**《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.

1. 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
2. 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
3. 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/qq125864004/article/details/109294338>

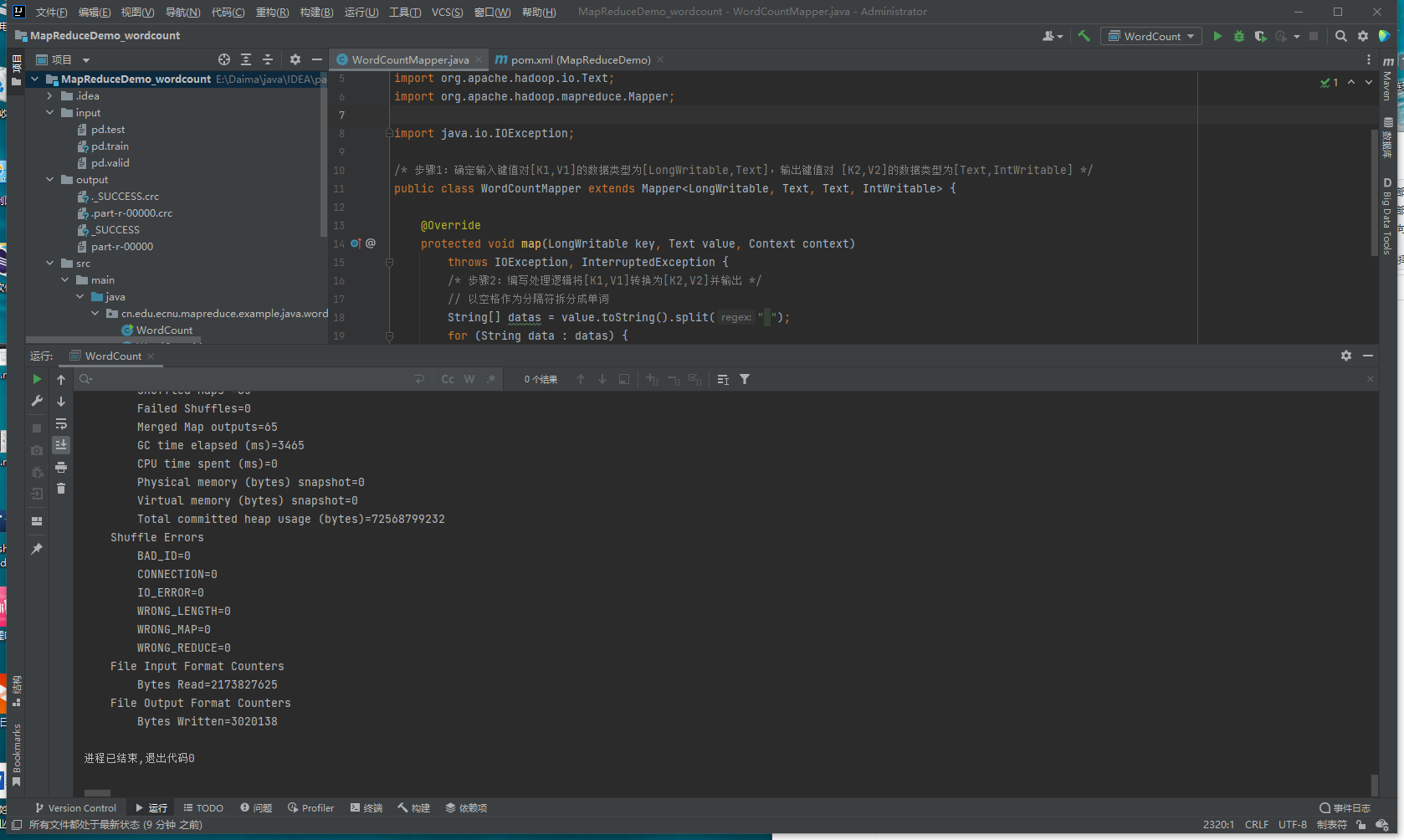
遇到

java.lang.ArrayIndexOutOfBoundsException: 1

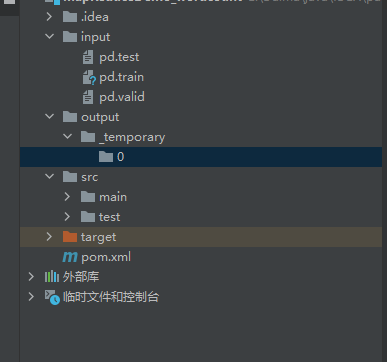
<https://blog.csdn.net/qq_45054784/article/details/115387017?spm=1001.2101.3001.6661.1&utm_medium=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc_relevant_multi_platform_whitelistv1&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1-115387017-blog-79281545.pc_relevant_multi_platform_whitelistv1&utm_relevant_index=1>

