大数据平台技术课程实验报告

**实验三：MapReduce操作实践**

姓名：\_\_ 杨佳森\_\_\_\_\_\_

学号：\_\_\_\_\_ 2112080106

专业：\_\_\_ 数据科学与大数据技术

年级：\_ 2021

主讲教师：\_ 刘春 \_

实验时间： 2023年3月3日 至 3月24日

实验内容与要求：

[1](#_Toc43287683) MapReduce基本知识总结

2 MapReduce的操作实践

2.1 基于MapReduce的WordCount实践：统计每个单词出现的次数

2.2 自定义Hadoop数据类型实践：自定义hadoop数据类型，在统计每个单词出现次数的过程中，同时计算每个单词的长度

2.3 多mapReduce任务的串联实践：基于MapReduce统计共有多少个单词，而不是每个单词出现的次数

**特别提醒：**

**（1）基本知识点的总结请使用自己的语言，根据自己的理解去总结，就像写课程笔记一样，不要去网上抄写**

**（2）实践需要有截图以及相应的文字介绍，欢迎大家把自己在实践过程中碰到的问题以及解决方法也记录下来**

1. **MapReduce基本知识的总结**
2. **概述**

MapReduce 是一个分布式运算程序的编程框架，是用户开发“基于 Hadoop 的数据分析应用”的核心框架。

MapReduce 核心功能是将用户编写的业务逻辑代码和自带默认组件整合成一个完整的分布式运算程序，并发运行在一个 Hadoop 集群上。

1）MapReduce 易于编程

它简单的实现一些接口，就可以完成一个分布式程序。也就是说写一个分布式程序，跟写一个简单的串行程序是一模一样的。用户只需要关心业务逻辑，实现一些简单的接口就可以了。

2）良好的扩展性

可以动态增加服务器，解决计算资源不够的问题。

3）高容错性

MapReduce 设计的初衷就是使程序能够部署在廉价的 PC 机器上，这就要求它具有很高的容错性。比如其中一台机器挂了，它可以把上面的计算任务转移到另外一个节点上运行，不至于这个任务运行失败，而且这个过程不需要人工参与，而完全是由 Hadoop 内部完成的。

4）适合 PB 级以上海量数据的离线处理

可以实现上千台服务器集群并发工作，提供数据处理能力。

**2、**

1）Hadoop下的一个负责分布式计算的组件

2）一个软件系统，运行于HDFS之上

3）定义了一种实现分布式计算的框架

4）负责计算任务在集群中的分配调度、负载均衡、容错处理、网络通信等一系列问题

5）方便编程人员在不熟悉分布式并行编程的情况下，能够编写程序对分布式环境下的大数据进行处理

6）借鉴了函数式编程

函数：集合之间的一种映射关系

不同于命令式编程，函数式编程关注集合之间的映射关系

函数可以作为另一个函数的输入和输出

**3、MapReduce框架**

将分布式环境下的并行大数据处理过程抽象为两个函数：map和reduce

Map: <key1,value1> 🡪 <key2, value2>

Reduce: <key2,value-list> 🡪 <key3, value3>

MapReduce是一个运行在Hadoop上的分布式计算框架，用于处理大规模的数据，计算的思想是“分治法”，将大量数据分为不互相依赖的数据块，存储在不同的datanode上，由map函数处理形成部分数据的处理结果，再由reduce函数整合形成最终的结果。

一个map函数处理一个<key1,value1>数据，map端的shuffle阶段将<key1,value1>转化为<key1,value-list>形式，减少数据传输的工作量，reduce端的shuffle阶段将map端对应分区数据取回reduce端，对数据按键值排序，reduce函数将<key1,value-list>的数据合并，转为<key2,value2>。

Hadoop的数据类型是对java数据类型的封装，便于进行序列化处理，使不同的数据以统一的格式在集群中存储传输，参考Hadoop的text类型源码，可以自定义wordcountandlen数据类型，输出每个单词的总数和长度。

