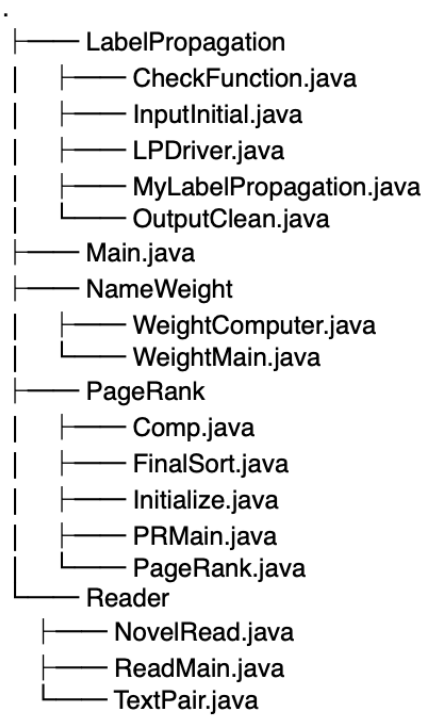
大数据课程设计实验报告

181850055 黄嘉怡 Task1 2 3

181860127 姚逸斐 Task4

181870290 周心瑜 Task5

项目代码结构



`Reader`：读取文本，提取名字，进行人物同现关系统计。

`NameWeight`：计算归一化的关系权重。

`PageRank`：根据人物关系权重计算PageRank值。

`LabelPropagation`：根据人物关系权重进行标签传播。

`Main.java`：整合函数。

任务1 数据预处理

任务简介

从原始的哈利波特系列小说文本中，抽取与人物互动相关数据，并屏蔽与人物姓名无关的文本内容。

设计思路

本次任务完成内容比较简单，为了提高效率，与任务2在一次mapreduce中完成。设计思路见任务2。

任务2 特征抽取：人物同现统计

任务简介

根据提取任务1提取的人名列表，统计在原文中姓名对的同现次数。

设计思路

- Setup

  读取用户自定义文本，形成自定义字典DicLibrary

- Map

  ```
  Input key : OFFSET
  Input value : 一行小说内容

  Output key : <name1,name2>
  Output value : 1
  ```

  调用`DicAnalysis.parse(String)`函数得到分词结果；

  对分词结果进行一次循环，提取并存储符合用户自定义词典中的内容；对同一人物的重名处理也在此进行，判断姓名重复时，将提取的内容替换为同一个；

  对任意两个不同姓名对对之间进行统计并传出。

- Combine

  由于一段文本中有联系紧密的人物关系，很有可能出现多次姓名对，因此为了提高分析效率加入Combiner，合并重复姓名对的次数。

- Reduce

  ```
  Input key : <name1,name2>
  Input value : times1,times2,...

  Output key : name1,name2,times
  Output value : NULL
  ```

  累加value中的次数，合并后传出。

程序运行截图

任务3 特征处理：人物关系图构建与特征归一化

任务简介

根据任务2统计的人物共现次数，生成归一化权重后的人物关系图的临接表表示。

在人物关系图中，人物是顶点，人物之间的互动关系是边，人物之间的互动关系通过人物的共现关系体现。

设计思路

- Map

```
Input key : OFFSET
Inout value : name1,name2,times

Output key : name1
Output value : name2,times
```

