**2020年全国高校计算机能力挑战赛**

**大数据应用赛答题卡第二部分**

**队长姓名： 队长学校：**

# 第二部分：大数据处理部分（Mapreduce）（20分）

数据文件 维修换件清单1.csv，维修换件清单2.csv，两个个的结构是一样的 分别为某工业公司维修部门的维修记录，有2个不同的组进行记录，2个文件里面有重复统计的数据，数据文件中的每行都是一个数据。 

## 任务一、请使用MapReduce对两个进行合并，合并后维修记录数据进行去重。去重后输出数据为weixiuhuanjian，并把重复的输出到 chongfudata,并对去除重复后的数据统计总行数,统计重复数据的行数（10分）

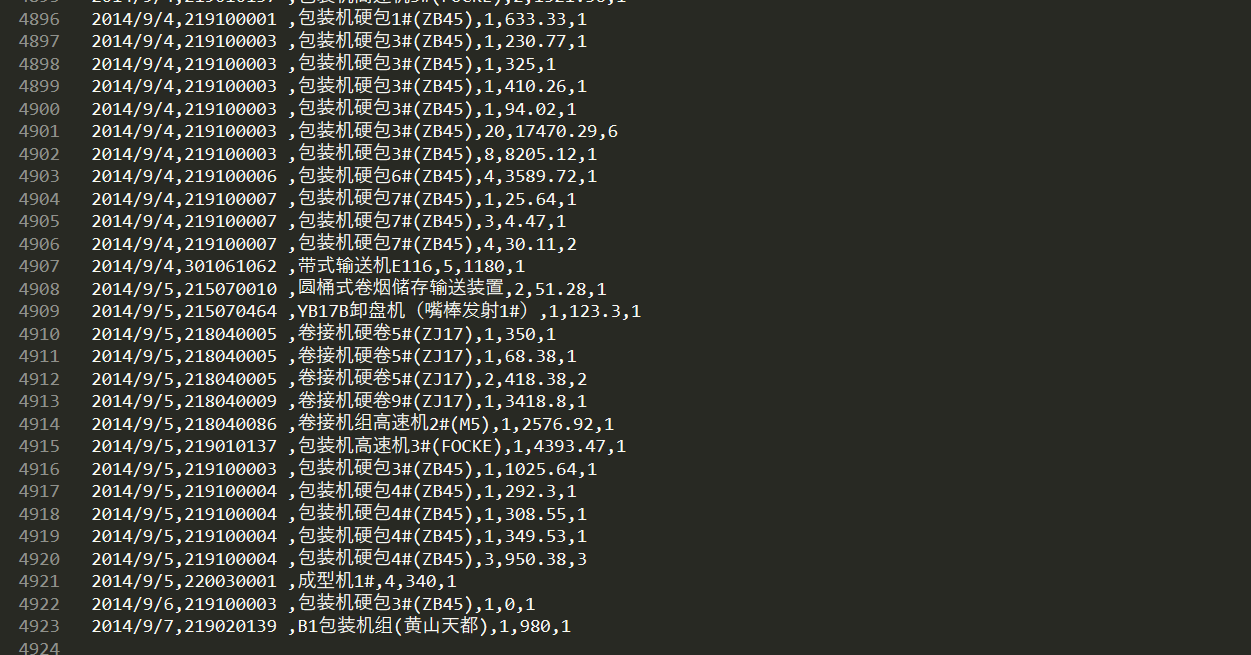
**本题我重写了OutPutFormatClass来输出两个文件**

时间,机台ID,机台名称,换件数量,维修费用,维修次数

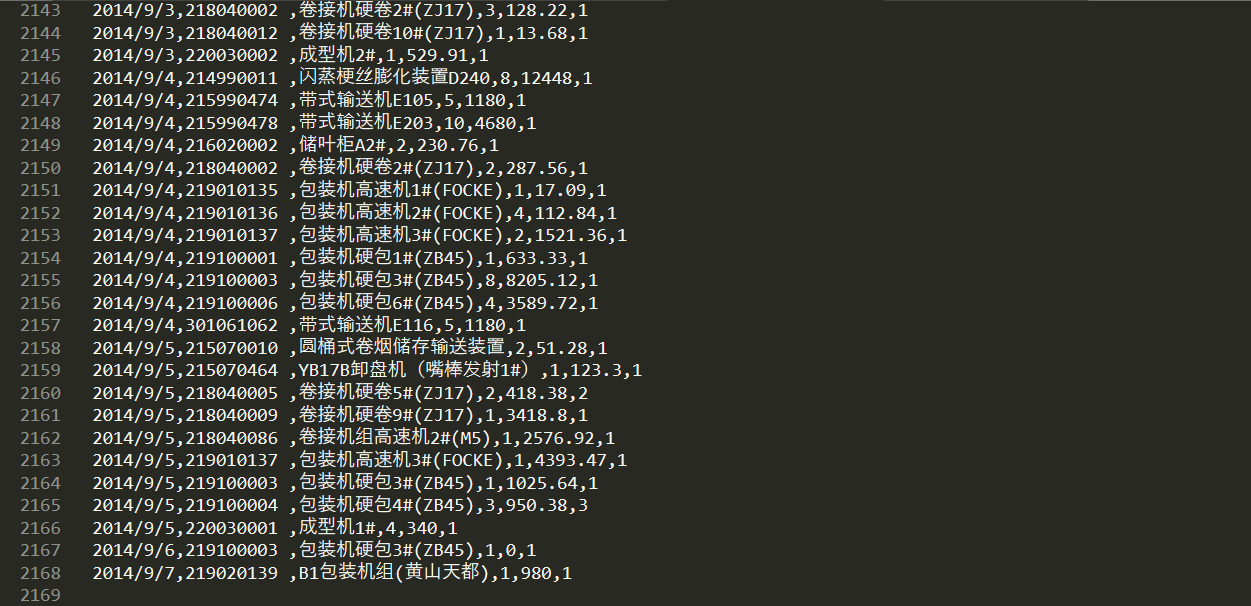
源码：



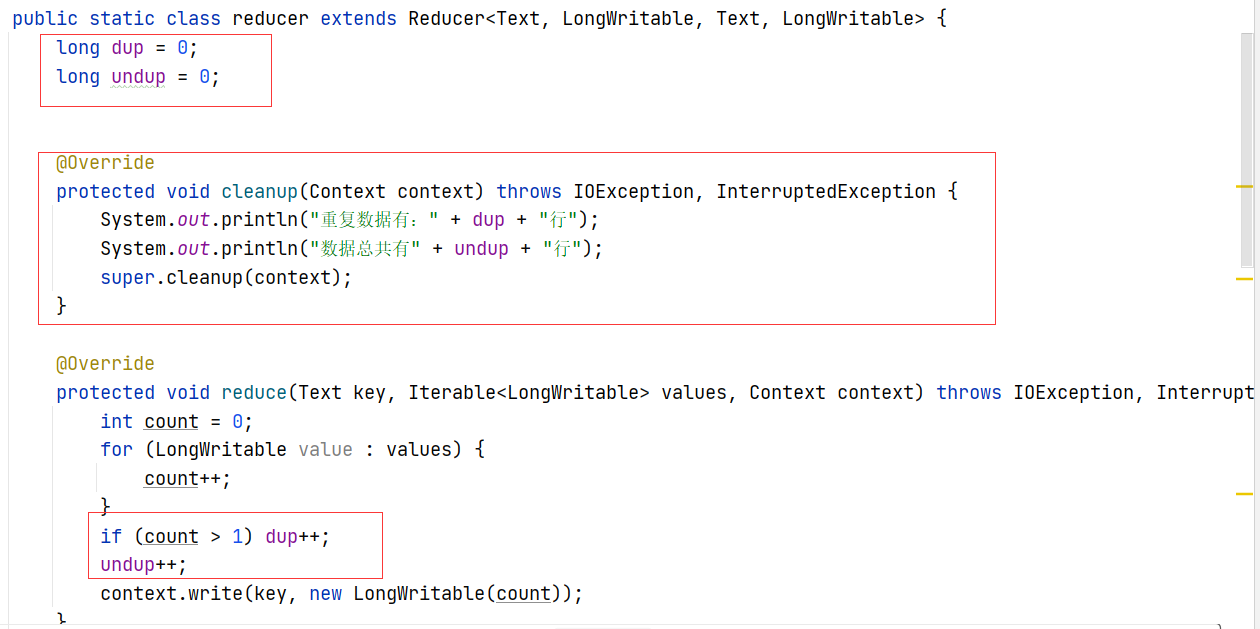
* + 1. 除重后文件中的数据内容截图：



