大数据处理综合实验：实验 2 —— 实验报告

小组 10

选做功能 1：排序

一、设计思路

（1）Mapper 类（SorterMapper）：

- Key 类型：输入的键是 `Object`，这通常是 Hadoop 的默认值，但实际上它不被使用。这里的关键是文件中的行号，而不是行内容。
- Value 类型：输入的值是 `Text`，代表了文件中的一行内容。

思路：

- 在 `map` 方法中，首先将输入的文本行按照指定的分隔模板进行分割。
- 分割后的第一个元素（`str_value[0]`）是待排序的值，第二个元素（`str_value[1]`）是该值对应的键。
- 将键值对中的值作为 `FloatWritable` 类型的键，键值对中的键作为 `Text` 类型的值进行输出。

（2）Reducer 类（SorterReducer）：

- Key 类型：输入的键是 `FloatWritable`，代表了待排序的值。
- Value 类型：** 输入的值是 `Text`，代表了该值对应的键。

思路：

- 在 `reduce` 方法中，直接遍历 values，即待排序的值对应的键的集合。
- 将每个值与其对应的键写入输出，这样输出的结果就是根据值进行了排序。

整个程序的目的是将输入的键值对按照值的大小进行排序，因此在 Map 阶段，键值对的键被作为值输出，值被作为键输出；在 Reduce 阶段，直接将值与键写入输出。这样，在 Reduce 阶段就可以按照键值对中的值进行排序。

（3）伪代码

MapReduce 中 Map 和 Reduce 的伪代码。

```
function map(key, value, context):
    split_template = "[\\s,]+" # 定义分隔模板
    str_value = value.split(split_template) # 根据模板分割文本行

    mykey = FloatWritable(parseFloat(str_value[1])) # 将第二个元素转换为浮点数作为键
    val = Text(str_value[0]) # 将第一个元素作为值

    context.write(mykey, val) # 输出键值对

```

```
function reduce(key, values, context):
    for value in values:
        context.write(value, key) # 将值和键写入输出

```

二、实验结果

*注：输入文件路径：/outputSS/part-r-00000（基础功能的输出 —— 先执行过倒排索引）

输出文件路径：/user/2024stu_10/sorted-2

（1）输出结果文件的部分截图

```
walks.  1.0
walking.        1.0
walked? 1.0
walked  1.0
walk?   1.0
walk;   1.0
walk:   1.0
walk.   1.0
walk'st 1.0
waking? 1.0
waking; 1.0
waking: 1.0
waking. 1.0
wakes]  1.0
wakes;  1.0
wakes.  1.0
wakes!  1.0
wakened 1.0
waken'd 1.0
waken   1.0
waked.  1.0
wake]   1.0
wake?   1.0
wake;   1.0
wake:   1.0
wake.'  1.0
wake.   1.0
```

```
entrails        1.29
head]   1.29
known;  1.29
must.   1.29
match'd 1.29
returns 1.29
amazement       1.29
True;   1.29
less;   1.29
knowledge.      1.29
know't. 1.29
swells  1.29
grieved 1.29
misfortune      1.29
tax     1.29
balls   1.29
speedy  1.29
coal    1.29
fence   1.29
nation  1.29
outrage 1.29
Best    1.29
elbow   1.29
forth?  1.29
wasteful        1.29
knives  1.29
[Throws 1.29
woo'd   1.29
sister? 1.29
master? 1.29
hurt.   1.29
```

```
CRESSIDA        181.0
DESDEMONA       182.0
not     186.38
ROMEO   194.0
ANTIPHOLUS      195.0
CORIOLANUS      197.0
that    198.25
ANDRONICUS      201.0
FORD    202.0
SYRACUSE        203.0
is      203.38
ROSALIND        213.0
LEAR    233.0
TIMON   238.0
MARK    243.0
in      246.65
MACBETH 252.0
CLEOPATRA       264.0
you     266.45
my      268.5
IAGO    286.0
OTHELLO 306.0
a       315.08
of      383.8
HAMLET  392.0
to      397.35
and     464.05
I       497.28
the     603.78
        3010.4
```

（2）Yarn Resource Manager 的 WebUI 执行报告内容

Logged in as: dr.who

Application application_1678703754602_9991

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Application Overview

User:	2024stu_10
Name:	Sorter
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	2024bigdataclass2
FinalStatus Reported by AM:	SUCCEEDED
Started:	Mon May 13 19:11:00 +0800 2024
Launched:	Mon May 13 19:11:00 +0800 2024
Finished:	Mon May 13 19:11:38 +0800 2024
Elapsed:	37sec
Tracking URL:	History
Log Aggregation Status:	SUCCEEDED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	97735 MB-seconds, 51 vcore-seconds, 0 yarn.io/gpu-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1678703754602_9991_000001	Mon May 13 19:11:00 +0800 2024	http://slave006:8042	Logs	0	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last