# 華中科技大學

# 课程报告

**课程名称：** 大数据导论

专业班级：CS1802

学　　号：U201814531

姓　　名：李响

指导教师：肖江

报告日期：2019. 5

计算机科学与技术学院

# 目 录

## 1.1 测试环境说明

### 1.1.1 CPU 配置：

CPU Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz

CPU 核心数 4

二级缓存 512 KB

三级缓存 3072 KB

数据宽度 64bit

### 1.1.2 虚拟机配置

虚拟机软件：VMware 14.0.0 ；

Linux 系统版本：Ubuntu 18.04.1 ；

虚拟机配置（Master，Slave1 及 Slave2 配置相同）：

内存：2GB，硬盘：20GB，处理器数：1，网络适配器：NAT 模式 ；

### 1.1.3 Hadoop 以及 java 配置

Hadoop 版本：Hadoop 2.9.1；

Hadoop 路径：/usr/local/hdaoop；

java 版本： jdk 1.8.0_211；

java 路径：/usr/java；

Java(TM) SE Runtime Environment (build 1.8.0_211-b12)

Java HotSpot(TM) 64-Bit Server VM (build 25.211-b12, mixed mode)

### 1.1.4 数据集特征说明

采用NLP的一个word2vector数据集，其大小为 100M左右，数据集中包含 17 万个英文单词，数据集较大，且具有普遍性。

## 1.2 测试应用说明

测试应用项目：WordCount（英文字符统计）；其意义是通过大数据集的字符出现频率的统计，来达成某种目的，其通过多节点的计算和存储来完成英文字符数的统计。

## 1.3 研究目的及意义

我们采用WordCount作为我们的课程作业，第一个原因是其入门简单，对于相关问题的解决程序已被打包，可以直接调用；第二个原因是，WordCount对文本的要求不高，可以对基础文本进行操作；第三个原因是，WordCount具有较大的实际应用价值，非常多的实验生活场景需要进行文本处理，需求市场广阔；除此之外，文本计数是NLP相关问题的基础，可

以为我们以后的学习奠定一定的基础。

## 1.4 问题挑战

由于是初次接触分布式的系统构架，甚至是第一次接触虚拟机，而且同学之中也没有特别擅长这方面的人，因此我们采用了网上的教程指导，由于版本和虚拟化环境的差异，以及各种教程品质的良莠不齐，我们在搭建过程中遇到了一系列问题（本次采用的虚拟化环境为Ubuntu18.04.1 版本）。但大多数问题是由于初次上手的原因导致的，并不十分典型，故在此挑选几个我认为比较重要的问题陈述。

**问题 1：**在设置ssh免密码访问时，总是无法互相访问，总是提示需要密码。

**解决方法：**发现一直在root用户下进行操作，而没有在Hadoop用户下进行操作，导致代码和文件信息不匹配，造成错误，调整用户进行操作，顺利完成。

**问题 2：**配置Hadoop文件时，遇到的问题异常多，由于不是十分了解代码含义，导致配置时常常“驴唇不对马嘴”的错配情况，对于这种错误，我们的解决方法是通过寻找更多的教程，了解代码含义，进行修改和配置，并通过阅读错误报告来找到问题，以下是其中一个具体的例子：Error: Java heap space，堆错误（空间不足导致溢出）。

**解决方案：**hadoop的文件配置错误，hadoop-env.sh文件中，export HADOOP_HEAPSIZE=的值过小，由于我给系统分配的内存是 2GB，根据内存大小来选择值，赋值 2000，然后再然后修改完修改mapred-site.xml文件，添加下面的语句：

```
<property>
    <name>mapred.child.java.opts</name>
    <value>-Xmx2000m</value>
</property>
```

成后，同步到集群其他节点上。（诸如此类问题还有许多，同类问题便不再赘述）

**问题 3：**start-all.sh命令下达后datanode未启动。

**解决方案：**多次hadoop namenode -format格式化namenode时，会在namenode数据文件夹中保存一个current/VERSION文件，记录clusterID，datanode中保存的current/VERSION文件中的clustreID的值是上一次格式化保存的clusterID，这样，datanode和namenode之间的ID不一致，通过删除hdfs文件夹下的data和name文件再重新hadoop namenode -format格式化后正常。

## 1.5 测试结果

本次测试使用的数据集是大小为 100MB 左右的 NLP 英文单词集，使用的测试程序为 WordCount 用于统计各个英文单词出现的频次。输出的文件有两个，分别为显示成功的文件“_SUCCESS”与统计文件“part-r-00000”其中有出现的单词以及各个单词的出现频数。具