  获取姓名对和对应一个互动关系的出现次数。

- Reduce

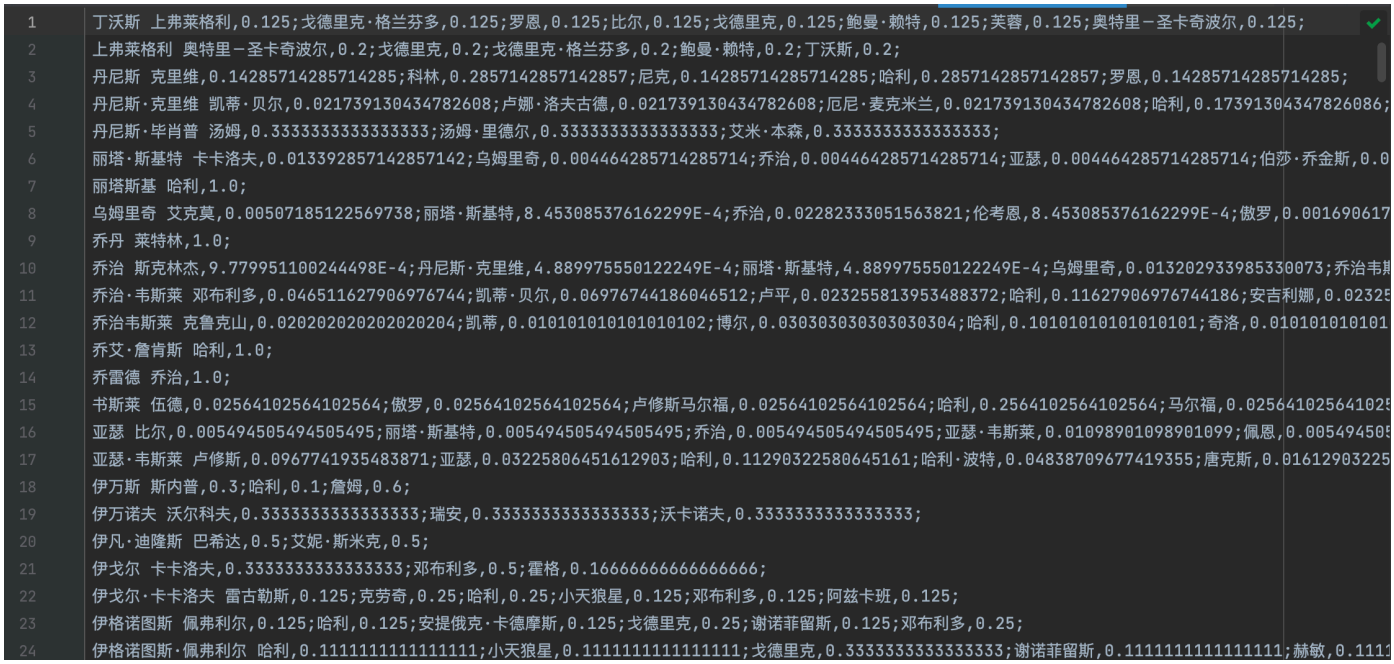
```
Input key : name
Input value : [name1,times],[name2,times]...

Output key : name name1,weight1;name2,weight2;...
Output value : NULL
```

  统计人物对应的互动关系总数；

重新计算所有互动关系占总数的比值并插入字符串。

程序运行截图



任务4 数据分析：基于人物关系图的PageRank计算

任务简介

对于任务3输出的归一化权重后的任务关系图，进行数据分析，计算PageRank值（后文简称为PR值），并对人物的PR值进行全局排序，从而定量地分析出哈利波特系列小说的“主角”们是哪些。

设计思路

本任务的解决方法可以划分为3个阶段，分别是初始化各人物的PR值，迭代计算各人物的PR值，排序最终的PR值并输出。

初始化各人物的PR值

任务3的输出格式为`name name,weight;name,weight;...`，不便于计算PR值，第一次初始化所有人物PR值为1。

- Map思路

  Map接收任务3的输入，发送的key为name，value为空格分隔符之后的list。

- Reduce思路

  Reduce将作为key的人物PR值设为1，发送的key依旧为原来的name，发送的value在原先value头部加入1和分隔符“#”。

迭代计算各人物的PR值

- Map思路

  遍历每条的value，计算每条作为key的人物对list中各人物的PR值的贡献值，也就是当前key人物的PR值乘以list中对应的weight，发送的key为list中的对应人物，value为原先作为key的人物对其的贡献值格式为`name*#contribution`，同时也需要发送`name name,weight;name,weight...`此格式的信息便于后续迭代。