有时一个MapReduce过程无法得到想要的结果，可以通过多mapre任务的串联完成，如要统计总共有多少个单词，可以使用两个MapReduce任务串联完成

**4、计算过程**

一个完整的MapReduce 程序在分布式运行时有三类实例进程：

（1）MrAppMaster：负责整个程序的过程调度及状态协调。 （MRAppMaster是MapReduce的ApplicationMaster实现）

（2）MapTask：负责 Map阶段的整个数据处理流程。

（3）ReduceTask：负责 Reduce 阶段的整个数据处理流程。

1. **MapReduce的操作实践**

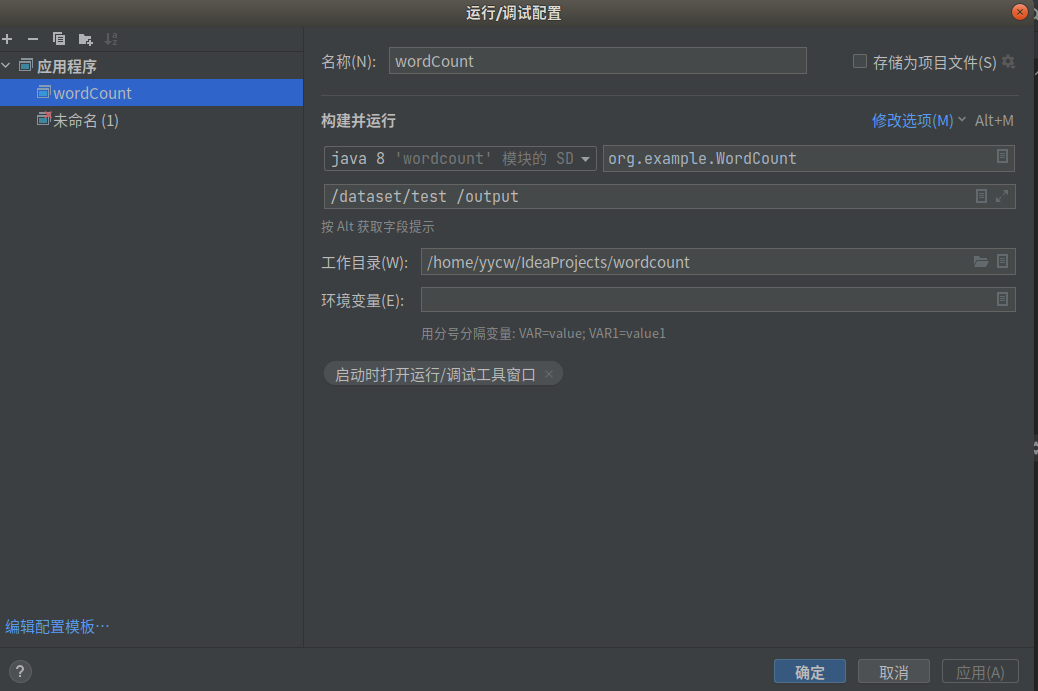