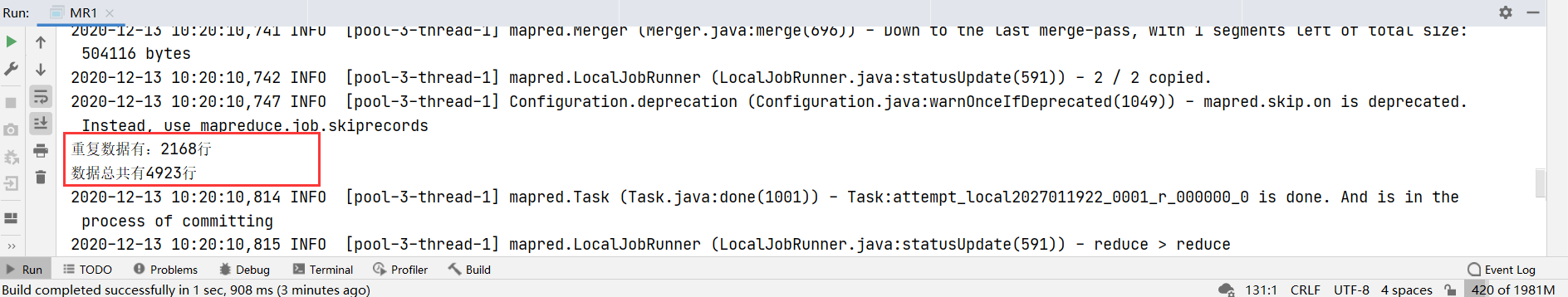
* + 1. 重复文件中的数据内容截图



* + 1. 统计除重后数据总行数 截图



输出结果：

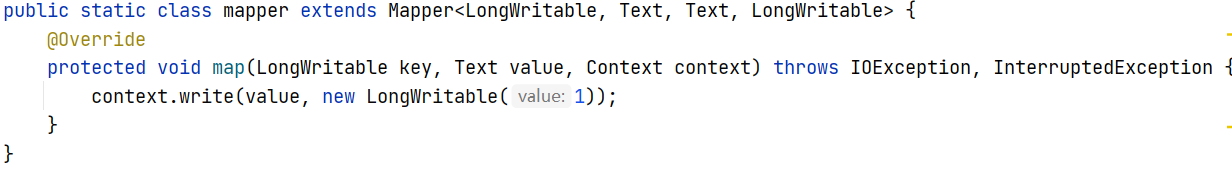


* + 1. 统计重复行数 截图

在上面一个问题中一起回答了

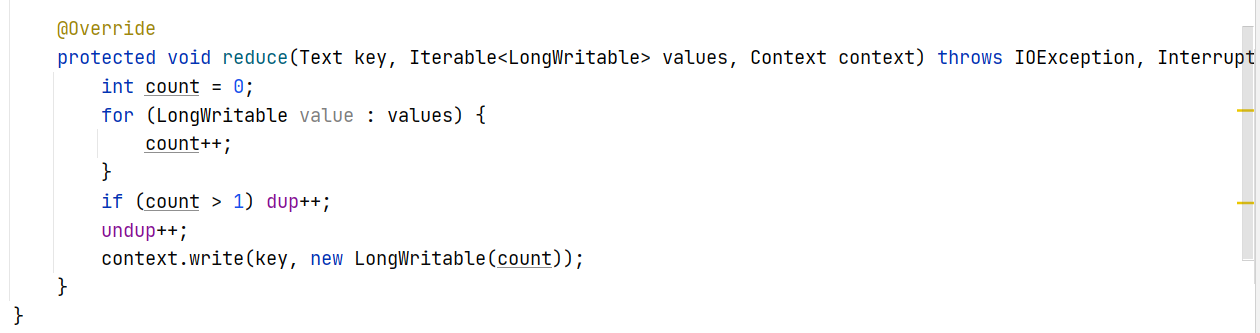
* + 1. 对数据文件合并的相关代码截图

**Mapper进行合并**

****

* + 1. 对合并后的数据进行去重的相关代码截图

