# 课程作业

说明：请从下面两个题目中选择1个，作为课程作业，在5月20日之前完成提交。

1. 生成1万个随机数，存储在HDFS文件系统中的data1.txt中，每个数字之间用“,”分隔。在Spark平台上实现如下的功能：
2. 从HDFS上读入data1.txt，生成RDD；
3. 找到这个数据集合的**中位数**（精确的）；
4. 只能使用RDD有关API，并且不能调用Spark提供的中位数计算的API；
5. 在Spark平台上实现中位数算法后，验证结果的正确性。
6. 生成1000个随机数，格式为KV格式，即（序号，浮点数）形式的二元组，其中序号为连续的整数（1~1000）。数据存储在HDFS文件系统中的data2.txt中。在Spark平台上实现如下的功能：
7. 从HDFS上读入data2.txt，生成RDD或者DataFrame；
8. 计算上述序列中浮点数值属性的5阶**移动平均数**。即按照序号顺序，从第5个二元组开始，依次计算第1个~第5个二元组浮点数的平均数，然后计算第2个~第6个二元组浮点数的平均数，......，直到第996个~第1000个二元组浮点数的平均数；
9. 可以使用RDD或者DataFrame的API；
10. 将计算得到的移动平均数序列按照序号顺序写到文件：MovingAverage.txt中。

## 作业提交要求：

源程序+算法说明文档+原数据文件+结果数据文件。