**2.1基于MapReduce的WordCount实践**

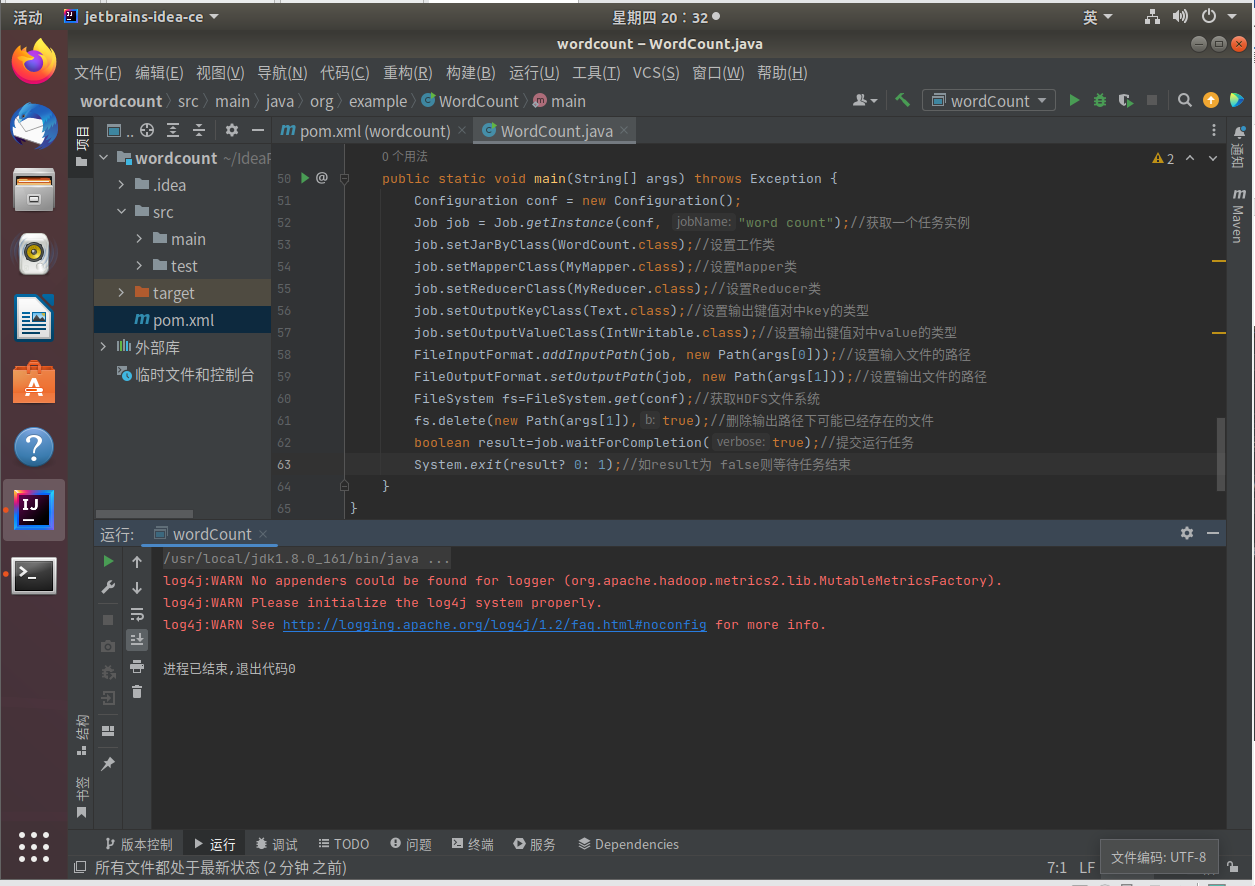
首先需要将Hadoop安装目录“etc/hadoop”路径下的core-site.xml和hdfs-site.xml两个文件复制项目的resource文件夹下

可以在idea终端在hadoop路径下使用以下命令来复制，hdfs-site.xml同理

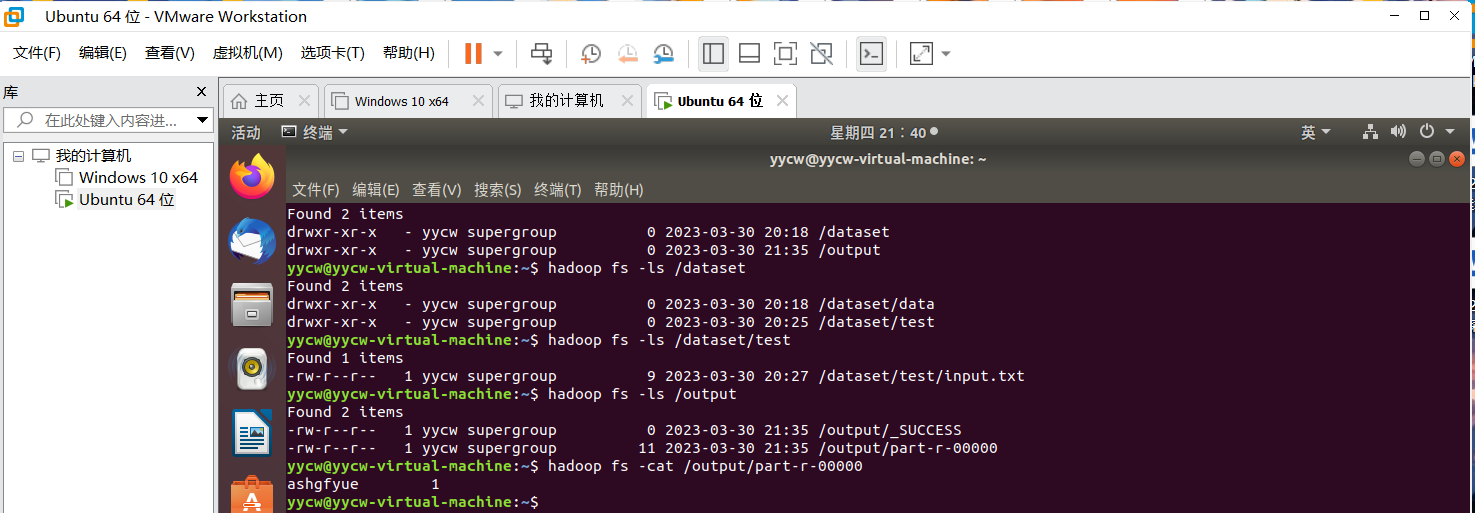
cp -r ./core-site.xml /home/yycw/IdeaProjects/wordcount/src/main/resources/core-site.xml