- Reduce思路

  如果接收到`name*#contribution`的信息，则累加起来得到sum，并在最后引入阻尼系数（默认为0.85）计算：

  > 1 - 0.85 + 0.85 * sum

  如果接收到`name name,weight;name,weight...`，则获取与该人物相关联的list。

  最后发送key为该人物，value为newPR+#+list，格式为`name newPR#name,weight;name,weight...`

排序最终的PR值并输出

- Map思路

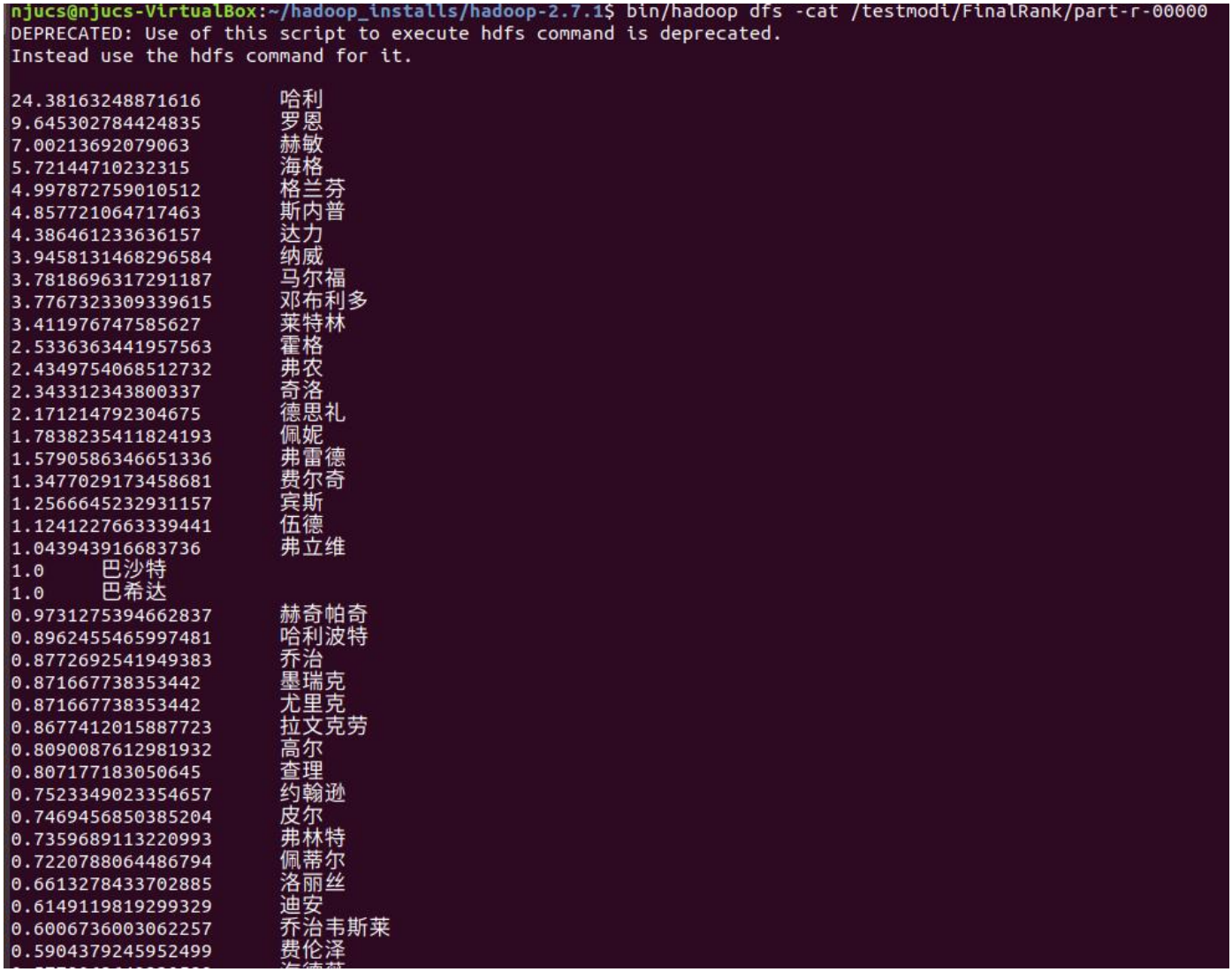
  负责将`name PR#name,weight;name,weight...`转换成`name PR`，发送的key为PR，value为name。

- Reduce思路

  接收的key为PR，value为相同的PR值的人物list，发送的key为PR，value为人物。

最后需要对key也就是PR进行全局排序，在这里自定义Comparator使之为降序排序。

程序运行截图



任务 5 数据分析：在人物关系图上的标签传播

任务介绍

标签传播(Label Propagation)是一种半监督的图分析算法，他能为图上的顶点打标签，进行图顶点的聚类分析，从而在一张类似社交网络图中完成社区发现(Community Detection)。

