计算机科学与技术学院 大数据管理与分析 课程实验报告

实验题目：并行化数据挖掘算法设计		学号：201605130116
日期：2019.5.15	班级：2016 级泰山学堂	姓名：杜洪超
Email：1503345074@qq.com		

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。

实验软件和硬件环境：

软件环境：

系统：Ubuntu16.04 LTS 64 位

软件：openjdk-7-jre,openjdk-7-jdk,java1.7.0_95

Python 2.7.12

Hadoop 2.9.2

Eclipse

硬件环境：

CPU: Intel® Core™ i5-6260U CPU @ 1.80GHz × 4

磁盘：121.8 GB

内存：7.7 GiB

实验原理和方法：

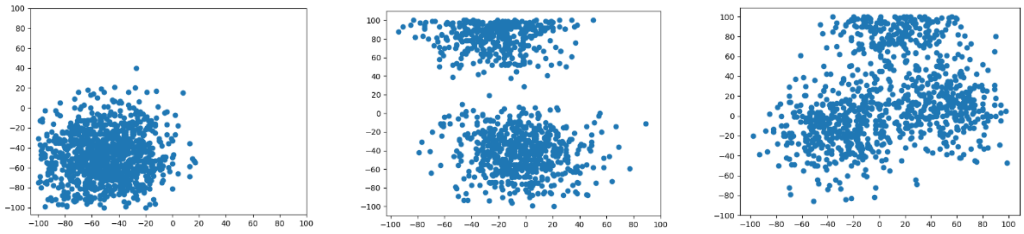
通过分别用非并行和并行方法实现 K-means 聚类算法，比较两种实现；其中非并行实现是用 python 语言编程，并行实现是使用 Java 语言实现的 MapReduce 编程。

实验步骤：（不要求罗列完整源代码）

1. 非并行 python 实现

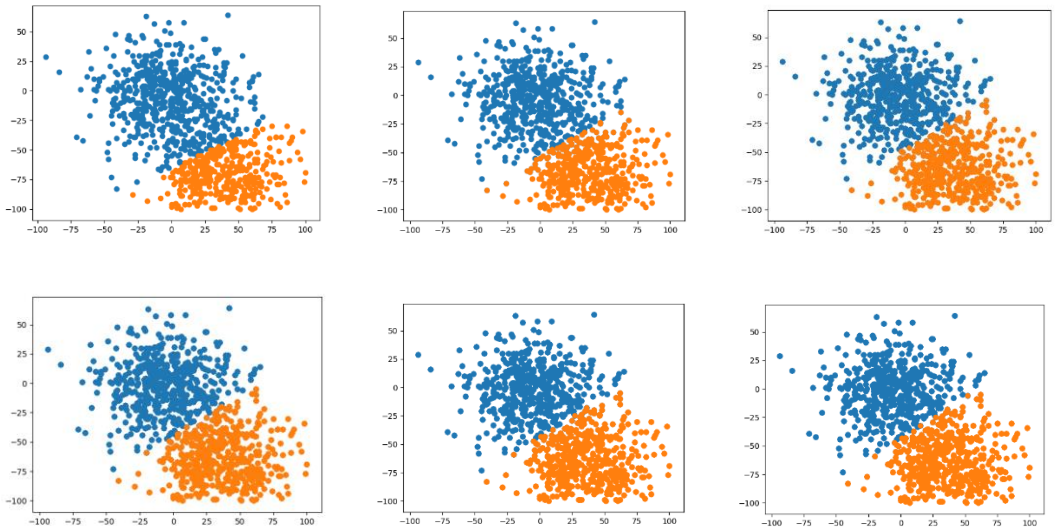
1.1. 用 python 构造随机数据

随机数据是一个点集，点的 x,y 坐标分布在（-100，100）之间，实验生成了 1000 个点，其中可选参数 K_1，代表初始数据类数，数据生成前先生成 K_1 个中心点，所有的随机数据都围绕其中一个中心点呈正态分布。下图分别是 K_1=1,2,3 时生成的数据：

