南华大学

计算机学院

实验报告

（ 2022 ~2023 学年度 第一学期 ）

|  |  |
| --- | --- |
| 课程名称 | 大数据平台及编程实践 |
| **实验名称** | 熟悉常用Spark操作 |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 刘纯龙 | 学号 | 20184351149 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 软卓01班 |

|  |  |  |  |
| --- | --- | --- | --- |
| 地点 |  | 教师 | 聂沛 |

1. **实验目的与要求**

实验目的：掌握使用spark访问本地文件和HDFS文件的方法

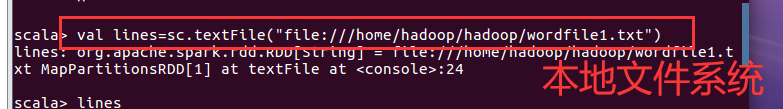
掌握spark应用程序的编写运行

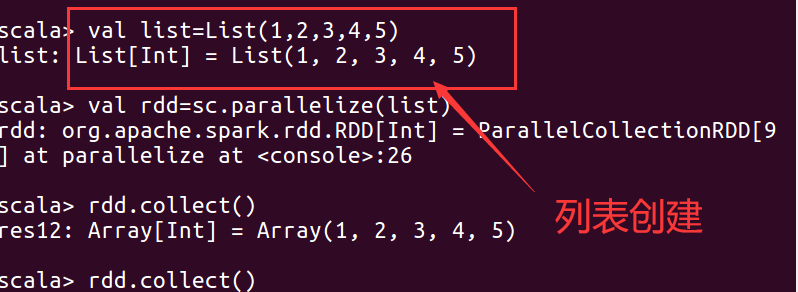
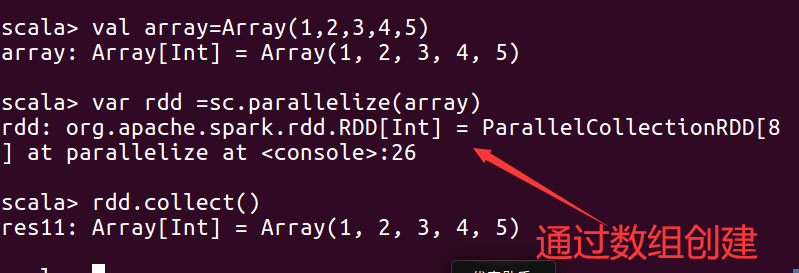
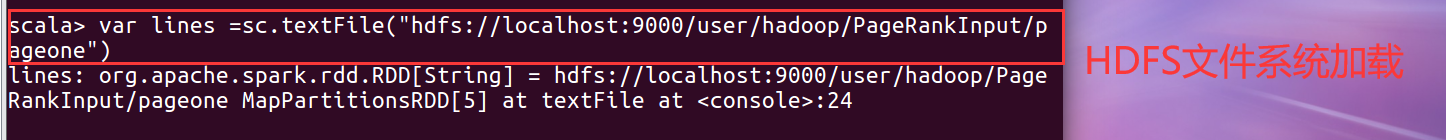
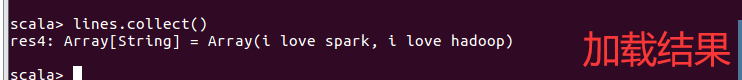
以及RDD的操作

