

## 《分布式编程模型与系统》期末考查作业

学号: 10195501415

姓名: 赵煜硕

成绩:

### 1. 实验目的

比较 Spark 和 MapReduce 的性能差异

### 2.

- a) MapReduce 执行迭代计算过程中会反复读写 HDFS, 因此可以在 HDFS 中观察到每一轮迭代的输出结果
- b) MapReduce 会提交一系列的作业, 而 spark 仅有一个应用, 在 Yarn 的 UI 显示会不一样
- c) 对于同样规模的数据集, spark 执行时间应当更短

### 3. 实验设置

JDK1.8

IDEA

Windows

Ubuntu18.04 2 核 4G

Hadoop2.6.1

### 4. 实验过程

<https://github.com/zhaoyushuo123/pagerank>

使用课本上给的 mapreduce 版本的 pagerank 代码

部分片段参考

<https://wenku.baidu.com/view/8a08e207ac45b307e87101f69e3143323968f5d6.html>

配置 input 目录和 output 目录

遇到

log4j:WARN No appenders could be found for logger (org.apache.zookeeper.ZooKeeper).

log4j:WARN Please initialize the log4j system properly.

log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.

参考

[https://blog.csdn.net/zhengzaifeidelushang/article/details/116239936?spm=1001.2101.3001.6661.1&utm\\_medium=distribute.pc\\_relevant\\_t0.none-task-blog-2%7Edefault%7ECTRLIST%7EPayColumn-1-116239936-blog-123576662.pc\\_relevant\\_default&depth\\_1-utm\\_source=distribute.pc\\_relevant\\_t0.none-task-blog-2%7Edefault%7ECTRLIST%7EPayColumn-1-116239936-blog-123576662.pc\\_relevant\\_default&utm\\_relevant\\_index=1](https://blog.csdn.net/zhengzaifeidelushang/article/details/116239936?spm=1001.2101.3001.6661.1&utm_medium=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7EPayColumn-1-116239936-blog-123576662.pc_relevant_default&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7EPayColumn-1-116239936-blog-123576662.pc_relevant_default&utm_relevant_index=1)

