浙江水利水电学院

实验报告

（ 2019-2020 学年 第一 学期）

|  |  |
| --- | --- |
| 课程名称： | 大数据分析与应用 |
| 班 级： | 软工S18-1 |
| 学 号： | 2018b31017 |
| 姓 名： | 阮杭建 |
| 实验室(中心) 名 称： | 软件实施分室 |
| 教学单位： 信艺学院 | |

2019年 10 月 19 日

实 验 名 称：RDD编程初级实践 指导教师：张运涛

实 验 日 期：2019.10.19 地 点：现南401

同组学生姓名：

|  |
| --- |
| 实验内容及要求： 1．spark-shell交互式编程 请到到ftp上下载“数据集”chapter5-data1.txt，该数据集包含了某大学计算机系的成绩，数据格式如下所示：  Tom,DataBase,80  Tom,Algorithm,50  Tom,DataStructure,60  Jim,DataBase,90  Jim,Algorithm,60  Jim,DataStructure,80  ……  请根据给定的实验数据，在spark-shell中通过编程来计算以下内容：  （1）该系总共有多少学生；  （2）该系共开设来多少门课程；  （3）Tom同学的总成绩平均分是多少；  （4）求每名同学的选修的课程门数；  （5）该系DataBase课程共有多少人选修；  （6）各门课程的平均分是多少；  （7）使用累加器计算共有多少人选了DataBase这门课。 2.编写独立应用程序实现数据去重 对于两个输入文件A和B，编写Spark独立应用程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新文件C。下面是输入文件和输出文件的一个样例，供参考。  输入文件A的样例如下：  20170101 x  20170102 y  20170103 x  20170104 y  20170105 z  20170106 z  输入文件B的样例如下：  20170101 y  20170102 y  20170103 x  20170104 z  20170105 y  根据输入的文件A和B合并得到的输出文件C的样例如下：  20170101 x  20170101 y  20170102 y  20170103 x  20170104 y  20170104 z  20170105 y  20170105 z  20170106 z 3.编写独立应用程序实现求平均值问题 每个输入文件表示班级学生某个学科的成绩，每行内容由两个字段组成，第一个是学生名字，第二个是学生的成绩；编写Spark独立应用程序求出所有学生的平均成绩，并输出到一个新文件中。下面是输入文件和输出文件的一个样例，供参考。  Algorithm成绩：  小明 92  小红 87  小新 82  小丽 90  Database成绩：  小明 95  小红 81  小新 89  小丽 85  Python成绩：  小明 82  小红 83  小新 94  小丽 91  平均成绩如下：  (小红,83.67)  (小新,88.33)  (小明,89.67)  (小丽,88.67) |
| 主要仪器名称及型号：   1. 计算机：linux，core-i5,8GB内存 2. Idea+sbt, 3. Scala-2.11.8 4. Spark-2.4.4 |
| 实验过程：（可附页）  1．spark-shell交互式编程      2.编写独立应用程序实现数据去重    3.编写独立应用程序实现求平均值问题 |
| 实验数据记录或图片：（可附页）   1. spark-shell交互式编程     图一.信息文档  （1）Tom同学的总成绩平均分是多少；    （2）求每名同学的选修的课程门数；    （3）各门课程的平均分是多少；    （4）该系总共有多少学生；  （5）该系共开设来多少门课程；  （6）该系DataBase课程共有多少人选修；  （7）使用累加器计算共有多少人选了DataBase这门课。     1. 编写独立应用程序实现数据去重     图二.输出文档   1. 编写独立应用程序实现求平均值问题     图三.输出文档 |
| 实验结论：  通过此次实验学习，使我进一步理解了有关scala语言的语法。并且在使用idea软件配合sbt编写RDD实验学习中，熟悉了Spark的RDD基本操作及键值对操作。特别是在键值对操作方面，充分体会到了该语言的简便性。  不过在实验过程中也遇到过一些问题，但经过文档参考和二次学习也是终于找出了原因所在，也暴露出了前期我在这方面的知识欠缺和经验不足。实践出真知，通过亲自动手调试和书籍知识查询，理解。使我们掌握的知识不再是纸上谈兵。  在这次实验学习过程中，我不断发现错误，不断改正，不断领悟，不断获取。终于顺利完成了。体会到了这门语言的独具特色，希望在今后的进一步学习实践中，能不知难而退，再接再厉。只有这样，才能成功的做成想做的事情，才能收获成功，收获知识和喜悦。 |
| 教师评语：  *从观察能力、实验能力、分析和解决问题的能力、实验结果、实验报告表达能力、实验结论等方面评价。*  *可手写或打印，若为电子稿则首行缩进两字符，宋体小四，1.5倍行距*  *字数不少于30字*  成绩： 批阅教师：  年 月 日 |