**北京工业大学**

毕业设计（论文）任务书

题目 一种基于统计推断的交互式大数据探索系统

专业 计算机科学与技术 学号 12070001 姓名 柯伟辰

主要内容、基本要求、主要参考资料等：

主要内容及基本要求：

1. 查询资料，对Hadoop HDFS分布式文件存储系统及Spark大数据分析框架建立初步的认知；
2. 设计出统计推断模块的核心算法；
3. 分模块设计并实现整个分析系统；
4. 对算法进行准确度分析，对系统进行性能分析，并作适当的改进；

主要参考资料

1. 王松桂，张忠占，程维虎. 概率论与数理统计（第三版）. 科学出版社, 2013.
2. Rui ding, Qiang Wang, Yingnong Dang, Qiang Fu, Haidong Zhang, Dongmei Zhang. YADING: Fast Clustering of Large-Scale Time Series Data. In proc of VLDB, 2015
3. Apache Spark: Lightning-Fast Cluster Computing, <http://spark.apache.org/>
4. Apache Hadoop Distributed File System.

<http://hadoop.apache.org/hdfs/>

1. S. Acharya, P. B. Gibbons, and V. Poosala. Congressional samples for

approximate answering of group-by queries. In ACM SIGMOD, May 2000.

完成期限：

指导教师签章：

专业负责人签章：

2015年11 月27日