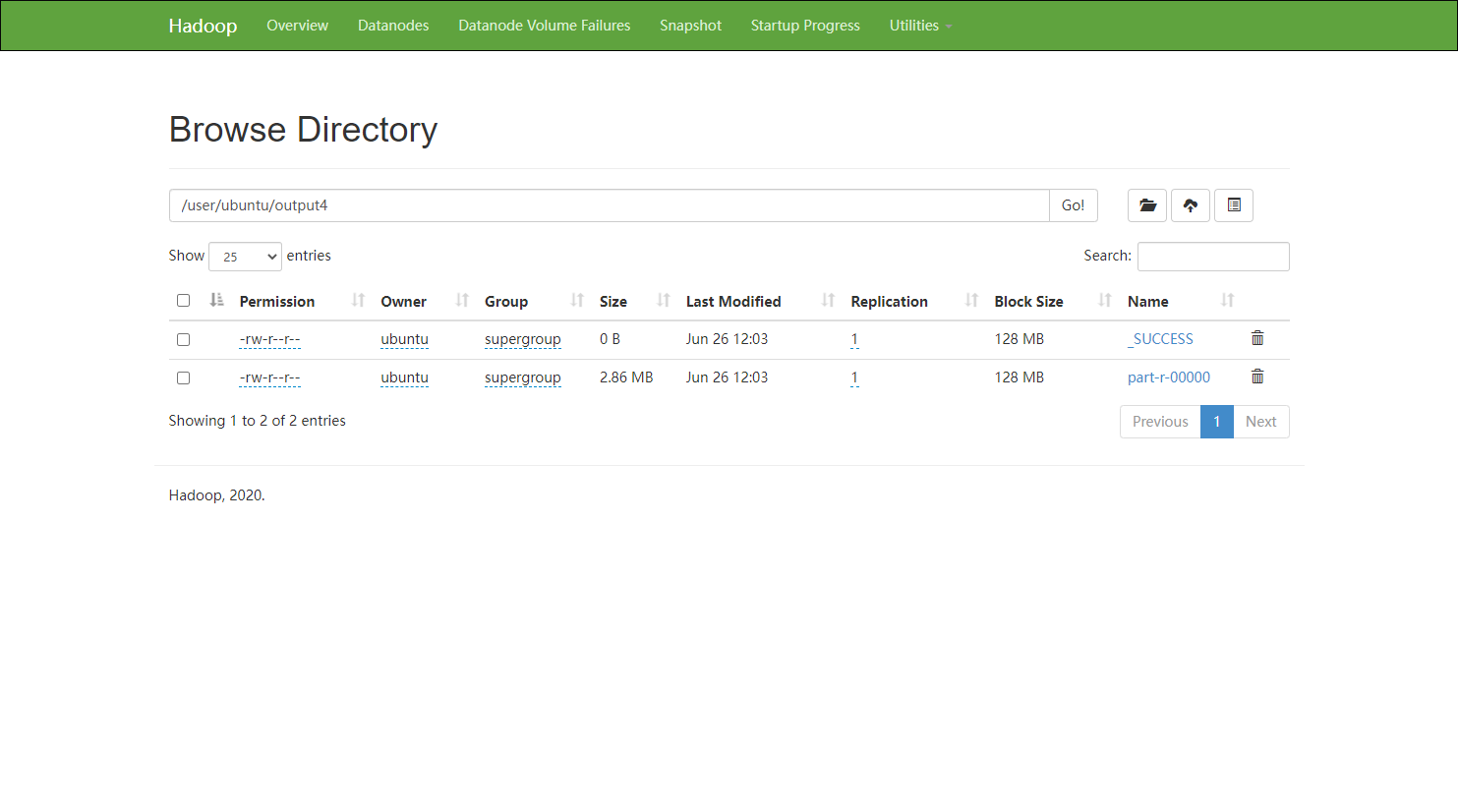
a)

