MapReduce练习题实验报告

201600301325赵虎

## 问题描述

在关系代数中，自然连接(Natural join)是一种特殊的等值连接，它要求两个关系中进行比较的分量必须是相同的属性组，并且在结果中把重复的属性列去掉。

根据给定的数据《student.xlsx》和《student\_course.xlsx》，请使用MapReduce计算模型，实现表student和表student\_course的自然连接，结果集包括学生学号、姓名、选修课程编号、考试成绩。

## 环境配置

实验用到的MapReduce计算模型需要用到Hadoop框架，首先我们来在windows上配置Hadoop。在Linux下配置Hadoop是比较轻松的，在windows上有些小坑，还是耗费了不少时间的。

首先下载Hadoop-2.7.3，之所以用这个版本是因为网上的资源提到该版本配置时出错较少，一开始下载的2.8.4，但无奈没有找到对应版本的eclipse插件，导致eclipse无法访问HDFS，所以退回到2.7.3版本。

下载后直接解压即可。然后就是配置环境变量，具体过程与JAVA配置环境变量过程类似，唯一要注意的是配置JAVA环境变量和Hadoop环境变量时路径不能有空格，如果jdk存在Program Files文件夹下，可以用PROGRA~1替代。

接着是下载hadooponwindows-master之后解压，替换掉hadoop下的bin文件（因为原来的bin文件夹下缺少相应的dll文件）。

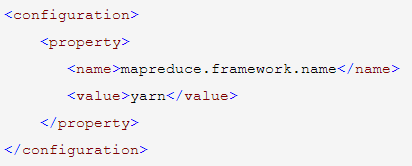
下一步是配置hadoop的四个XML文件和一个环境配置脚本：

1. D:\hadoop\hadoop-2.7.3\etc\hadoop\core-site.xml

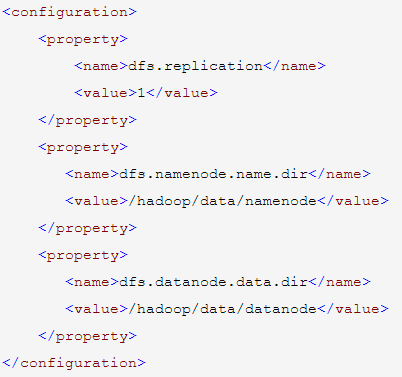


该文件配置hdfs的ip和端口号。

1. D:\hadoop\hadoop-2.7.3\etc\hadoop\mapred-site.xml



1. D:\hadoop\hadoop-2.7.3\etc\hadoop\hdfs-site.xml

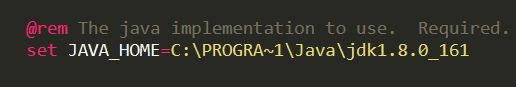


该文件配置namenode和datanode地址。

1. D:\hadoop\hadoop-2.7.3\etc\hadoop\yarn-site.xml



1. D:\hadoop\hadoop-2.7.3\etc\hadoop\hadoop-env.cmd



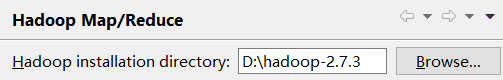
这是windows下的hadoop环境配置脚本，在里面修改jdk路径为本机jdk路径。

在初次启动hadoop时用hdfs namenode -format命令格式化HDFS文件系统，然后以管理员身份（划重点！）将目录切换到sbin下，键入”start-all”启动hadoop，接着会启动四个进程Namenode、Datanode、YARN resourcemanager、YARN nodemanager。

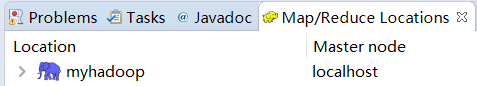
由于我是在eclipse上编写MapReduce代码，所以这里还需要在eclipse上集成Hadoop；