体运行以及结果图如下，图 1-图 3。

```
hadoop@master:~$ hadoop fs -ls /test/output6
Found 2 items
-rw-r--r--   3 hadoop supergroup          0 2019-05-03 23:56 /test/output6/_SUCCESS
-rw-r--r--   3 hadoop supergroup    2806033 2019-05-03 23:56 /test/output6/part-r-00000
```

图 1 WordCount 输出文件位置以及名称图

```
hadoop@master:~$ hadoop dfsadmin -report
DEPRECATED: Use of this script to execute hdfs com
Instead use the hdfs command for it.

Configured Capacity: 63004459008 (58.68 GB)
Present Capacity: 31909691392 (29.72 GB)
DFS Remaining: 31585619968 (29.42 GB)
DFS Used: 324071424 (309.06 MB)
DFS Used%: 1.02%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (3):

Name: 192.168.100.15:50010 (master)
Hostname: master
Decommission Status : Normal
Configured Capacity: 21001486336 (19.56 GB)
DFS Used: 108023808 (103.02 MB)
Non DFS Used: 9738260480 (9.07 GB)
DFS Remaining: 10064789504 (9.37 GB)
DFS Used%: 0.51%
DFS Remaining%: 47.92%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun May 05 05:03:02 PDT 2019
Last Block Report: Sun May 05 04:47:30 PDT 2019

Name: 192.168.100.16:50010 (slave1)
Hostname: slave1
Decommission Status : Normal
Configured Capacity: 21001486336 (19.56 GB)
DFS Used: 108023808 (103.02 MB)
Non DFS Used: 9042755584 (8.42 GB)
DFS Remaining: 10760294400 (10.02 GB)
DFS Used%: 0.51%
DFS Remaining%: 51.24%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun May 05 05:03:04 PDT 2019
Last Block Report: Sun May 05 04:47:14 PDT 2019


Name: 192.168.100.17:50010 (slave2)
Hostname: slave2
Decommission Status : Normal
Configured Capacity: 21001486336 (19.56 GB)
DFS Used: 108023808 (103.02 MB)
Non DFS Used: 9042513920 (8.42 GB)
DFS Remaining: 10760536064 (10.02 GB)
DFS Used%: 0.51%
DFS Remaining%: 51.24%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Sun May 05 05:03:04 PDT 2019
Last Block Report: Sun May 05 04:47:14 PDT 2019
```

图 2 Master/Slave1/Slave2 节点状态报告示意图

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| moderado | 1 | moderados | 3 | moderate | 430 | moderated | 54 |
| moderately | 97 | moderates | 33 | moderating | 21 | moderation | 57 |
| moderations | 4 | moderatism | 2 | moderative | 1 | moderato | 2 |
| moderator | 38 | moderators | 13 | moderland | 1 | modern | 7790 |
| moderner | 2 | modernes | 6 | moderni | 3 | modernisation | 25 |
| modernisations | 1 | modernise | 8 | modernised | 16 | modernisers | 3 |

图 3 WordCount 部分测试结果示意图

## 1.6 角色分工

我们小组是 CS1802 的第一组，成员包括：吴晨，柳昕，杨彪，张峥，李响，魏子清，黎奕辰。杨彪，柳昕，黎奕辰三人负责打包程序并测试数据，李响（本人）负责搭建系统和测试系统运行状态，吴晨和魏子清负责制作 PPT，张峥负责答辩和演讲。

## 1.7 心得体会与总结

由于我们小组中并没有曾经做过大数据分布式系统的大佬，大家对分布式系统的搭建都不太熟悉，所以我们采取了广撒网的策略，让每个人都单独配置系统，然后采用最快完成同学的系统，而我有幸最快完成系统的配置。但是作为一个 Hadoop 和 HDFS 方面完全的小白，这次搭建系统和测试工作确实给了我极大的挑战，甚至是困扰。由于没有任何经验，而且同学之中也没有特别擅长这方面的人，所以我采用了网上的教程指导，但是由于版本和虚拟化环境的差异，以及各种教程品质的良莠不齐，我在搭建过程中遇到了一系列问题。虽然很多问题都带给我长时间的困扰，但是在我耐心的一遍遍尝试下，问题都得到了较好的解决，不得不说的是，尝试一门新的技术确实是让人“痛并快乐着”，时间总是在心流状态下悄悄流逝，最终在我的不懈努力下，我在凌晨三点完成了系统的配置，虽然确实让人疲惫不堪，但也带给了我满满的成就感。经过这一次的学习和体验，我了解到了大数据的些许魅力，我也希望这次的经历可以为我之后的学习打开新的方向和思路。