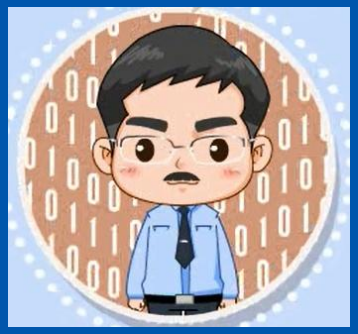


《大数据基础编程、实验和案例教程（第2版）》

教材官网：
http://dblab.xmu.edu.cn/post/bigdatapractice2/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第3章 Hadoop的安装和使用

（PPT版本号：2020年12月版本）

扫一扫访问教材官网

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

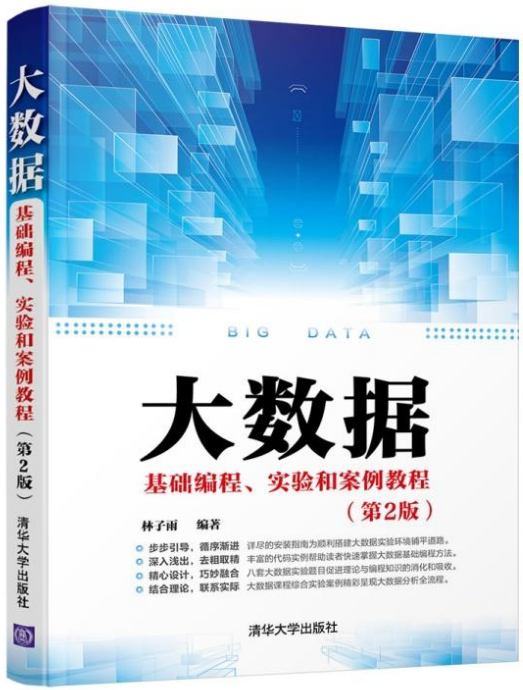
主页：http://dblab.xmu.edu.cn/linziyu

教材简介

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

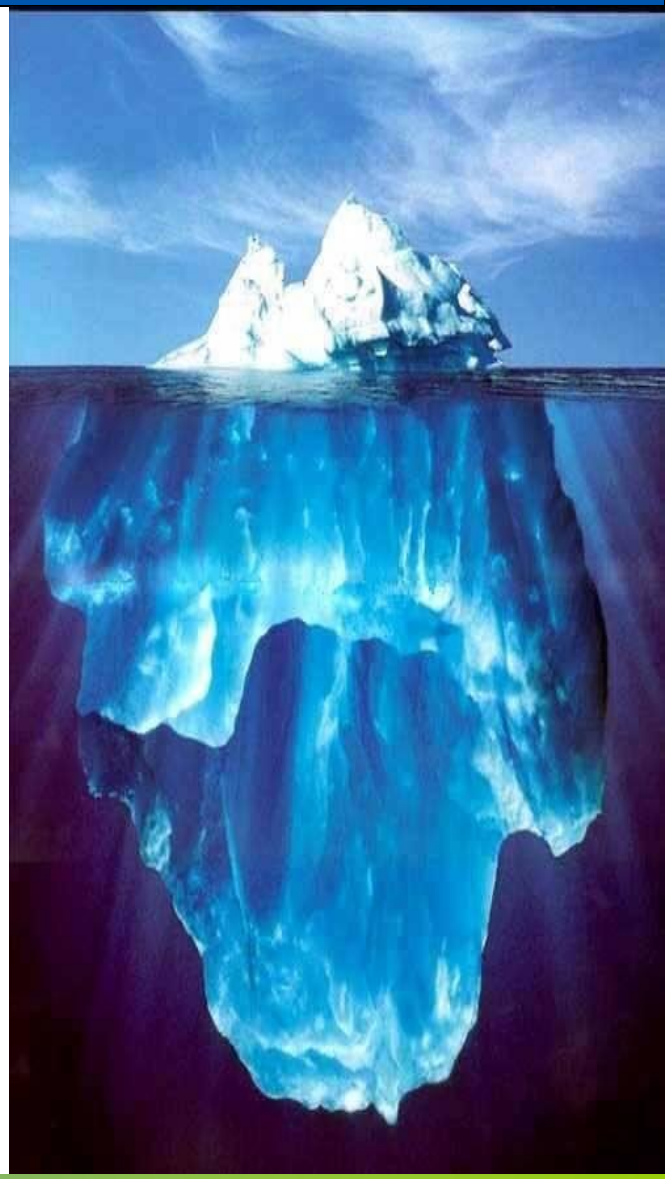
林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元，2020年10月第2版
教材官网： http://dblab.xmu.edu.cn/post/bigdatapractice2/

扫一扫访问
教材官网



- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

提纲



高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

3.1 Hadoop简介

•Hadoop是Apache软件基金会旗下的一个开源分布式计算平台，为用户提供了系统底层细节透明的分布式基础架构。Hadoop是基于Java语言开发的，具有很好的跨平台特性，并且可以部署在廉价的计算机集群中。Hadoop的核心是分布式文件系统（Hadoop Distributed File System，HDFS）和MapReduce。

•Apache Hadoop版本分为三代，分别是Hadoop 1.0、Hadoop 2.0和Hadoop3.0。

•除了免费开源的Apache Hadoop以外，还有一些商业公司推出Hadoop的发行版。2008年，Cloudera成为第一个Hadoop商业化公司，并在2009年推出第一个Hadoop发行版。此后，很多大公司也加入了做Hadoop产品化的行列，比如MapR、Hortonworks、星环等。2018年10月，Cloudera和Hortonworks宣布合并。一般而言，商业化公司推出的Hadoop发行版也是以Apache Hadoop为基础，但是前者比后者具有更好的易用性、更多的功能以及更高的性能。

3.2 安装Hadoop前的准备工作

3.2.1 创建hadoop用户
3.2.2 更新APT
3.2.3 安装SSH
3.2.4 安装Java环境

3.2.1 创建hadoop用户

本教程全部采用hadoop用户登录Linux系统，并为hadoop用户增加了管理员权限。在前面的“第2章 Linux系统的安装和使用”内容中，已经介绍了hadoop用户创建和增加权限的方法，请一定按照该方法创建hadoop用户，并且使用hadoop用户登录Linux系统，然后再开始下面的学习内容。本教程所有学习内容，都是采用hadoop用户登录Linux系统。

3.2.2 更新APT

本教程第2章介绍了APT软件作用和更新方法，为了确保Hadoop安装过程顺利进行，建议按照第2章介绍的方法，用hadoop用户登录Linux系统后打开一个终端，执行下面命令更新APT软件：

```
$ sudo apt-get update
```

3.2.3 安装SSH

Ubuntu默认已安装了SSH客户端，因此，这里还需要安装SSH服务端，请在Linux的终端中执行以下命令：

```
$ sudo apt-get install openssh-server
```