<https://blog.csdn.net/q125864004/article/details/109294338>

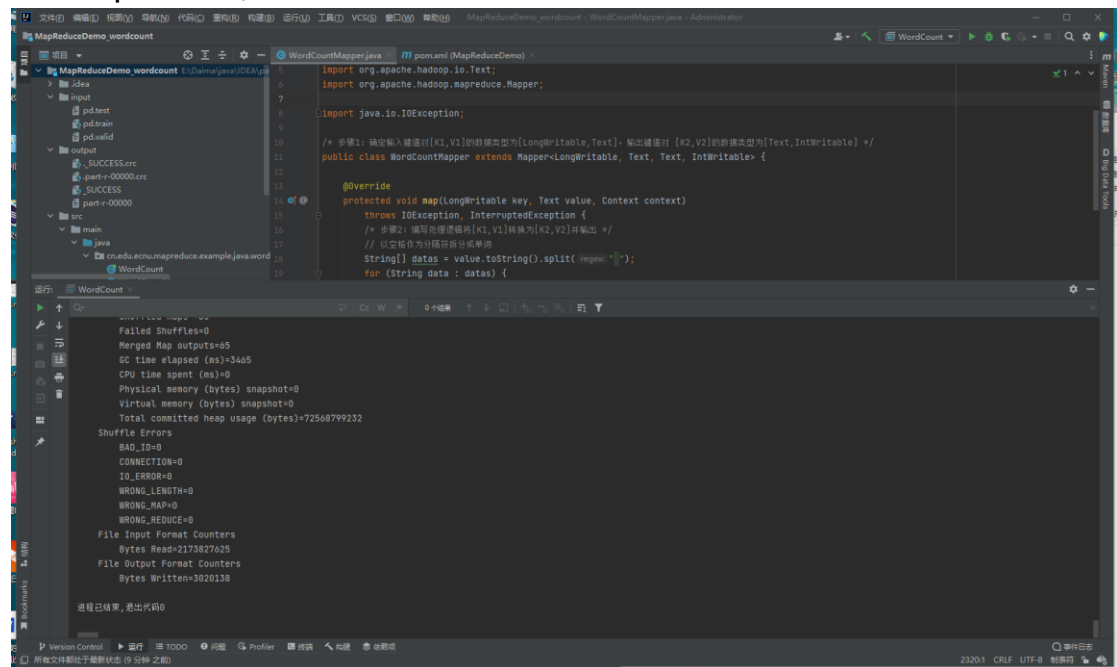
遇到

java.lang.ArrayIndexOutOfBoundsException: 1

[https://blog.csdn.net/qg\\_45054784/article/details/115387017?spm=1001.2101.3001.6661.1&utm\\_medium=distribute.pc\\_relevant\\_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc\\_relevant\\_multi\\_platform\\_whitelistv1&depth\\_1-utm\\_source=distribute.pc\\_relevant\\_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc\\_relevant\\_multi\\_platform\\_whitelistv1&utm\\_relevant\\_index=1](https://blog.csdn.net/qg_45054784/article/details/115387017?spm=1001.2101.3001.6661.1&utm_medium=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc_relevant_multi_platform_whitelistv1&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc_relevant_multi_platform_whitelistv1&utm_relevant_index=1)

a)

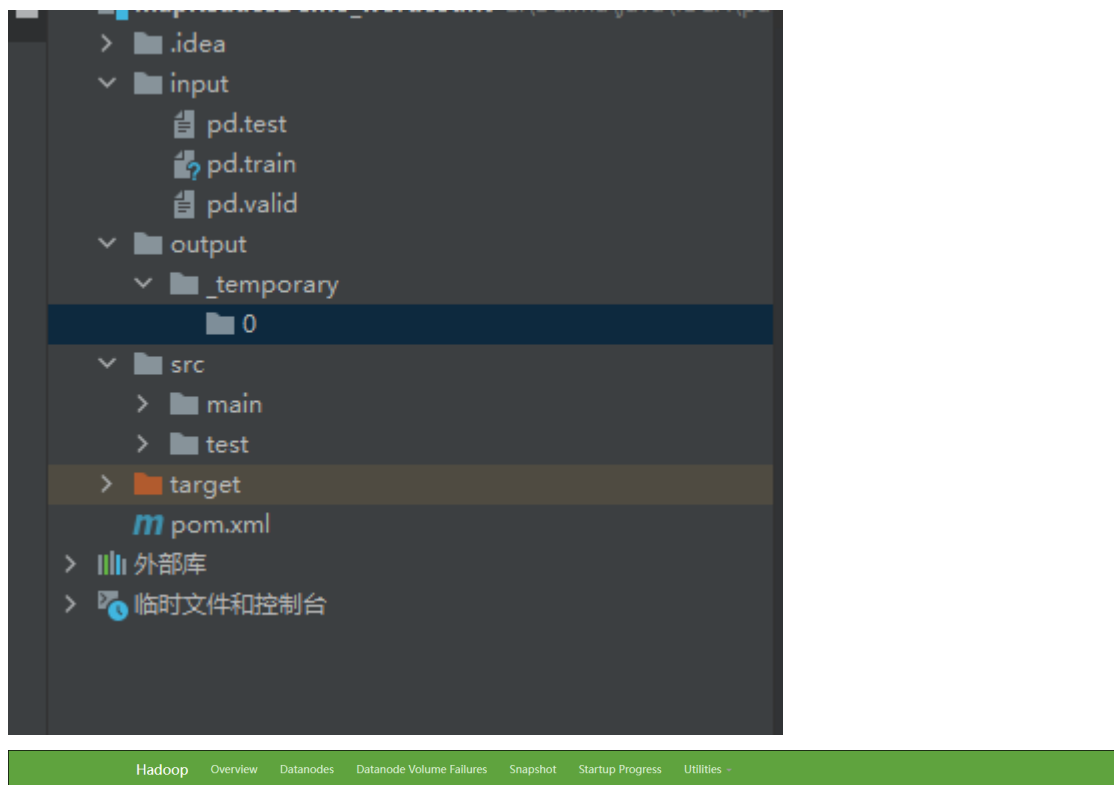
## 执行 mapreduce 的 wordcount



The screenshot shows an IDE with the following components:

- Project Explorer:** Shows a project named 'MapReduceDemo\_wordcount' with subfolders 'input', 'output', 'src', and 'main'. The 'src' folder contains 'WordCountMapper.java'.
- Editor:** Displays the code for 'WordCountMapper.java'. The code imports necessary classes and implements the 'map' method of the 'Mapper' interface. It splits the input text into words and outputs them as key-value pairs.
- Run Console:** Shows the output of the 'WordCount' command. It displays various statistics such as 'Failed Shuffles=0', 'Merged Map outputs=05', 'GC time elapsed (ms)=3465', 'CPU time spent (ms)=0', 'Physical memory (bytes) snapshot=0', 'Virtual memory (bytes) snapshot=0', 'Total committed heap usage (bytes)=72568799232', 'Shuffle Errors', 'File Input Format Counters', and 'File Output Format Counters'.