设计思路

在设计算法时，标签传播任务可以分为三个子任务：初始化信息处理、标签传播算法计算出最终信息、处理人物最终的标签输出最终格式。

初始化信息处理

- Map思路

  Map将任务3的输入划分为每一条边的单独信息并给邻居结点附带上初始化标签信息。Map的输出key为本结点的name，value为邻居节点的信息，可以记为`<neighbor_info>`，格式如下：

  ```
  <name>,<label>,<weight>;
  ```

其中name为人物的名称，每一个人物的label初始化值为自身的name，weight为任务三中输入的权重。

- Reduce思路

  Reduce把所有邻居结点（与本结点人物有关系的人物）信息进行整合，把无向图的信息以以下格式保存：

  ```
  <label>&<neighbor_info><neighbor_info>...&<max_weight>
  ```

  其中label为本结点的初始化标签，即自身的name；neighbor_info为Map中的输出的value；初始化的最大权重表示的含义为此结点标签接管本结点的权重，初始值为0。

标签传播算法

因为异步标签传播的算法对于处理标签信息的步骤有较强的依赖关系，所以我们采用同步算法设计Hadoop框架下的具体实现。同时，还需要考虑到同步算法在一定的图中会产生振荡而无法收敛，故引入 `MAX_EPOCH_NUM` 限制最大迭代次数。其中，当epoch为t时，停止迭代的条件为：

$$\forall\, i, weight_t^{(i)} \leq weight_{t-1}^{(i)}$$

其中，$weight_t^{(i)}$表示第t次迭代，人物i标签所占的权重。

- Map思路

  计算当前人物关系中所有标签的权重记录在HashMap中。找到最大支配的标签和标签权重，将人物信息中label与max_weight两项值更新，将新的无向图信息输出给当前结点。key为当前人物name，value为更新后的无向图信息。

  同时，将自身结点更新的标签信息输出给所有邻居结点，便于在reduce过程中更新邻居结点标签信息。key为所有邻居结点的name，value格式如下：

  ```
  <name>/<new_label>
  ```

