\

**<<Hadoop-MapReduce>>**

**软件开发计划书**



项目成员

SY1506107 陈昆度

SY1506112 马宇晴

SY1506105 杨云龙

SY1506423 李嘉艺

北京航空航天大学

2016-3-16

目录

[1引言 1](#_Toc446272314)

[1.1编写目的 1](#_Toc446272315)

[1.2 背景 1](#_Toc446272316)

[1.3特点 1](#_Toc446272317)

[1.4本项目计划解决问题 2](#_Toc446272318)

[1.5参考资料 2](#_Toc446272319)

[2 项目概述 3](#_Toc446272320)

[2.1工作内容 3](#_Toc446272321)

[2.2主要参加人员 3](#_Toc446272322)

[2.3产品 3](#_Toc446272323)

[2.3.1程序 3](#_Toc446272324)

[2.3.2文件 3](#_Toc446272325)

[2.4验收标准 4](#_Toc446272326)

[2.5完成项目的最迟期限 4](#_Toc446272327)

[3实施计划 4](#_Toc446272328)

[3.1 组织形式 4](#_Toc446272329)

[3.2任务分解与计划进度 4](#_Toc446272330)

[3.2.1 第一周：软件需求分析 4](#_Toc446272331)

[3.2.2 第二-三周：软件需求评审 4](#_Toc446272332)

[3.2.3第四周：软件进度计划/工作量估计/配置管理 5](#_Toc446272333)

[3.2.4 第五-七周：软件产品改进与展示 5](#_Toc446272334)

[3.2.5 第八-十二周：软件测试 5](#_Toc446272335)

[3.2.6第十三周：综合实验分析 6](#_Toc446272336)

[3.2.7第十四周：综合实验总结 6](#_Toc446272337)

[4支持条件 7](#_Toc446272338)

[4.1计算机系统支持 7](#_Toc446272339)

**软件开发计划书**

# 1引言

## 1.1编写目的

本次项目，我们组主要解决MapReduce计算模型对于数据的处理问题。我们计划通过对该模型详细的研究，重点学习java环境下该模型的类的机制，编程实现Map和Reduce函数，用来处理大规模的数据，同时在Hadoop下，使用该模型对数据进行并行处理。在研究过程中，针对当下对于数据处理的新需求，做出MapReduce的需求分析和需求报告，对map和reduce函数及机制进行改进，实现功能改进和扩展。

## 1.2 背景

MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算。"Map（映射）"和"Reduce（归约）"，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。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（归约）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。