**Reduce自动去重**



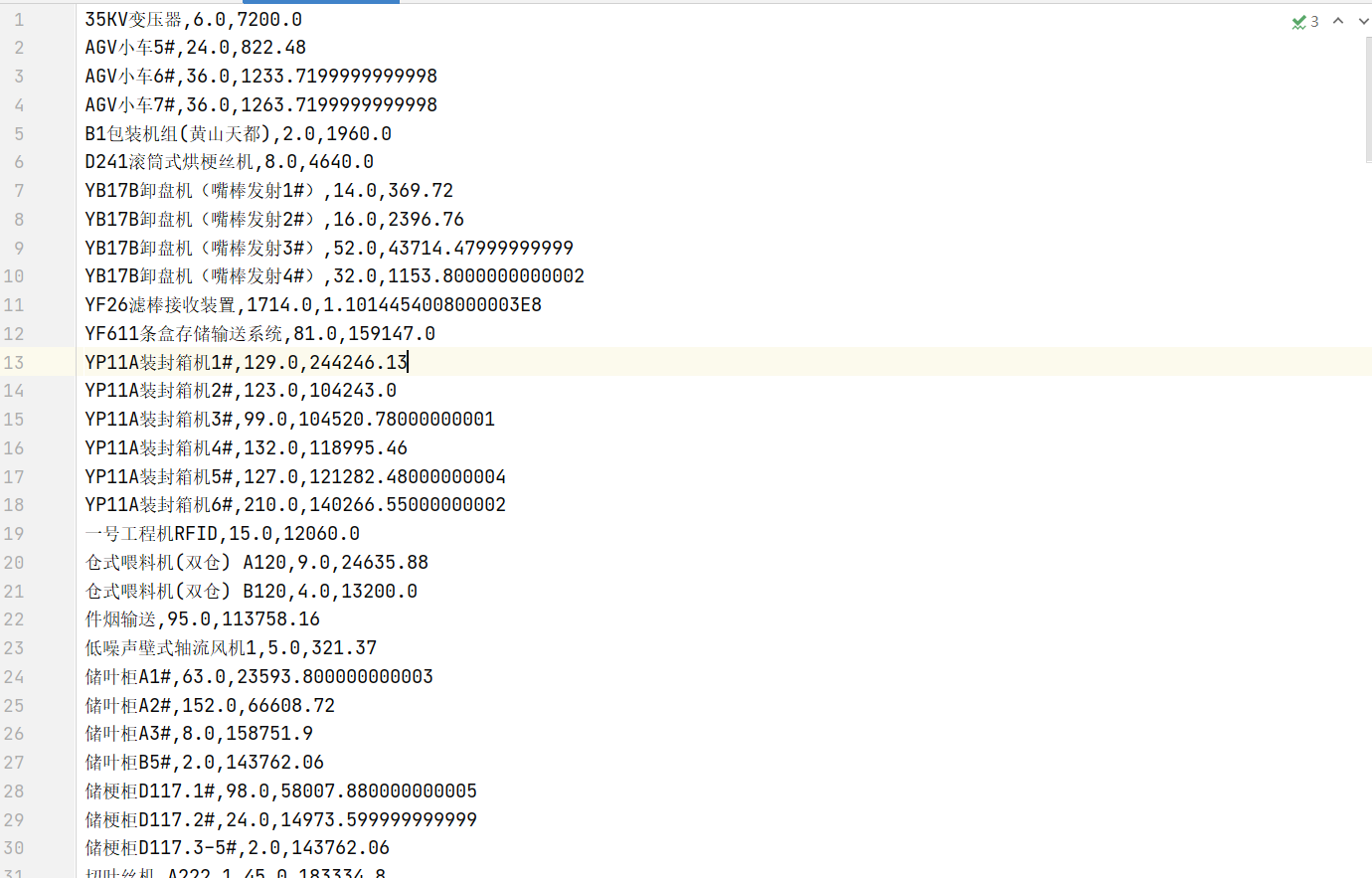
* + 1. 统计行数的相关代码截图



## 任务二、请使用MapReduce统计 维修换件清单1.csv中的每个设备，维修的换件数量和维修费用 ，并输出格式 为 设备名称，换件数量，维修费用（5分）。

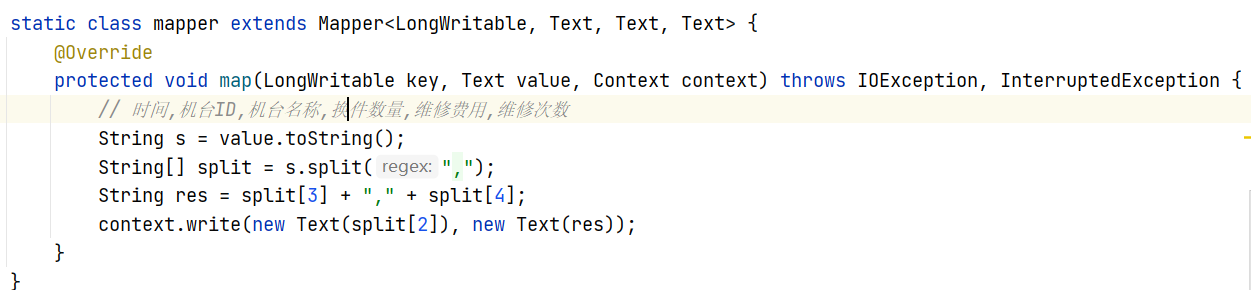
源码：

1）结果截图：

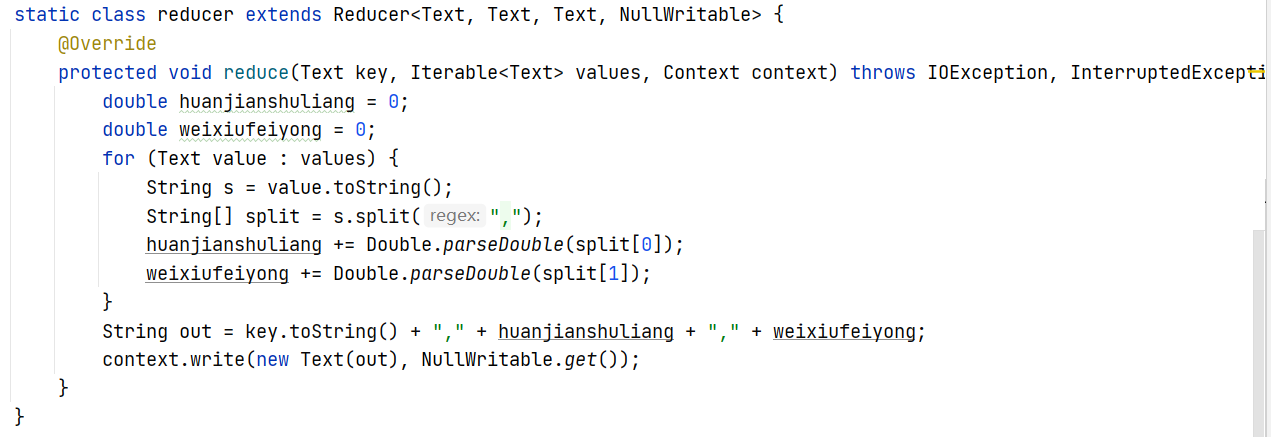