实验要求：配置好SPARK环境，,读取本地和HDFS数据，对rdd进行操作

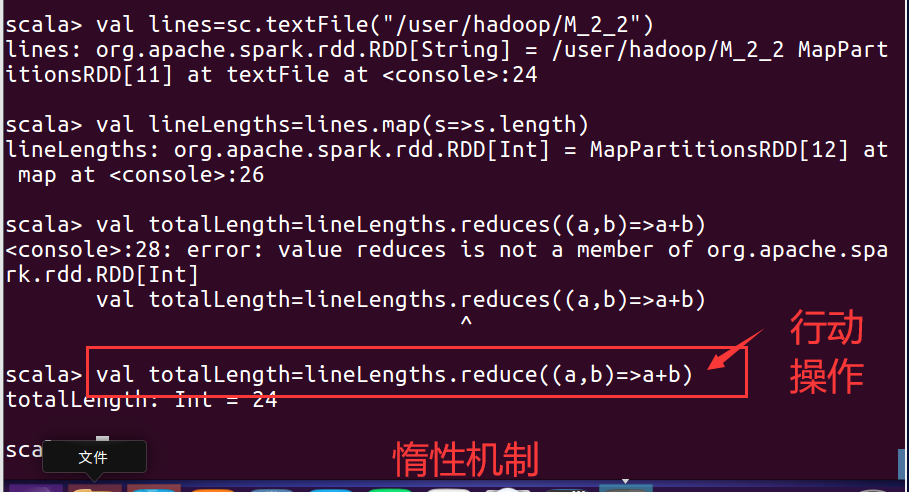
运行程序等

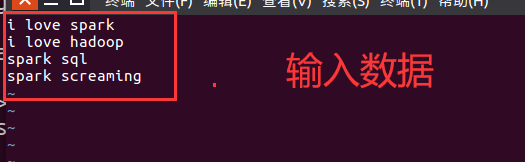
1. **实验内容与步骤**
2. **创建RDD**

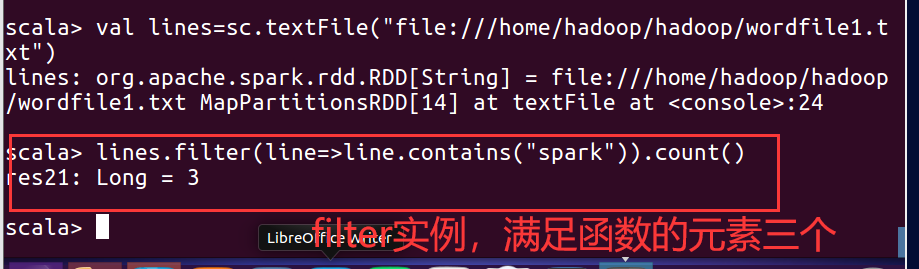