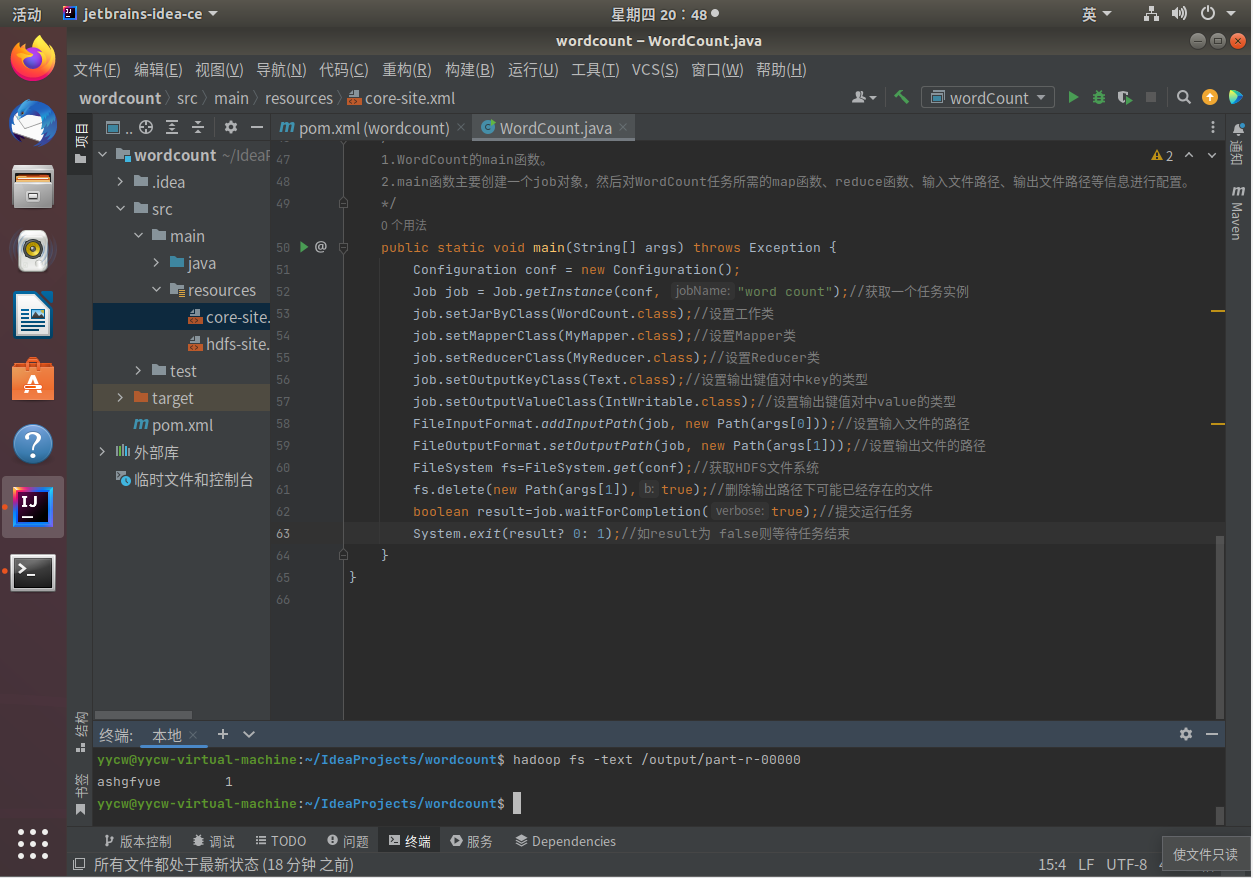
然后在项目的configuration的“program arguments”选项中填写输入文件和输出文件在HDFS中的目录或者文件路径，要注意dataset/test里面要有数据。

