**实验内容与要求**

1. 每人在自己本地电脑上正确安装和运行伪分布式Hadoop系统

   安装操作手册和本课程课件请从MapReduce课程目录下载。

2. 安装完成后,自己寻找一组英文网页数据,在本机上运行Hadoop系统自带的WordCount可执行程序文件,并产生输出结果

3. 实验结果提交：要求书写一个实验报告，其中包括：

（1）系统安装运行的情况

（2）实验数据说明（下载的什么网页数据，多少个HTML或text文件)

（3）程序运行后在Hadoop Web作业状态查看界面上的作业运行状态屏幕拷贝

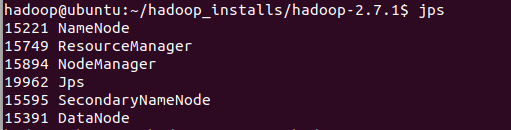
（4）实验输出结果开头部分的屏幕拷贝

（5）实验体会

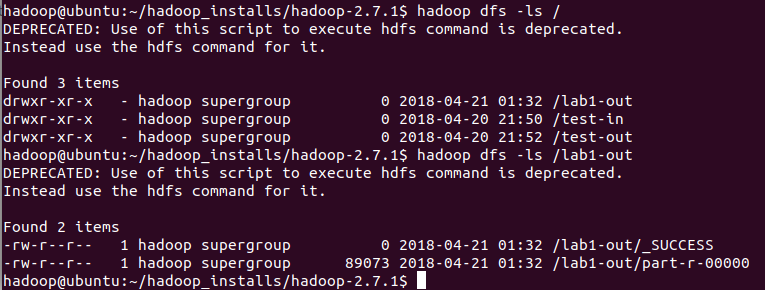
**实验报告**

**151220129 计科 吴政亿 nju\_wzy@163.com**

1. **系统安装运行的情况**



通过jps指令得到hadoop已正确运行



通过ls指令查询，其中test-in为我的test数据集，test-out是测试样例输出，lab1-out为本次实验输出

1. **实验数据说明**

选自维基百科的闪电侠、神盾局特工、超女三个页面进行统计，地址如下：

<https://en.wikipedia.org/wiki/Flash_(comics)>

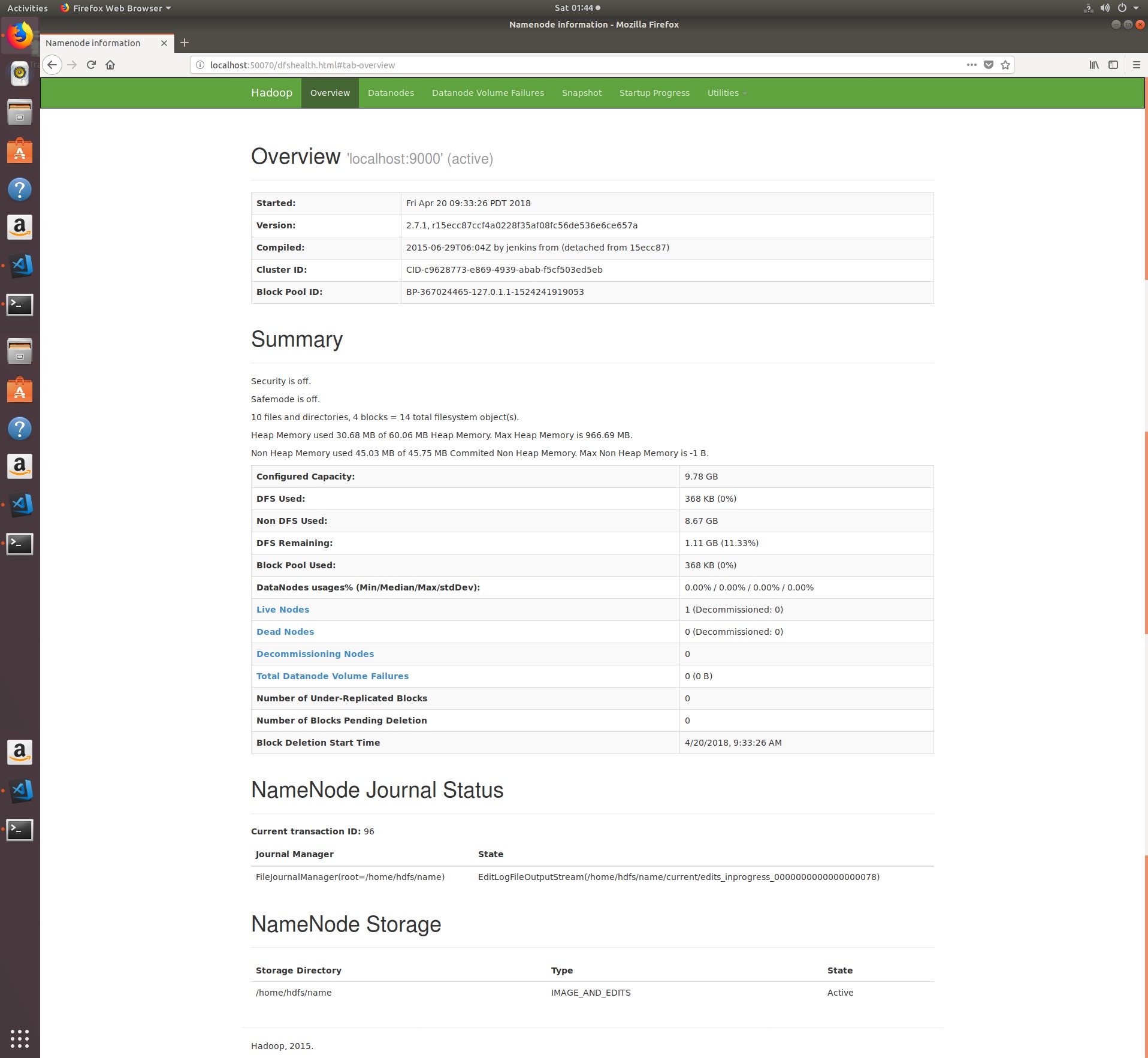
<https://en.wikipedia.org/wiki/Agents_of_S.H.I.E.L.D>.

