# 实验四 云-边协同计算实验

实验目的与要求

1. 了解MapReduce计算模型的原理；
2. 学会使用编程语言和工具实现MapReduce的Map和Reduce功能，并行实现特定任务的高效计算；

实验内容

1. 阅读经典论文“MapReduce: Simplified data processing on large clusters，OSDI 2004”，学习掌握MapReduce计算框架的基本原理；
2. 以熟悉的编程语言（如Java、Python等）编程实现MapReduce功能，进而完成对给定文本文件的词频统计。针对给定文件（如A.txt），统计里面文章中单词的出现频次并降序输出统计结果到文件Sta\_A.txt。Sta\_A.txt文件包含A.txt文件里面每个单词及其出现的频次。
   1. 数据集包含两个文件，1个小文件约20 MB左右，1个大文件约1.5 GB左右。
3. **探索：**除了编程实现MapReduce之外，也可以在电脑按照本地的Apache Spark计算框架（下载链接：<https://spark.apache.org/downloads.html>，Spark安装参考链接：<https://dblab.xmu.edu.cn/blog/4322/> ），并利用Spark来实现词频统计任务。

实验要求

1. 指定文件夹里面每个文件的单词频次都得统计，并保存到对应输出文件。输出文件中单词按频次降序排列。

2. 统计程序运行时间，包括总体运行时间和各个文件处理所花时间。

3. 改进优化程序，使其运行效率提高，并分析程序运行慢和MapReduce实现的困难。