执行过程中



## Browse Directory

/user/ubuntu/output4 Go! 🏠 🔍 📄

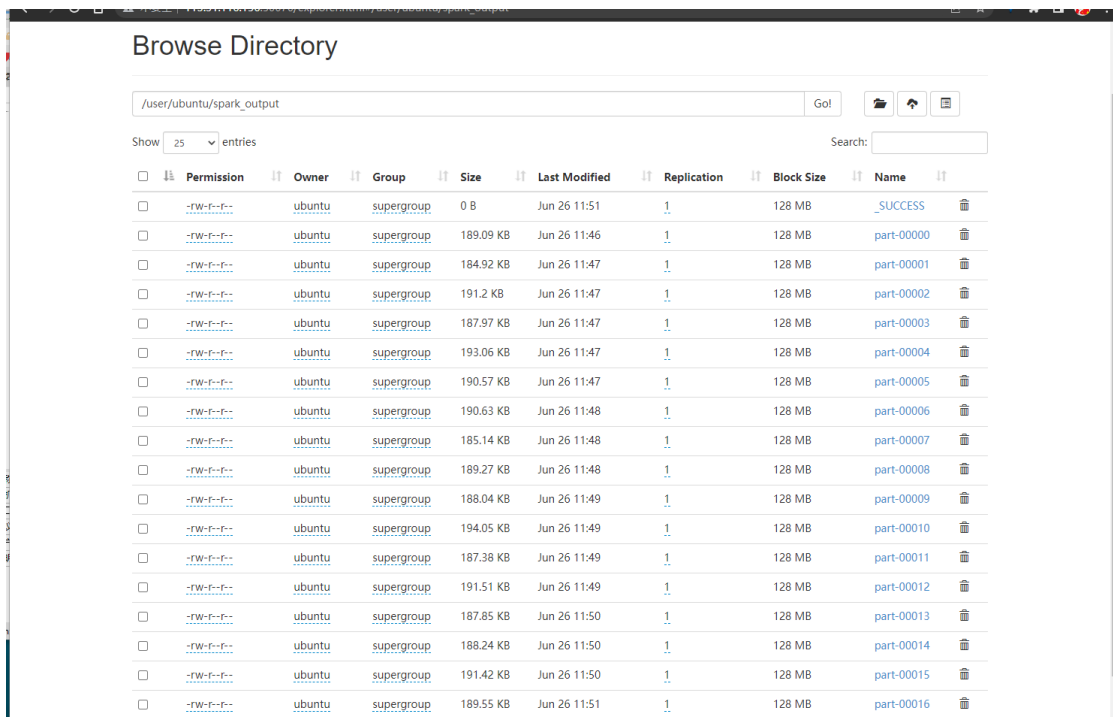
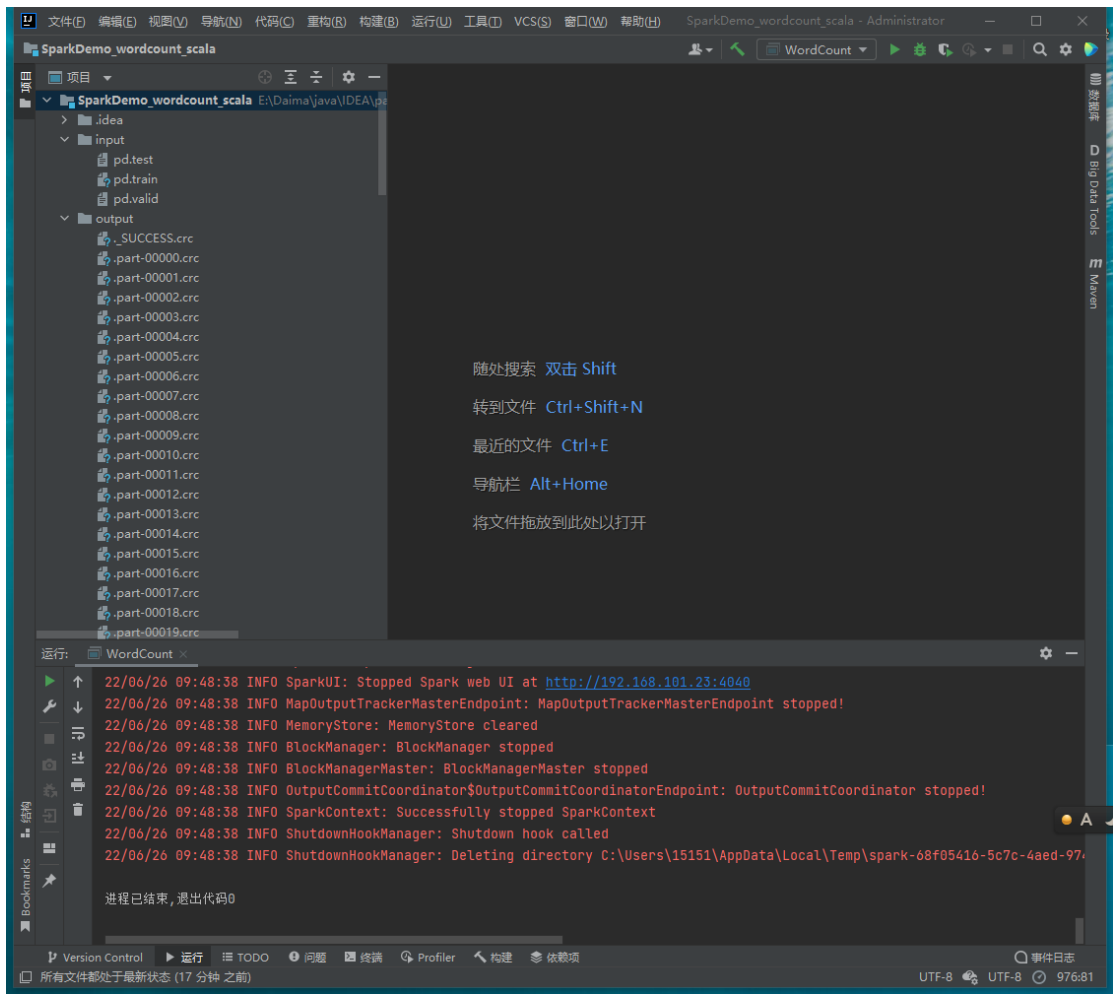
Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	ubuntu	supergroup	0 B	Jun 26 12:03	1	128 MB	._SUCCESS	🗑️
<input type="checkbox"/>	-rw-r--r--	ubuntu	supergroup	2.86 MB	Jun 26 12:03	1	128 MB	part-r-00000	🗑️

