**分布式操作系统实验课设计**

实验课设计共分为4课：

1. Hadoop环境搭建
2. HDFS文件系统熟悉和使用
3. MapReduce编程模型的使用
4. MapReduce的高级应用

一、Hadoop环境搭建

**基础：**

建议Fedora，Ubuntu也可以

建议jkd1.6之上版本，OpenJDK不行，需先卸载

**过程：**

5位同学组成小组，用自己的笔记本共同组成一个Hadoop机群，本实验按照小组计分。

按照Hadoop官方网站的操作顺序（助教可先打印一份以指导操作）

**需要注意的问题和难点**

按照之前的经验，问题和难点主要出现在下面几个地方：

1. 未为Hadoop设置专门用户，或者使用root用户。这样会在不同的操作系统之上产生不同的错误，很难解决。建议所有同学都专为Hadoop设立用户
2. ssh配置错误非常多。注意几点，首先使用专用用户进行操作；其次使用scp命令拷贝公钥；authorized\_keys文件模式是644
3. 机器之间保证互联，如果是DHCP配置，每次启动Hadoop都需要先确认ip
4. 需先执行命令，格式化文件系统方能启动 Hadoop namenode –format
5. 所有从节点中文件和主节点保持一致。

二、HDFS文件系统的熟悉和使用

**基础：**

安装好的Hadoop机群，至少5台机器构成的机群。

**过程：**

1. 了解日常操作HDFS的基本命令。上传、下载、列目录、移动等
2. 了解Hadoop配置参数中与HDFS相关的参数，**备份系数为3**
3. 通过进入和离开安全模式，了解安全模式的概念
4. 5台服务器中，先启动其中4台，上传大量文件，之后启动第五台加入机群，运行balance命令，观察结果
5. 机群运行过程中，撤掉其中一台机器，观察结果

**难点和问题：**

一般没有太大的问题

三、MapReduce编程框架的熟悉和使用

**基础：**

安装好的Hadoop机群，熟悉HDFS操作之后

**过程：**

1. 助教讲解MapReduce模型的基本理念和在Hadoop中的执行流程
2. 简单讲解word count和sort的原理
3. 本机安装Eclipse Hadoop集成开发环境（可选）
4. 同学在本机编写word count和sort程序，使用eclipse编译或者通过ant脚本编译
5. 使用HDFS上传输入文件
6. 执行自己编写的word count和sort程序

**难点和问题：**

一般没有太大的问题

四、MapReduce的高级应用

**基础：**

准备两个大文件（1G左右），其中一个文件的格式的id,name。

另外一个文件是id,location。利用 mapreduce程序编写找出具有相同id的对应的name和location,输出文件格式为id,name,location。简单起见，id不重复。

比如：第一个文件为1，zhang

2, li

3, wang

第二个文件为1，shanghai

2，guangzhou

3, beijing

输出的结果：1,zhang，shanghai

2,li,guangzhou

3,wang,beijing

**过程：**

1）把两个文件通过map程序分成key-value对。

2）具有相同key的值发送到一个reduce中。

3）在reduce中处理相应的关联操作。

**难点和问题：**

1）编写map和reduce函数。

2）理解hash based partition 的方法。