大数据处理综合实验课程（2025）

课程实验4-MapReduce高级编程技术

南京大学　　计算机学院

任务1：社交网络中的互相关注好友

- 输入数据为某社交平台关注列表。该文件由若干行组成，每一行由user_name: 表示用户user_name的关注列表，关注列表通过’ ’进行分隔。

```
7: 216 142 104 194 233 217 203 219 197 195 67 23 121 180 58 85 42
8: 146 189 25 204 111 60 13 215 122 35 141 132 88 28 166 6 33 184 20 1 40 221 39 234 74 116 109 84 15 31 161 81 147 155 140 12 127 227 47 65 105 62 79 90 148 137 170 76 158 97 58 123 167 45 51 230 162 37
9: 185 188 22 222 213 193 198 181 17 36 66 209 78 119 169 200
10: 25 192 12 205 236 167 162
11: 72 61 146 163 4 177 138 212 151 226 117 133 18 154 176 33 44 89 232 224 114 145 106 199 34 168 3 92 225 180 112 42 174 120
12: 146 25 201 204 60 215 115 10 122 35 153 95 91 38 132 88 28 166 218 156 6 2 8 231 165 184 20 164 1 40 221 128 39 48 84 15 31 161 81 147 155 140 160 127 227 47 65 205 105 62 182 79 90 100 148 137 170 76 158 97 58 134 123 167 45
```

- 根据每个用户的关注列表找到互相关注的用户对，形式为’user_A-user_B’

```
1-100
1-105
1-108
1-109
1-12
1-123
1-127
1-128
1-134
1-137
1-139
1-140
1-146
1-147
1-148
1-15
```

任务2：社交网络寻找共同关注

- 任务1中得到了互相关注的用户对，根据这两个用户的关注列表，找到两个用户的共同关注。
  要求：通过configuration传递参数x，并且将共同关注数量小于等于x的共同关注列表输出到一个文件，将共同关注数量大于x的共同关注列表输出到另一个文件。(x可以取[10,20])
- 如图为两个文件（部分）

```
1-189: 45 85 21 140 8
1-238: 140 88 15 161 48 105 21 84 45
10-12: 167 25 205
10-162: 25 167 236
10-167: 25 205 12 162
10-192: 236 205
10-205: 236 12 192 25 167
10-236: 162 205 192 25
10-25: 167 236 12 162 205
```

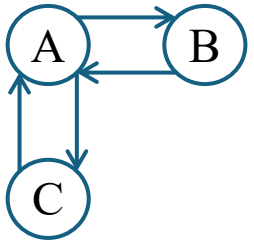
```
1-100: 179 12 60 128 231 204 148 88 156 182 62 85 218 65 215 229 2 39 54 91 221 139 140 51 108 184 47 147 21 170 48 79 207 15 160 158 155 123 90 201 109 146 75 165 20 6 161 227 134 167 84 58 105
1-105: 91 2 166 128 231 41 84 167 204 20 134 161 6 60 148 156 88 165 47 146 45 100 109 201 90 123 155 137 238 12 62 15 85 207 218 8 127 227 79 48 170 21 147 184 108 35 65 140 215 139 221 229 54 39
1-108: 158 88 146 182 21 45 218 54 79 215 221 51 100 62 155 207 85 65 231 167 128 229 201 109 184 6 90 20 48 139 15 170 156 127 91 140 160 137 2 39 58 165 204 161 179 60 105 84
1-109: 137 207 156 218 8 39 127 227 229 79 88 21 48 170 2 134 58 105 108 161 146 51 45 60 47 100 231 201 184 6 85 90 128 155 215 167 140 182 148 179 91 158 166 160 84 15 54 139
1-12: 137 134 65 161 6 148 155 127 88 28 51 75 146 45 100 201 35 47 85 90 123 215 140 182 165 156 158 160 15 62 2 8 218 227 79 48 170 39 221 147 184 166 91 128 105 60 231 84 58 167 204 20
1-123: 60 105 231 204 137 148 88 45 100 182 62 65 215 91 39 166 221 140 51 127 147 184 170 48 227 8 15 160 158 165 47 90 201 75 28 155 161 134 20 167 58 84 41 12 128
1-127: 51 109 179 148 91 139 167 54 128 184 170 123 15 158 6 84 182 88 166 28 45 58 221 65 108 146 160 39 85 41 20 105 227 47 62 147 12 165 204 75 207 2 79 60 140 137 90 48 201 155 229 35 215 8 134 218 156
161
1-128: 137 179 12 166 165 156 91 221 146 231 201 229 47 123 139 15 88 184 28 62 160 105 2 140 170 215 218 51 75 167 148 207 182 127 84 90 6 39 20 108 45 54 58 227 147 65 100 155 79 204 134 21 48 109 85 161
60
```

