南 京 大 学

本科生毕业论文（设计、作品）指导情况记录

开题简况

题目：MPI/Hadoop在K-means聚类分析算法上的性能比较分析

1、选题质量（简述选题与专业培养目标、专业要求关系、题目难度、工作量、创新性、理论性、实用性）

基于学生对未来职业的选择和个人发展的考量批准此题。大数据分析方法过去一直用于地球科学领域（数字地球），如今更是成为全社会的热门话题。对此领域的重要算法-聚类分析在云同平台上进行性能分析评价，可以成为大数据分析平台发展的重要参考。中上难度，需要系统构建、并行程度设计等多方面的综合知识。

2、开题意见：

同意开题。重点放在海量数据、高精度分析的性能比较上。

指导教师签名：周会平

2015年 1 月 17 日

中期检查

指导教师检查论文的进展情况：（指导和培养学生查阅文献资料、综合运用知识、研究方案设计、研究方法和手段运用和外文应用等能力简况）

已完成文南大调研工作。必要时启补充外文文献的阅读。已完成计算集群的集成（主控服务器1加计算服务器4），并已完成基础软件的安装和调试。正在两种平台上，聚类算法的并行程序设计。注意算法的边设计边调优。

指导教师签名：周会平

2015年 4 月 5 日

南 京 大 学

本科生毕业论文（设计、作品）指导教师评阅意见

指导教师评语： 大数据分析的理论基础是概率论中的中心极限定理，方法则归属于多元统计分析。过去以及现在大量应用于科学研究领域，其目的是探明随机现象，即不确定性现象的统计规律。近年来，大数据分析的概念被过度地高出包装，相当多的分析平台偏离了正确的发展方向。蒋鑫同学的论文以高性能计算平台为基础，利用过去几十年科学计算领域积累下来的经验、技术和软件资产，与目前非常流行的大数据分析平台 Hadoop 作了比较研究。通过搭建两种不同分析平台，对机器学习中最常用的 K-means 算法，利用相同的数据，进行了细致的性质评估。以显见的事实说明大数据分析的未来发展方向就是高性能计算。论文立意正确、分析合理、实验设计周到，结论有重要意义。是一篇优秀的大学毕业论文。

指导教师签名： 周志成

2015 年 6 月 12 日

南 京 大 学

本科生毕业论文（设计、作品）评阅教师评阅意见

评阅教师评语：

蒋鑫同学的毕业论文“MPI/Hadoop 在 K-means 算法上的性能比较分析”是对地球科学中的海量计算应用研究与计算机科学中流行的并行计算、分布式计算的一次结合。选题十分新颖，是地球科学的前言领域，应用性较强。该生在毕业论文完成期间，态度端正，仔细认真，能够独立设计出合理实验并进行了相关项目的测试，表现出了较强的专业素养和创新能力。从最终的论文中可以看出，该生查阅了国内外的大量相关文献，具备了一定的文献综述和资料整理能力。同时论文内容比较完整，思路清晰，观点突出，逻辑性强，在论文最后对所得到的结果进行了比较详尽的分析解释，并得出了令人信服的结论。表明该生具备了一定的独立工作能力，达到了南京大学地科院本科培养的目标。

评阅教师签名： 刘国东

2015 年 6 月 1 日

南 京 大 学

本科生毕业论文（设计、作品）答辩记录、成绩评定

答辩记录：

1. 进行测试的软件是哪一款？

使用 Hadoop-2.6 和 openmpi-1.8.4 进行试验，实验数据使用随机算法进行生成。运行环境是 CentOS6.5。

2.测试结果在地质学模拟上的应用前景和潜力。

从测试结果来开，对于地学中常见的复杂迭代性计算，使用 MPI 相对于 Hadoop 有着比较大的性能优势，所以对于今后的地质学模拟中的海量计算可以更多的考虑基于 MPI 的高性能计算模型，例如论文中所提的油气储藏量预测，高精度地震资料处理等。

答辩记录人签名： 黄晓路

答辩小组评语：

该生的研究方向比较新颖，希望对地学研究和计算机科学中的并行计算进行结合，通过答辩可以看出该生进行了比较深入的研究，并取得了一定的实验成果。对于答辩老师提出的问题，能够比较合理的解释说明，思路比较清晰，有理有据，达到了我院对本科毕业论文答辩的要求。

答辩小组成员： [illegible] 董少春 [illegible]

成绩 87

组长签名： [illegible]

答辩时间： 2015年 6月 8日