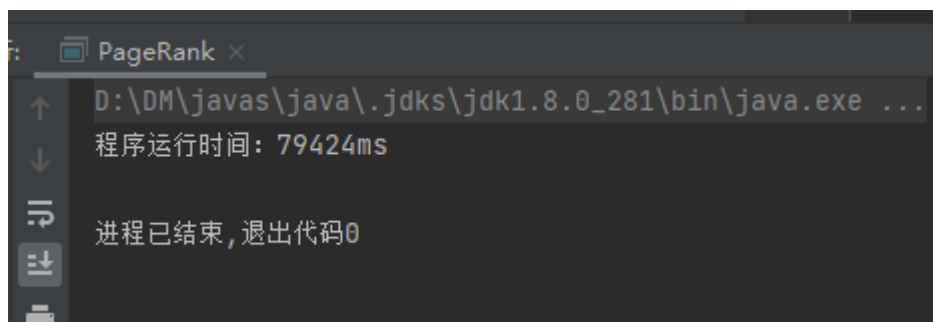
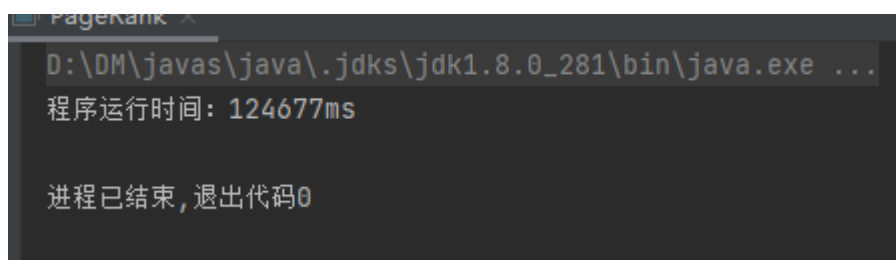
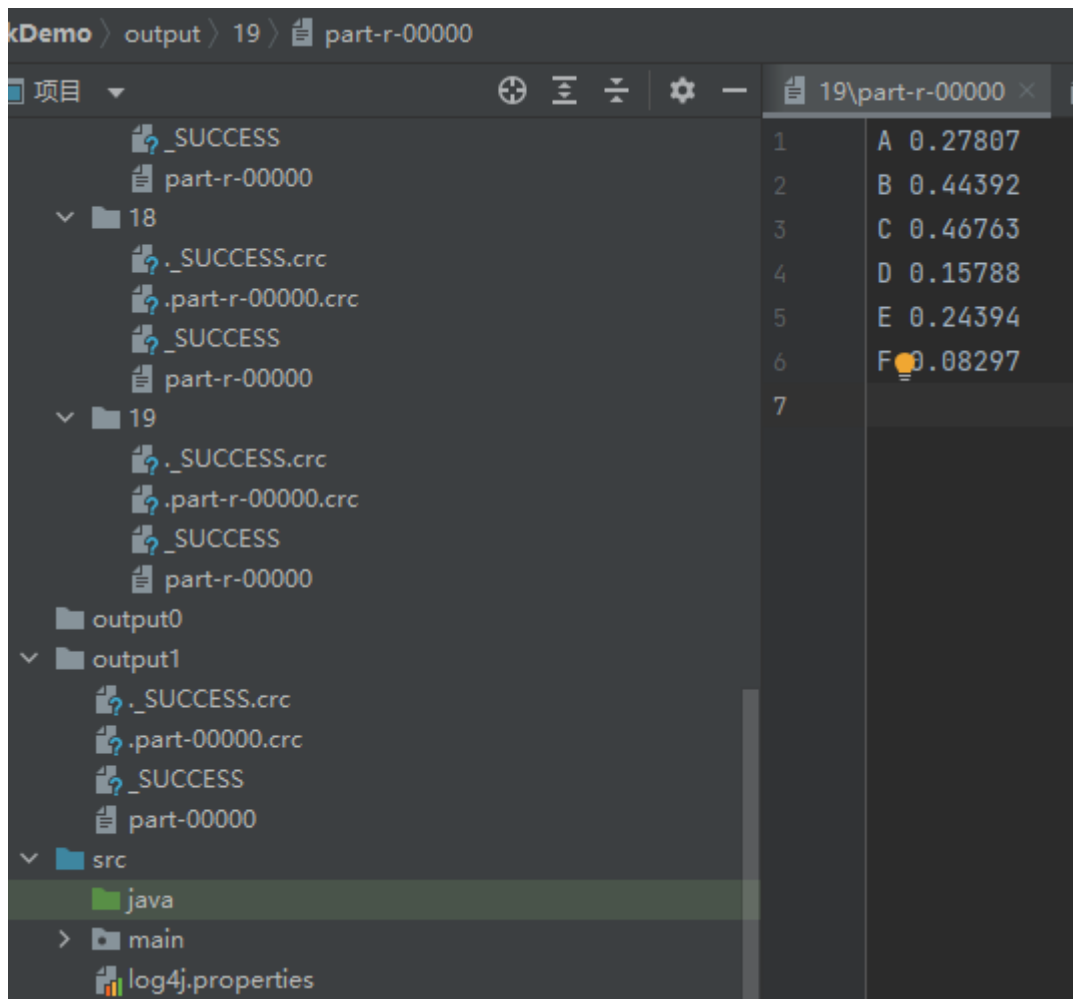
Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2020.

