

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级：**

**学 号：**

**姓 名：**

**指导教师：**

**报告日期：**

**计算机科学与技术学院**

**目录**

[实验一 wordCount算法及其实现 1](#_Toc89439814)

[**1.1实验目的** 1](#_Toc89439815)

[**1.2 实验内容** 1](#_Toc89439816)

[**1.3 实验过程** 1](#_Toc89439817)

[1.3.1 编程思路 1](#_Toc89439818)

[1.3.2 遇到的问题及解决方式 1](#_Toc89439819)

[1.3.3 实验测试与结果分析 1](#_Toc89439820)

[**1.4 实验总结** 1](#_Toc89439821)

# 实验一 wordCount算法及其实现

## **1.1实验目的**

1、理解map-reduce算法思想与流程；

2、应用map-reduce思想解决wordCount问题；

3、（可选）掌握并应用combine与shuffle过程。

## **1.2 实验内容**

提供9个预处理过的源文件（source01-09）模拟9个分布式节点，每个源文件中包含一百万个由英文、数字和字符（不包括逗号）构成的单词，单词由逗号与换行符分割。

要求应用map-reduce思想，模拟9个map节点与3个reduce节点实现wordCount功能，输出对应的map文件和最终的reduce结果文件。由于源文件较大，要求使用多线程来模拟分布式节点。

学有余力的同学可以在map-reduce的基础上添加combine与shuffle过程，并可以计算线程运行时间来考察这些过程对算法整体的影响。

提示：实现shuffle过程时应保证每个reduce节点的工作量尽量相当，来减少整体运行时间。

## **1.3 实验过程**

### 1.3.1 编程思路

### 1.3.2 遇到的问题及解决方式

### 1.3.3 实验测试与结果分析

## **1.4 实验总结**