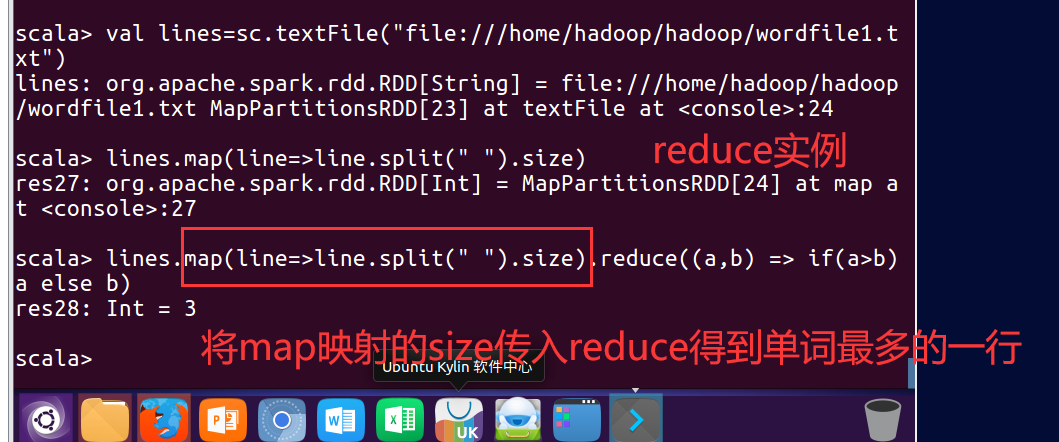


1. **RDD操作**

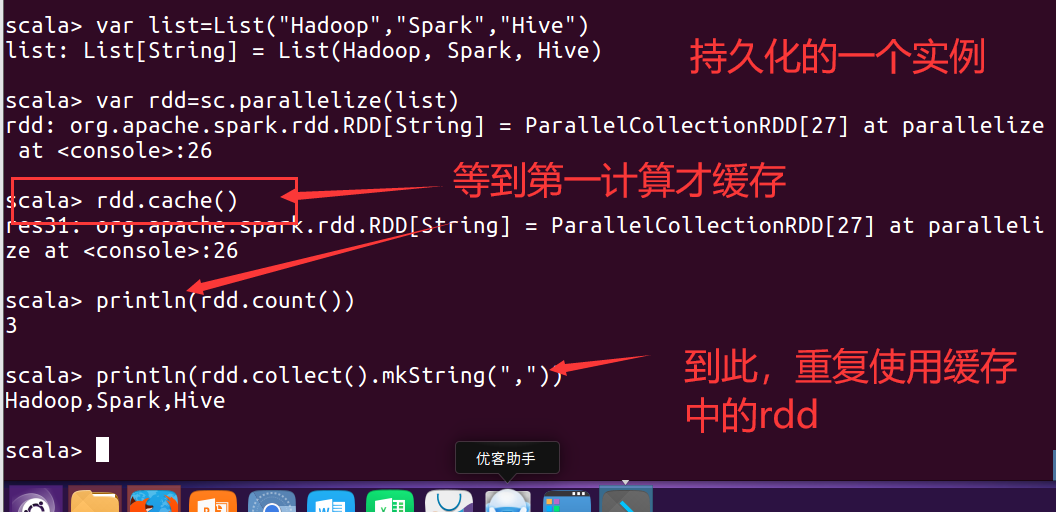




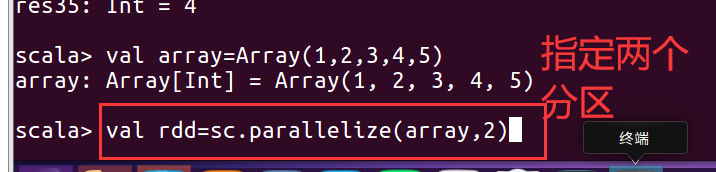
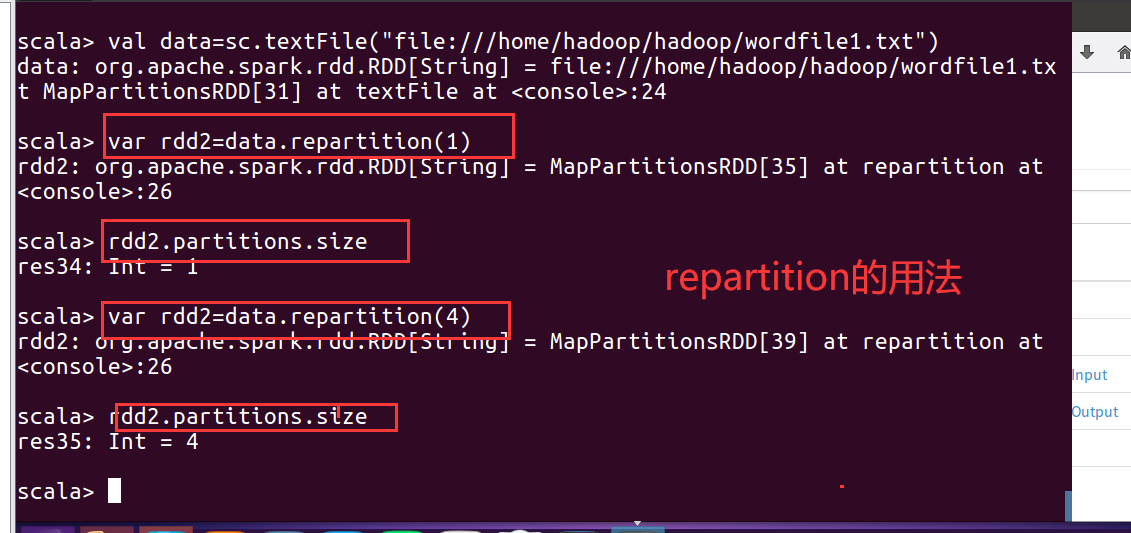