任务3：好友推荐

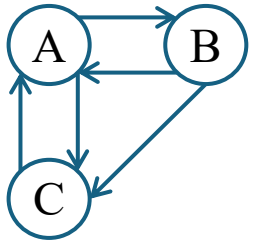
- 若userA和userB为互相关注的好友， userA和userC为互相关注好友但是userB没有关注userC，则社交网站会推荐userB关注userC，请设计程序以完成社交网络的好友推荐，输出每个用户推荐列表中前5的好友，若数量不足5个则全部输出。（注：可能会出现以下三种情况）

- 输出：

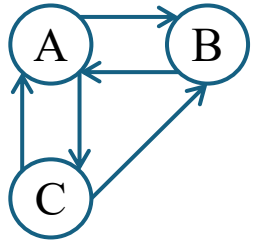
```
13   88,195,230,110,198
131  88,46,47,190,191
132  192,196,199,233,234
133  49,116,237,119,98
134  192,195,111,199,233
135  66,13,36,69,17
136  67,159,206
137  89,190,192,110,112
138  116,237,118,119,98
139  192,194,111,5,7
14   45,190,151,195,110
140  190,196,199,233,119
141  192,194,195,234,116
142  45,193,151,152,197
143  194,233,237,50,53
144  77,89,58,190,175
```



向B推荐C，
且向C推荐B



不向B推荐C，
但向C推荐B



向B推荐C，但
不向C推荐B

实验数据

- 本次实验使用的是社交朋友圈，每个文件代表一个朋友圈，其中的数据格式为:

  user_name: friend1 friend2 friend3 ....

- 单机测试样例：单机测试样例为部分数据集，可在“本科教学支撑平台”中下载。该数据集主要供本地调试使用。
- 全部数据集：全部数据集位于集群的 HDFS 存储上，HDFS 存储位置为：/user/root/Exp4