1. 代码截图：

Mapper



Reducer



## 任务三、请使用MapReduce统计 维修换件清单1.csv中的每年 维修费用最高的前三条记录（5分）。返回格式：年 ，机台名称，维修费用

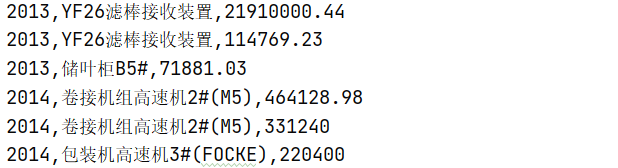
源码：

* 1. 结果截图：

带重复数据：

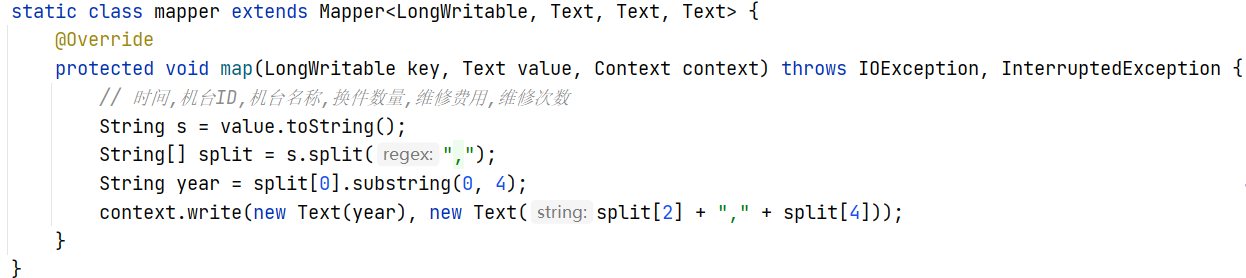


不带重复数据：



2）代码截图：

Mapper



Reducer