安装后，可以使用如下命令登录本机：

```
$ ssh localhost
```

执行该命令后会出现如图3-1所示的提示信息(SSH首次登录提示)，输入“yes”，然后按提示输入密码hadoop，就登录到本机了。

```
hadoop@DBLab-XMU:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is a9:28:e0:4e:89:40:a4:cd:75:8f:0b:8b:57:79:67:86.
Are you sure you want to continue connecting (yes/no)? yes
```

3.2.3 安装SSH

首先，请输入命令“exit”退出刚才的SSH，就回到了原先的终端窗口；然后，可以利用ssh-keygen生成密钥，并将密钥加入到授权中，命令如下：

```
$ cd ~/.ssh/       # 若没有该目录，请先执行一次ssh localhost
$ ssh-keygen -t rsa    # 会有提示，都按回车即可
$ cat ./id_rsa.pub >> ./authorized_keys  # 加入授权
```

此时，再执行ssh localhost命令，无需输入密码就可以直接登录了，如图所示。

```
hadoop@ubuntu:~$ ssh localhost
Welcome to Ubuntu 16.04 LTS (GNU/Linux 4.4.0-171-generic
 x86_64)

 * Documentation:  https://help.ubuntu.com/

399 packages can be updated.
19 updates are security updates.

Last login: Sun Jan 26 10:59:05 2020 from 192.168.20.1
```

3.2.4 安装Java环境

执行如下命令创建“/usr/lib/jvm”目录用来存放JDK文件：

```
$cd /usr/lib
$sudo mkdir jvm #创建/usr/lib/jvm目录用来存放JDK文件
```

执行如下命令对安装文件进行解压缩：

```
$cd ~ #进入hadoop用户的主目录
$cd Downloads
$sudo tar -zxvf ./jdk-8u162-linux-x64.tar.gz -C /usr/lib/jvm
```

3.2.4 安装Java环境

下面继续执行如下命令，设置环境变量：

```
$vim ~/.bashrc
```

上面命令使用vim编辑器打开了hadoop这个用户的环境变量配置文件，请在这个文件的开头位置，添加如下几行内容：

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_162
export JRE_HOME=${JAVA_HOME}/jre
export CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
export PATH=${JAVA_HOME}/bin:$PATH
```

3.2.4 安装Java环境

保存.bashrc文件并退出vim编辑器。然后，继续执行如下命令让.bashrc文件的配置立即生效：

```
$source ~/.bashrc
```

这时，可以使用如下命令查看是否安装成功：

```
$java -version
```

如果能够在屏幕上返回如下信息，则说明安装成功：

```
java version "1.8.0_162"
Java(TM) SE Runtime Environment (build 1.8.0_162-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.162-b12, mixed mode)
```

3.3 安装Hadoop

Hadoop包括三种安装模式：

•单机模式：只在一台机器上运行，存储是采用本地文件系统，没有采用分布式文件系统HDFS；

•伪分布式模式：存储采用分布式文件系统HDFS，但是，HDFS的名称节点和数据节点都在同一台机器上；

•分布式模式：存储采用分布式文件系统HDFS，而且，HDFS的名称节点和数据节点位于不同机器上。

3.3.1 下载安装文件

本教程采用的Hadoop版本是3.1.3，可以到Hadoop官网下载安装文件（http://mirrors.cnnic.cn/apache/hadoop/common/）

请使用hadoop用户登录Linux系统，打开一个终端，执行如下命令：

```
$ sudo tar -zxf ~/下载/hadoop-3.1.3.tar.gz -C /usr/local    # 解压到/usr/local中
$ cd /usr/local/
$ sudo mv ./hadoop-3.1.3/ ./hadoop      # 将文件夹名改为hadoop
$ sudo chown -R hadoop ./hadoop       # 修改文件权限
```

Hadoop解压后即可使用，可以输入如下命令来检查 Hadoop是否可用，成功则会显示 Hadoop版本信息：

```
$ cd /usr/local/hadoop
$ ./bin/hadoop version
```

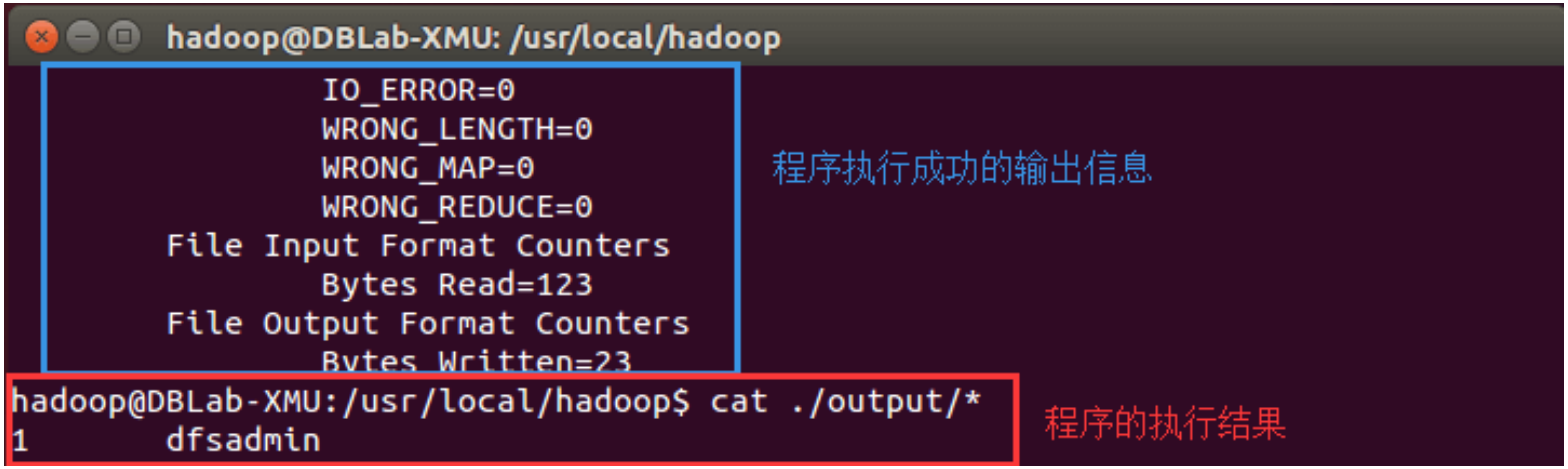
3.3.2 单机模式配置

Hadoop默认模式为非分布式模式（本地模式），无需进行其他配置即可运行。Hadoop附带了丰富的例子，运行如下命令可以查看所有例子：

```
$ cd /usr/local/hadoop
$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.3.jar
```

这里选择运行grep例子

```
$ cd /usr/local/hadoop
$ mkdir input
$ cp ./etc/hadoop/*.xml ./input   # 将配置文件复制到input目录下
$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar grep ./input ./output 'dfs[a-z.]+'
$ cat ./output/*          # 查看运行结果
```



3.3.3 伪分布式模式配置

1. 修改配置文件

修改以后，core-site.xml文件的内容如下：

```
<configuration>
    <property>
        <name>hadoop.tmp.dir</name>
        <value>file:/usr/local/hadoop/tmp</value>
        <description>Abase for other temporary directories.</description>
    </property>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