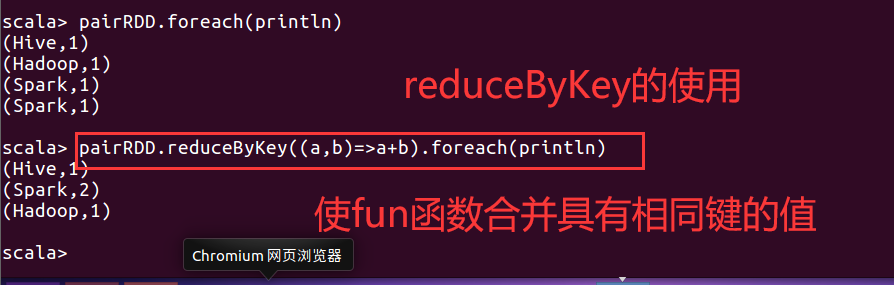
1. **RDD分区**

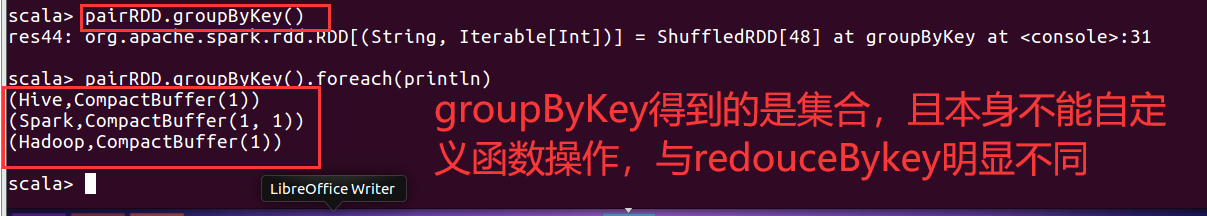


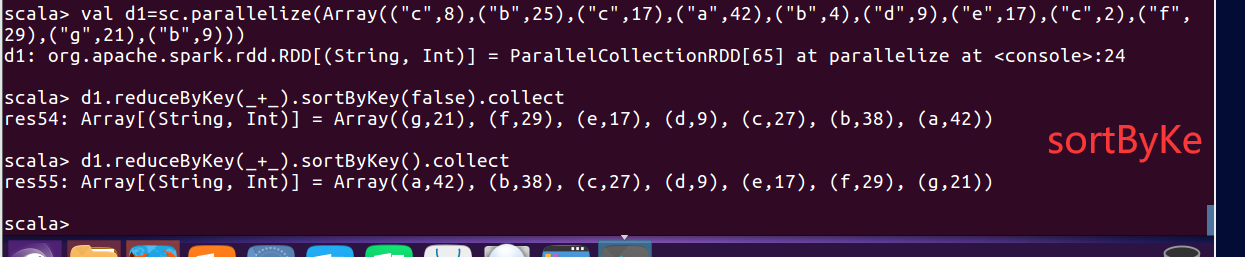
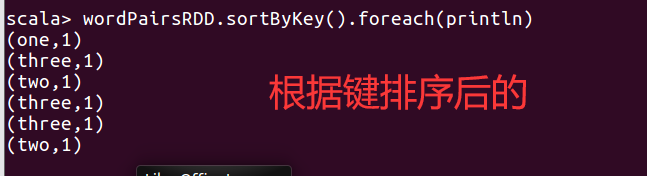
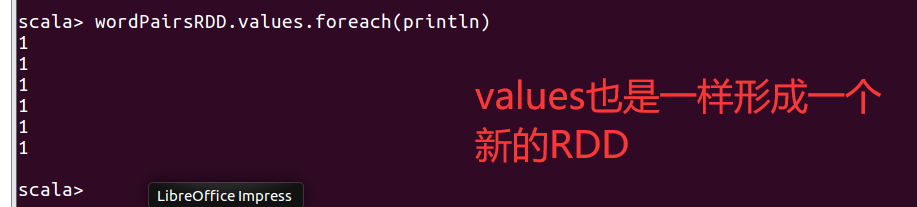
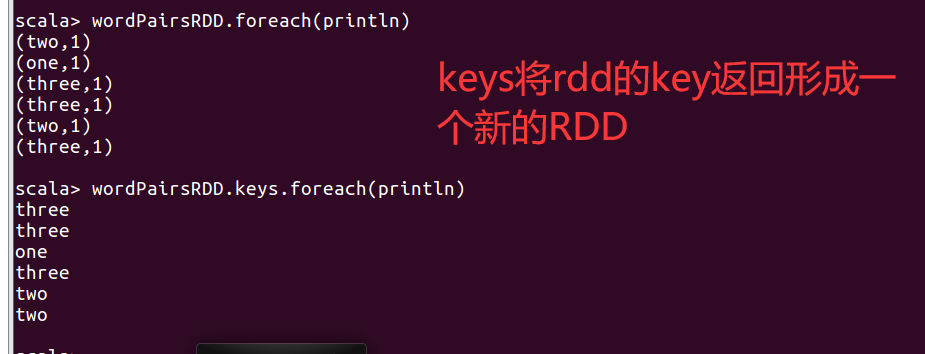
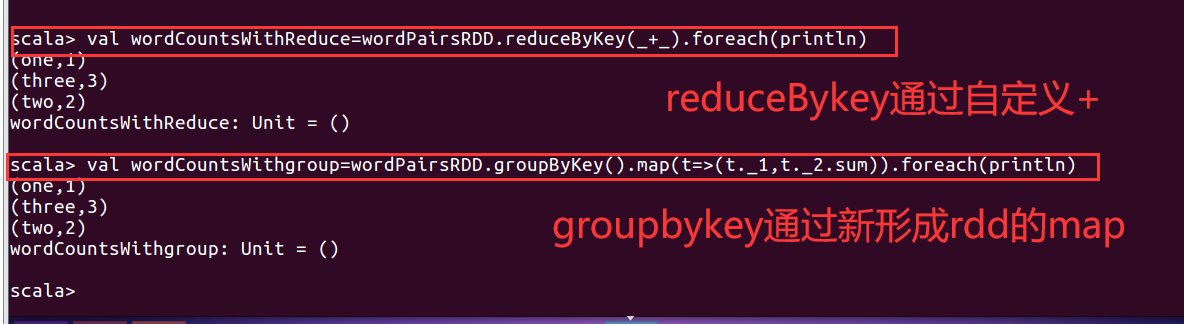
1. **Pair RDD**

