

实验六 MapReduce的使用

一、实验目的

1. 熟悉MapReduce的编程规范；
2. 熟悉Hadoop使用的数据类型；
3. 熟悉 Map 阶段的实现方法；
4. 熟悉 Reduce 阶段的实现方法；
5. 熟悉 Driver 的实现方法。

二、实验内容

基于伪分布式运行模式或完全分布式运行模式，使用Maven，通过编写 Mapper、Reducer 和 Driver 进行如下实验：

尽量在一个项目下进行实验，不同的功能通过运行时指定不同主类实现。

2.1 WordCount

1. 下载<https://gitee.com/lueyoung/hadoop-empirical-data>项目中的[fail2ban.log](#)文件；
2. 进行需求分析；
3. 编写 Mapper；
4. 编写 Reducer；
5. 编写 Driver；
6. 使用 Maven，创建 Jar 形式 Package；
7. 运行项目；
8. 观察并分析结果。

2.2 排序

1. 下载<https://gitee.com/lueyoung/hadoop-empirical-data>项目中的[nums.txt](#)文件；
2. 进行需求分析；

要求输出的形式为 序号 数值 的形式，如：

```
1    5
2   10
3   56
4   77
5  105
6  220
7 1300
...
```

3. 编写 Mapper；
4. 编写 Reducer；
5. 编写 Driver；
6. 使用 Maven，创建 Jar 形式 Package；

7. 运行项目；

8. 观察并分析结果。