## 1.3特点

1. **并行性**

通过Map-Reduce的框架，能够充分的利用廉价的计算机集群，共同工作来提高对数据的处理速度和质量，完成分布式的并行处理。

1. **处理大数据**

通过对大量数据进行分块的处理，并且用HDFS进行合理的存储和分布，就可以在Map-Reduce模型下进行快速稳定的数据处理。

1. **非结构化数据**

通过对数据进行键值的保存和分析，能够很好的对非二维的数据进行快速查询和处理。

## 1.4本项目计划解决问题

目前的MaoReduce会存在如下问题，我们将通过阅读文献，研究MapRecuce的机制，通过编程，尽量解决这些问题。

1. MapReduce运行速度较慢，map和reduce之间的信息交互是造成计算慢的主要原因。

待定问题：

2. 扩展性差：MapReduce中，jobTracker兼备资源管理和作业控制两个功能，这是系统最大瓶颈，制约了Hadoop集群扩展性。

3. 可靠性差：MapReduce采用master/slave结构，其中master存在单点故障问题，一旦故障，将使集群不可用。

4. 资源利用率低：MapReduce采用基于槽位的资源分配模型，一个任务不会用完槽位对应的资源，其他任务也无法使用这些资源。

5. 无法支持多种计算框架：随着互联网高速发展，MapReduce这种基于磁盘的离线集散框架已经不能满足应用需求，MapReduce不能支持多种计算框架并存。

解决方案：

我们将尽量结合硬软件进行改进，尝试解决这些问题。

## 1.5参考资料

《Hadoop权威指南》

《Hadoop中文文档》

# 2 项目概述

## 2.1工作内容

1.部署Hadoop

2.研究调查MapReduce的工作机制，完成MapReduce需求文档

3.在java下实现一个map和reduce函数，处理数据

4.对于新需求的研究，制定改进map和reduce机制方案。

5.更新需求文档，改进Map和reduce函数。

## 2.2主要参加人员

SY1506107 陈昆度

SY1506112 马宇晴

SY1506105 杨云龙

SY1506423 李嘉艺

## 2.3产品

### 2.3.1程序

MapReduce代码。

### 2.3.2文件

1. 项目计划书。

2. 需求文档。

3. 测试文档。

4. 项目总结报告。

## 2.4验收标准

MapReduce代码、Hadoop部署、实验所需的相关文档和测试用例等等。

## 2.5完成项目的最迟期限

本学期第十七周。

# 3实施计划

## 3.1 组织形式

整个项目推进过程中的工作方法：敏捷开发，采用迭代、循序渐进的方法进行软件开发。通过通讯软件及时沟通，推进项目的进度，完善解决方案。

## 3.2任务分解与计划进度

### 3.2.1 第一周：软件需求分析

1软件及相关资料收集-李嘉艺

2获取并说明软件需求-陈昆度

3需求验证-马宇晴

4需求修改和完善-杨云龙

### 3.2.2 第二-三周：软件需求评审

**评审**

1评审方式指定-陈昆度

2参与评审-所有人

3评审报告-陈昆度，杨云龙

4问题报告-马宇晴，李嘉艺

**复评审**

1评审方式指定-陈昆度

2参与复评审-所有人

3复评审报告-陈昆度，杨云龙

4复问题报告-马宇晴，李嘉艺

### 3.2.3第四周：软件进度计划/工作量估计/配置管理

1变更与管理分析报告-马宇晴

2工作量分析报告-杨云龙

3工作日志、进度控制分析报告-陈昆度

4整理和改进，以及跟踪-李嘉艺

### 3.2.4 第五-七周：软件产品改进与展示

**方案设计：1周**

1产品对比-陈昆度

2产品准备-李嘉艺

3产品研究-杨云龙

4产品修改和矫正-马宇晴

**进展和改进：2周**

1产品进一步对比-马宇晴

2产品修改-杨云龙

3产品扩展-陈昆度

4产品应用-李嘉艺

### 3.2.5 第八-十二周：软件测试

**软件测试分析：1周**

**被测软件执行结果分析报告：**

1测试需求定义和测试用例设计-陈昆度

2测试需求和测试用例评审-杨云龙

**软件问题报告：**

3选择并学习使用测试工具，生成并执行测试脚本-李嘉艺

4充分性分析与增强测试-马宇晴

**软件测试评审：1周**

1参与评审-所有人

2测试需求和测试用例评审/软件问题报告-陈昆度，杨云龙

3测试结果分析与评审/软件评审报告-李嘉艺，马宇晴

**软件测试复评审：1周**

1参与评审-所有人

2测试需求和测试用例评审/软件复问题报告-陈昆度，杨云龙

3测试结果分析与评审/软件复评审报告-李嘉艺，马宇晴

**软件演示与测评：1周**

1软件演示-马宇晴

2软件讲解-陈昆度

3文稿ppt准备-李嘉艺

4软件测评-杨云龙

### 3.2.6第十三周：综合实验分析

1软件功能需求分析-李嘉艺，马宇晴

2文稿撰写-陈昆度，杨云龙

### 3.2.7第十四周：综合实验总结

1问题报告：李嘉艺

2需求规格说明书：马宇晴

3改进计划：杨云龙

4总结报告：陈昆度

# 4支持条件

## 4.1计算机系统支持

Linux/Windows7,java,Hadoop