## Spark 版本的 wordcount



执行 MapReduce 的 pagerank,发现 output 中会给出每轮迭代的结果



在 spark 上会快一点

#### Job Counters

Launched map tasks=1  
Launched reduce tasks=1  
Data-local map tasks=1  
Total time spent by all maps in occupied slots (ms)=2373  
Total time spent by all reduces in occupied slots (ms)=2571  
Total time spent by all map tasks (ms)=2373  
Total time spent by all reduce tasks (ms)=2571  
Total vcore-milliseconds taken by all map tasks=2373  
Total vcore-milliseconds taken by all reduce tasks=2571  
Total megabyte-milliseconds taken by all map tasks=2429952  
Total megabyte-milliseconds taken by all reduce tasks=2632704

#### Map-Reduce Framework

Map input records=6  
Map output records=18  
Map output bytes=557  
Map output materialized bytes=599  
Input split bytes=121  
Combine input records=0  
Combine output records=0  
Reduce input groups=6  
Reduce shuffle bytes=599  
Reduce input records=18  
Reduce output records=6  
Spilled Records=36  
Shuffled Maps =1  
Failed Shuffles=0  
Merged Map outputs=1  
GC time elapsed (ms)=118  
CPU time spent (ms)=1160  
Physical memory (bytes) snapshot=459915264  
Virtual memory (bytes) snapshot=3890864128  
Total committed heap usage (bytes)=291504128

#### Shuffle Errors

BAD\_ID=0  
CONNECTION=0  
IO\_ERROR=0  
WRONG\_LENGTH=0  
WRONG\_MAP=0  
WRONG\_REDUCE=0

#### File Input Format Counters

Bytes Read=207

#### File Output Format Counters

Bytes Written=60

程序运行时间: 439545ms

ubuntu@10-24-13-173:~/hadoop-2.10.1\$

<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:31	<a href="#">0</a>	0 B	<a href="#">output60</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:53	<a href="#">0</a>	0 B	<a href="#">output70</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:54	<a href="#">0</a>	0 B	<a href="#">output71</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:57	<a href="#">0</a>	0 B	<a href="#">output710</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:57	<a href="#">0</a>	0 B	<a href="#">output711</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:58	<a href="#">0</a>	0 B	<a href="#">output712</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:58	<a href="#">0</a>	0 B	<a href="#">output713</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:58	<a href="#">0</a>	0 B	<a href="#">output714</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:59	<a href="#">0</a>	0 B	<a href="#">output715</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:59	<a href="#">0</a>	0 B	<a href="#">output716</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 13:00	<a href="#">0</a>	0 B	<a href="#">output717</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 13:00	<a href="#">0</a>	0 B	<a href="#">output718</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 13:00	<a href="#">0</a>	0 B	<a href="#">output719</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:54	<a href="#">0</a>	0 B	<a href="#">output72</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:54	<a href="#">0</a>	0 B	<a href="#">output73</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:55	<a href="#">0</a>	0 B	<a href="#">output74</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:55	<a href="#">0</a>	0 B	<a href="#">output75</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:56	<a href="#">0</a>	0 B	<a href="#">output76</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:56	<a href="#">0</a>	0 B	<a href="#">output77</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:56	<a href="#">0</a>	0 B	<a href="#">output78</a>	
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">ubuntu</a>	<a href="#">supergroup</a>	0 B	Jun 26 12:57	<a href="#">0</a>	0 B	<a href="#">output79</a>	

b)

