**MapReduce实验报告**

****

**学号：201600301079**

**姓名：崔玉峰**

**班级：2016级**

## 问题描述

1. **实验题目**

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。

1. **实验思路：**

搭建Hadoop平台，然后通过java连接hadoop并且执行Mapreduce操作，将student和student\_course表做自然连接。下面给出Mapreduce实现自然连接的思路：

**Map：**

读数据并且对数据进行筛选和标记：

1. 来自student的数据，设置key为学号，value为姓名+**标记**；
2. 来自student\_course的数据，设置key为学号，value 为选课编

号+考试成绩+**标记**。

**Reduce**

读入已经根据key值学号分好组的数据，values可能来自student

或者student\_course， 根据标记进行判断：

1. 如果来自实体表student，输入的value值的数据为姓名，就将Reduce函数输出key 设置为学号+姓名
2. 如果来自关系表student\_course，输入的value值的数据为选课编号+考试成绩 ，并存入数组。（因为是一对多的关系）
3. 最后遍历数组，将Reduce函数输出key为学号+姓名，输出Value设置为选课编号+考试成绩，然后输出。

## 具体实现：

1. **搭建Hadoop平台**
2. 下载Hadoop,解压到本地目录
3. 配置环境变量
4. 修改hadoop的配置文件

core-sit.xml,

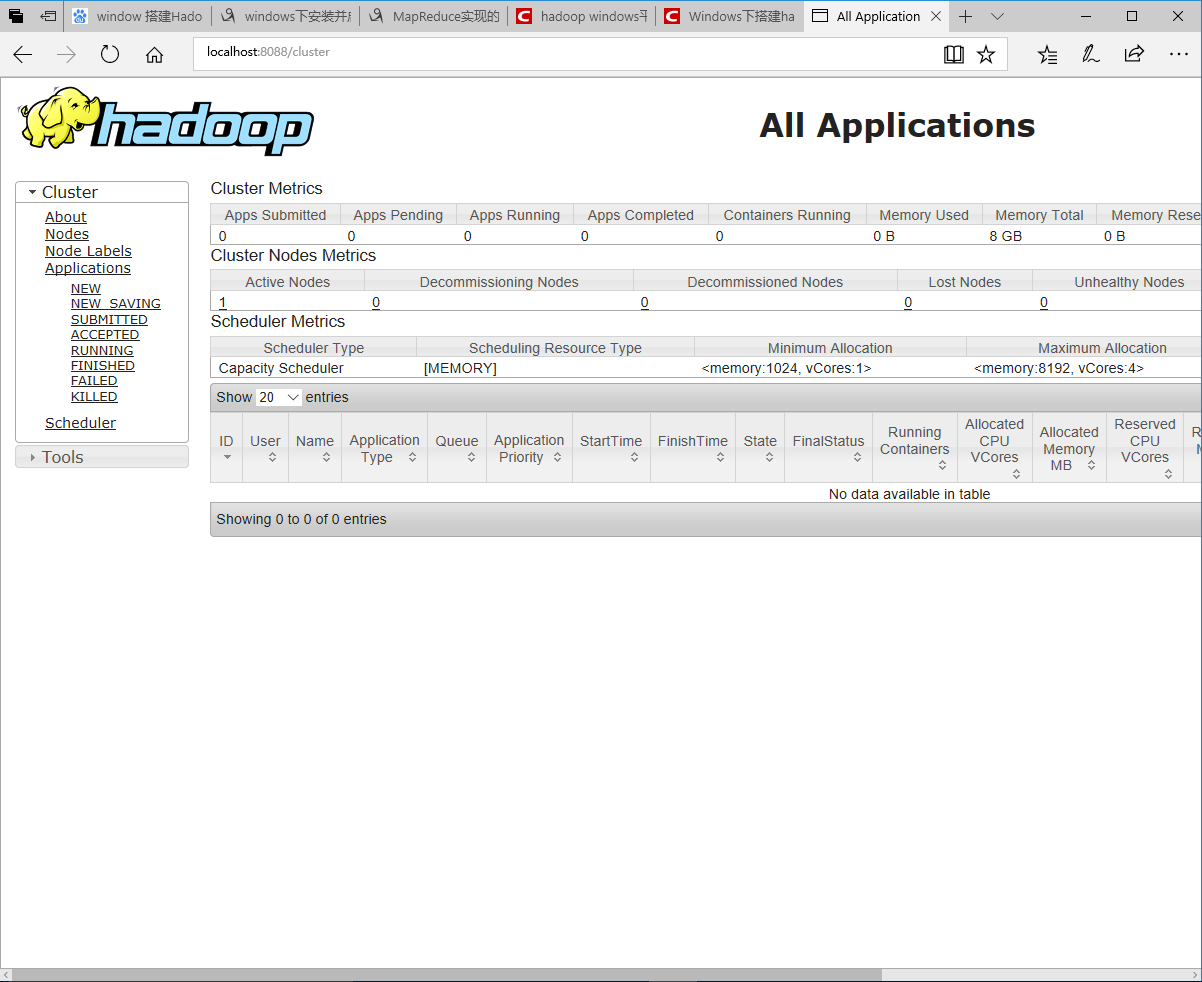
mapred-site.xml,

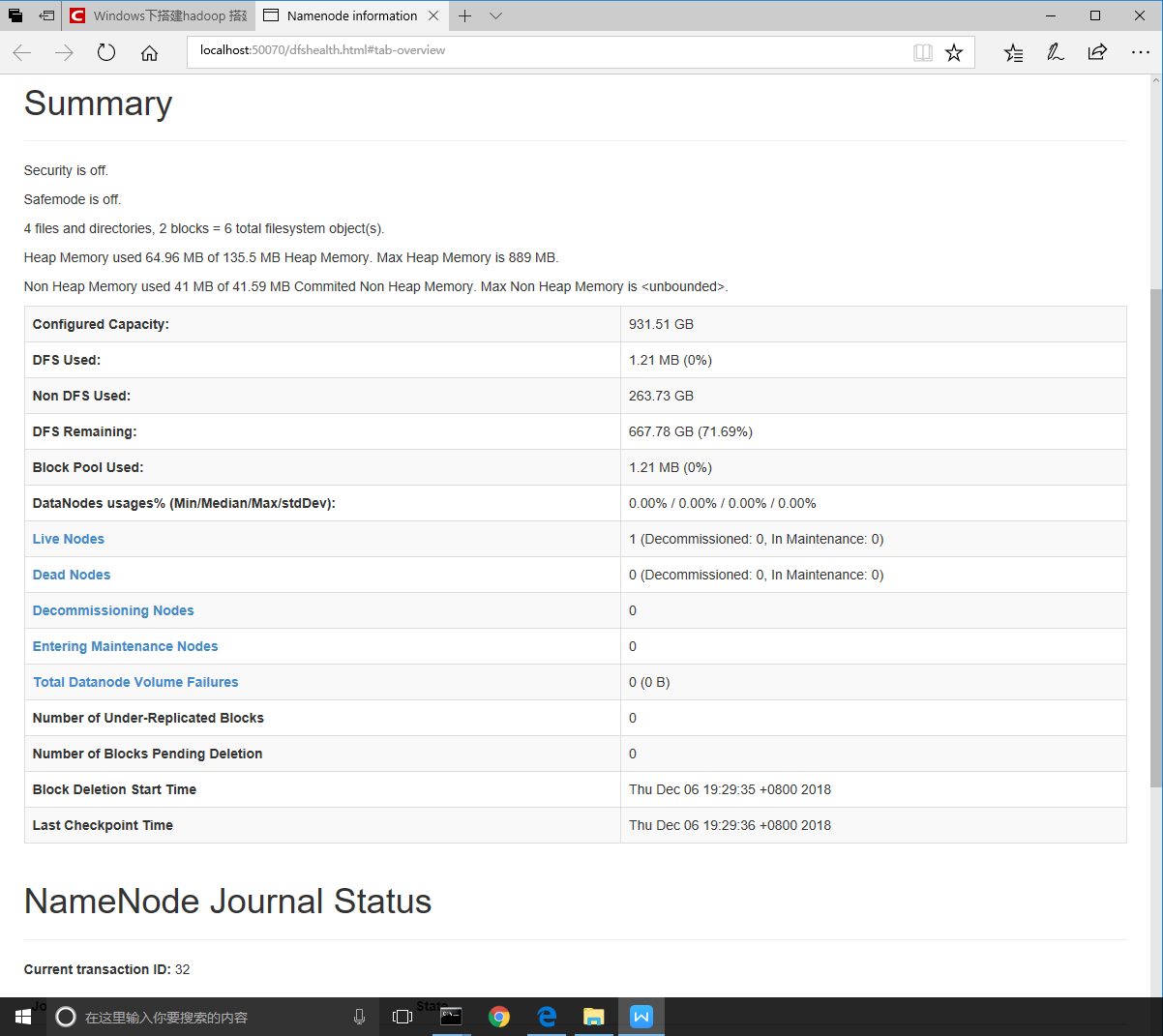
hdfs-site.xml,

yarn-site.xml,

hadoop-env.cmd

1. 通过命令start-all.cmd运行Hadoop平台
2. 通过浏览器访问：





1. **处理数据**
2. 给定的两个文件是xlsx格式的，通过java很难进行处理，所以先将他们转换成csv格式在进行处理。CSV格式直接通过逗号就可以分割开各个字段方便处理。



