# 6.22学习日志

## ——1813079 那孜亚·吐尔逊

任务：了解Hadoop 、 安装Apache Hadoop，尝试运行Hadoop自带示例WordCount

完成情况：全部完成

感想：Hadoop主要是分布式计算和存储的框架，所以Hadoop工作过程主要依赖于HDFS(Hadoop Distributed File System)分布式存储系统和Mapreduce分布式计算框架。分布式存储系统HDFS中工作主要是一个主节点namenode(master)（hadoop1.x只要一个namenode节点，2.x中可以有多个节点）和若干个从节点Datanode(数据节点)相互配合进行工作,HDFS主要是存储Hadoop中的大量的数据，namenode节点主要负责的是：1、接收client用户的操作请求，这种用户主要指的是开发工程师的Java代码或者是命令客户端操作。2、维护文件系统的目录结构，主要就是大量数据的关系以及位置信息等。3、管理文件系统与block的关系，Hadoop中大量的数据为了方便存储和管理主要是以block块（64M）的形式储存。一个文件被分成大量的block块存储之后，block块之间都是有顺序关系的，这个文件与block之间的关系以及block属于哪个datanode都是有namenode来管理。Datanode的主要职责是：1、存储文件。2、将数据分成大量的block块。3、为保证数据的安全，对数据进行备份，一般备份3份。当其中的一份出现问题时，将由其他的备份来对数据进行恢复。MapReduce主要也是一个主节点JOPtracker和testtracker组成，主要是负责hadoop中的数据处理过程中的计算问题。joptracker主要负责接收客户端传来的任务，并且把计算任务交给很多testtracker工作，同时joptracker会不断的监控testtracker的执行情况。testtracker主要是执行joptracker交给它的任务具体计算，例如给求大量数据的最大值，每个testtracker会计算出自己负责的数据中的最大值，然后交给joptracker。Hadoop的主要两个框架组合成了分布式的存储和计算，使得hadoop可以很快的处理大量的数据。