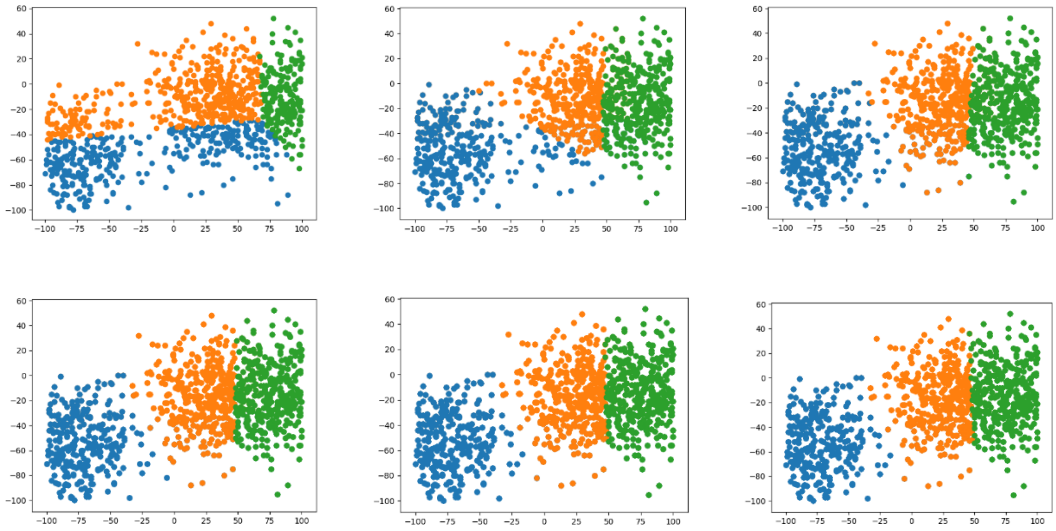


1.2. Python 实现 K-Means 算法

输入类数 K_2,算法把前 K_2 个数据作为初始中心点，因为数据是随机生成的，所以前 K_2 个数据也相当于随机选取 K_2 个点，重复执行 K-Means 算法，直到每一类的点数固定。下图是对于一个 K_1=K_2=2,迭代了七次结束的情况，第七次相比第六次没有变化，没有列出；初始中心坐标为[[-6, -2], [37, -66]]，算法生成的中心为[[-7, -1], [36, -63]]：



下图是 K1=K2=3 的情况，初始中心为[[86, -13], [31, -17], [-79, -58]]，算法生成的中心为[[74.7, -14.86], [22.26, -16.42], [-70.42, -53.35]]：



2. 并行 Java MapReduce 实现

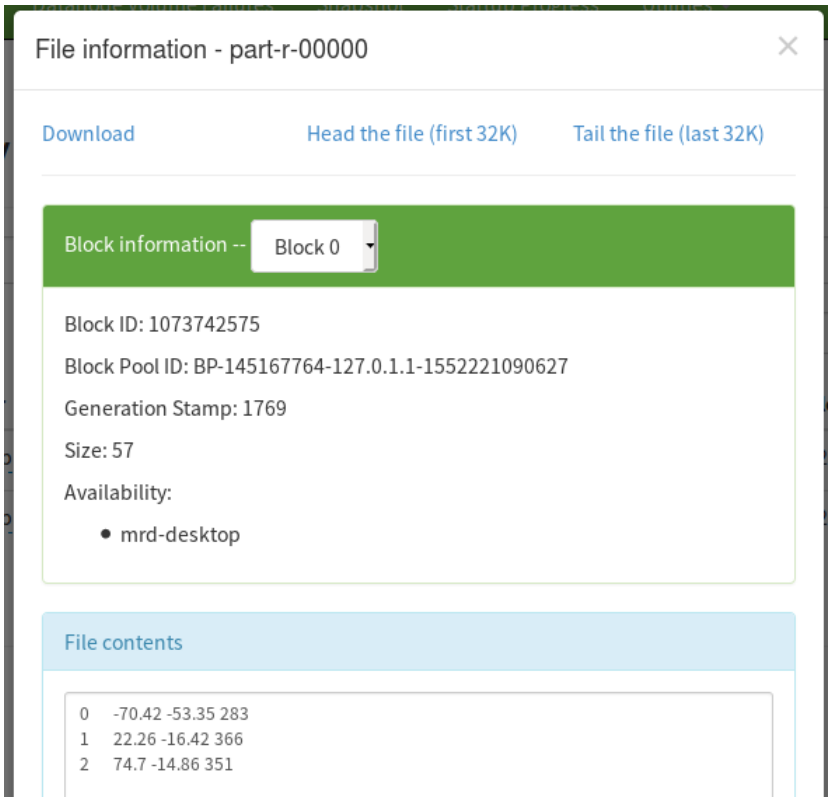
2.1. MapRuduce 实现

在初始化中读入前 K_2 个点作为全局的中心变量

Map 函数对于每一个点，计算这个点到每个中心的距离，输出 key 为此中心的索引，value 为点加上权值 1；

Combiner 和 Reduce 使用同一个类，对于输入的 key，把所有的 value 中的点集按照权值加权平均，求出此次迭代新的中心；

2.2. 在 K_1=K_2=3 的情况下，MapReduce 求出的结果和非并行化的中心相同：



但因为数据集不够大，导致运行效率上差别不大。

结论分析与体会：

通过对并行化数据挖掘算法 K-Means 的实现，掌握了并行化处理问题的分析方法和编程思想，学会了根据实际情况定制并行化的算法解决问题。

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：

1. 生成的随机数据聚类不明显

先生成几个中心点，再利用 python 库生成围绕中心点成正态分布的数据