

实验 4 SecondSort

1. 实验要求

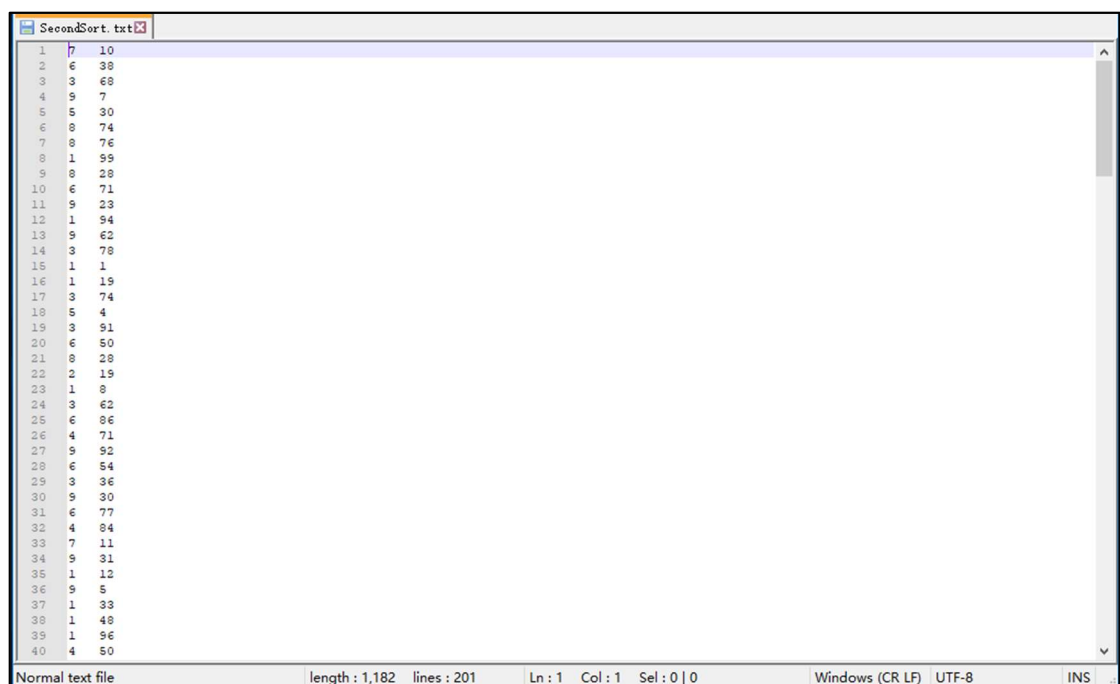
实验任务

使用 MapReduce 完成对数据的二次排序。

实验数据在“实验集群”/data/exercise_4 的实验目录下。

输入数据为 SecondSort.txt。

实验数据分为两列，第一列为 1 到 10 之间的随机数字，第二列为 1 到 100 的随机数字。本次实验我们先对第一列数字按照升序排列，即所谓的分组，再在每组中按照第二列数组进行降序排序完成二次排序。



要求提交二次排序后的结果文件。

输出格式

```
1 99
1 96
1 95
1 94
1 94
1 93
1 84
1 83
1 79
1 68
1 65
1 64
1 50
```

2. 实验数据

文本文件均使用 UTF-8 字符编码，数据之间使用 \t 分隔。

输入数据的情况如下图所示：



```
1 7 10
2 6 38
3 3 68
4 9 7
5 5 30
6 8 74
7 8 76
8 1 99
9 0 28
10 6 71
11 9 23
12 1 94
13 9 62
14 3 78
15 1 1
16 1 19
17 3 74
18 5 4
19 3 91
20 6 50
21 8 28
22 2 19
23 1 8
24 3 62
25 6 86
26 4 71
27 9 92
28 6 54
29 3 36
30 9 30
31 6 77
32 4 94
33 7 11
34 9 31
35 1 12
36 9 5
37 1 33
38 1 48
39 1 96
40 4 50
```