**执行mapreduce的wordcount**

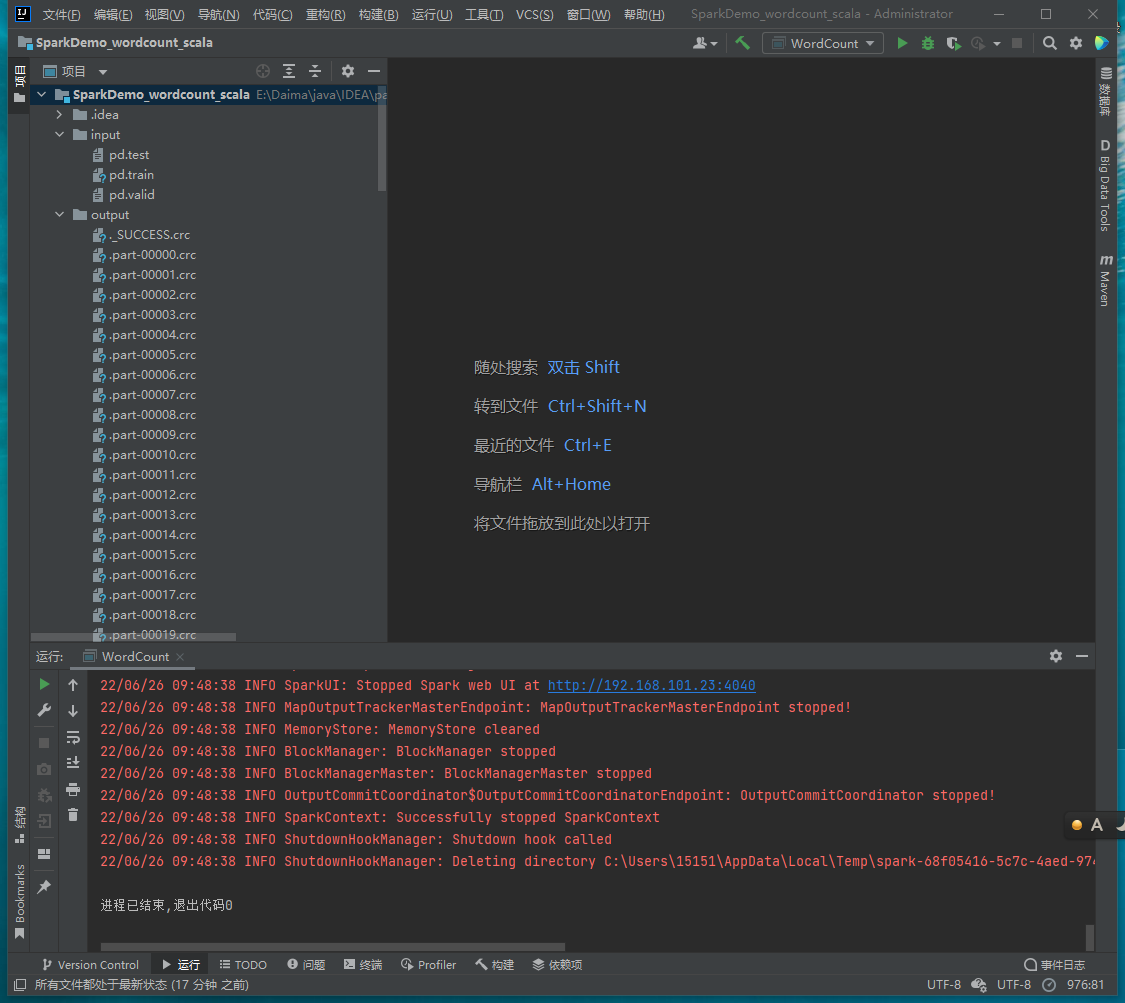


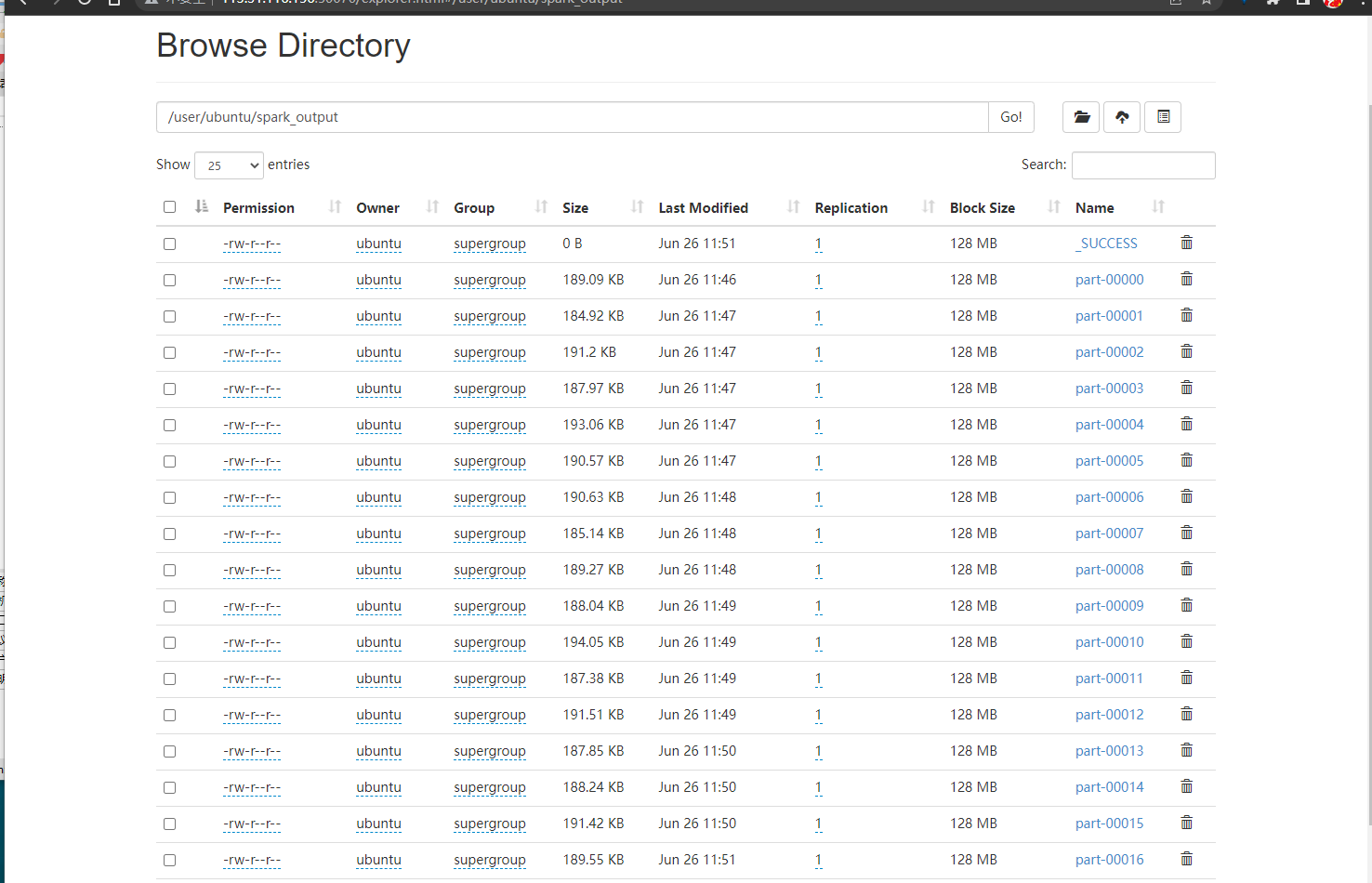
执行过程中



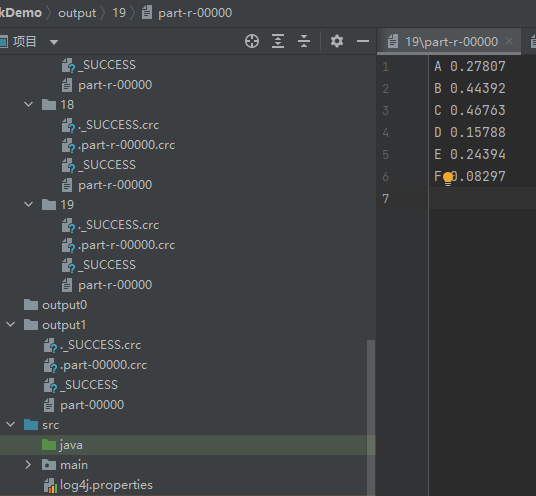