3.3.3 伪分布式模式配置

同样，需要修改配置文件hdfs-site.xml，修改后的内容如下：

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:/usr/local/hadoop/tmp/dfs/name</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:/usr/local/hadoop/tmp/dfs/data</value>
    </property>
</configuration>
```

3.3.3 伪分布式模式配置

2. 执行名称节点格式化

修改配置文件以后，要执行名称节点的格式化，命令如下：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs namenode -format
```

如果格式化成功，会看到“successfully formatted”的提示信息

```
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = hadoop/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:  version = 3.1.3
************************************************************
*****/
......
2020-01-08 15:31:35,677 INFO common.Storage: Storage dir
ectory /usr/local/hadoop/tmp/dfs/name has been successfu
lly formatted.
......
/************************************************************
*****
SHUTDOWN_MSG: Shutting down NameNode at hadoop/127.0.1.1
************************************************************
*****/
```

3.3.3 伪分布式模式配置

3. 启动Hadoop

执行下面命令启动Hadoop：

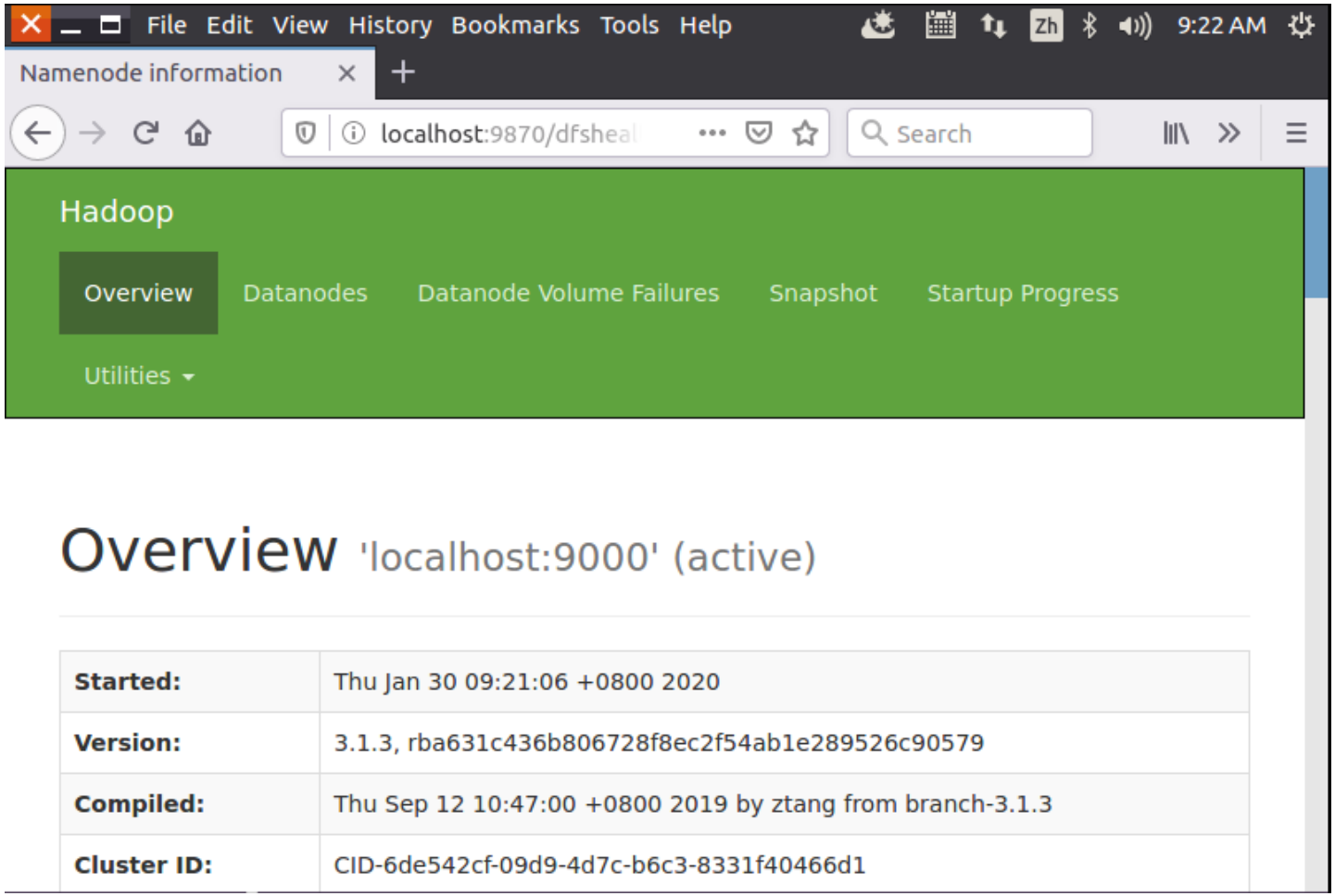
```
$ cd /usr/local/hadoop
$ ./sbin/start-dfs.sh  #start-dfs.sh是个完整的可执行文件，中间没有空格
```

如果出现如图3-5所示的SSH提示，输入yes即可：

```
hadoop@DBLab-XMU:/usr/local/hadoop$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-na
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-da
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is a9:28:e0:4e:89:40:a4:cd:75:8f:0b:8b:57:79:67:86.
Are you sure you want to continue connecting (yes/no)? yes
```

3.3.3 伪分布式模式配置

5. 使用Web界面查看HDFS信息



3.3.3 伪分布式模式配置

6. 运行Hadoop伪分布式实例

要使用HDFS，首先需要在HDFS中创建用户目录（本教程全部统一采用hadoop用户名登录Linux系统），命令如下：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -mkdir -p /user/hadoop
```

接着需要把本地文件系统的“/usr/local/hadoop/etc/hadoop”目录中的所有xml文件作为输入文件，复制到分布式文件系统HDFS中的“/user/hadoop/input”目录中，命令如下：

```
$ cd /usr/local/hadoop
$ ./bin/hdfs dfs -mkdir input  #在HDFS中创建hadoop用户对应的input目录
$ ./bin/hdfs dfs -put ./etc/hadoop/*.xml input  #把本地文件复制到HDFS中
```

3.3.3 伪分布式模式配置

现在就可以运行Hadoop自带的grep程序，命令如下：

```
$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.3.jar grep input output 'dfs[a-z.]+'
```

运行结束后，可以通过如下命令查看HDFS中的output文件夹中的内容：

```
$ ./bin/hdfs dfs -cat output/*
```

执行结果如图所示

```
hadoop@DBLab-XMU:/usr/local/hadoop$ bin/hdfs dfs -cat output/*
1       dfsadmin
1       dfs.replication
1       dfs.namenode.name.dir
1       dfs.datanode.data.dir
```

3.3.3 伪分布式模式配置

7. 关闭Hadoop

如果要关闭Hadoop，可以执行下面命令：

