

实验 1 191220154 张涵之

1. 系统安装运行情况

- a) 虚拟机为计算机网络实验的 njucs-VirtualBox，版本为 Ubuntu 18.04.3
- b) 安装的 Java 版本为 1.7.0_79（之前听说版本太高会不支持，再加上开始没仔细看课件，不小心下了个 1.6.x，装 Hadoop 的时候不兼容报错，又老老实实卸了重装。另外发现每次解压以后 source /etc/profile 在同一个终端里 java -version 显示正常，开个新终端又 command not found，于是试了一般解压、直接执行 .bin、rpm，算是对软件的不同安装方式有了全新的了解，最后发现不管哪种只要重启一下虚拟机就好了）。
- c) IntelliJ: idea-IC-213.7172.25
- d) Maven: apache-maven-3.8.5
- e) Java、IntelliJ、Maven 都安装在 /usr 目录下
- f) Hadoop: hadoop-2.7.1，安装在 /home/hadoop/hadoop_installs 下

java	2 items	10:48
IntelliJ	2 items	15:00
bin	1,668 items	19:53
maven	2 items	20:11
hadoop-2.7.1	11 items	11:18
hadoop-2.7.1.tar.gz	210.6 MB	Yesterday
hdfs	2 items	09:28
mapred	2 items	10:02

Overview 'njucs-VirtualBox:9000' (active)

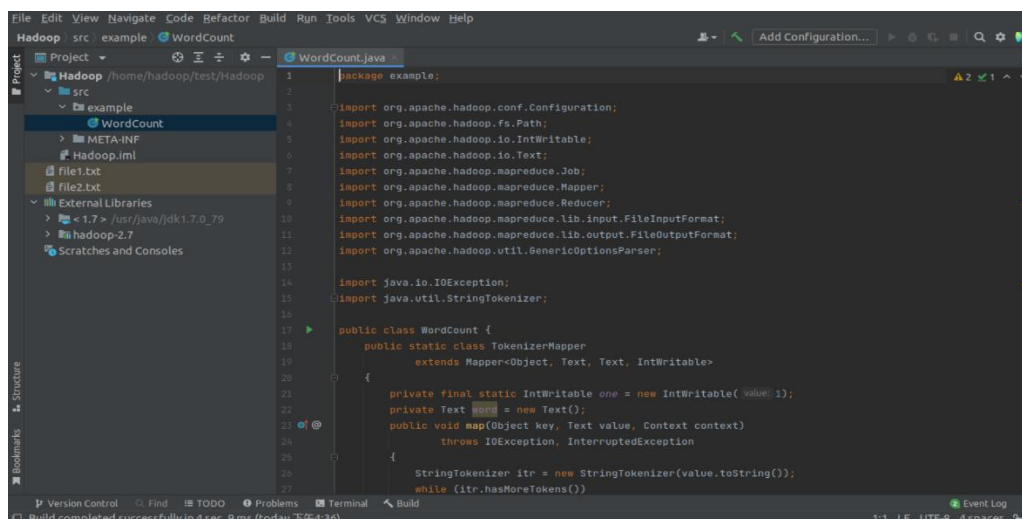
Started:	Sun Mar 27 16:19:41 CST 2022
Version:	2.7.1, r15ecc87ccf4a0228f35af08fc56de536e6ce657a
Compiled:	2015-06-29T06:04Z by jenkins from (detached from 15ecc87)
Cluster ID:	CID-395371a2-18ac-4a56-909a-2c2b4d48ee5d
Block Pool ID:	BP-1403897165-127.0.1.1-1648350564855

Directory: /logs/

SecurityAuth-hadoop.audit	0 bytes Mar 27, 2022 11:18:58 AM
hadoop-hadoop-datanode-njucs-VirtualBox.log	135467 bytes Mar 27, 2022 11:54:51 PM
hadoop-hadoop-datanode-njucs-VirtualBox.out	721 bytes Mar 27, 2022 11:49:49 PM
hadoop-hadoop-datanode-njucs-VirtualBox.out.1	721 bytes Mar 27, 2022 4:19:46 PM
hadoop-hadoop-datanode-njucs-VirtualBox.out.2	721 bytes Mar 27, 2022 11:26:32 AM
hadoop-hadoop-datanode-njucs-VirtualBox.out.3	721 bytes Mar 27, 2022 11:19:03 AM
hadoop-hadoop-namenode-njucs-VirtualBox.log	203388 bytes Mar 27, 2022 11:57:34 PM
hadoop-hadoop-namenode-njucs-VirtualBox.out	4965 bytes Mar 27, 2022 11:51:02 PM

遇到的问题:

