Assignment 3.1:

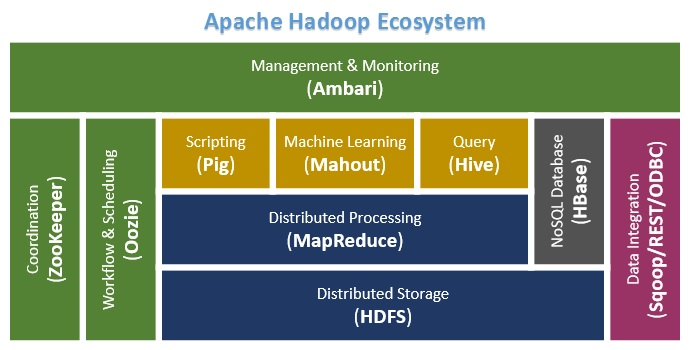
List the Components of Hadoop 2.x and explain each component in detail.

Ans:

Core Hadoop 2.x consists of the following components:

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

Below image shows the architecture of Hadoop ecosystem.



Apache Hadoop 2.x (MRv2) consists of the following daemons:

* NameNode
* Secondary NameNode
* Data Node
* Resource Manager
* Node Manager

All these Daemons are nothing but a piece of code. Java code is running at the background. In order to run Java Code, we need JVM, So each daemon service need some JVM service to run.

**Name Node (Hadoop File System Component):** Stores Metadata only.

The Namenode is the centrepiece 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.

NameNode contains two important files on its hard disk:

1. fsimage (file system image)

It contains:

• All directory structure of HDFS

• Replication level of file

• Modification and access times of files

• Access permissions of files and directories

• block size of files

2. Edits

• When any write operation takes place in HDFS, the directory structure gets modified.

• These modifications are stored in memory as well as in edits files (edits files are stored on

Hard disk).

• If existing fsimage file gets merged with edits, we’ll get updated fsimage file.

• This process is called checkpointing and is carried out by Secondary Namenode. It takes

fsimage and edits files from Namenode and returns updated fsimage file after merging.

**Secondary NameNode:** Performs house-keeping activities for NameNode, like periodic merging of namespace and edits.

• This is not a back up for NameNode.

**DataNode (Hadoop FileSystem Component):** It is also known as Slave. HDFS Datanode is responsible for storing actual data in HDFS. DataNode performs read and write operation as per the request of the clients. Replica block of DataNode consists of 2 files on the file system. The first file is for data and second file is for recording the block’s metadata. HDFS Metadata includes checksums for data. At start up, each DataNode connects to its corresponding Namenode and does handshaking. Verification of namespace ID and software version of DataNode take place by handshaking. At the time of mismatch found, DataNode goes down automatically.

Tasks of HDFS DataNode

• DataNode performs operations like block replica creation, deletion, and replication according to the instruction of NameNode.

• DataNode manages data storage of the system.

**Resource Manager (YARN Component):**

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.  
The Scheduler is responsible for allocating resources to the various running applications.  
The ApplicationsManager is responsible for accepting job-submissions, negotiating the first container for ApplicationMaster and provides the service for restarting the Application Master container on failure.

**Node Manager (YARN Component):**

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.