```
$ cd /usr/local/hadoop
$ ./sbin/stop-dfs.sh
```

3.3.3 伪分布式模式配置

8. 配置PATH变量

首先使用vim编辑器打开“~/.bashrc”这个文件，然后，在这个文件的最前面位置加入如下单独一行：

```
export PATH=$PATH:/usr/local/hadoop/sbin
```

在后面的学习过程中，如果要继续把其他命令的路径也加入到PATH变量中，也需要继续修改“~/.bashrc”这个文件。当后面要继续加入新的路径时，只要用英文冒号“:”隔开，把新的路径加到后面即可，比如，如果要继续把“/usr/local/hadoop/bin”路径增加到PATH中，只要继续追加到后面，如下所示：

```
export PATH=$PATH:/usr/local/hadoop/sbin:/usr/local/hadoop/bin
```

添加后，执行命令“source ~/.bashrc”使设置生效。设置生效后，在任何目录下启动Hadoop，都只要直接输入start-dfs.sh命令即可，同理，停止Hadoop，也只需要在任何目录下输入stop-dfs.sh命令即可。

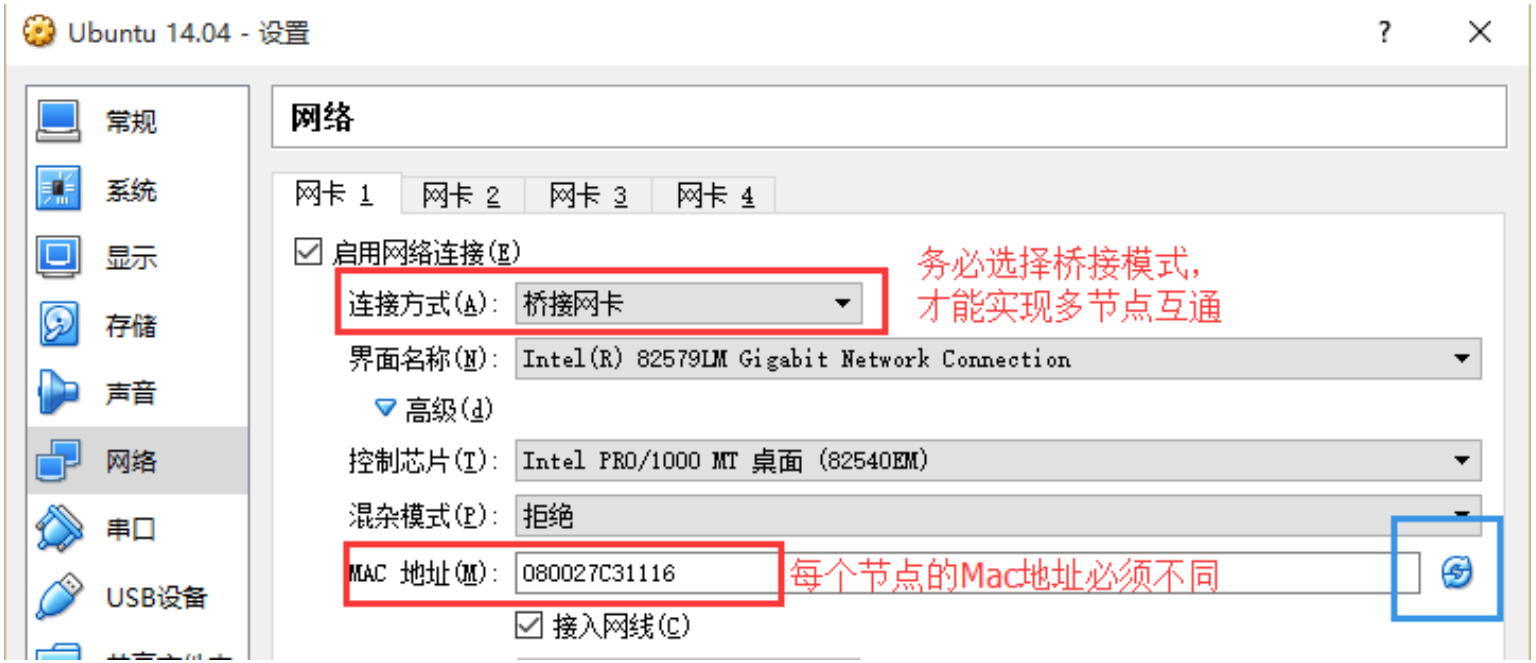
3.3.4 分布式模式配置

Hadoop 集群的安装配置大致包括以下步骤：

- 步骤1：选定一台机器作为 Master；
- 步骤2：在Master节点上创建hadoop用户、安装SSH服务端、安装Java环境；
- 步骤3：在Master节点上安装Hadoop，并完成配置；
- 步骤4：在其他Slave节点上创建hadoop用户、安装SSH服务端、安装Java环境；
- 步骤5：将Master节点上的“/usr/local/hadoop”目录复制到其他Slave节点上；
- 步骤6：在Master节点上开启Hadoop；

3.3.4 分布式模式配置

1. 网络配置



3.3.4 分布式模式配置

在Ubuntu中，我们在 Master 节点上执行如下命令修改主机名：

```
$ sudo vim /etc/hostname
```

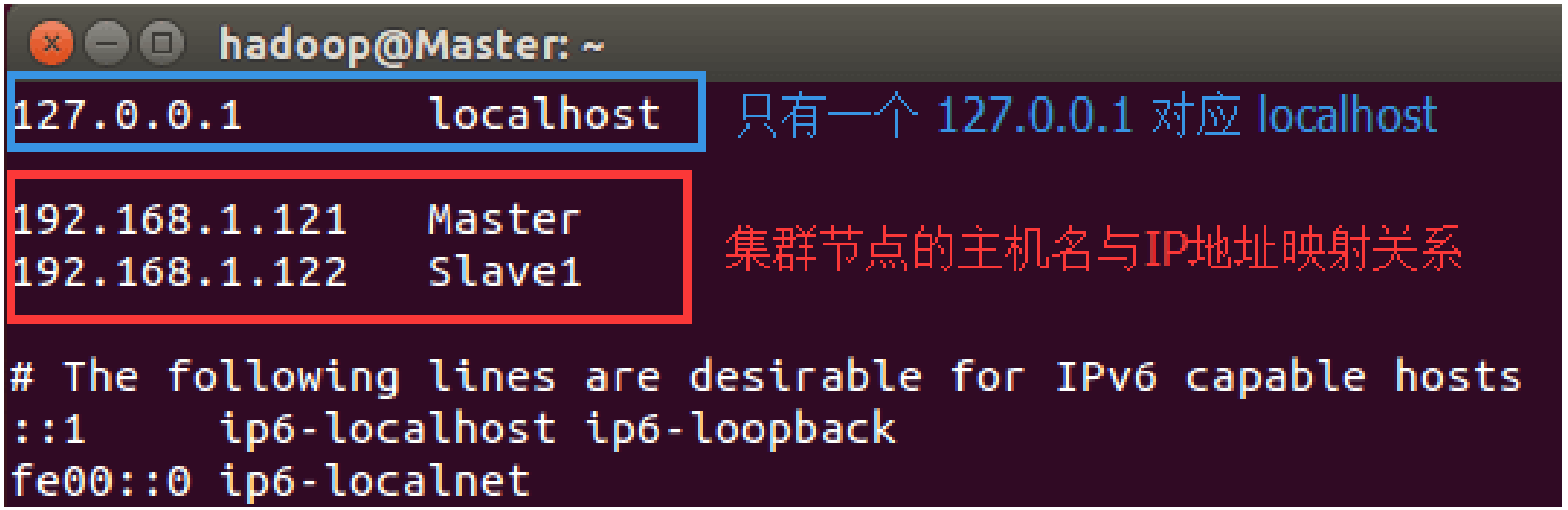
打开这个文件以后，里面就只有“dblab-VirtualBox”这一行内容，可以直接删除，并修改为“Master”（注意是区分大小写的），然后，保存退出vim编辑器，这样就完成了主机名的修改，需要重启Linux系统才能看到主机名的变化。

执行如下命令打开并修改Master节点中的“/etc/hosts”文件：