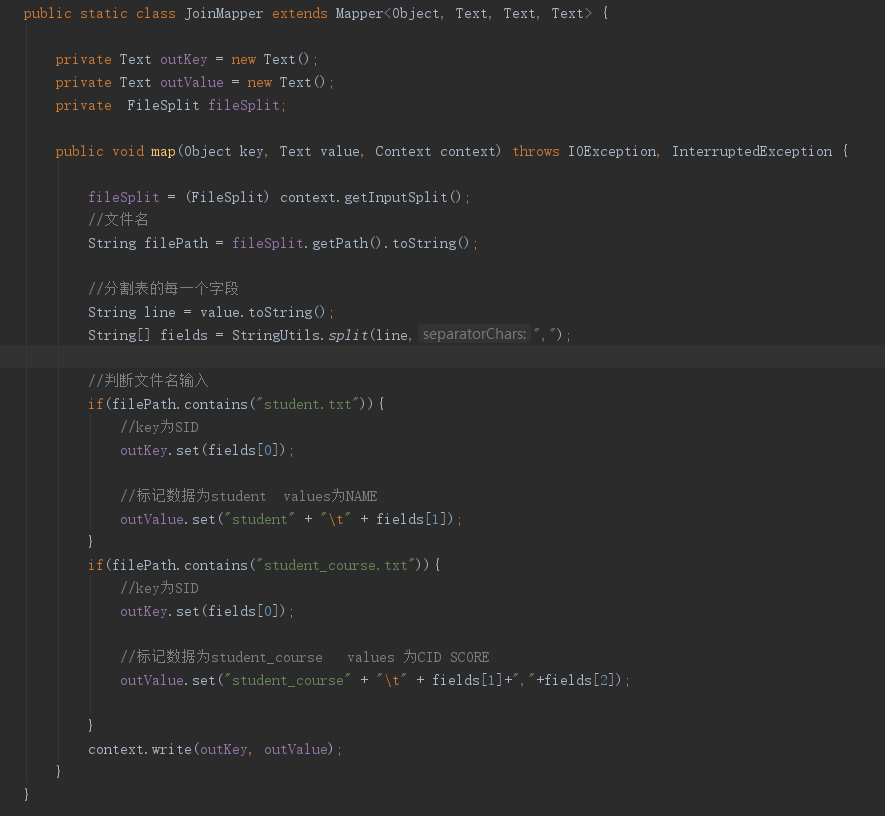
1. 然后通过命令将上述文件上传到hadoop文件平台HDFS上

fs -put student.txt

Fs - put student-course.txt

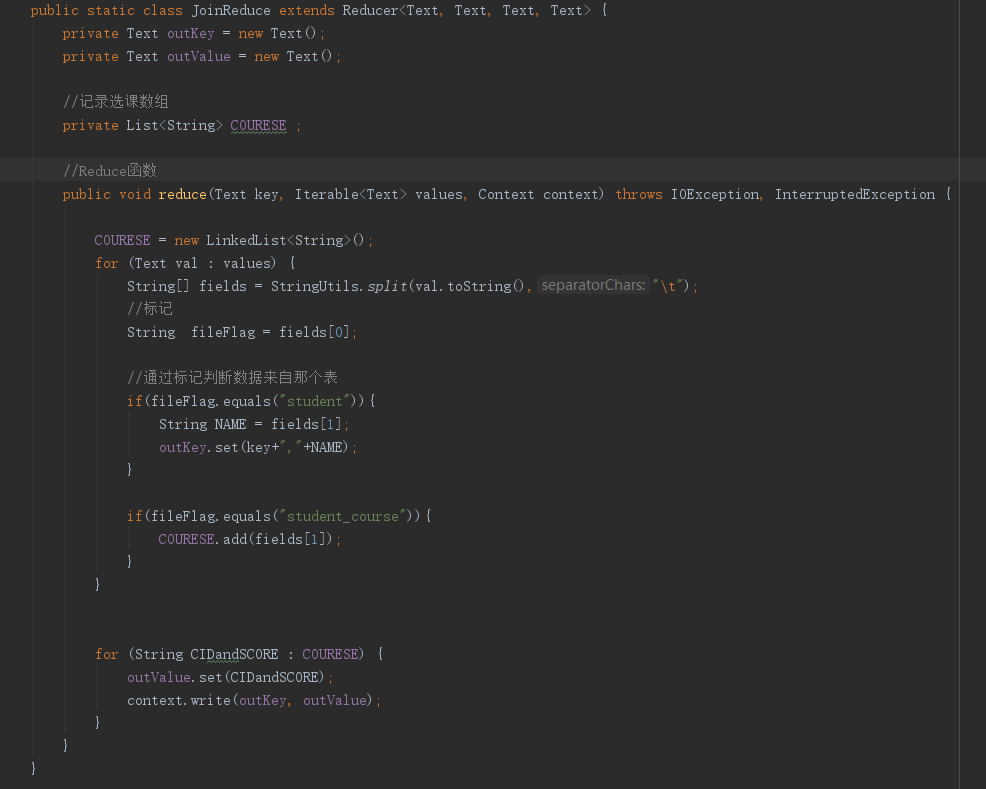
1. **Map函数的实现：**

根据上述的思路就可以通过Hadoop提供的接口编写Map函数具体实现代码如下：



1. **Reduce函数的实现：**

根据上述的思路就可以通过Hadoop提供的接口编写Reduce函数具体实现代码如下：

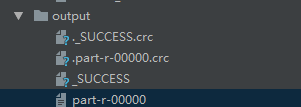


1. **执行Mapreduce做自然连接：**

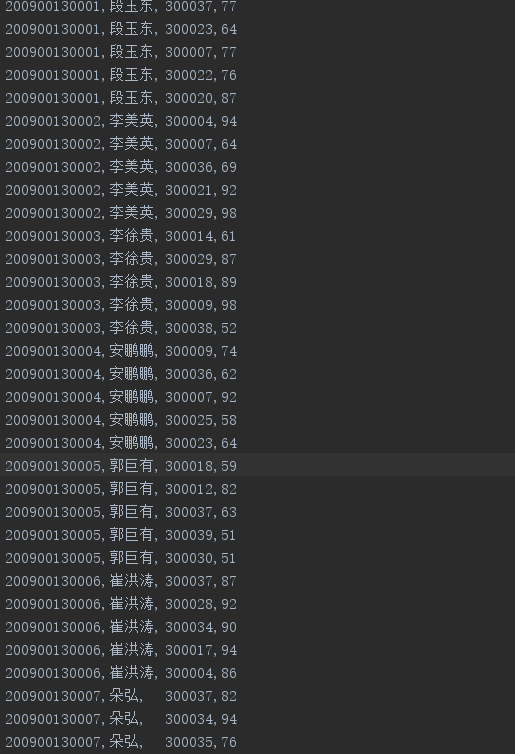
