

**<<Hadoop-MapReduce>>**

**项目实施方案修改**



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

# 1.原始计划

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.计划变更

在计划的实际进行中，我们发现有诸多不可行或者存在疑问的地方，所以我们对之前的实施方案进行了比较大幅度的修改。其具体的情况如下所示。

1. 资料收集

我们最早收集的原始数据存在比较多的问题，数据过于非格式化，难以处理，同时缺少很多参数值。所以我们对原始数据进行了改进，使其更加容易被程序阅读。同时增加了价格参数用于数据比对。

1. 环境配置

由于考虑到了现在Hadoop的更新，所以我们同时完成Hadoop V1.x和Hadoop V2.x的配置。

1. 数据的展示

由于可视化的目的在于方便MapReduce的性能改进，所以我们决定不用网站呈现，直接使用计算得到的数据表明。。

1. 排名

由于我们之前制定的排名没有依据，所以我们计划加入价格作为比较标准，通过参数计算出计算机的性价比，然后排序。

1. MapReduce函数

目前由于进度速度不够，我们还在考虑除了修改MapReducer之外，是否还要进行源码修改