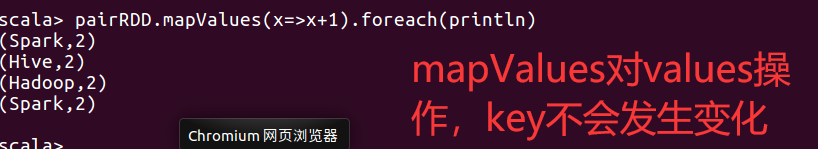


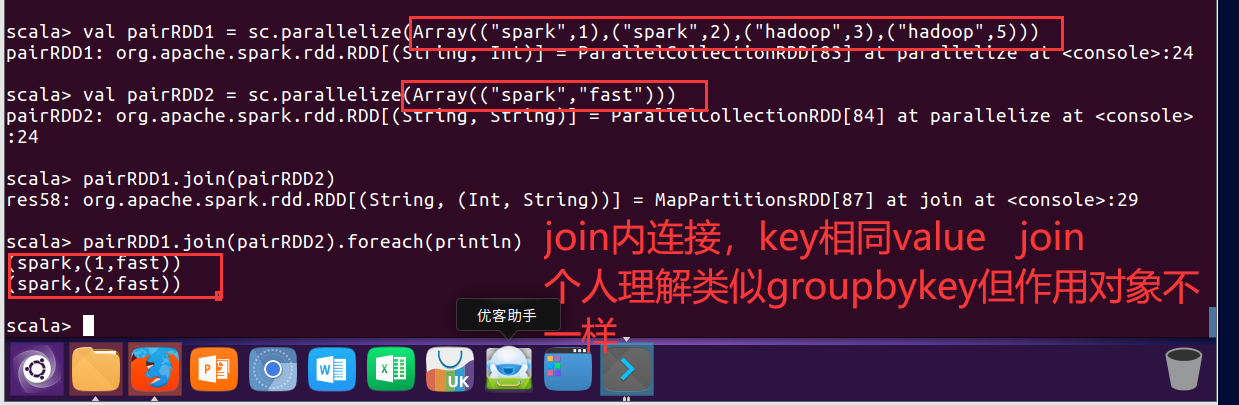


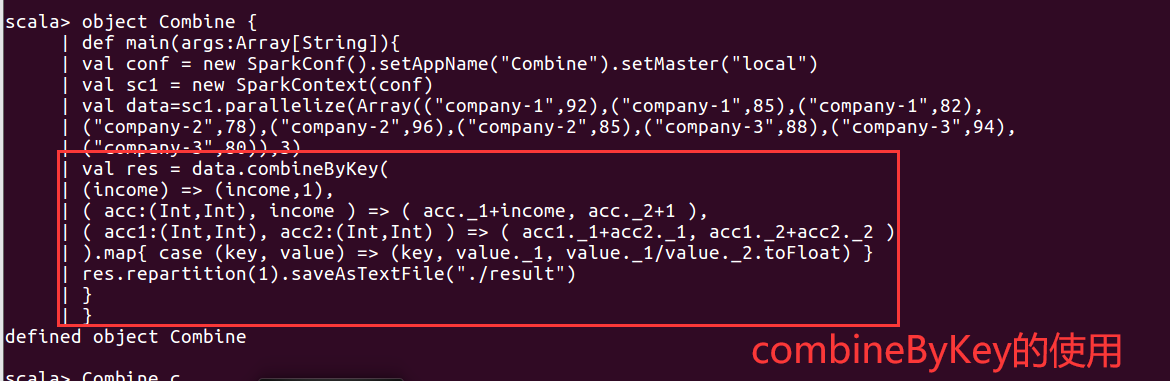


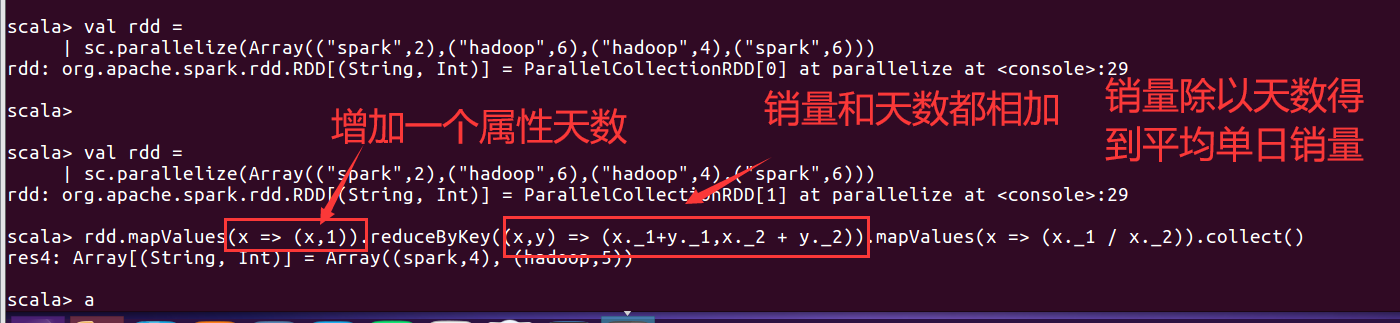


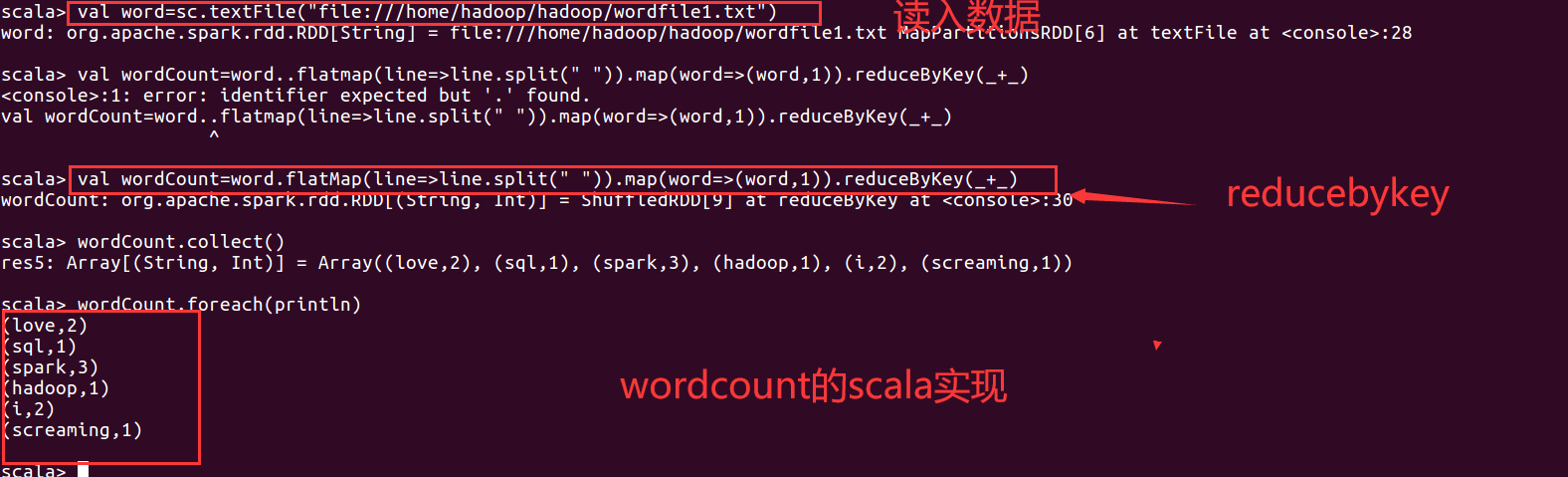




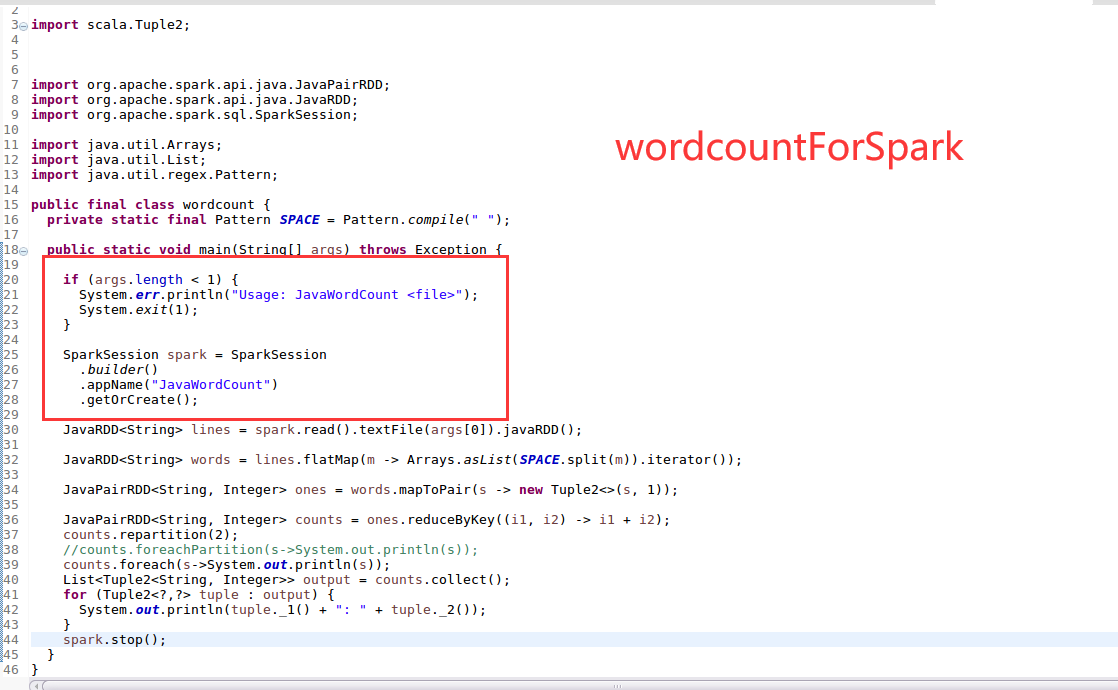


