学习日志

姓名：辛昊洋　　学号：1813090　　日期：6.23

学习内容安排：

了解 MapReduce 以及 Hadoop 分布式文件系统（HDFS）

学习反馈：

已掌握知识：

MapReduce 是 Google 提出的一个软件架构，用于大规模数据集（大于 1TB）的并行运算。概念"Map（映射）"和"Reduce（化简）"，和他们的主要思想，都是从函数式编程语言借来的，还有从矢量编程语言借来的特性。

当前的软件实现是指定一个 Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce（化简）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

√hadoop 的四大组件：

HDFS：分布式存储系统

MapReduce：分布式计算系统

YARN： hadoop 的资源调度系统

Common： 以上三大组件的底层支撑组件，主要提供基础工具包和 RPC 框架等

Mapreduce 是一个分布式运算程序的编程框架，是用户开发“基于 hadoop 的数据分析 应用”的核心框架

Mapreduce 核心功能是将用户编写的业务逻辑代码和自带默认组件整合成一个完整的 分布式运算程序，并发运行在一个 hadoop 集群上

为什么需要 MapReduce？

(1) 海量数据在单机上处理因为硬件资源限制，无法胜任

(2) 而一旦将单机版程序扩展到集群来分布式运行，将极大增加程序的复杂度和开发难度

(3) 引入 MapReduce 框架后，开发人员可以将绝大部分工作集中在业务逻辑的开发上，而将 分布式计算中的复杂性交由框架来处理

HDFS，是 Hadoop Distributed File System 的简称，是 Hadoop 抽象文件系统的一种实现。Hadoop 抽象文件系统可以与本地系统、Amazon S3 等集成，甚至可以通过 Web 协议（webhsfs）来操作。HDFS 的文件分布在集群机器上，同时提供副本进行容错及可靠性保证。例如客户端写入读取文件的直接操作都是分布在集群各个机器上的，没有单点性能压力。

未掌握知识：

现在只是在理论层面上对 Hadoop 有了一定的了解，实操欠缺。

学习心得：
Hadoop 是非常功能强大的集群，还需要进一步的学习与掌握。