  其中name为当前人物名称，new_label为当前人物新的标签。

- Reduce的思路

  通过特殊字符 `/` 来区分接受的消息为标签更新消息还是无向图边信息。更新标签后输出key为当前人物name，value为更新过后的无向图边信息。

最终的标签输出

- Map思路

  将最终的无向图信息转变为最终输出格式，格式如下：

  ```
  <name> <label>
  ```

  输出key为name，value为label。

- Reduce思路

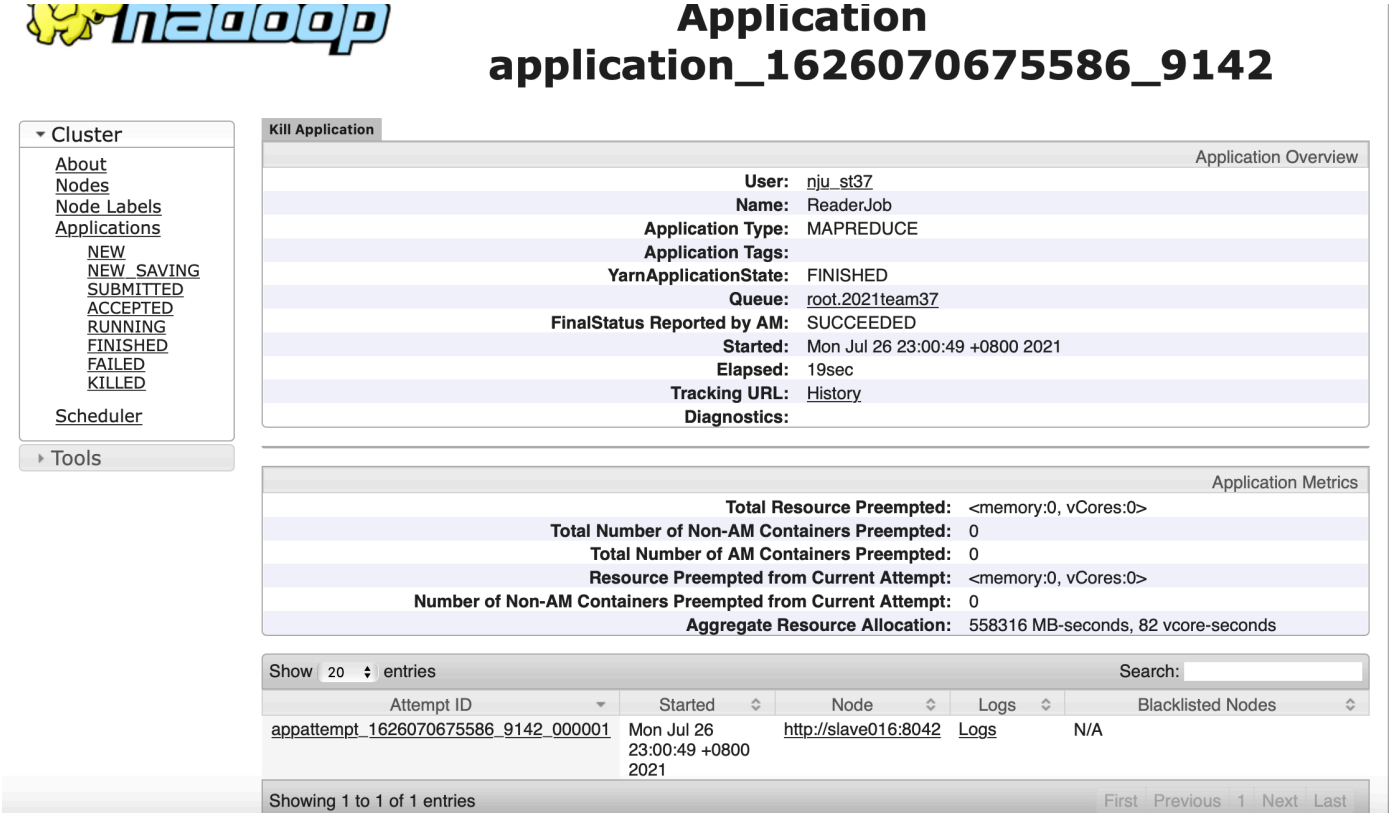
  将接收到key直接输出，value为list但只有一个元素即为当前人物的最终label。