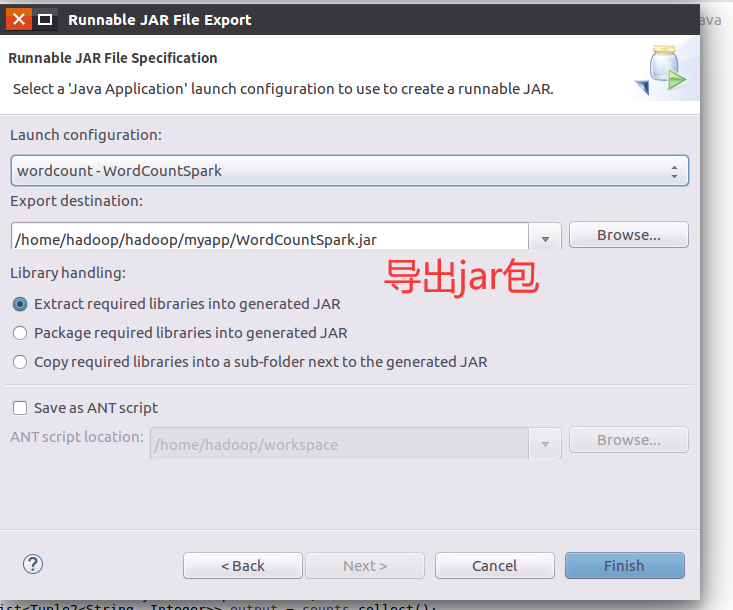


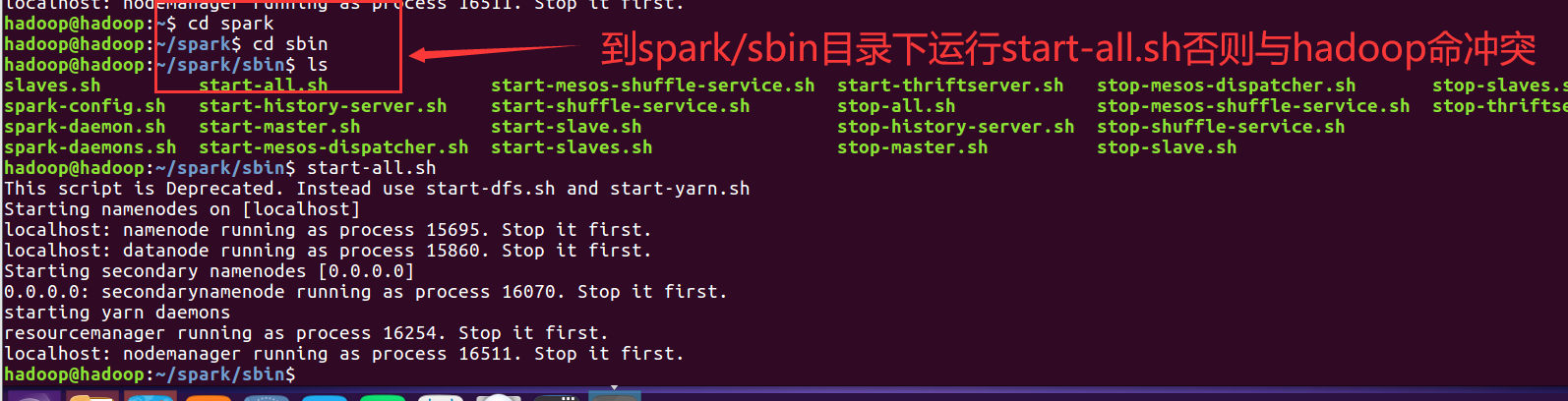


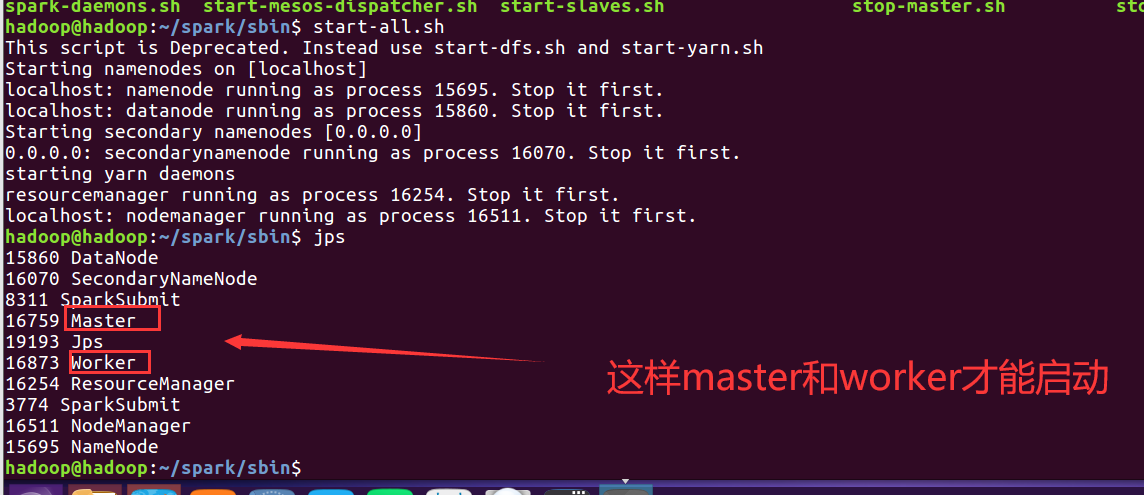


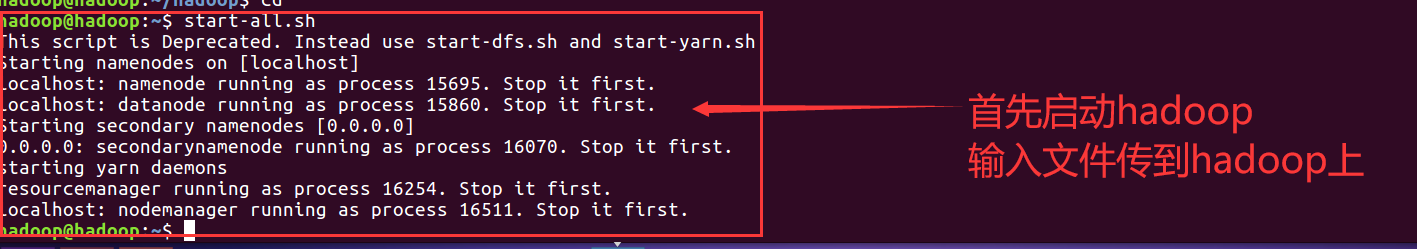
1. **Wordcount**

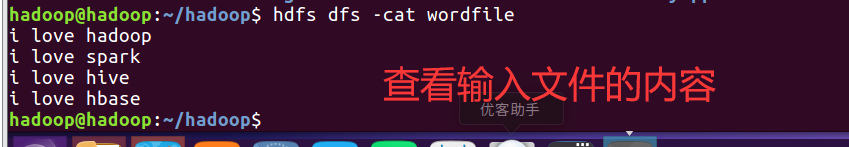


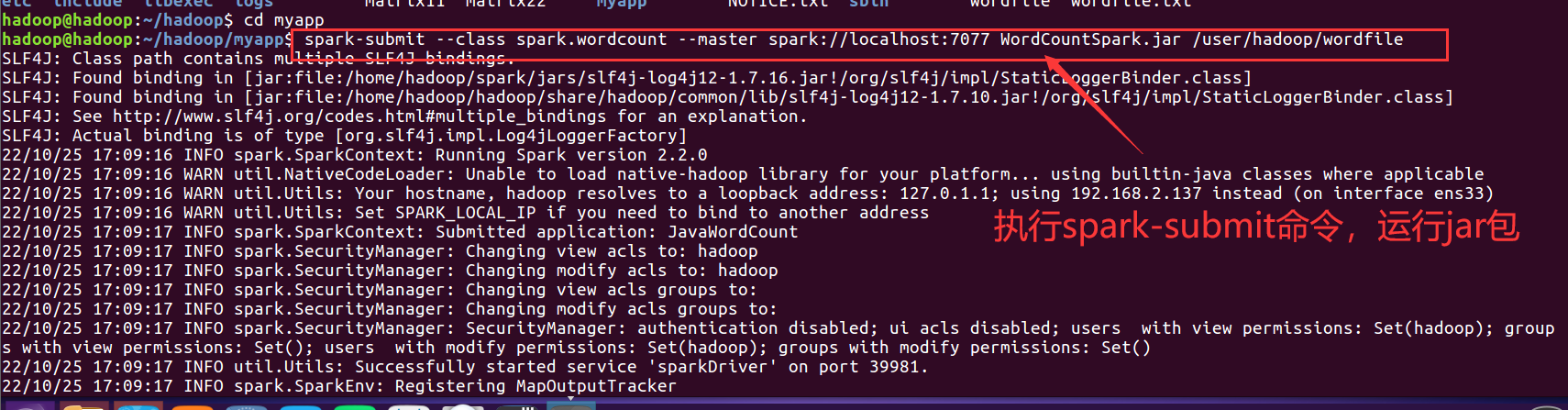
















1. **实验总结**

首先第一个内容关于RDD的创建，通过本地文件系统或者HDFS文件系统加载数据或者并行数组创建RDD.其次通过具体的实例了解RDD的惰性机制，紧接着发现惰性机制的弊端：每次行动操作需要大量的重复计算。由此了解到持久化cashe()rdd来避免需要重复计算.不过cashe()并不是直接加载，是第一次计算后再加载。

其次了解到RDD分区的概念，以及默认分区和如何指定分区.

然后是PairRDD的创建与操作，创建主要是通过加载后的数据形成一个rdd,再使用map函数创建pairRDD,然后了解了reduceByKe和groupByKey的区别，以及keys，values的特性和sortBy和sortByKey的区别等等。

最后是关于单词计数器wordcount使用scala编写和java编写，scala编写的话通过map和reduceBykey两行就可以写完，使用java编写，spark-submit运行jar包，其中遇到了一些问题，首先开启worker和matser节点需要在spark/sbin里面运行start-all.sh否则与hadoop的命令冲突。

除此之外，还学会了如何在scala里编写一个类和方法，以及如何创建实例和运行方法，其次还了解了Object的特性。