 2.x. Also known as MR V2.
* MapReduce is a Batch Processing or Distributed Data Processing Module. It is also known as MR V1 as it was also a part of Hadoop 1x. In Hadoop 2.x it has few updated features.
* Remaining all Hadoop EcoSystem components work on top of the **three major components: HDFS, YARN and MapReduce**.

**Hadoop 2.x Core Components:**

**Hadoop 2.x Core Components**

**Node Manager**

**Resource Manager**

**Secondary NameNode**

**DataNode**

**NameNode**

**YARN**

**HDFS**

1. **HDFS:** Hadoop Data File System (HDFS) has 2 components:
2. **NameNode:** This is a center piece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster files data is kept.
3. **DataNode:** This stores the actual data in the HDFS. A functional filesystem typically have more than one DataNode in the cluster, with data replicated across them.

On startup DataNode connects to the NameNode, spinning until the service comes up.

All namespace edits logged to shared NFS storage

Reads edit logs and applies to its own namespace

**Shared Edit Logs**

**Secondary NameNode**

**StandBy NameNode**

**Active NameNode**

DataNode

DataNode

NodeManager

NodeManager

AppMaster

Container

AppMaster

Container

1. **YARN:** Hadoop 2.0 is popularly known as YARN (Yet Another Resource Negotiator), it was introduced on October 2013.

YARN has taken an edge over the cluster management responsibilities from MapReduce, so that now MapReduce just takes care of the Data Processing and other responsibilities are taken care by YARN.

Resource Manager

**x**

DataNode

NodeManager

AppMaster

Container

DataNode

AppMaster

NodeManager

Container

The job tracker in YARN mainly depends on 3 components:

* **Resource Manager** component, which is considered as negotiator of all the resources in the cluster.
* **Node Manager** component, which is the job history server component of YARN which will furnish the information about all the complete jobs. The Node Manager keeps a track of all the user’s jobs and their workflow particular given node.
* **Application Master** component, which is responsible for managing each and every Map Reduce Job and is concluded once the job completes processing.

1. **MapReduce V1:** MapReduce V1 (MRv1) is a part of Hadoop 1.x and is an implementation of the MapReduce Programming paradigm.

Map Reduce itself can be broken into following parts:

1. **End-User MapReduce:** This is the API needed to develop the MapReduce applications.
2. **MapReduce Frameowrk:** This is the run-time implementation of various phases, such as map phase , the sort/shuffle/merge aggregation phase and the reduce phase.
3. **MapReduce System:** This is the backend infrastructure required to run the MapReduce application and includes things such as cluster resource management, scheduling of jobs and so on.

Since it runs on cluster, its cluster management components was also tightly coupled with the MR Programming paradigm. The **only** thing that could run on Hadoop 1.x was an MR Job.

In MRv1 the cluster was managed by a single Job Tracker and multiple Task Tracker running on the Data Nodes.

In Hadoop 2.x, the old MRv1 framework was written to run on top of YARN. This application was named MRv2 or YARN.