实验报告要求

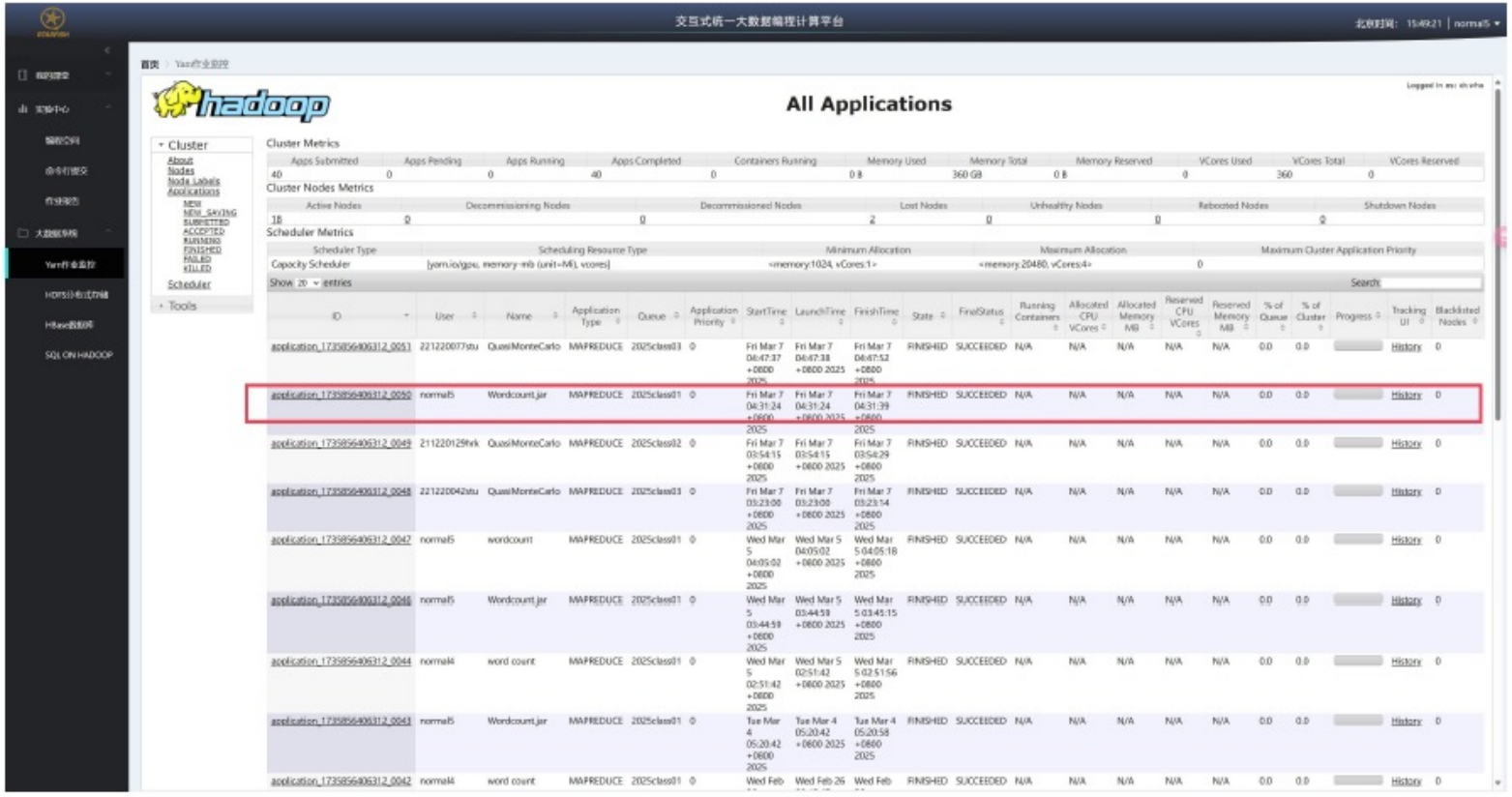
- 实验报告要求提交一个压缩包，压缩包内除了包含源代码、JAR 包、JAR 包执行方式说明，还需要包含一个实验报告。实验报告中包含：
  - Map 和 Reduce 的设计思路（含 Key、Value 类型）。
  - MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
  - 输出结果文件的部分截图。输出结果文件在 HDFS 上的路径（某些情况下助教会检查 HDFS 上的输出文件）。
  - 在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容（请完整包括执行报告内容，以表明该 Job 是在集群上实际执行过的，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。

- 实验报告文件命名规则： 学号_姓名_BG_4
- 实验报告提交至：本科教学支撑平台 http://cslabcms.nju.edu.cn/
- 实验提交截止日期：5月6日（包含当天，实验时长共3周）

