# 实验四 并行化数据挖掘算法设计

## 一、实验目的

机器学习和数据挖掘算法是大数据分析处理领域的重要内容，随着数据规模的不断扩大，设计面向大数据处理的并行化机器学习和数据挖掘算法越来越有必要。通过对并行化数据挖掘算法的实现，掌握并行化处理问题的分析方法和编程思想方法，能够根据实际情况定制并行化的算法解决问题。

## 二、实验平台

1）操作系统：Linux（实验室版本为 Ubuntu17.04）；

2）Hadoop 版本：2.9.0；

3）JDK 版本：1.8；

4）Java IDE：Eclipse 3.8；

5 ）Spark 版本：2.1.0。

## 三、实验内容

自行准备数据集，设计两种数据挖掘算法（聚类、分类、频繁项集挖掘或其他主题）对数据集进行信息提取，要求分别使用并行化和非并行化的方式实现该算法。实验环境可选择 Hadoop 或者 Spark，程序语言可选用 Java、Python、Scala 等，在伪分布式环境下完成并行化算法的编写和测试，在单机环境下完成非并行化算法的编写和测试。

## 四、实验要求

自行对比并行化和非并行化实现方法的数据挖掘结果，两种结果需完全一致。

## 五、实验报告

## 计算机科学与技术学院 大数据管理与分析 课程实报告

| | | |
|---|---|---|
| 实验题目： | | 学号：201500000000 |
| 日期：2018.3.20 | 班级：2015 级 1 班/菁英班 | 姓名：张三 |
| Email：zhangsan@qq.com | | |
| 实验目的： | | |
| 实验软件和硬件环境： | | |

| |
|---|
| 实验原理和方法： |
| 实验步骤：（不要求罗列完整源代码） |
| 结论分析与体会： |
| 就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1－3 道问答题： |