<https://en.wikipedia.org/wiki/Supergirl>

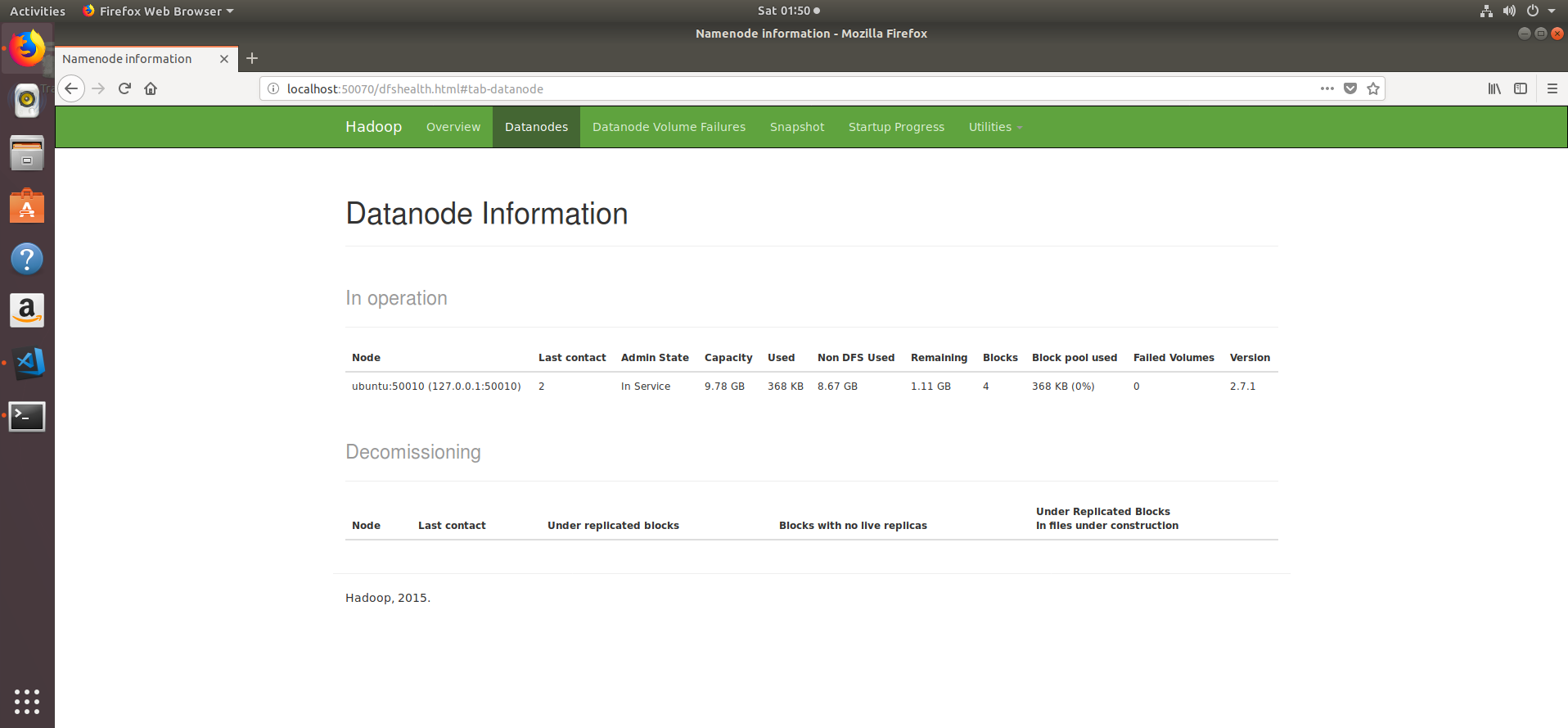
其中我将他们文字部分ctrl + A复制到了三个记事本里并简单的处理了一下，把html标签去除了。