将mapreduce文件输入地址设置为HDFS上的地址，调用代码执行MapReduce操作



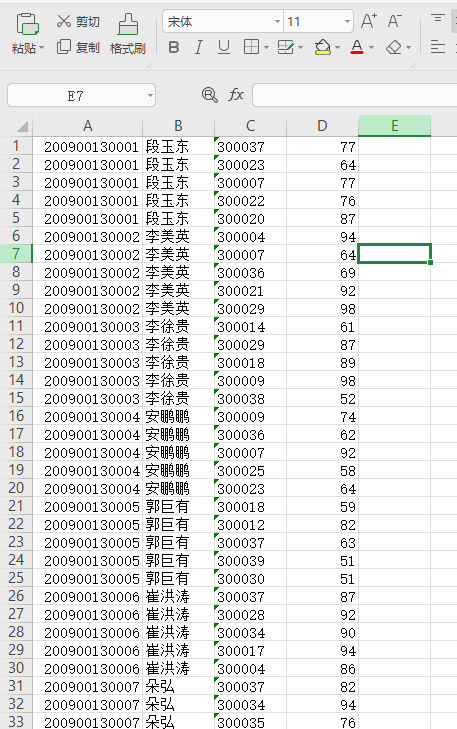
1. **执行结果：**
2. 运行成功会生成几个文件



1. 查看自然连接结果：



1. 可以看到运行成功，为方便查看可以将文件格式改为csv文件方便查看



## 实验结论和感悟：

通过实验学习到了Hadoop平台的搭建和使用，和Mapreduce的基本使用方式，以及如何用JAVA api编写Mapreduce函数；因为对于Hadoop平台和Mapreduce操作都是为了处理分布式存储的海量数据而存在的所以完成此次实验更多的是让我们熟悉MapReduce的基本书写操作。虽然执行的速度没有

本次实验会将连接结果csv文件共同上交