

Hadoop编程

1752124 朴 雪



2020-1-10

同济大学

软件学院

# 平台

运行环境：部署在Dadoop集群

单个主机性能：

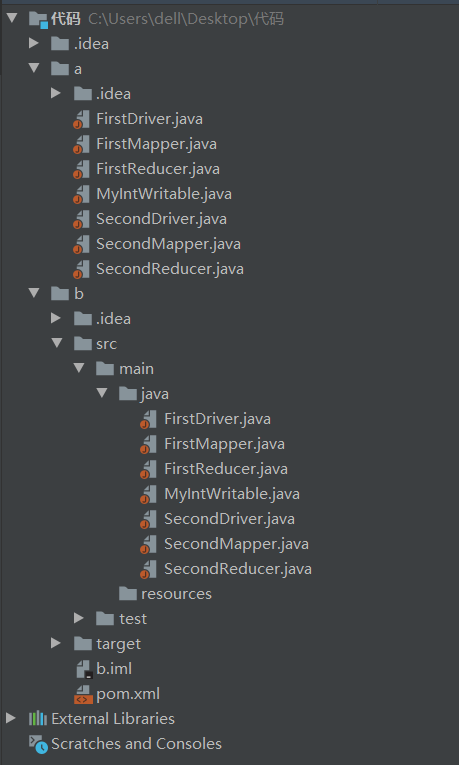
|  |  |  |
| --- | --- | --- |
| CPU | 内存 | 硬盘 |
| 1核 | 1GB | 20GB |

开发环境：IDEA

# 使用说明

将项目代码打包到Hadoop平台上运行。

# 代码结构



文件夹a为任务a的实现代码；

文件夹b为任务b的实现代码。

Driver是主类，Mapper和Reducer是自定义的MapReducer程序，MyIntWritable是自定义序列化类型。

每个程序都分为两个MapReducer，要先执行FirstDriver，然后将FirstReducer的输出作为SecondReducer的输入执行SecondDriver.

其中，MapReducer中的Shuffle默认为升序，因此为了实现降序排列（计数量由大到小）--自定义MyIntWritable使词频按降序排列。

# 提交说明

代码及在Hadoop平台上的运行结果已经在压缩包中存储，这里不再赘述。a, b文件夹分别为a, b任务生成的.txt格式词频排序结果，c文件夹为两个不同节点程序运行的时间统计结果，作图，及分析