平台运行截图

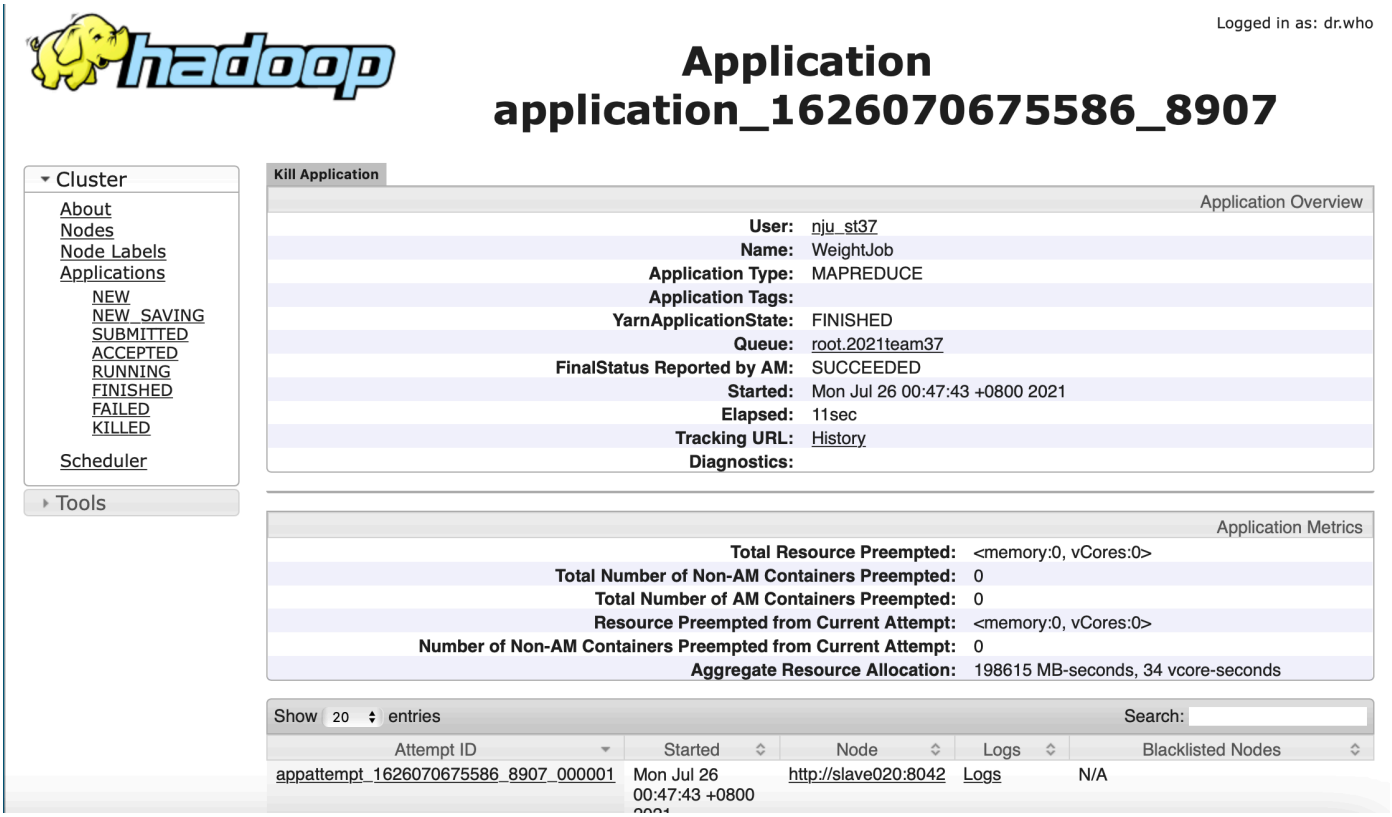
MapReduce运行记录：

	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tr
_8921	nju_st37	MyLabelPropagation	MAPREDUCE	root.2021team37	Mon Jul 26 00:52:18 +0800 2021	Mon Jul 26 00:52:34 +0800 2021	FINISHED	SUCCEEDED		Hi
_8920	nju_st37	LPInitialize	MAPREDUCE	root.2021team37	Mon Jul 26 00:52:01 +0800 2021	Mon Jul 26 00:52:16 +0800 2021	FINISHED	SUCCEEDED		Hi
_8919	nju_st37	PageRank Final Sort	MAPREDUCE	root.2021team37	Mon Jul 26 00:51:40 +0800 2021	Mon Jul 26 00:51:57 +0800 2021	FINISHED	SUCCEEDED		Hi
_8918	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data10	MAPREDUCE	root.2021team37	Mon Jul 26 00:51:20 +0800 2021	Mon Jul 26 00:51:36 +0800 2021	FINISHED	SUCCEEDED		Hi
_8917	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data9	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:59 +0800 2021	Mon Jul 26 00:51:16 +0800 2021	FINISHED	SUCCEEDED		Hi
_8916	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data8	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:39 +0800 2021	Mon Jul 26 00:50:54 +0800 2021	FINISHED	SUCCEEDED		Hi
_8915	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data7	MAPREDUCE	root.2021team37	Mon Jul 26 00:50:18 +0800 2021	Mon Jul 26 00:50:33 +0800 2021	FINISHED	SUCCEEDED		Hi
_8914	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4//Data6	MAPREDUCE	root.2021team37	Mon Jul 26	Mon Jul 26 00:50:14	FINISHED	SUCCEEDED		Hi
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	26 23:03:53 +0800 2021	23:04:12 +0800 2021	[illegible]	[illegible]		His
6_9148	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4/Data2	MAPREDUCE	root.2021team37	Mon Jul 26 23:03:32 +0800 2021	Mon Jul 26 23:03:50 +0800 2021	FINISHED	SUCCEEDED		His
6_9147	nju_st37	PRIter,out:/user/nju_st37/FinalLab/Task4/Data1	MAPREDUCE	root.2021team37	Mon Jul 26 23:03:12 +0800 2021	Mon Jul 26 23:03:29 +0800 2021	FINISHED	SUCCEEDED		His
6_9146	nju_st37	Initialize for PageRank	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:52 +0800 2021	Mon Jul 26 23:03:09 +0800 2021	FINISHED	SUCCEEDED		His
6_9145	nju_st37	WeightJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:31 +0800 2021	Mon Jul 26 23:02:49 +0800 2021	FINISHED	SUCCEEDED		His
6_9144	nju_st37	ReaderJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:02:08 +0800 2021	Mon Jul 26 23:02:27 +0800 2021	FINISHED	SUCCEEDED		His
6_9143	nju_st37	WeightJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:01:11 +0800 2021	Mon Jul 26 23:01:28 +0800 2021	FINISHED	SUCCEEDED		His
6_9142	nju_st37	ReaderJob	MAPREDUCE	root.2021team37	Mon Jul 26 23:00:49 +0800 2021	Mon Jul 26 23:01:09 +0800 2021	FINISHED	SUCCEEDED		His

