**2020年安徽省大数据与人工智能应用初赛**

**学校名称： 皖西学院 队伍名称： ZZW 队伍编号： wxxy3**

# 第二部分：大数据预处理部分（20分）

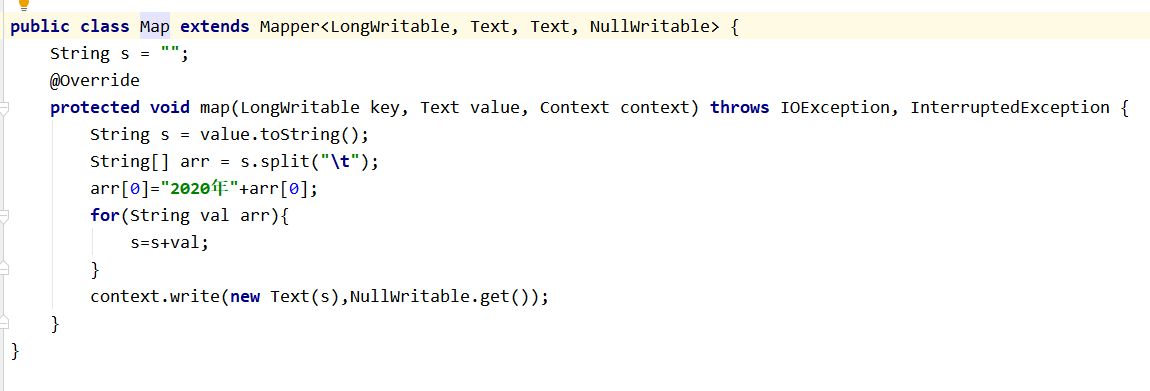
2020年新冠肺炎对我国社会各方面影响巨大，大数据技术在抗击疫情过程中发挥了巨大作用，尤其在新增、确认等相关病例数据的采集及统计上应用颇广，下面有一份数据是今年1月20-4月29日的全国各省市及国外的疫情数据，请你按照要求使用MapReduce程序完成相关数据预处理。

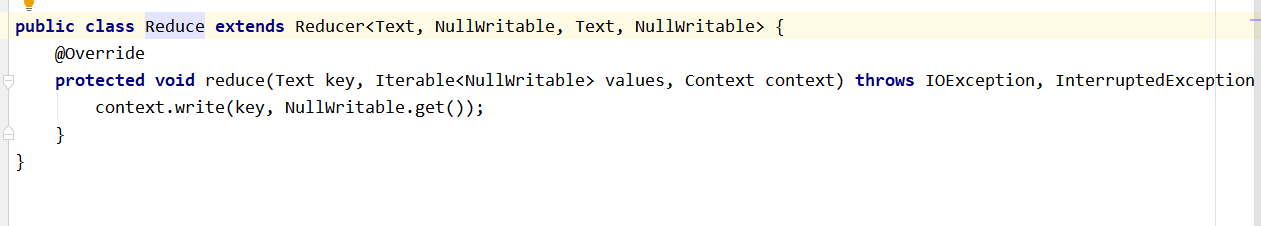


### 数据转换：请将数据中日期字段格式，替换成日期格式为xxxx年xx月xx日（5分）。

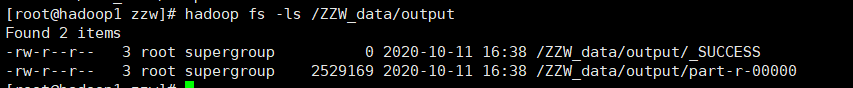
请粘贴代码和运行结果截图（在hadoop环境中查看hdfs结果数据截图）

代码截图：





结果截图:

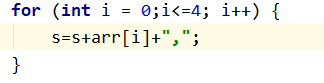




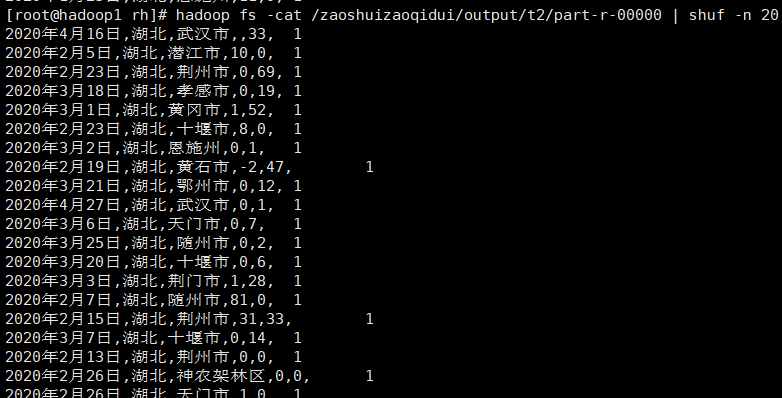
1. 数据清洗：以下规则同时进行（请粘贴代码和运行结果截图）

规则1 从上述小题中，截取前5个字段。（5分）

代码截图：

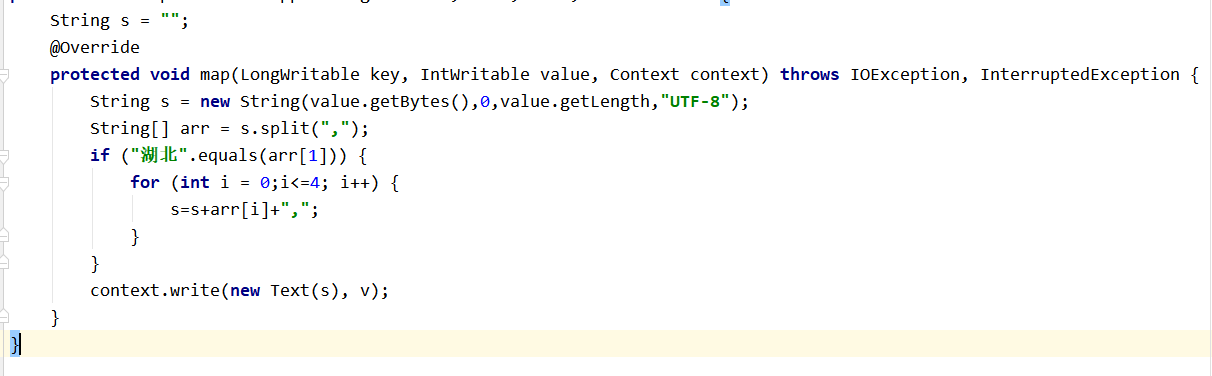


结果截图：

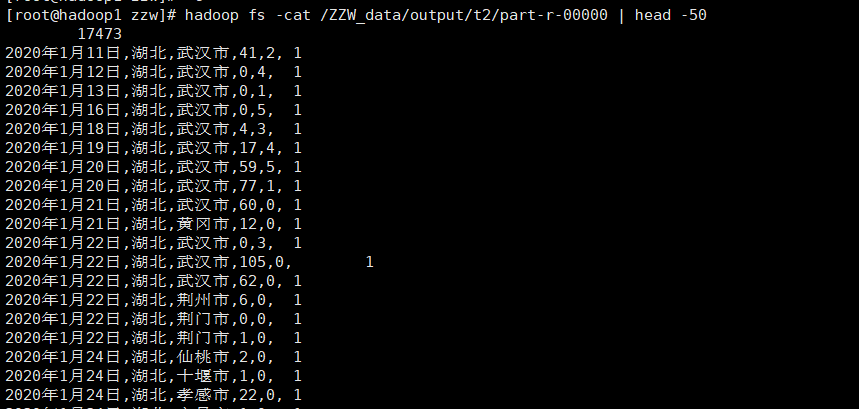


规则2 过滤出省份为湖北省的数据。（5分）

代码截图：

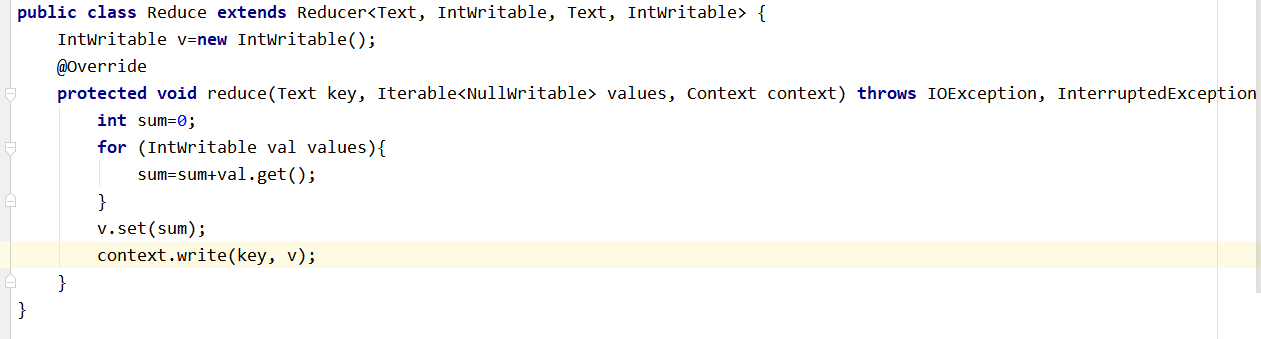


结果截图：



规则3 对5个字段去重,生成新的数据，将结果数据输出到hdfs。（5分）

代码截图：



结果截图：