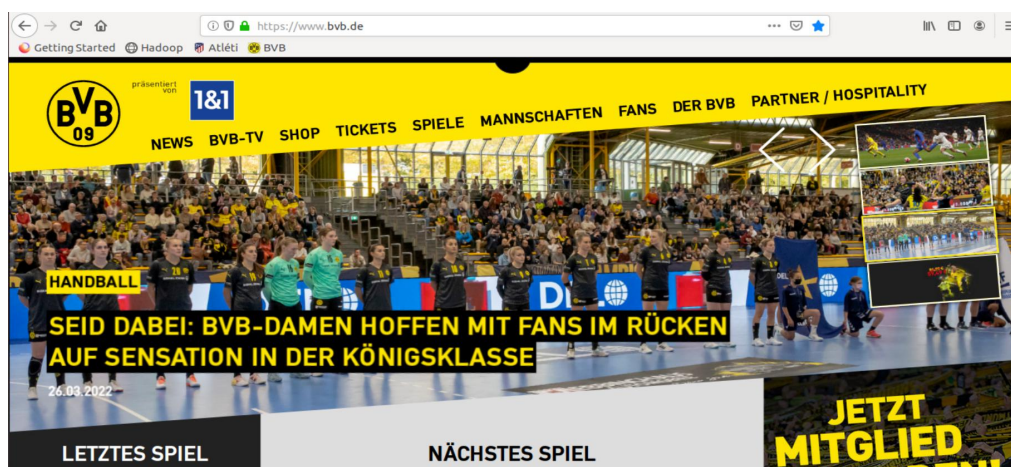
- 开始装的时候,对绝对和相对路径的理解一塌糊涂,时而加“/”时而不加,还经常纳闷为什么找不到文件,像一个绝望的电脑盲。经过一系列操作,终于明白了平时打开终端“:”和“\$”中间的“~”是啥。
- 又听说装机配环境的时候最好经常备份,装一个软件来一个快照,于是最后弄出了十几个备份,跟玩游戏似的,一言不合就读档。
- 刚开始装还做一步截一个图,做一步截一个图,再后来发现老是做错改来改去又读档,就完全懒得截图了,只想赶紧装完。
- 复制报错信息到搜索引擎,得到的结果非常良莠不齐。
- 不知道要导入哪些外部依赖 jar 包才能正常编译,觉得很麻烦。
- 生成 jar 包的时候可以指定一个主类 MainClass 作为默认执行的类,然后使用 `hadoop jar xxx.jar wordcount /input /output` 运行的时候参数报错,去掉 wordcount 可以正常运行,猜测 Hadoop 包提供的 example jar 含有很多类且没有指定主类,所以运行的时候才需要加上 wordcount。



```
1 package example;
2
3 import org.apache.hadoop.conf.Configuration;
4 import org.apache.hadoop.fs.Path;
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Job;
8 import org.apache.hadoop.mapreduce.Mapper;
9 import org.apache.hadoop.mapreduce.Reducer;
10 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
11 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
12 import org.apache.hadoop.util.GenericOptionsParser;
13
14 import java.io.IOException;
15 import java.util.StringTokenizer;
16
17 public class WordCount {
18     public static class TokenizerMapper
19         extends Mapper<Object, Text, Text, IntWritable>
20     {
21         private final static IntWritable one = new IntWritable(1);
22         private Text word = new Text();
23         public void map(Object key, Text value, Context context)
24             throws IOException, InterruptedException
25         {
26             StringTokenizer itr = new StringTokenizer(value.toString());
27             while (itr.hasMoreTokens())
```

2. 实验数据说明

从德甲足球队多特蒙德的官网 <https://www.bvb.de/> 选取最近的十余条新闻,试图直接下载 HTML 文件,但发现那样会混杂很多与新闻正文无关的文本如页面顶部菜单、侧边栏、底部的其他链接等。所以最后还是采取了原始的复制文章粘贴到纯文本 (.txt) 文件里,最后得到 11 个.txt 文件。



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	3.18 KB	3/27/2022, 7:46:07 PM	1	128 MB	Borussia_Dortmund_holt_Harma_van_Kreij_zurück.txt
-rw-r--r--	hadoop	supergroup	655 B	3/27/2022, 7:46:07 PM	1	128 MB	DFL_legt_Spieltermine_bis_Saisonende_fest.txt
-rw-r--r--	hadoop	supergroup	2.57 KB	3/27/2022, 7:46:07 PM	1	128 MB	Heimspiel_gegen_Bensheim_Auerbach_abgesagt.txt
-rw-r--r--	hadoop	supergroup	963 B	3/27/2022, 7:46:07 PM	1	128 MB	Jude_Bellingham_gewinnt_NXGN-Award_2022.txt
-rw-r--r--	hadoop	supergroup	2.28 KB	3/27/2022, 7:46:07 PM	1	128 MB	Kartenvorverkauf_für_Heimspiel_gegen_Leipzig_gestartet.txt
-rw-r--r--	hadoop	supergroup	523 B	3/27/2022, 7:46:07 PM	1	128 MB	Neuer_Spieltermin_der_BVB-Frauen_für_das_Kreispokal-Halbfinale.txt
-rw-r--r--	hadoop	supergroup	2.49 KB	3/27/2022, 7:46:07 PM	1	128 MB	T-T_T_Was_Rose_an_Wolf_liebt.txt
-rw-r--r--	hadoop	supergroup	2.71 KB	3/27/2022, 7:46:07 PM	1	128 MB	U23_verliert_Heimspiel_am_Montagabend_gegen_Eintracht_Braunschweig.txt