```
$ sudo vim /etc/hosts
```

```
192.168.1.121   Master
192.168.1.122   Slave1
```

3.3.4 分布式模式配置


hadoop@Master: ~
127.0.0.1 localhost
只有一个 127.0.0.1 对应 localhost
192.168.1.121 Master
192.168.1.122 Slave1
集群节点的主机名与IP地址映射关系
The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet

3.3.4 分布式模式配置

把Slave节点上的“/etc/hostname”文件中的主机名修改为“Slave1”，同时，修改“/etc/hosts”的内容，在hosts文件中增加如下两条IP和主机名映射关系：

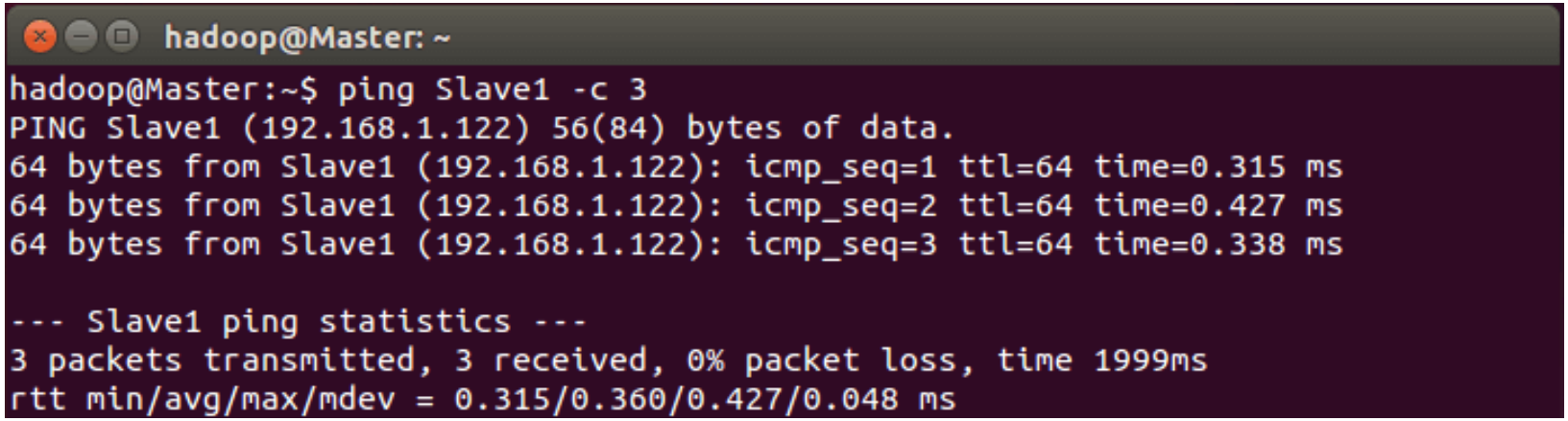
192.168.1.121 Master
192.168.1.122 Slave1

修改完成以后，请重新启动Slave节点的Linux系统。

3.3.4 分布式模式配置

需要在各个节点上都执行如下命令，测试是否相互ping得通，如果ping不通，后面就无法顺利配置成功：

```
$ ping Master -c 3   # 只ping 3次就会停止，否则要按Ctrl+c中断ping命令
$ ping Slave1 -c 3
```



3.3.4 分布式模式配置

2. SSH无密码登录节点

必须要让Master节点可以SSH无密码登录到各个Slave节点上。首先，生成Master节点的公匙，如果之前已经生成过公钥，必须要删除原来生成的公钥，重新生成一次，因为前面我们对主机名进行了修改。具体命令如下：

```
$ cd ~/.ssh               # 如果没有该目录，先执行一次ssh localhost
$ rm ./id_rsa*            # 删除之前生成的公匙（如果已经存在）
$ ssh-keygen -t rsa       # 执行该命令后，遇到提示信息，一直按回车就可以
```

为了让Master节点能够无密码SSH登录本机，需要在Master节点上执行如下命令：

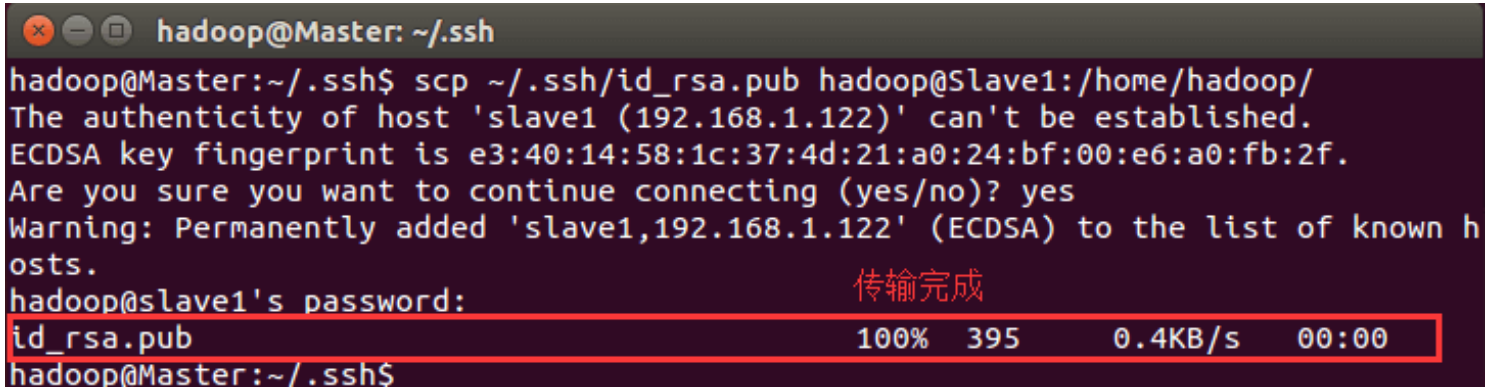
```
$ cat ./id_rsa.pub >> ./authorized_keys
```

完成后可以执行命令“ssh Master”来验证一下，可能会遇到提示信息，只要输入yes即可，测试成功后，请执行“exit”命令返回原来的终端。

3.3.4 分布式模式配置

接下来，在Master节点将上公匙传输到Slave1节点：