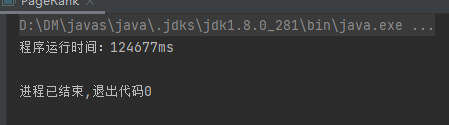
**Spark版本的wordcount**

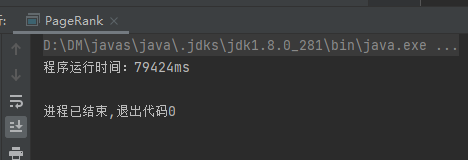




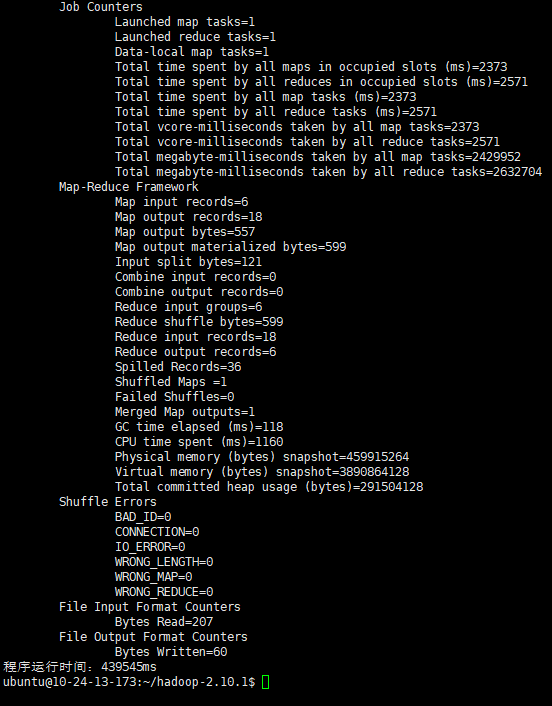
**执行MapReduce的pagerank**,发现output中会给出每轮迭代的结果