****

****

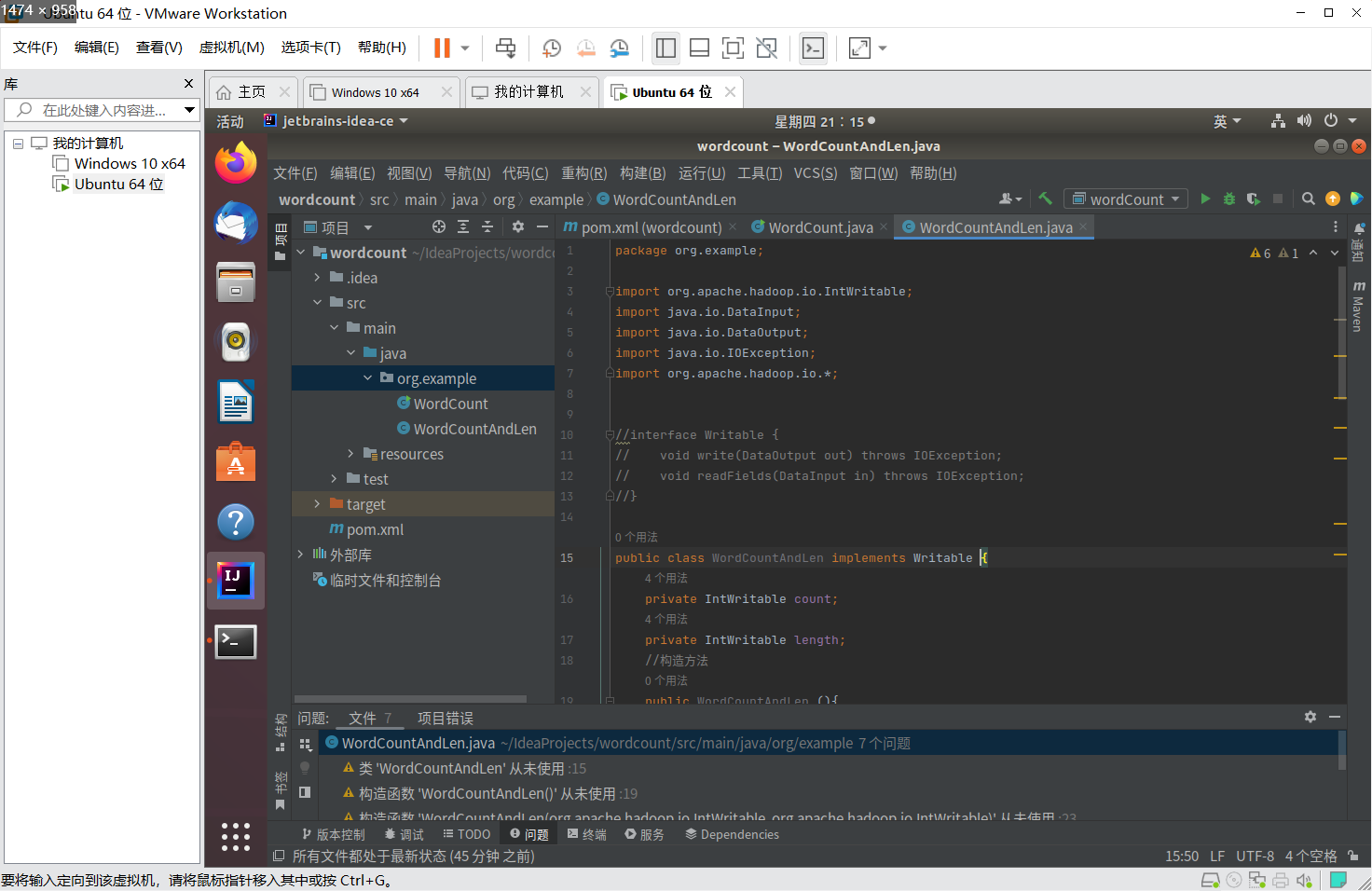
**查看文件输出内容（建立文件时输入的）**

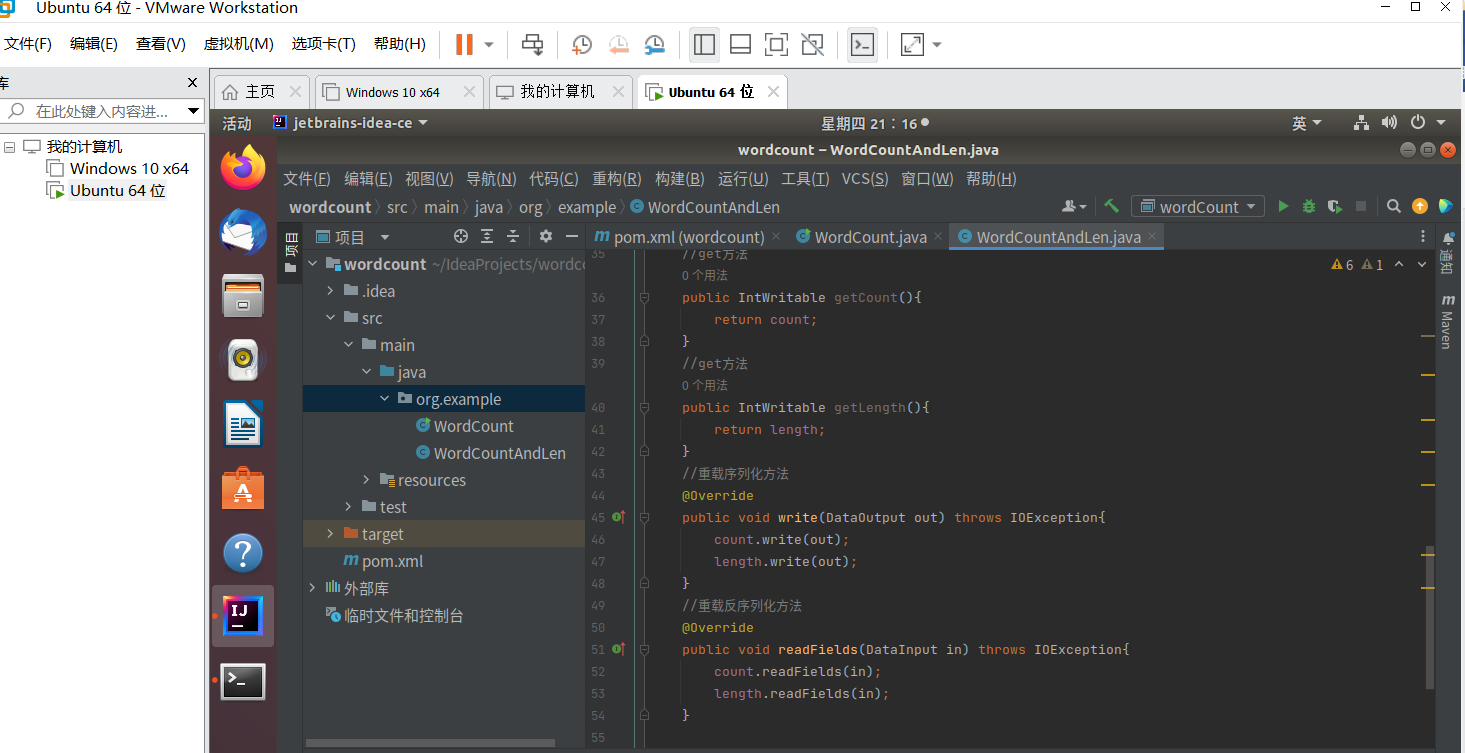
