

**<<Hadoop-MapReduce>>**

**项目实施方案及实施情况**



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.项目介绍 3](#_Toc448953677)

[1．1项目背景 3](#_Toc448953678)

[1．2项目目标 3](#_Toc448953679)

[2.实施方案 3](#_Toc448953680)

[2．1实施步骤 3](#_Toc448953681)

[2．2实施详情 4](#_Toc448953682)

[3.工作展望 5](#_Toc448953683)

# 1.项目介绍

## 1．1项目背景

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。

## 1．2项目目标

我们的项目将试图通过在Hadoop环境下对MapReduce计算方式（Yarn）进行改进。本项目将会收集并且使用MapReduce模型处理一组比较庞大的半结构化数据集，这组数据是由大量笔记本电脑的性能参数组成的。我们的目标是从数据中抽取几个关键的参数，例如cpu，内存和显存等，之后生成一个排名，并且抽取位于前列的数据。在完成了这个目标之后，我们将会改进MapReduce函数和工作函数，来优化产品的计算速度和准确性。

# 2.实施方案

## 2．1实施步骤

我们的项目计划分为4个阶段进行实施。

1. 资料收集阶段

在该阶段中，我们将收集Hadoop和MapReduce有关的各类参考文档，了解Hadoop的运行机理，熟悉Linux环境以及其命令，了解Yarn的结构。同时比较深入的了解MapReduce机理，为改进做准备。

落实情况：本阶段的有关参考文档已经收集完毕，Hadoop运行机制和Yarn的结构也已经熟悉，Linux环境下的相关命令已经熟练应用。

1. 环境配置阶段

我们将在这个阶段布置好Linux环境，同时完成Hadoop V1.x和Hadoop V2.x的配置。同时安装编程环境。这里我们选择的开发环境是带有Hadoop插件的Eclipse Mars。

落实情况：Linux环境已经给配置好，Hadoop V1.x和Hadoop V2.x的配置完成，Eclipse Mars安装配置完成。

1. 函数编写阶段

这个阶段主要是实际的编写数据处理的函数，其中包括数据爬虫文件，数据的预处理文件，MapReduce函数，以及最后的可视化函数。

落实情况：数据爬虫文件，已经用python完成，数据预处理文件也已经用java完成，MapReduce的测试函数已经写完，实际数据处理函数还在进行，可视化展示还没有进行。

1. 程序改进阶段

我们将根据运行结果对MapReduce函数进行修改，通过运行时间等参数判断最后的改进是否成功。

落实情况：该部分的工作还未开始。

## 2．2实施详情

###### 2.2.1资料收集阶段

我们查找了若干Hadoop技术文档作为项目的参考书，其中包括《深入Hadoop-MapReduce的原理和解析》，《深入Yarn-MapReduce的原理和解析》，《Hadoop权威指南》，《Yarn技术手册》。

经过资料查阅，我们确定MapReduce的计算速度是可以提高的，通过对数据进行预处理和减少数据在网络上传递的时间，计算速度可以有效的提升。

同时我们确定了从中关村在线获取我们的基础数据。

###### 2.2.2环境配置阶段

我们成功的配置了Hadoop的运行环境。

# 3.工作展望

在接下来一周半时间里，我们需要完成MapReduce函数对真实数据的处理，目前的数据处理程序还未完善，程序还存在一些bug需要调试，我们就将会在接下来时间里进行bug的修改。数据处理结束后，由于数据哭的效率问题，我们会以文件的形式保存处理后的数据，将这些进行可视化展示，对于数据的运行时间进行MapReduce函数的修改。