执行时间

Pg

A B D

B C

C A B E

D B C F

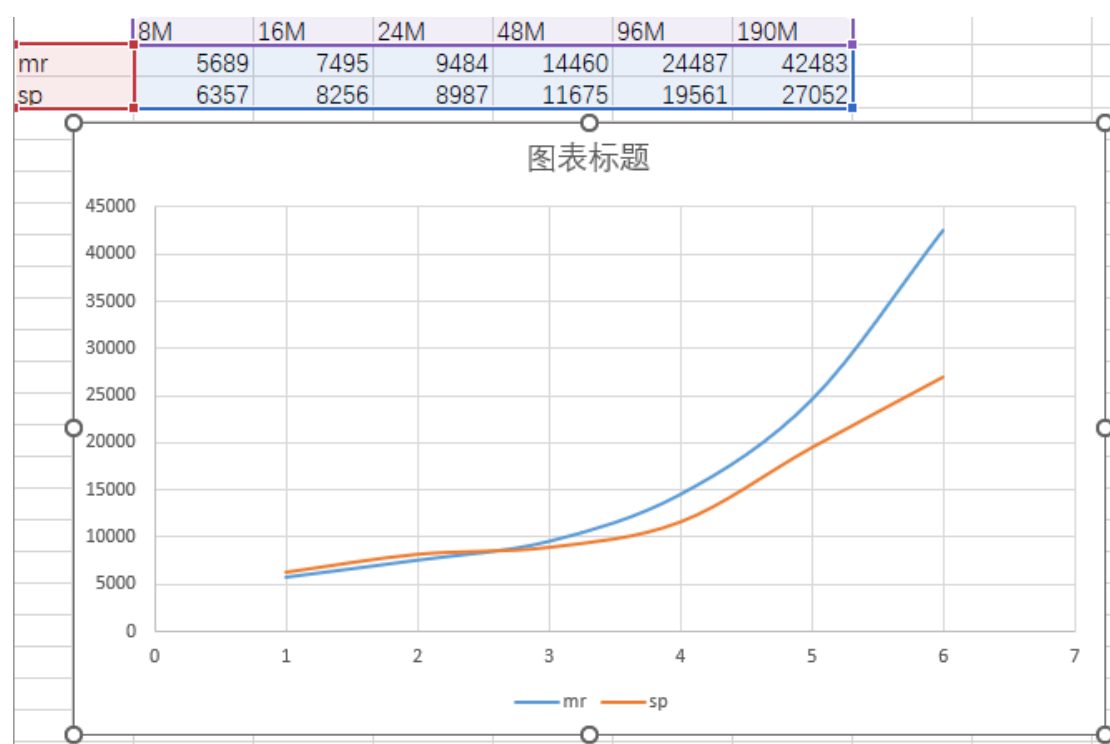
E A B

F E

Wc pd.train

	Mapreduece	Spark
Pg	439545	279424
Wc	441370	365000

Wourdcount

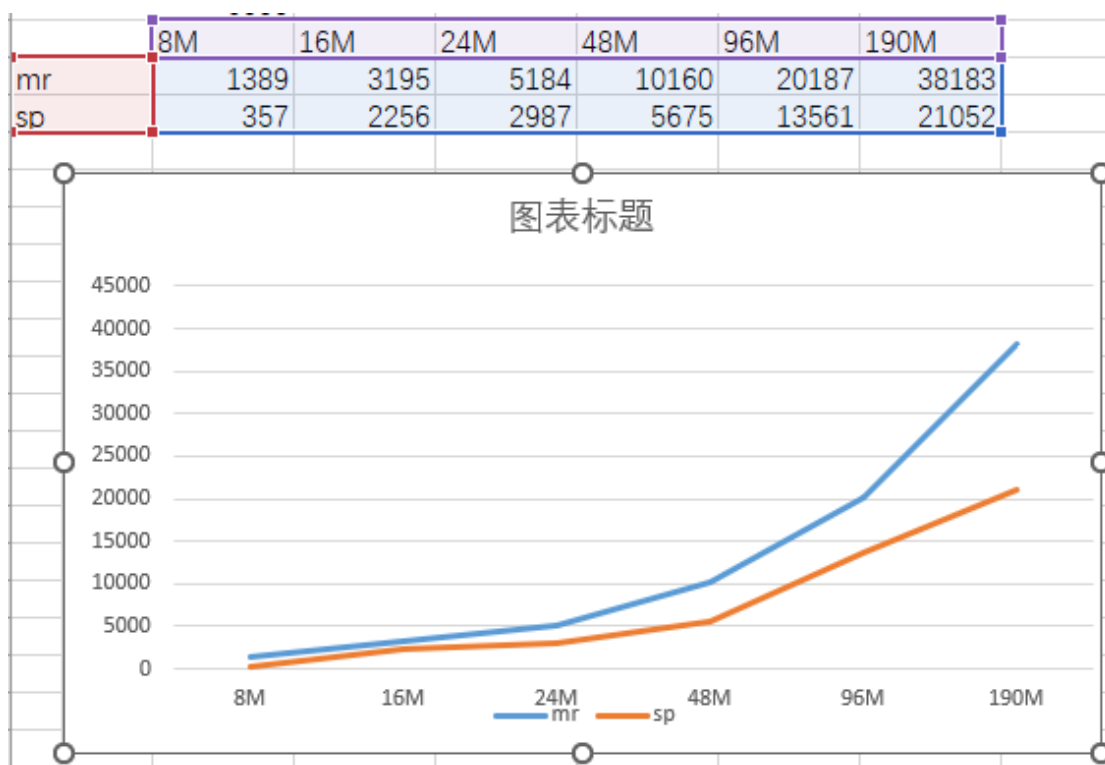


追踪

	640kb	1282kb	2600kb	空
mr	4386	4530	4398	4337
sp	6012	6324	6257	5837

去掉启动时间





MR 的 Wordcount 卡死或强制暂停

37%

Search:

Replication	Block Size	Name	
0	0 B	_temporary	

Previous 1 Next