首先下载插件hadoop-eclipse-plugin-2.7.3.jar，然后将该jar包放到eclipse下的plugins目录下，重启eclipse。

接着在windows-Preferences中的Hadoop/Reduce选项设置Hadoop安装路径：

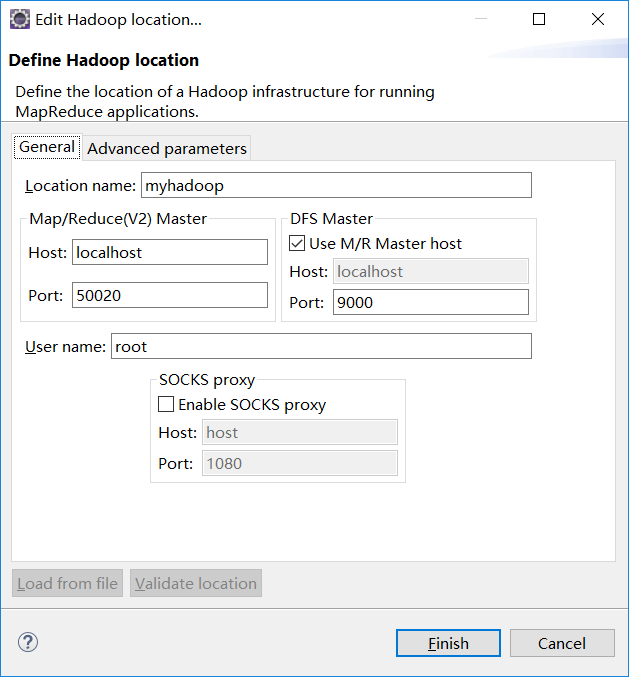


打开Windows—Open Perspective—Other，选择Map/Reduce，点击OK，控制台会出现：

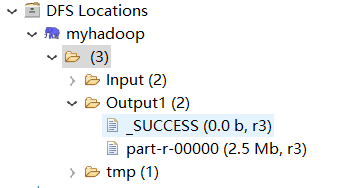


右键 new Hadoop location 配置hadoop：输入Location Name，任意名称即可。

配置Map/Reduce Master和DFS Mastrer，Host和Port配置成与core-site.xml的设置一致即可。



在左侧若能看到如下界面，则配置成功：



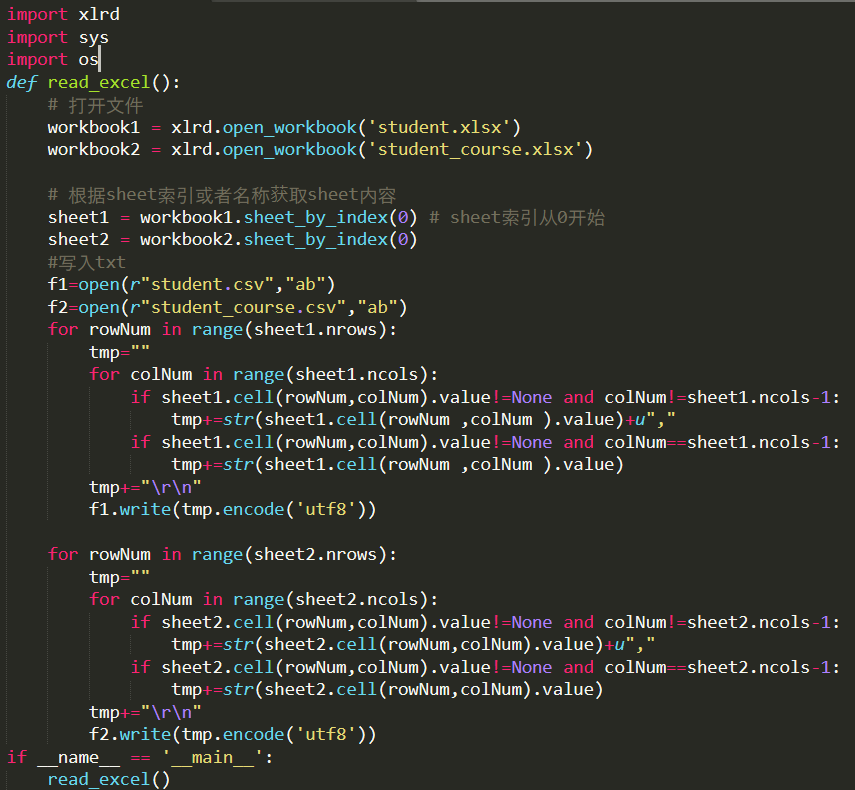
## 问题解决

在以前的学习中，我们进行自然连接是在mysql中。一旦数据量过大就会造成计算时间过长，并且在内存中进行连接的话还会造成内存不够用，此时我们可以用到MapReduce计算模型。

MapReduce是一个计算框架，他有一个输入，然后通过我们写的程序操作这个Input，通过本身定义好的模型得到一个Output，即为结果。

在运行一个MapReduce任务时，任务被分为map阶段和reduce阶段，针对这个问题，我给出以下的解决方案：

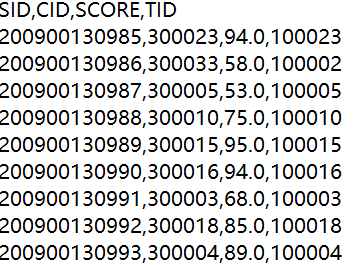
首先我们观察已有数据，为方便数据操作，提前用Python程序将xlsx文件转为csv文件：



