# 6.25学习日志

## ——1813079 那孜亚·吐尔逊

任务：MapReduce工作原理、类型与格式

完成情况： 全部完成。

感想：MapReduce处理数据的最佳速度最好与数据在集群中的传输速度相同，而处理小文件将增加运行作业而必需的寻址次数。在HDFS中集群中存储大量的小文件会浪费namenode的内存。一个可以减小大量小文件的方法是使用顺序文件将这些小文件合并成一个或多个大文件，可以将文件名作为键，文件名作为值，如果HDFS中已经有大量的小文件，CombineFileInputFormat方法比较适合(将多个小文件打包到一个分片中以便每个mapper可以处理更多的数据)