```
$ scp ~/.ssh/id_rsa.pub hadoop@Slave1:/home/hadoop/
```



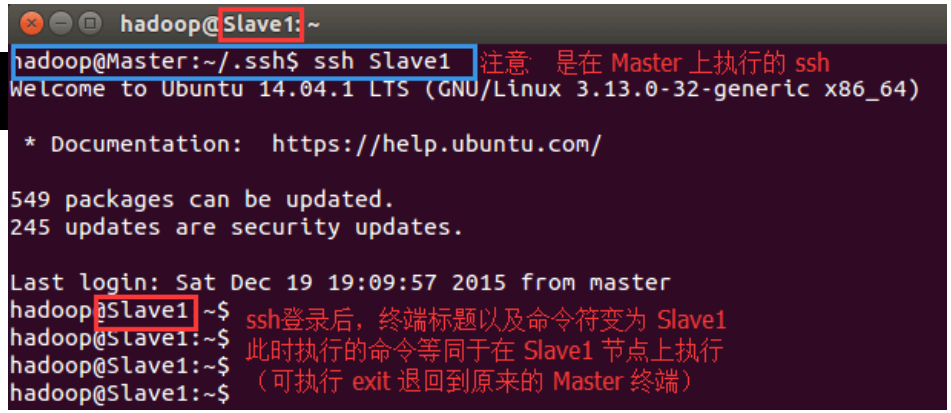
3.3.4 分布式模式配置

接着在Slave1节点上，将SSH公匙加入授权：

```
$ mkdir ~/.ssh       # 如果不存在该文件夹需先创建，若已存在，则忽略本命令
$ cat ~/id_rsa.pub >> ~/.ssh/authorized_keys
$ rm ~/id_rsa.pub    # 用完以后就可以删掉
```

如果有其他Slave节点，也要执行将Master公匙传输到Slave节点以及在Slave节点上加入授权这两步操作。
这样，在Master节点上就可以无密码SSH登录到各个Slave节点了，可在Master节点上执行如下命令进行检验：

```
$ ssh Slave1
```



3.3.4 分布式模式配置

3. 配置PATH变量

首先执行命令“vim ~/.bashrc”，也就是使用vim编辑器打开“~/.bashrc”文件，然后，在该文件最上面的位置加入下面一行内容：

```
export PATH=$PATH:/usr/local/hadoop/bin:/usr/local/hadoop/sbin
```

3.3.4 分布式模式配置

4. 配置集群/分布式环境

在配置集群/分布式模式时，需要修改“/usr/local/hadoop/etc/hadoop”目录下的配置文件，这里仅设置正常启动所必须的设置项，包括workers 、core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml共5个

（1）修改文件workers

本教程让Master节点仅作为名称节点使用，因此将workers文件中原来的localhost删除，只添加如下一行内容：

```
Slave1
```

3.3.4 分布式模式配置

（2）修改文件core-site.xml

请把core-site.xml文件修改为如下内容：

```
<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://Master:9000</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>file:/usr/local/hadoop/tmp</value>
                <description>Abase for other temporary directories.</description>
        </property>
</configuration>
```

3.3.4 分布式模式配置

（3）修改文件hdfs-site.xml

```
<configuration>
        <property>
                <name>dfs.namenode.secondary.http-address</name>
                <value>Master:50090</value>
        </property>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/usr/local/hadoop/tmp/dfs/name</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/usr/local/hadoop/tmp/dfs/data</value>
        </property>
</configuration>
```

3.3.4 分布式模式配置

（4）修改文件**mapred-site.xml**

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>mapreduce.jobhistory.address</name>
        <value>Master:10020</value>
    </property>
    <property>
        <name>mapreduce.jobhistory.webapp.address</name>
        <value>Master:19888</value>
    </property>
    <property>
            <name>yarn.app.mapreduce.am.env</name>
            <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
        </property>
        <property>
            <name>mapreduce.map.env</name>
            <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
        </property>
        <property>
            <name>mapreduce.reduce.env</name>
            <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
        </property>
</configuration>
```

3.3.4 分布式模式配置

（5）修改文件 **yarn-site.xml**

```
<configuration>
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>Master</value>
        </property>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>
```

3.3.4 分布式模式配置

首先在Master节点上执行如下命令：

```
$ cd /usr/local
$ sudo rm -r ./hadoop/tmp     # 删除 Hadoop 临时文件
$ sudo rm -r ./hadoop/logs/*   # 删除日志文件
$ tar -zcf ~/hadoop.master.tar.gz ./hadoop   # 先压缩再复制
$ cd ~
$ scp ./hadoop.master.tar.gz Slave1:/home/hadoop
```

然后在Slave1节点上执行如下命令：

```
$ sudo rm -r /usr/local/hadoop    # 删掉旧的（如果存在）
$ sudo tar -zxf ~/hadoop.master.tar.gz -C /usr/local
$ sudo chown -R hadoop /usr/local/hadoop
```