在student表中，数据的格式如下：



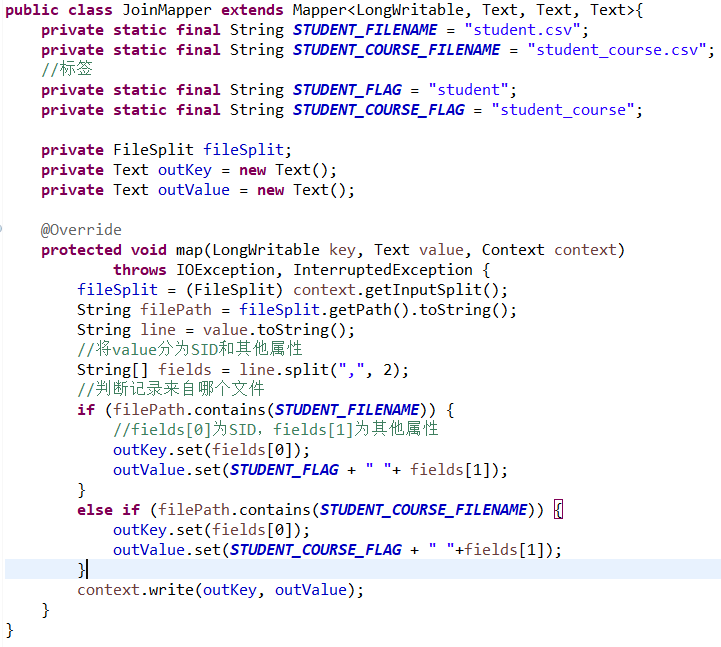
在student\_course表中，数据的格式为：



我们观察到两张表相同的属性值为SID，都在表的第一列。

于是在模型的map函数中，我们要做的事情是将以后数据存为<Key,Value>形式，Key即为SID属性的值，Value是当前行中除SID以外的属性。在存储Value时还要注意给数据打标签，即标明该Value是来源于哪张表，至此map的工作就完成了。

对应代码如下：



这里打标签的方式为在Value的头部加上对应的FLAG并用空格隔开（因为csv文件用逗号隔开，这里用空格加以区分，在reduce程序中好分离FLAG），之后将SID作为Key值，打过标签的其他属性作为Value值写入。

MapReduce框架中map函数的输入value即为文件读入内容，map阶段之后是reduce阶段，对于reduce方法：

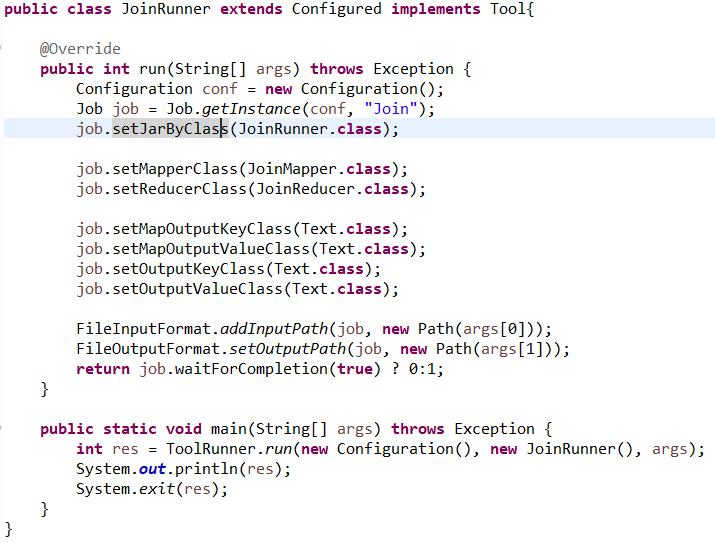


它的输入key是map阶段从context中写入的key，value是一个key对应的一组值，在reduce阶段我们需要做的事情是首先判断value中的值是来自哪张表，如果是来自student中（value的头部为STUDENT\_FLAG），那么就将它存入一个变量，如果是来自student\_course中，那么就将它们存入一个列表中。



对于来自student\_course中的多条数据，它们都对应相同SID的一条数据，于是最后一步就是进行自然连接并写出结果。

最后是Runner代码：



在运行MapReduce之前都要初始化Configuration，该类的主要作用是读取系统配置信息，然后创建一个job，也就是一个计算任务，它的名字是“Join”，接着通过job设置Map类和Reduce类还有输入输出的类型，然后是输入参数，也就是main方法中的args参数，通过eclipse中Run Configuration设置。第一个参数是输入路径（必须存在），第二个是输出路径（不可提前创建）：