1. Web**作业运行状态**

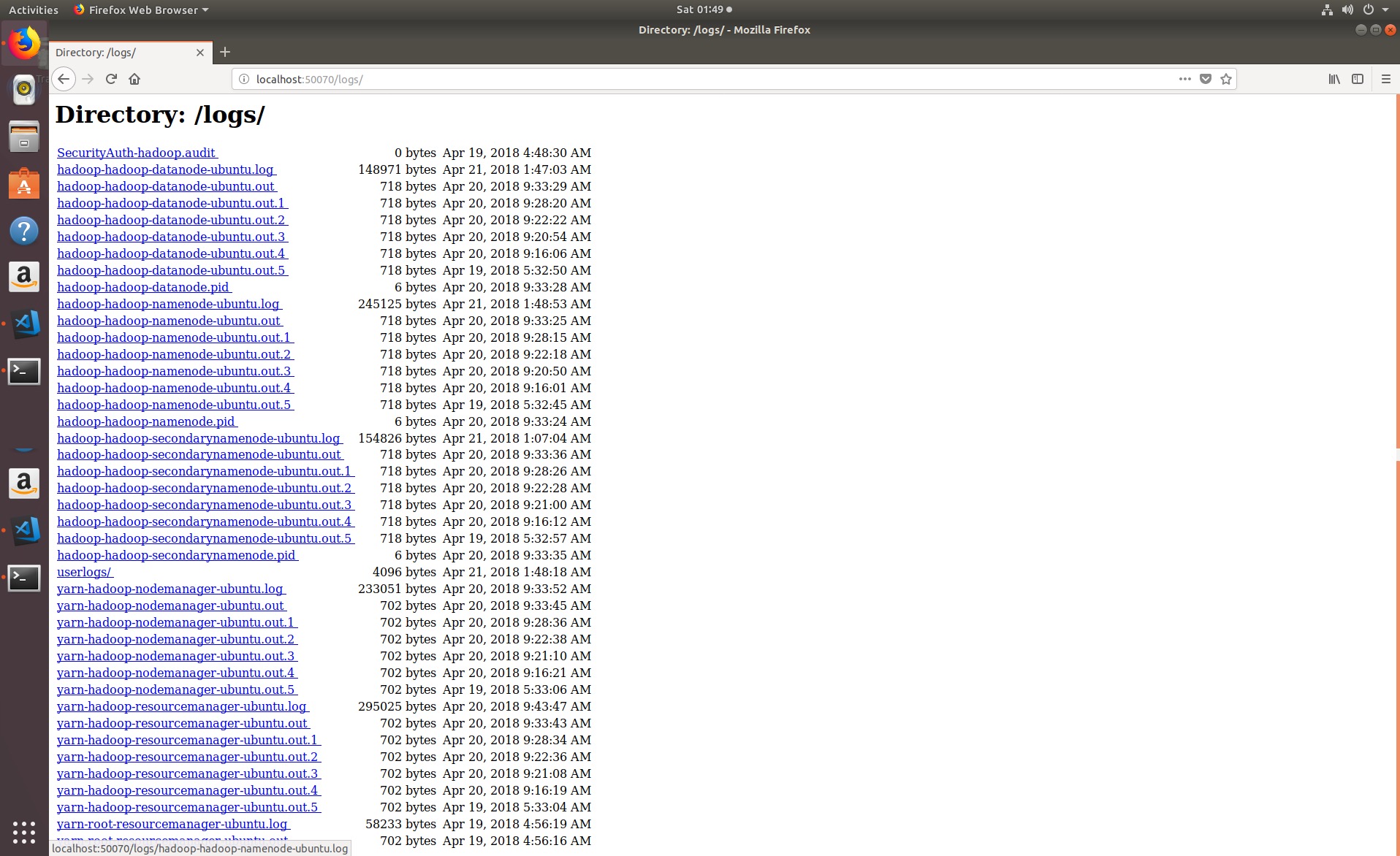
**Summary:**



**DataNode:**



**Logs:**



1. **实验输出结果**

Wordcount:

""4,722 1

""Agents    8

"'Agents    27

"'Arrival,' 1

"'Black 1

"'Blue  4

"'CHiPs,'   1

"'Flash'    1

"'Gravity,' 1

"'Hawaii    1

"'Marvel's  12

"'NCIS' 1

"'S.H.I.E.L.D.' 1

"'The   1

"'Twilight' 1

"'a 1

"0-8-4" 1

"10 2

"12th   1

"1960s".    1

"2013". 1

"2013-14:   1

"2014-15:   1

"2016   1

"4,722  4

"5  1

"66th   1

"A  5

"ABC    16

"ABC's  2

"Adrianne   2

"After  1

"Aftershocks".[42][99]  1

"Agent  1

"Agents 52

"All    3

"And    1

"Andromeda" 1

"Animation  1

"Ann    1

"Any    1

"Archived   2

"Artisans   1

"Ascension" 1

"Aussie 1

"Bart"  1

"Bear   1

"Beginning  1

"Broken 2

"CTV    1

"Can    1

"Catch-22", 1

"Chinese-American   1

"Chris  1

"Claire 1

"Clark  1

"Comic-Con: 1

"Composers  1

"Coulson    1

"Creative   1

"Credits".  1

"Crisis 1

"DC 3

"Deals  1

"Declassifying  3

"Disney 1

"Do 1

"EXCLUSIVE  2

"EXCLUSIVE: 1

"Earth-2".  1

"Earth-Born 1

"Earth-C-Minus",    1

"Edgar  1

"Emancipation", 1

"Emily  1

"End    3

"Essential  1

"Exclusive  2

"Exclusive! 1

"Exclusive: 5

"Ezra   1

"Fall   1

"Fallen 3

"Faultless  1

"Fig.   1

"Film   1

"Final  1

"Flash  2

"Flash's    1

"For    1

"Foxtel 1

"Full   3

"FuseFX 1

"Gamechanger"". 1

"Ghost  1

"Girls  1

"Go 1

"God's  1

"HD 1

"He's   1

"Helen  1

"Henry  3

"Here's 2

"Hiatus 1

"Hot    1

"Hotly  1

"How    1

"I  1

"I'm    1

"IFTA   1

"IMDB   1

执行wordcount的部分logs:

18/04/21 01:32:56 INFO output.FileOutputCommitter: Saved output of task 'attempt\_local1611856096\_0001\_r\_000000\_0' to hdfs://localhost:9000/lab1-out/\_temporary/0/task\_local1611856096\_0001\_r\_000000

18/04/21 01:32:56 INFO mapred.LocalJobRunner: reduce > reduce

18/04/21 01:32:56 INFO mapred.Task: Task 'attempt\_local1611856096\_0001\_r\_000000\_0' done.

18/04/21 01:32:56 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1611856096\_0001\_r\_000000\_0

18/04/21 01:32:56 INFO mapred.LocalJobRunner: reduce task executor complete.

18/04/21 01:32:57 INFO mapreduce.Job: map 100% reduce 100%

18/04/21 01:32:57 INFO mapreduce.Job: Job job\_local1611856096\_0001 completed successfully

18/04/21 01:32:57 INFO mapreduce.Job: Counters: 35

    File System Counters

        FILE: Number of bytes read=1388757

        FILE: Number of bytes written=2818414

        FILE: Number of read operations=0

        FILE: Number of large read operations=0

        FILE: Number of write operations=0

        HDFS: Number of bytes read=778459

        HDFS: Number of bytes written=89073

        HDFS: Number of read operations=33

        HDFS: Number of large read operations=0

        HDFS: Number of write operations=6

    Map-Reduce Framework

        Map input records=1777

        Map output records=35862

        Map output bytes=369283

        Map output materialized bytes=145633

        Input split bytes=360

        Combine input records=35862

        Combine output records=10405

        Reduce input groups=8617

        Reduce shuffle bytes=145633

        Reduce input records=10405

        Reduce output records=8617

        Spilled Records=20810

        Shuffled Maps =3

        Failed Shuffles=0

        Merged Map outputs=3

        GC time elapsed (ms)=151

        Total committed heap usage (bytes)=782516224

    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=227908

    File Output Format Counters

        Bytes Written=89073

1. **实验体会**
2. PPT指引部分指令有些老，希望可以更新一下，一些部分全靠自由发挥和百度google，由于和PPT流程不太一致所以一开始做的有点怀疑自己
3. 实验中经常遇到权限问题，需要提前给一下文件夹权限
4. 配置文件部分感觉可以在PPT中多强调一下，一开始直接掠过那页PPT了以为缺省就可以😂
5. 实验后的单词统计通过观察可以发现混入了一些英文单词+符号，不过因为是wordcount实现问题，并未理会。
6. 感谢群里面同学们的答疑，对我的帮助很大。