**Components of Hadoop 1.x**

**NameNode**

1. Contains Hadoop File System Tree and other metadata information about files and directories.
2. Contains in-memory mapping of which blocks are stored in which datanode.

**Secondary Namenode**

1. Performs house-keeping activities for namenode, like periodic merging of namespace and edits.
2. This is not a back up for namenode.

**DataNode**

1) Stores actual data blocks of file in HDFS on its own local disk.

2) Sends signals to NameNode periodically (called as Heartbeat) to verify it is active.

3) Sends block reporting to the nameode on cluster startup as well as periodically at every 10th Heartbeat.

4)The data node are the workhorse of the system.

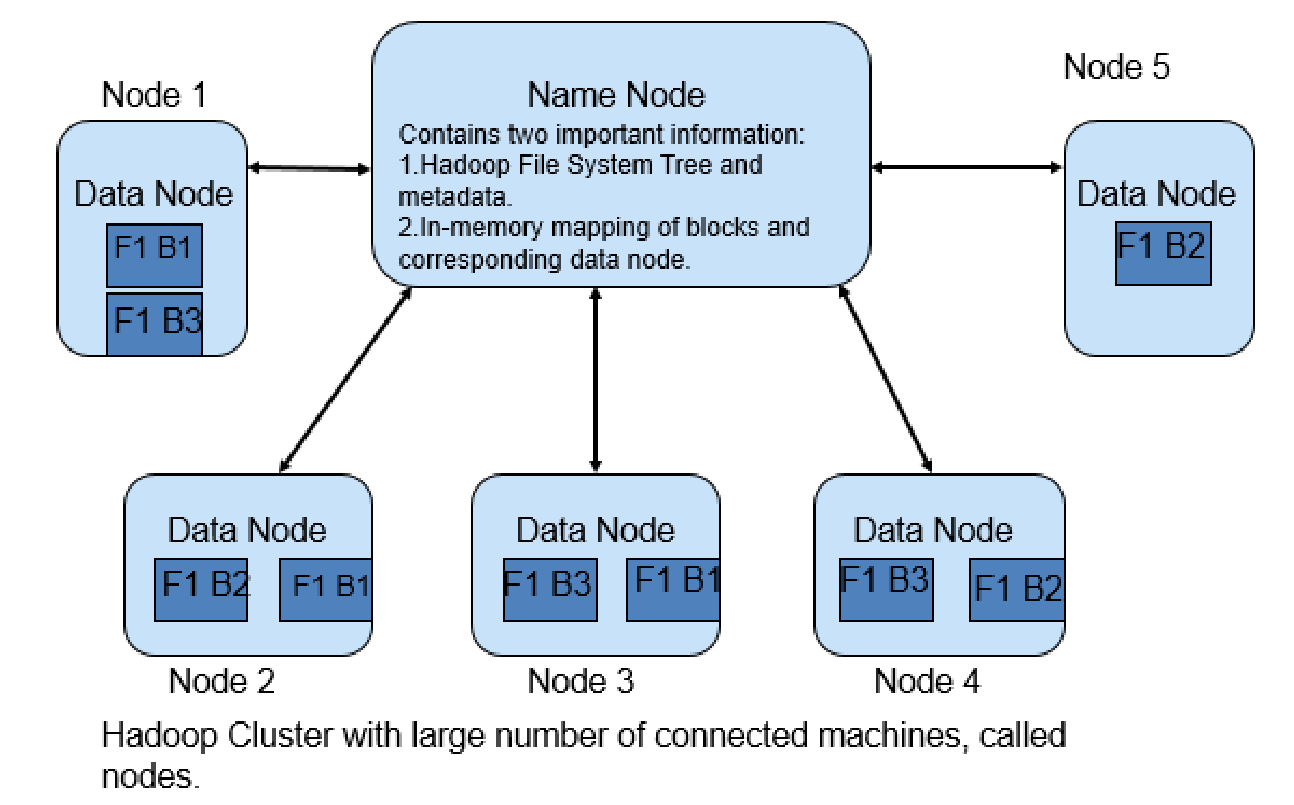
5) They perform all the block operation including periodic checksum. They receive instructions from the name node of where to put the blocks and how to put the blocks.