3.3.4 分布式模式配置

首次启动Hadoop集群时，需要先在Master节点执行名称节点的格式化（只需要执行这一次，后面再启动Hadoop时，不要再次格式化名称节点），命令如下：

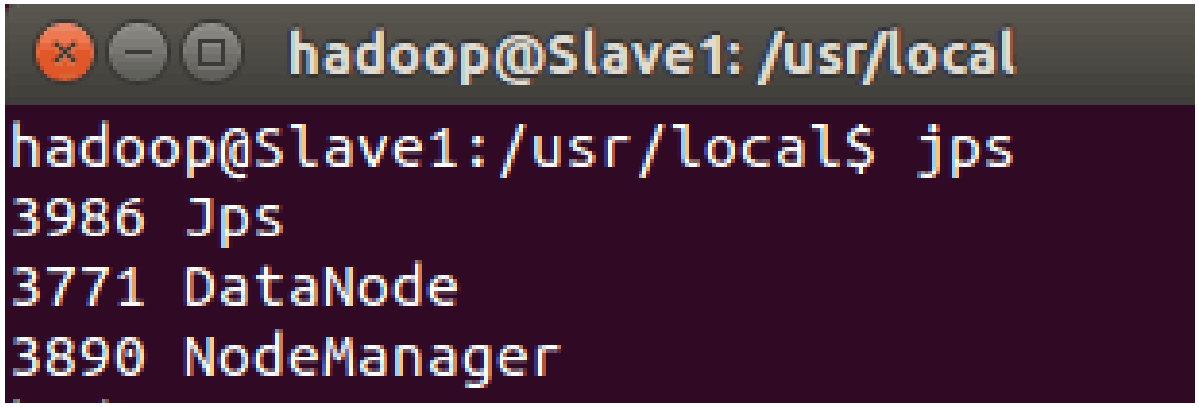
```
$ hdfs namenode -format
```

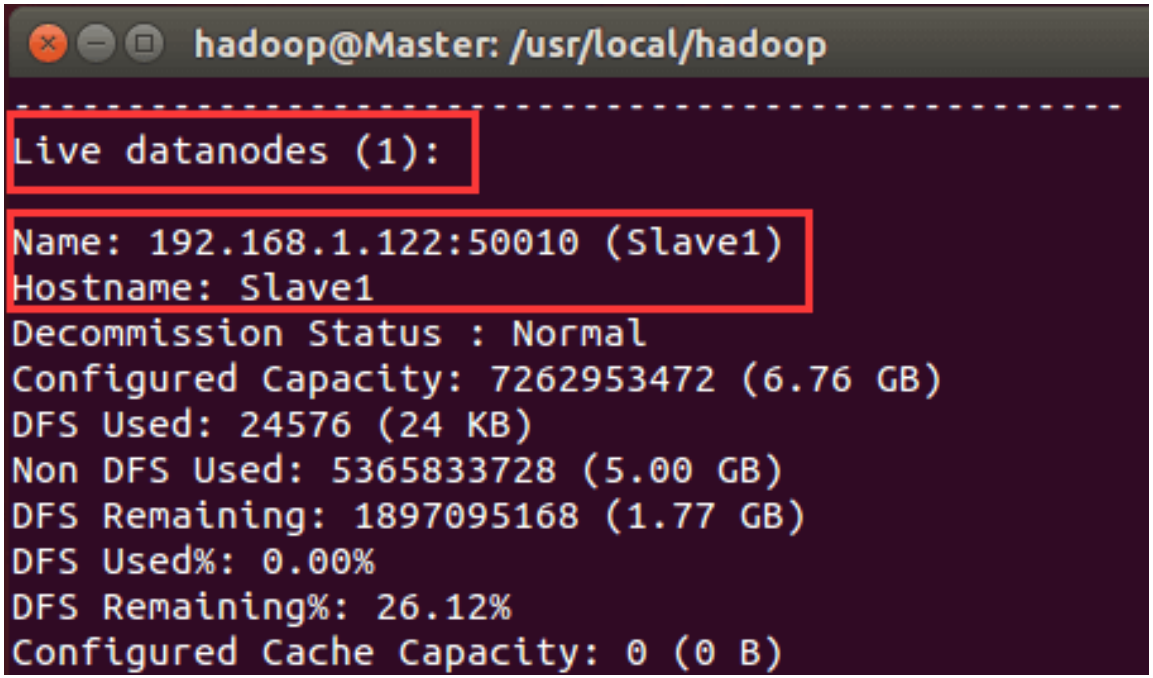
现在就可以启动Hadoop了，启动需要在Master节点上进行，执行如下命令：

```
$ start-dfs.sh
$ start-yarn.sh
$ mr-jobhistory-daemon.sh start historyserver
```

```
hadoop@Master: /usr/local/hadoop
hadoop@Master:/usr/local/hadoop$ jps
4871 JobHistoryServer
3750 SecondaryNameNode
4902 Jps
3899 ResourceManager
3554 NameNode
```

3.3.4 分布式模式配置


hadoop@Slave1: /usr/local
hadoop@Slave1:/usr/local$ jps
3986 Jps
3771 DataNode
3890 NodeManager


hadoop@Master: /usr/local/hadoop
Live datanodes (1):
Name: 192.168.1.122:50010 (Slave1)
Hostname: Slave1
Decommission Status : Normal
Configured Capacity: 7262953472 (6.76 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 5365833728 (5.00 GB)
DFS Remaining: 1897095168 (1.77 GB)
DFS Used%: 0.00%
DFS Remaining%: 26.12%
Configured Cache Capacity: 0 (0 B)

3.3.4 分布式模式配置

5. 执行分布式实例

执行分布式实例过程与伪分布式模式一样，首先创建HDFS上的用户目录，命令如下：

```
$ hdfs dfs -mkdir -p /user/hadoop
```

然后，在HDFS中创建一个input目录，并把“/usr/local/hadoop/etc/hadoop”目录中的配置文件作为输入文件复制到input目录中，命令如下：

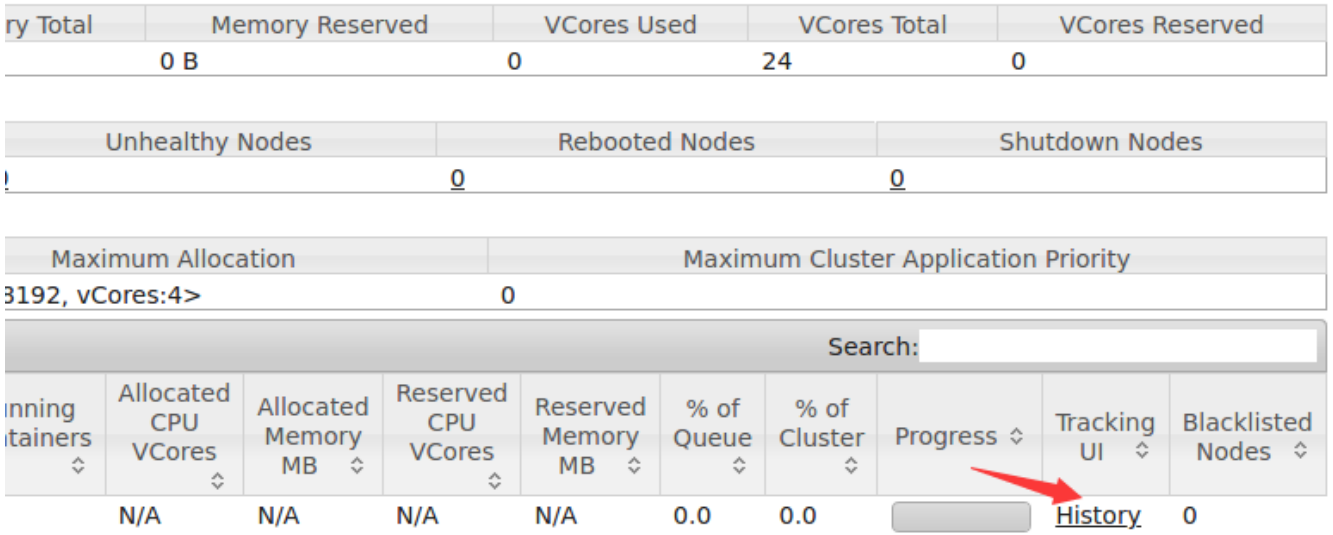
```
$ hdfs dfs -mkdir input
$ hdfs dfs -put /usr/local/hadoop/etc/hadoop/*.xml input
```

