机房上机实验指南

实验 3

MapReduce 编程初级实践

（版本号：2016 年 11 月 14 日版本）

目录

实验 3

MapReduce 编程初级实践

1. 实验目的

1.通过实验掌握基本的 MapReduce 编程方法；

2.掌握用 MapReduce 解决一些常见的数据处理问题，包括数据去重、数据排序和数据挖掘等。

2. 实验平台

已经配置完成的 Hadoop 伪分布式环境。

3. 实验内容和要求

1.编程实现文件合并和去重操作

对于两个输入文件，即文件 A 和文件 B，请编写 MapReduce 程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新的输出文件 C。下面是输入文件和输出文件的一个样例供参考。

输入文件 A 的样例如下：

```
20150101 x
20150102 y
20150103 x
20150104 y
20150105 z
20150106 x
```

输入文件 B 的样例如下：

```
20150101 y
20150102 y
20150103 x
20150104 z
20150105 y
```

根据输入文件 A 和 B 合并得到的输出文件 C 的样例如下：

```
20150101 x
20150101 y
20150102 y
20150103 x
20150104 y
```

```
20150104        z
20150105        y
20150105        z
20150106        x
```

2. 编写程序实现对输入文件的排序

现在有多个输入文件，每个文件中的每行内容均为一个整数。要求读取所有文件中的整数，进行升序排序后，输出到一个新的文件中，输出的数据格式为每行两个整数，第一个数字为第二个整数的排序位次，第二个整数为原待排列的整数。下面是输入文件和输出文件的一个样例供参考。

输入文件 1 的样例如下：

```
33
37
12
40
```

输入文件 2 的样例如下：

```
4
16
39
5
```

输入文件 3 的样例如下：

```
1
45
25
```

根据输入文件 1、2 和 3 得到的输出文件如下：

```
1 1
2 4
3 5
4 12
5 16
6 25
7 33
```

```
8 37
9 39
10 40
11 45
```

3. 对给定的表格进行信息挖掘

下面给出一个 child-parent 的表格，要求挖掘其中的父子辈关系，给出祖孙辈关系的表格。

输入文件内容如下：

```
child parent
Steven Lucy
Steven Jack
Jone Lucy
Jone Jack
Lucy Mary
Lucy Frank
Jack Alice
Jack Jesse
David Alice
David Jesse
Philip David
Philip Alma
Mark David
Mark Alma
```

输出文件内容如下：

```
grandchild grandparent
Steven Alice
Steven Jesse
Jone Alice
Jone Jesse
Steven Mary
```

Steven	Frank
Jone	Mary
Jone	Frank
Philip	Alice
Philip	Jesse
Mark	Alice
Mark	Jesse

4. 实验报告

《大数据工程导论》 实验报告				
题目：		姓名		日期
实验环境：				
解决问题的思路：				
实验内容与完成情况：				
出现的问题：				
解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：				