****

****

**2.2 自定义Hadoop数据类型实践**

**重新建WordCountAndLen类**



****

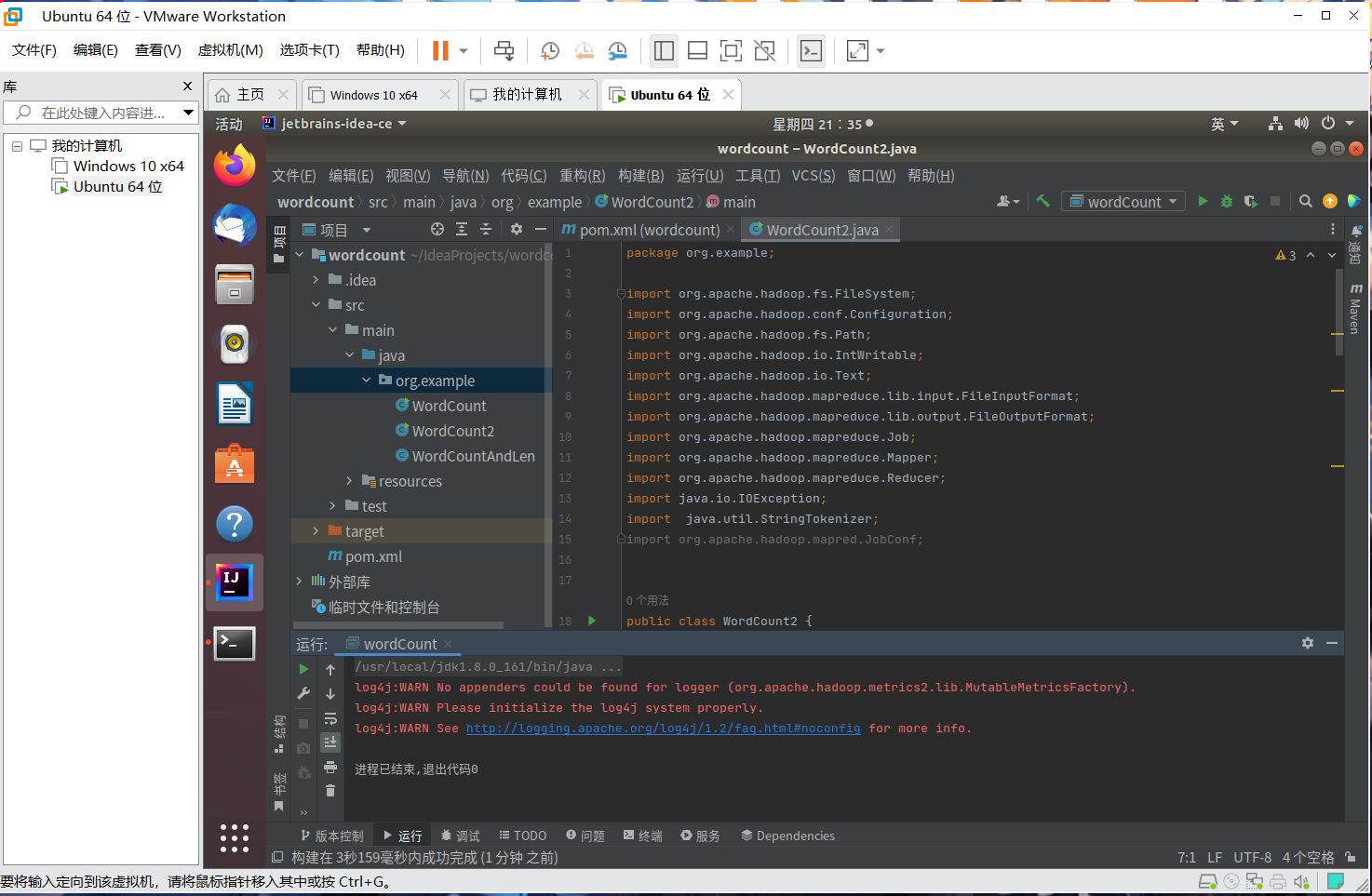
* 1. **多mapReduce任务的串联实践**

**编辑配置：**

图形用户界面, 文本

描述已自动生成

**写入代码：**

****