接着就可以运行 MapReduce 作业了，命令如下：

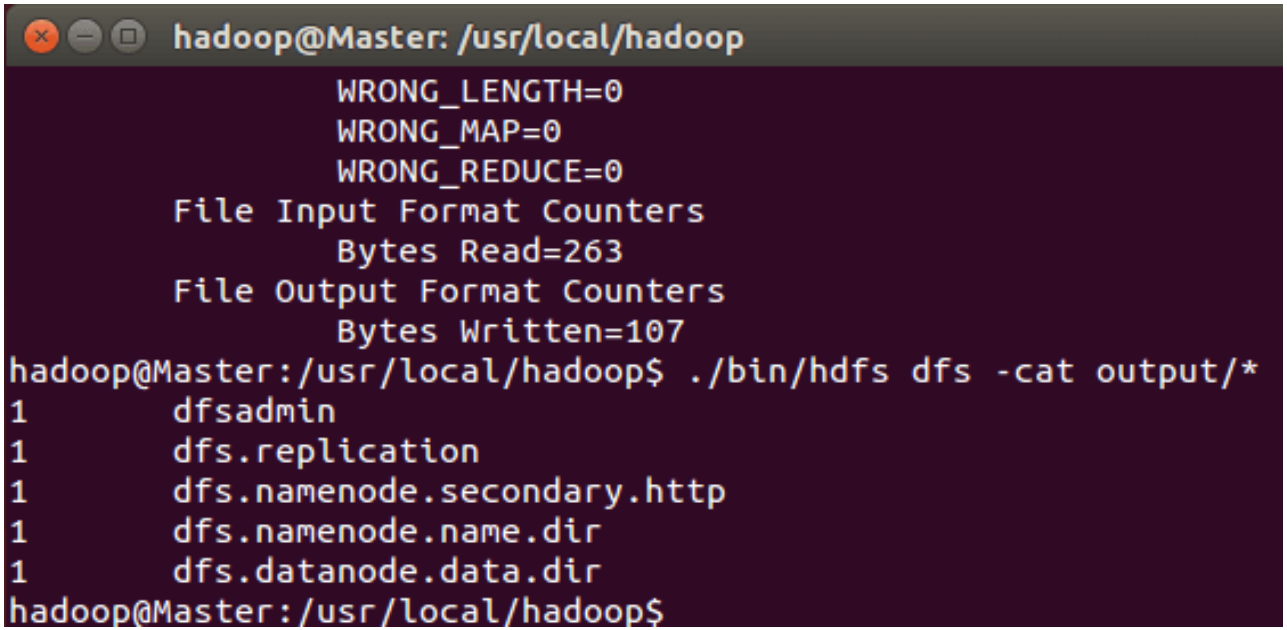
```
$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.3.jar grep input output 'dfs[a-z.]+'
```

3.3.4 分布式模式配置

```
-rw-r--2020-02-13 19:48:42,438 INFO mapreduce.Job: Running job: job_1581651767093_0003
duler.x2020-02-13 19:48:52,728 INFO mapreduce.Job: Job job_1581651767093_0003 running i
-rw-r--n uber mode : false
-rw-r--2020-02-13 19:48:52,728 INFO mapreduce.Job:  map 0% reduce 0%
.xml   2020-02-13 19:49:01,860 INFO mapreduce.Job:  map 0% reduce 100%
-rw-r--2020-02-13 19:49:01,876 INFO mapreduce.Job: Job job_1581651767093_0003 completed
-rw-r-- successfully
```



3.3.4 分布式模式配置



最后，关闭Hadoop集群，需要在Master节点执行如下命令：

```
$ stop-yarn.sh
$ stop-dfs.sh
$ mr-jobhistory-daemon.sh stop historyserver
```

3.4 本章小结

Hadoop是当前流行的分布式计算框架，在企业中得到了广泛的部署和应用。本章重点介绍如何安装Hadoop，从而为后续章节开展HDFS和MapReduce编程实践奠定基础。

Hadoop是基于Java开发的，需要运行在JVM中，因此，需要为Hadoop配置相应的Java环境。Hadoop包含三种安装模式，即单机模式、伪分布式模式和分布式模式。本章分别介绍了三种不同模式的安装配置方法。在初学阶段，建议采用伪分布式模式配置，这样可以快速构建起Hadoop实战环境，有效开展基础编程工作。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过200万次，累计访问量超过1000万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元

教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用》教材

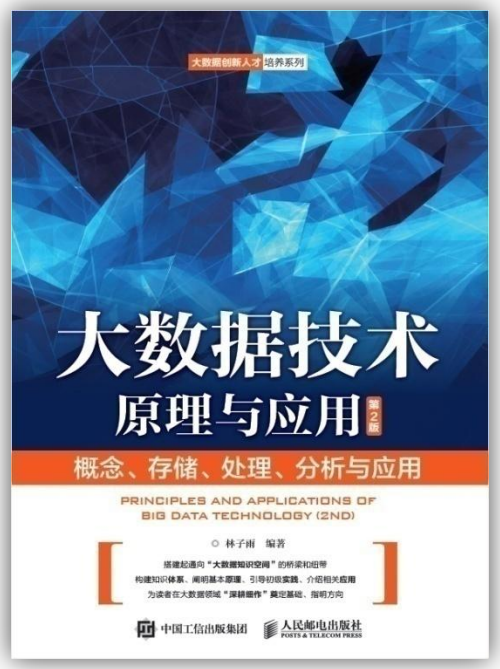
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

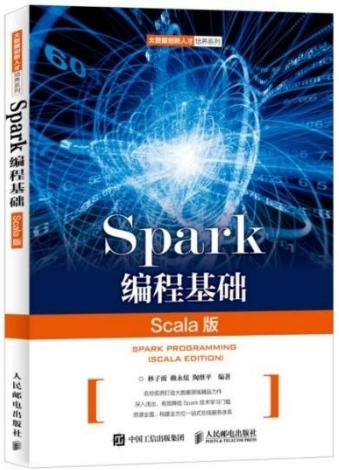
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

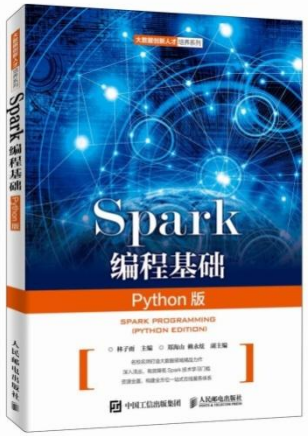
厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3
教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

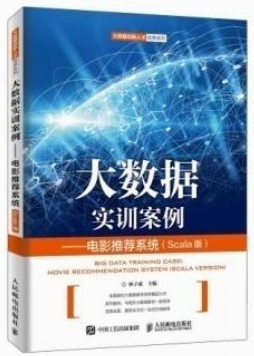
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

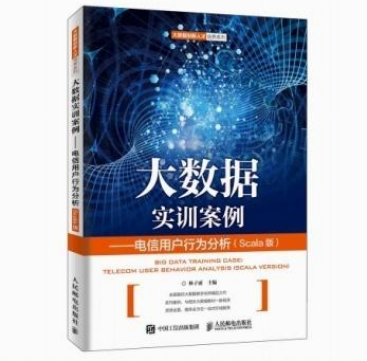
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!