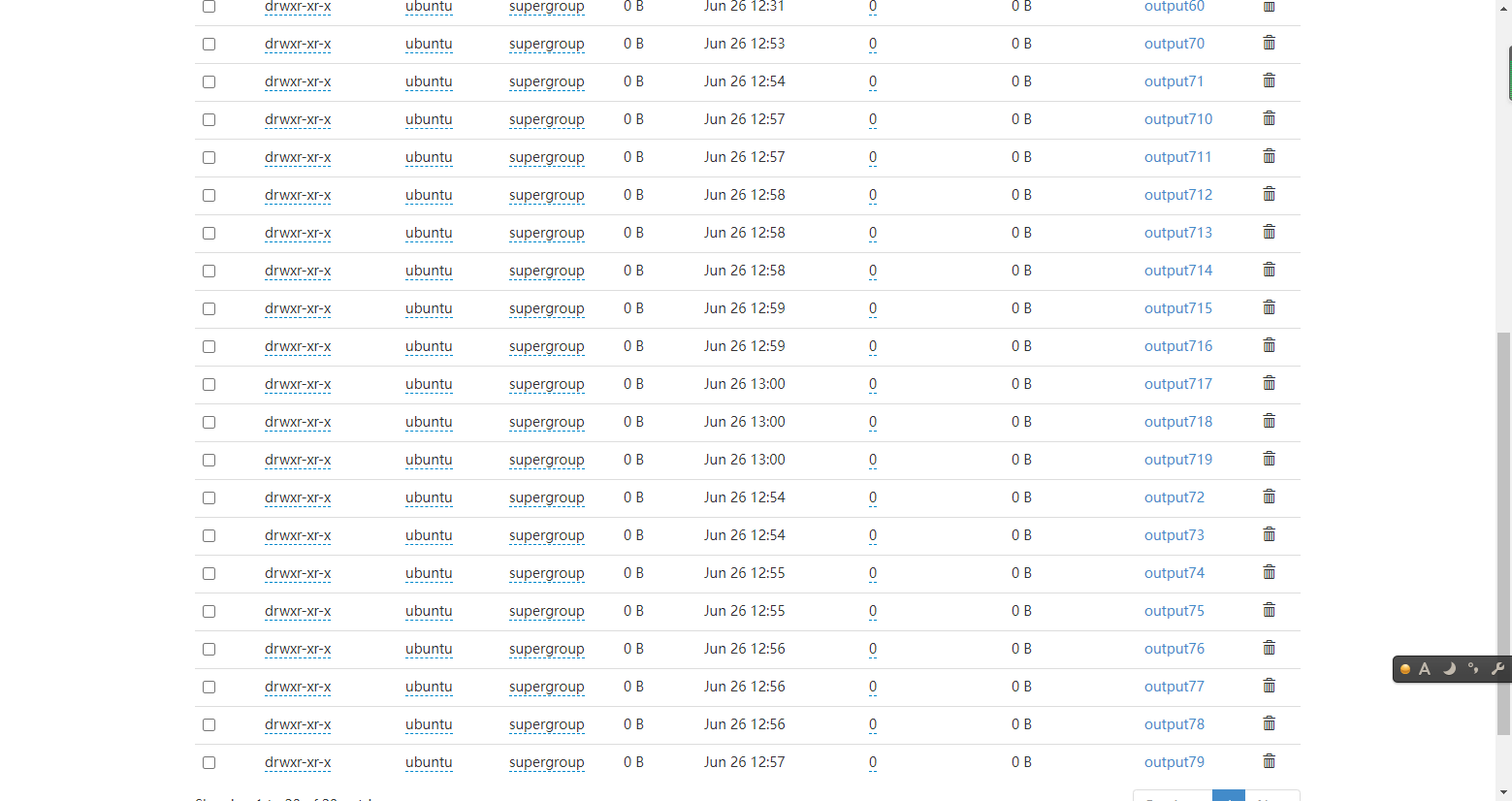






在spark上会快一点





b)

**执行时间**

Pg

A B D

B C

C A B E

D B C F

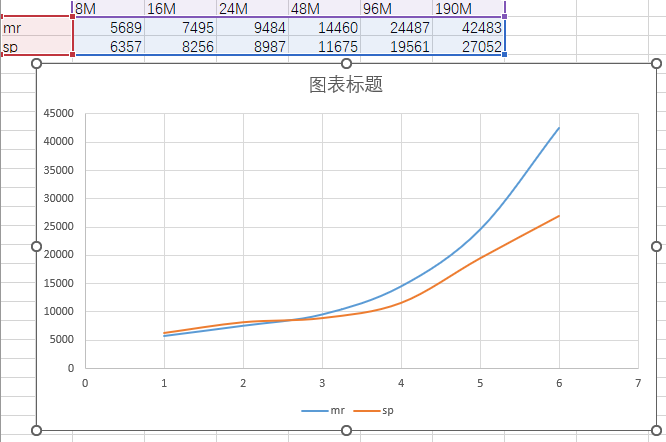
E A B

F E

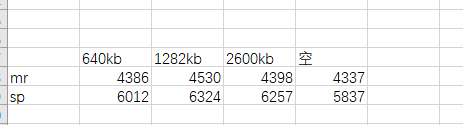
Wc pd.train

|  |  |  |