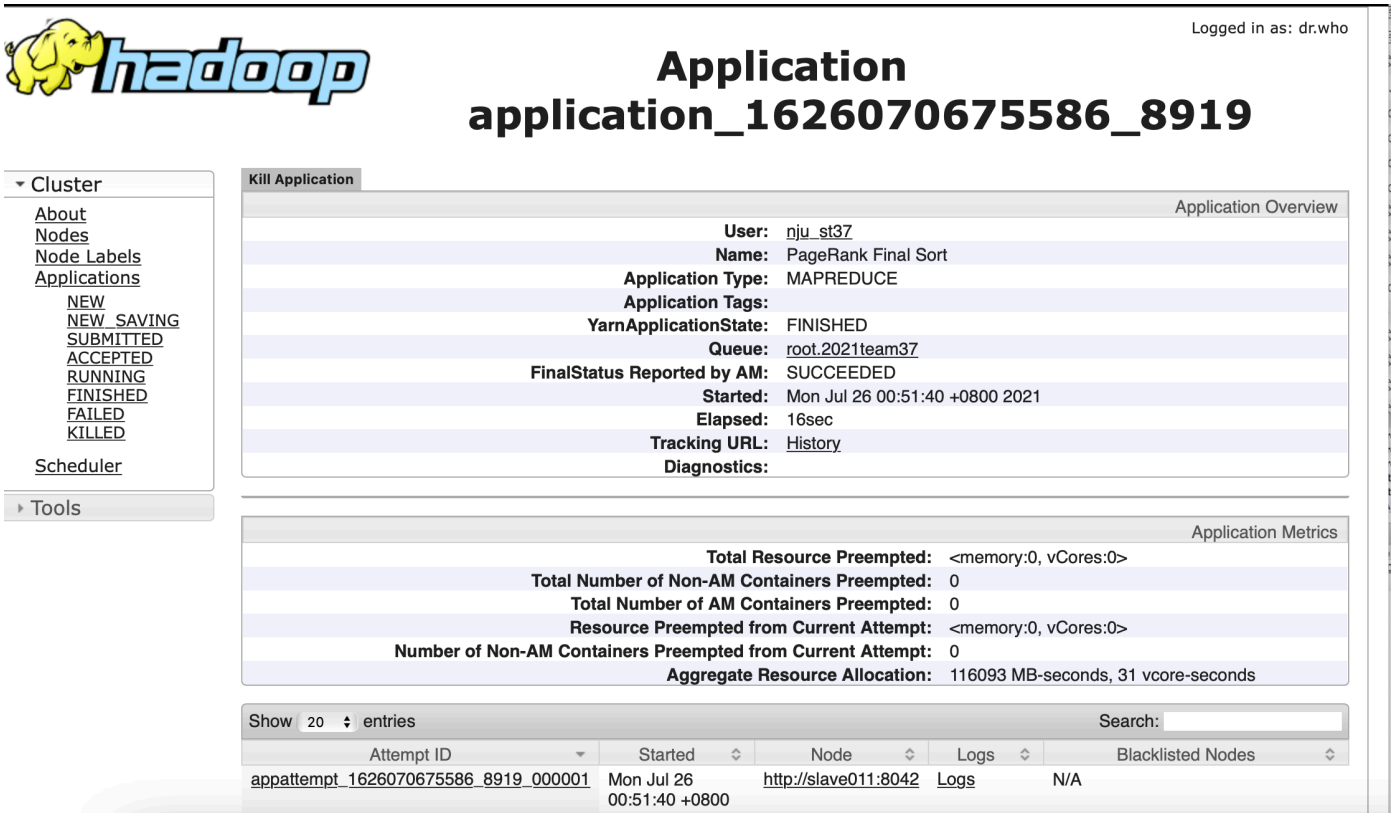
ReaderJob运行记录：



WeightJob运行记录：



PageRank Final Sort运行截图：



MyLabelPropagation运行截图：