学习日志

姓名：辛昊洋　　学号：1813090　　日期：6.22

学习内容安排：（以 6.15 为例）

上午：
了解 Hadoop 、 安装 Apache Hadoop
下午：
尝试运行 Hadoop 自带示例 WordCoun

学习反馈：

已掌握知识：

Hadoop就是存储海量数据和分析海量数据的工具。

Hadoop是由java语言编写的，在分布式服务器集群上存储海量数据并运行分布式分析应用的开源框架，其核心部件是HDFS与MapReduce。

HDFS是一个分布式文件系统：引入存放文件元数据信息的服务器Namenode和实际存放数据的服务器Datanode，对数据进行分布式储存和读取。

MapReduce是一个计算框架：MapReduce的核心思想是把计算任务分配给集群内的服务器里执行。通过对计算任务的拆分（Map计算/Reduce计算）再根据任务调度器（JobTracker）对任务进行分布式计算。

HDFS为海量的数据提供了存储，则MapReduce为海量的数据提供了计算。把HDFS理解为一个分布式的，有冗余备份的，可以动态扩展的用来存储大规模数据的大硬盘。把MapReduce理解成为一个计算引擎，按照MapReduce的规则编写Map计算/Reduce计算的程序，可以完成计算任务。

大数据存储：分布式存储

日志处理：擅长日志分析

ETL:数据抽取到 oracle、mysql、DB2、mongdb 及主流数据库

机器学习：比如 Apache Mahout 项目

搜索引擎:Hadoop + lucene 实现

数据挖掘：目前比较流行的广告推荐，个性化广告推荐

Hadoop 是专为离线和大规模数据分析而设计的，并不适合那种对几个记录随机读写的在线事务处理模式。

未掌握知识：

在安装 Hadoop 的过程中，出现了很多情况，在网上没有找到相应解答。

学习心得：

在安装过程中出现了重重问题，而且也浪费了很多时间。