WebUI执行结果：

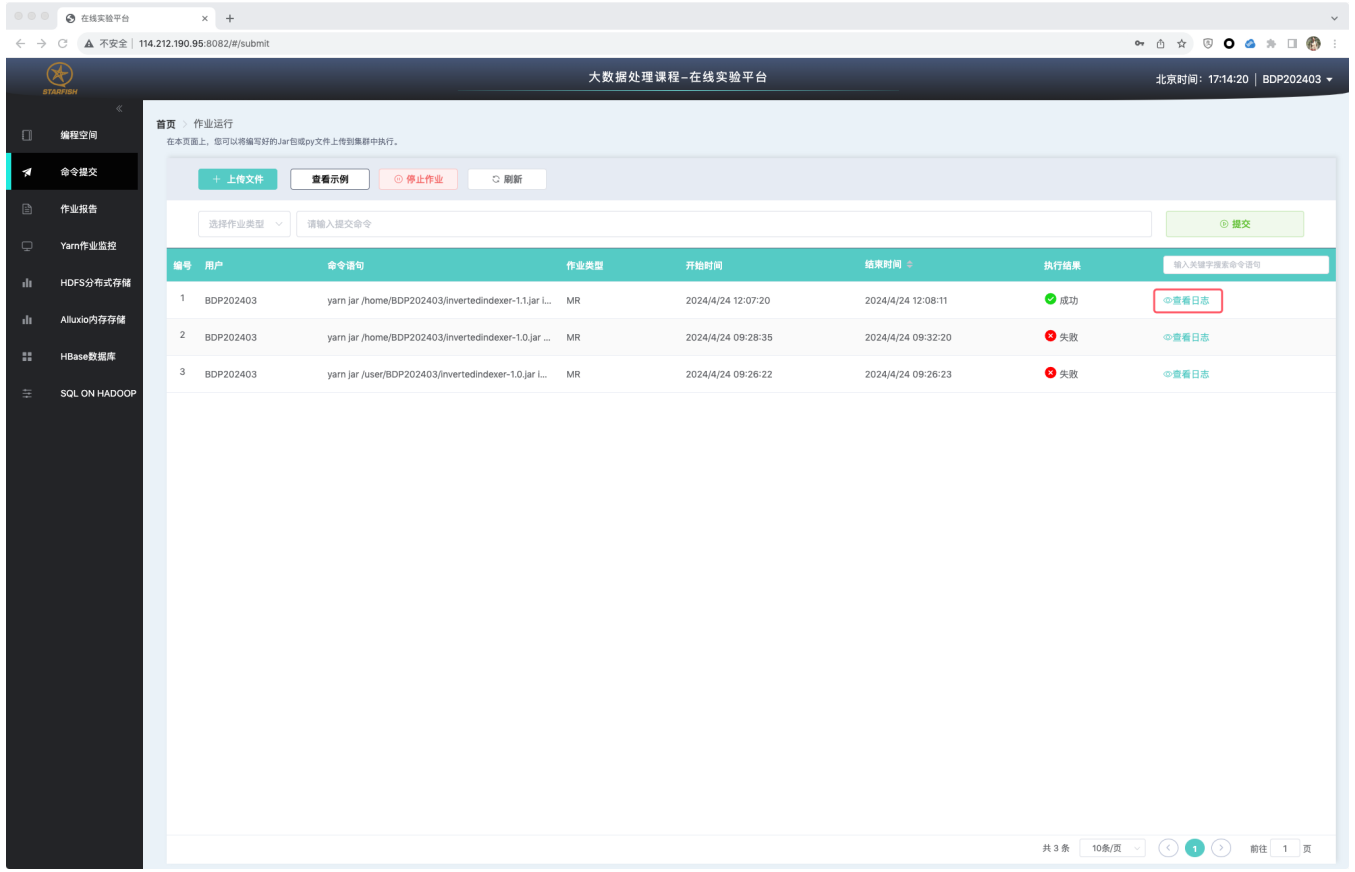
- 在以后的实验报告中，若需要在集群上执行 MapReduce Job，请在报告中附带上相关的 MapReduce Job 的执行报告，以作为评分依据。否则在评分时将认为该 MapReduce Job 没有在集群上执行，会影响实验得分。
- 校园网访问实验平台 http://114.212.130.202:8082/
- 参考资料：《大数据平台使用手册.pdf》
- 输入账户和密码，点击左侧栏“Yarn作业监控”，可以进入集群监控页面（见下图）。

WebUI执行结果：


All Applications
Cluster
Cluster Metrics
Cluster Nodes Metrics
Scheduler Metrics

WebUI执行结果：

- 或者在“命令行提交”页面“查看日志”，并截图



WebUI执行结果：

- 或者在“命令提交”页面“查看日志”，并截图