**JobTracker (Not present in Hadoop 2.x)**

1) Controls overall execution of map reduce jobs

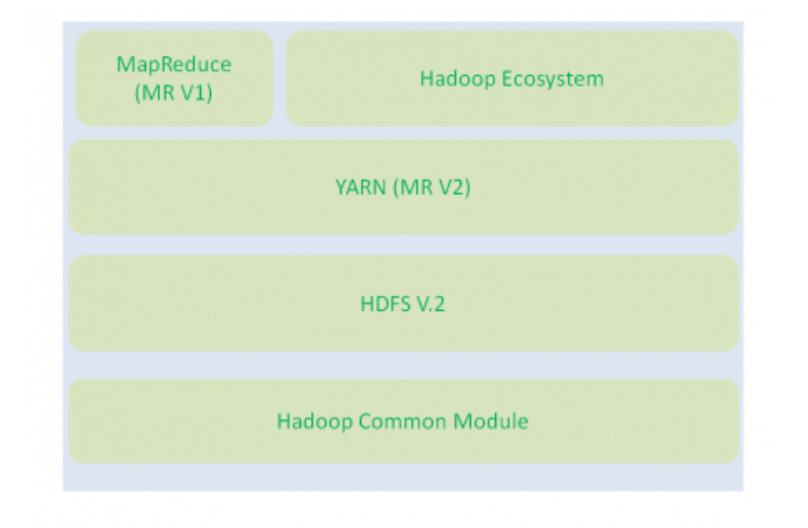
**TaskTracker (Not present in Hadoop 2.x)**

1. Runs individual map-reduce jobs on datanodes.
2. Periodically communicates with jobtracker to give updates and receive instructions



**Hadoop 2.x Architecture**

Apache Hadoop 2.x or later versions are using the following Hadoop Architecture. It is a Hadoop 2.x High-level Architecture



**Components of the Hadoop 2.x in depth**

• MapReduce, Yarn, HDFS, Pig, Hive, Hbase, and Oozie

* Hadoop Common Module is a Hadoop Base API (A Jar file) for all Hadoop Components. All other components works on top of this module.
* HDFS stands for Hadoop Distributed File System. It is also know as HDFS V2 as it is part of Hadoop 2.x with some enhanced features. It is used as a Distributed Storage System in Hadoop Architecture.
* YARN stands for Yet Another Resource Negotiator. It is new Component in Hadoop 2.x Architecture. It is also know as “MR V2”.
* Apache Hadoop YARN is the resource management and job scheduling technology in the open source [Hadoop](https://searchcloudcomputing.techtarget.com/definition/Hadoop) distributed processing framework. One of Apache Hadoop's core components, YARN is responsible for allocating system resources to the various applications running in a [Hadoop cluster](https://searchbusinessanalytics.techtarget.com/definition/Hadoop-cluster) and scheduling tasks to be executed on different cluster nodes.
* MapReduce is a Batch Processing or Distributed Data Processing Module. It is also know as “MR V1” as it is part of Hadoop 1.x with some updated features.
* Remaining all Hadoop Ecosystem components work on top of these three major components: HDFS, YARN and MapReduce.

**Apache Pig**

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

* **Ease of programming.** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
* **Optimization opportunities.** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
* **Extensibility.** Users can create their own functions to do special-purpose processing.

**Hive**

Apache Hive is a **data warehouse software project** built on top of Apache Hadoop for providing data summarization, query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop

**HBASE**

HBase - Is a **NoSQL database on top of HDFS** (hadoop distributed file system which is the file system and core component of Hadoop). Hive - Is an **ecosystem/component of Hadoop** which is similar to SQL to handle mostly structured data. Pig - Is an ecosystem/component of Hadoop which is a kind of scripting language.

**Oozie**

**Oozie** is a workflow scheduler system to manage Apache **Hadoop** jobs. **Oozie** Workflow jobs are Directed Acyclical Graphs ([DAGs](https://www.bing.com/search?q=Directed+acyclic+graph&filters=sid%3afb03b0f7-7fef-e71a-f72c-bf0e1eec409f&form=ENTLNK)) of actions. **Oozie** Coordinator jobs are recurrent **Oozie** Workflow jobs triggered by time (frequency) and data availability.