数据集位于集群 HDFS 存储上，HDFS 存储位置为：`hdfs://master001:9000/data/exercise_4/`

注意 最终每个小组的程序必须在课程指定集群上运行，而且输入数据集是全部数据集。结果输出到集群的 HDFS 上。

3. 实验报告要求

在最后提交的压缩包中，除了包含结果文件，源代码、JAR 包、JAR 包执行方式说明，还需要包含一个实验报告。实验报告中请包含：

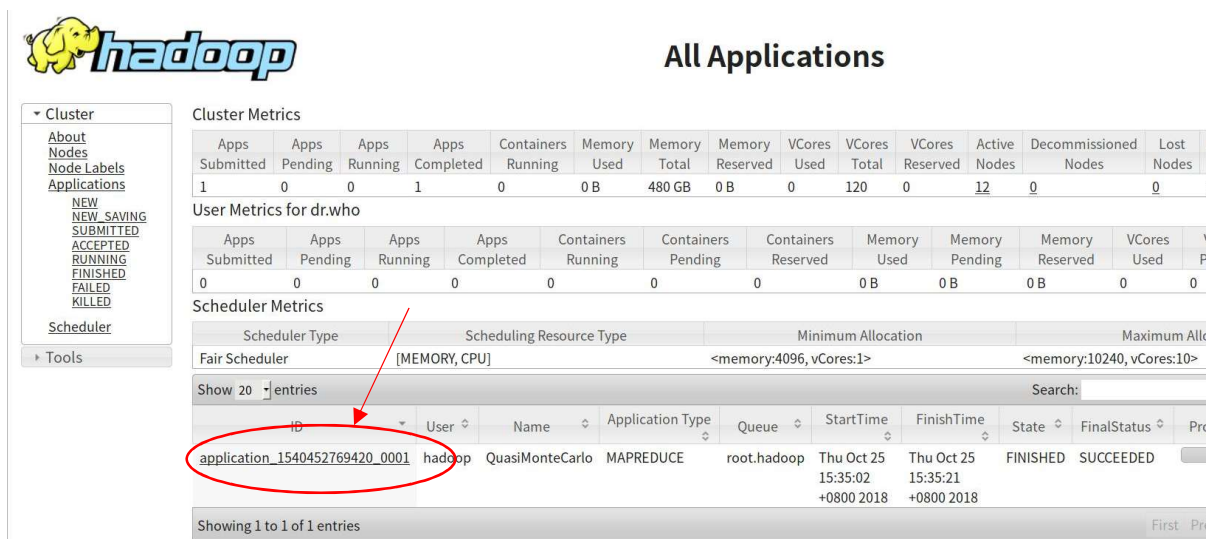
1. Map 和 Reduce 的设计思路（含 Key、Value 类型）。
2. MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
3. 输出结果文件开头部分截图。
4. 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。请完整包括执行报告内容，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。

4. WebUI 执行报告

在以后的实验报告中，如果需要在集群上执行 MapReduce Job，请在实验报告中附上相关的 MapReduce Job 的执行报告，以作为评分依据。如果没有执行报告，在评分时将会认为该 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 **114.212.190.95:8082**

输入小组账户和密码，点击左侧栏“大数据并行计算平台”，再点击“MapReduce 并行计算”可以进入集群监控页面（见下图）。



The screenshot displays the Hadoop WebUI interface for monitoring applications. The left sidebar contains navigation links for Cluster, Nodes, Node Labels, Applications, and Tools. The main content area is titled 'All Applications' and includes several sections:

- Cluster Metrics:** A table showing overall cluster status.
- User Metrics for dr.who:** A table showing metrics for the user 'dr.who'.
- Scheduler Metrics:** A table showing scheduler configuration.
- Applications Table:** A table listing individual applications. The application 'application_1540452769420_0001' is highlighted with a red circle. It is a MAPREDUCE job by user 'hadoop' that has finished successfully.

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	1	0	0 B	480 GB	0 B	0	120	0	12	0	0

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used
0	0	0	0	0	0	0	0 B	0 B	0 B	0

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Fair Scheduler	[MEMORY, CPU]	<memory:4096, vCores:1>	<memory:10240, vCores:10>

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress
application_1540452769420_0001	hadoop	QuasiMonteCarlo	MAPREDUCE	root.hadoop	Thu Oct 25 15:35:02 +0800 2018	Thu Oct 25 15:35:21 +0800 2018	FINISHED	SUCCEEDED	

图 1. 集群监控页面

在该页面上，每个 MapReduce Job 都有一项记录，在记录最右侧“Tracking UI”一栏可以访问到该 Job 的执行情况（见上图画圈的位置）。在执行情况页面（见下图）记录的有 Job 的执行时间、执行状态（是否 SUCCEEDED）等信息。

请在实验报告中附上 MapReduce Job 的执行情况页面截屏，以表明该 Job 是在集群上实际执行过的。

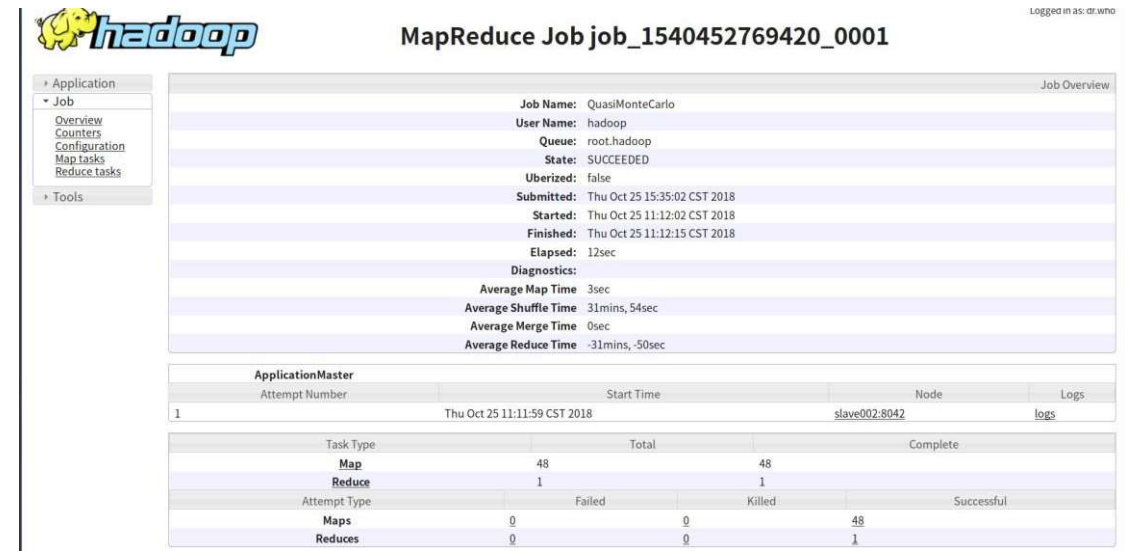


图 2. Job 执行情况页面