| --- | --- | --- |
|  | Mapreduece | Spark |
| Pg | 439545 | 279424 |
| Wc | 441370 | 365000 |

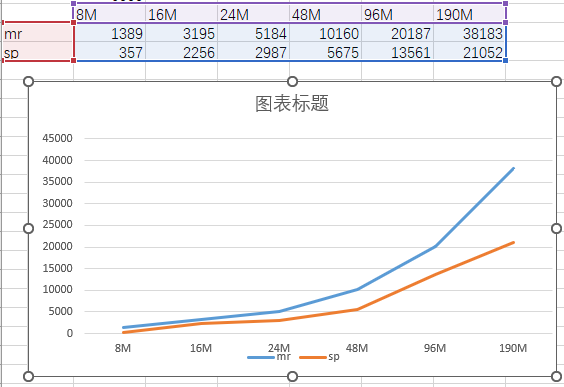
Wourdcount



追测

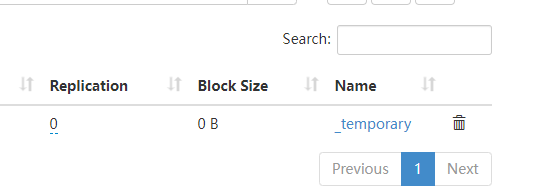


去掉启动时间



**MR的Wordcount卡死或强制暂停**

37%

****