Lab 0 Map-Reduce Task

中国人民大学 王大林 sxwangdalin@ruc.edu.cn

实验结果

1. 完成 Map-Reduce 框架

运行结果截图如下

```
Case7 PASS, dataSize=1GB, nMapFiles=60, cost=57.2940855s
Case8 PASS, dataSize=1GB, nMapFiles=60, cost=52.2311004s
Case9 PASS, dataSize=1GB, nMapFiles=60, cost=44.4832323s
Case10 PASS, dataSize=1GB, nMapFiles=60, cost=46.9629053s
--- PASS: TestExampleURLTop (1059.75s)
PASS
ok      talent  1059.919s
```

测试通过！总用时1059.919s

2. 基于 Map-Reduce 框架编写 Map-Reduce 函数

运行结果截图如下

```
Case5 PASS, dataSize=1GB, nMapFiles=60, cost=2.8324922s
Case6 PASS, dataSize=1GB, nMapFiles=60, cost=2.2833331s
Case7 PASS, dataSize=1GB, nMapFiles=60, cost=3.0798512s
Case8 PASS, dataSize=1GB, nMapFiles=60, cost=3.8696515s
Case9 PASS, dataSize=1GB, nMapFiles=60, cost=4.3931049s
Case10 PASS, dataSize=1GB, nMapFiles=60, cost=2.865301s
PASS
ok      talent  139.656s
```

测试通过！总用时139.656s

实验总结

1. debug总结

- 预留磁盘空间不足。实测本测试生成的tmp文件夹需要47.7GB磁盘空间，开始时因为磁盘空间不足而报错

```
Case10 PASS, dataSize=500MB, nMapFiles=40, cost=31.1902837s
Case0 PASS, dataSize=1GB, nMapFiles=60, cost=1m19.3000663s
2022/01/09 15:30:13 write /tmp/mr_homework/case1-1GB-60/mrtmp.Case1-Round0-44-3: There is not enough space on the disk.
exit status 1
FAIL    talent  528.985s
```

- 运行时间超出go test的默认最大等待时间

开始时遇到这样的报错信息

```
Case0 PASS, dataSize=1GB, nMapFiles=60, cost=52.8369335s
Case1 PASS, dataSize=1GB, nMapFiles=60, cost=50.6658313s
Case2 PASS, dataSize=1GB, nMapFiles=60, cost=49.1173202s
Case3 PASS, dataSize=1GB, nMapFiles=60, cost=48.3126884s
panic: test timed out after 10m0s

goroutine 1480 [running]:
testing.(*M).startAlarm.func1()
        D:/codeEnv/Go/src/testing/testing.go:1618 +0xe6
created by time.goFunc
        D:/codeEnv/Go/src/time/sleep.go:167 +0x4b

goroutine 1 [chan receive, 10 minutes]:
testing.(*T).Run(0xc000029680, 0x764334, 0x11, 0x76e170, 0x663001)
        D:/codeEnv/Go/src/testing/testing.go:1169 +0x2da
testing.runTests.func1(0xc000029500)
        D:/codeEnv/Go/src/testing/testing.go:1439 +0x7f
testing.tRunner(0xc000029500, 0xc00006fde0)
```

调查发现是因为go test 设定了默认的最大运行时间为10min，而由于机器性能的限制，本次测试的最大运行时间可能会超过10min，因此，需要通过`-testout 30m`来指定时间，防止因为超时而报错。

2. Map-Reduce总结

Example版本的方法如下：

- 第一轮
  - map：把<url，“”>这样的组合全部存储下来
  - reduce：把url相同的“”字符的个数记为value
- 第二轮
  - map：<"", "url count">
  - reduce：构建map，把用空格分隔的url和count写入，随后排序取top10

对于上面的做法可以做如下改进：

- 在第一轮map中就可以开始进行局部count，然后在reduce中把各个局部和累加为最终的count
- 在第二轮map中，只输出局部top10，在reduce时仍就能保证结果不变