****

云计算技术及应用

课程实践报告

|  |  |
| --- | --- |
| 标 题 | Spark与大数据处理技术实践 |
| 学 号 |  |
| 姓 名 |  |

东南大学计算机科学与工程学院

二零二二年五月

## 实验环境

1. 虚拟机操作系统：
2. 虚拟机硬件配置：CPU、内存、磁盘
3. 虚拟机主机名/IP：

## Spark配置环境截图

1. Spark若干关键配置文件的截图（Java, Scala, Python）
2. Spark成功启动的截图（./spark-shell）

## GMM执行结果截图

1. 任务提交以后，Spark执行过程的截图
2. 任务完成后，打印结果的截图

## 思考（注：画图说明，并分析原因）

1. Spark Local模式和Standalone模式执行同一任务时间比较（画图或列表）
2. 理想情况下N台机器的执行时间为单机的1/N倍。对照自己的实验结果，并分析原因。
3. 根据前150个数据的预测结果推断聚类的表现（将预测结果与原始数据标签作对比），并尝试分析造成这种表现的原因。