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

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus
application_1648441419677_0003	hadoop	word count	MAPREDUCE	default	Mon Mar 28 13:32:11 +0800 2022	Mon Mar 28 13:32:32 +0800 2022	FINISHED	SUCCEEDED
application_1648441419677_0002	hadoop	word count	MAPREDUCE	default	Mon Mar 28 12:57:58 +0800 2022	Mon Mar 28 12:58:21 +0800 2022	FINISHED	SUCCEEDED
application_1648441419677_0001	hadoop	word count	MAPREDUCE	default	Mon Mar 28 12:24:41 +0800 2022	Mon Mar 28 12:24:55 +0800 2022	FINISHED	SUCCEEDED

Showing 1 to 3 of 3 entries

三个作业分别是安装过程中测试时用 Hadoop 自带的 example jar 包统计少量简单的文件（课件中的 file1.txt 和 file2.txt），用 Hadoop 自带 example 统计德语文章，和用自己编写的 WordCount 统计德语文章，后面两个结果相同。

4. 实验输出结果开头部分的屏幕拷贝

```
hadoop@njucs-VirtualBox:~$ hadoop jar test/wordcount.jar /input /output
22/03/28 13:32:10 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/03/28 13:32:10 INFO input.FileInputFormat: Total input paths to process : 11
22/03/28 13:32:10 INFO mapreduce.JobSubmitter: number of splits:11
22/03/28 13:32:11 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648441419677_0003
22/03/28 13:32:11 INFO impl.YarnClientImpl: Submitted application application_1648441419677_0003
22/03/28 13:32:11 INFO mapreduce.Job: The url to track the job: http://njucs-VirtualBox:8088/track/1648441419677_0003
22/03/28 13:32:11 INFO mapreduce.Job: Running job: job_1648441419677_0003
22/03/28 13:32:15 INFO mapreduce.Job: Job job_1648441419677_0003 running in uber mode : false
22/03/28 13:32:15 INFO mapreduce.Job: map 0% reduce 0%
22/03/28 13:32:24 INFO mapreduce.Job: map 36% reduce 0%
22/03/28 13:32:25 INFO mapreduce.Job: map 55% reduce 0%
22/03/28 13:32:31 INFO mapreduce.Job: map 64% reduce 0%
22/03/28 13:32:32 INFO mapreduce.Job: map 91% reduce 0%
22/03/28 13:32:33 INFO mapreduce.Job: map 100% reduce 0%
22/03/28 13:32:34 INFO mapreduce.Job: map 100% reduce 100%
22/03/28 13:32:34 INFO mapreduce.Job: Job job_1648441419677_0003 completed successfully
22/03/28 13:32:34 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=28343
  FILE: Number of bytes written=1443988
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=23106
  HDFS: Number of bytes written=16108
  HDFS: Number of read operations=36
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
```

注意到这个 word count 程序不是很智能，仅以空格作为单词之间切分的标准而没有考虑标点符号和大小写，所以前后带有标点的同一个单词会分开统计多次，如“Dortmund”、“Dortmund,”、“Dortmund.”和“(Dortmund”。

```
Dortmund      13
Dortmund,     2
Dortmund.     3
```



The screenshot shows a text editor window titled 'part-r-00000' with a file path of '~/Downloads'. The editor contains a list of German words and their corresponding frequencies, separated by spaces. The words are listed in ascending order of frequency. The status bar at the bottom indicates 'Plain Text', 'Tab Width: 8', 'Ln 1, Col 1', and 'INS'.

Word	Frequency
Aber	2
Abgleich	1
Abschlüsse	1
Abstände	1
Abteilungsleiter	1
Akanji	1
Aktion	1
Alle	1
Allrounderin,	1
Alters	1
Am	4
An	1
Anastasiya	1
Andersson	1
Andreas	1
André	3
Anlass	1
Anreise.	1
Anschluss	1
Anstoßzeit	1
Anstoßzeiten	1
Ansu	1
Antigenschnelltest	1
Antl,	1
Antonios	1
Apple	1
April	7

5. 实验体会

- 课件和网络教程里的命令和配置文件别急着复制，看清楚干啥的，免得有些文件路径和 IP 地址也无脑复制了，后续白天找不出原因。
- 配环境的时候，多搞几个快照存档非常有用，之前配置文件没弄好只在本地跑网站上找不到作业记录，改了一会 DataNode 消失了，又改了一会索性连 NodeManager 也消失了，赶紧读档从头再来。