# 《Hadoop 大数据》实训指导书

院系名称：经管信息学院

课程代码：31091020

总学时数：64

适用专业：大数据技术与应用专业

编 制 人：韦 祥、孙 检

编制日期：2020 年 7 月

审 核 人：

审 定 人：

# 《Hadoop 大数据》实训指导书

## 一、实训目的与要求

《Hadoop 大数据》实训主要目的是让学生通过这门实践技能课程的学习了解和掌握 Hadoop 框架在大数据领域的应用，Hadoop 作为处理大数据的分布式存储和计算框架，得到了国内外大小型企业广泛的应用。Hadoop 是一个可以搭建在廉价服务器上的分布式集群系统架构，它具有可用性高、容错性高和可扩展性高等优点。由于它提供了一个开放式的平台，用户可以在完全不了解底层实现细节的情形下，开发适合自身应用的分布式程序。经过十多年的发展，目前 Hadoop 已经成长为一个全栈式的大数据技术生态圈，并在事实上成为应用最广泛最具有代表性的大数据技术。因此，学习 Hadoop 技术是从事大数据行业工作所必不可少的一步。

## 二、实训内容

（一）理论实训

学习并掌握大数据和 Hadoop 相关概念，理解 Hadoop 基础框架。

（二）实例实训

学习官方 Grep 和 WordCount 案例，学生根据项目内容搭建不同 Hadoop 平台，完成对文件的上传、下载等操作。

## 三、参考课时

| 标题 | 实训内容 | 实训课时 |
|---|---|---|
| 实训一 | 大数据概论 | 2 |
| 实训二 | 从 Hadoop 框架讨论大数据生态 | 2 |
| 实训三 | 虚拟机环境准备 | 4 |
| 实训四 | 安装 Hadoop | 4 |
| 实训五 | Hadoop 本地运行模式 | 4 |
| 实训六 | Hadoop 伪分布式运行模式 | 8 |
| 实训七 | Hadoop 完全分布式运行模式 | 8 |
| 实训八 | HDFS 客户端操作 | 12 |

| | | |
|---|---|---|
| 实训九 | 总结 | 4 |
| 总计 | | 48 |

**四、实训材料准备**

（一）软件准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

（二）硬件准备

网络条件：与因特网连接的局域网。

教师用机：Windows 7 及以上版本，Centos6 以上版本。

学生用机：Windows 7 及以上版本，Centos6 以上版本。

**五、综合实训考核办法：**

系统文档　20 分

项目功能　40 分

Bug 调试　10 分

实训出勤　20 分

效果展示　10 分

# 目　　录

## 实训一 大数据概论

### 一、实训目的和要求

通过学习大数据概念、大数据应用等了解大数据特点，了解大数据发展前景和工作模式。

### 二、实训内容

大数据概念、大数据特点、大数据应用场景、发展前景等。

### 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

### 四、实训步骤

1. 大数据概念
2. 大数据特点（4V）
3. 大数据应用场景
4. 大数据发展前景
5. 大数据部门业务流程分析
6. 大数据部门组织结构

### 五、思考和练习

1. 什么是大数据？
2. 大数据有哪些特点。

## 实训二 从 Hadoop 框架讨论大数据生态

### 一、实训目的和要求

通过学习 Hadoop 框架组成了解大数据技术生态体系。

### 二、实训内容

1. 了解 Hadoop 概念及发展历史；
2. 掌握 Hadoop 三大发行版本；
3. 掌握 Hadoop 组成；
4. 理解并掌握大数据技术生态体系。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

Hadoop 三大发行版本：Apache、Cloudera、Hortonworks。

Apache 版本最原始（最基础）的版本，对于入门学习最好。

Cloudera 在大型互联网企业中用的较多。

Hortonworks 文档较好。

Apache Hadoop

官网地址：http://hadoop.apache.org/releases.html

下载地址：https://archive.apache.org/dist/hadoop/common/

Cloudera Hadoop

官网地址：https://www.cloudera.com/downloads/cdh/5-10-0.html

下载地址：http://archive-primary.cloudera.com/cdh5/cdh/5/

（1）2008 年成立的 Cloudera 是最早将 Hadoop 商用的公司，为合作伙伴提供 Hadoop 的商用解决方案，主要是包括支持、咨询服务、培训。

（2）2009 年 Hadoop 的创始人 Doug Cutting 也加盟 Cloudera 公司。Cloudera 产品主要为 CDH，Cloudera Manager，Cloudera Support

（3）CDH 是 Cloudera 的 Hadoop 发行版，完全开源，比 Apache Hadoop 在兼容性，安全性，稳定性上有所增强。

（4）Cloudera Manager 是集群的软件分发及管理监控平台，可以在几个小时内部署好一个 Hadoop 集群，并对集群的节点及服务进行实时监控。Cloudera Support 即是对 Hadoop 的技术支持。

（5）Cloudera 的标价为每年每个节点 4000 美元。Cloudera 开发并贡献了可实时处理大数据的 Impala 项目。

3. Hortonworks Hadoop

官网地址：https://hortonworks.com/products/data-center/hdp/

下载地址：https://hortonworks.com/downloads/#data-platform

（1）2011 年成立的 Hortonworks 是雅虎与硅谷风投公司 Benchmark Capital

合资组建。

（2）公司成立之初就吸纳了大约 25 名至 30 名专门研究 Hadoop 的雅虎工程师，上述工程师均在 2005 年开始协助雅虎开发 Hadoop，贡献了 Hadoop80%的代码。

（3）雅虎工程副总裁、雅虎 Hadoop 开发团队负责人 Eric Baldeschwieler 出任 Hortonworks 的首席执行官。

（4）Hortonworks 的主打产品是 Hortonworks Data Platform（HDP），也同样是 100%开源的产品，HDP 除常见的项目外还包括了 Ambari，一款开源的安装和管理系统。

（5）HCatalog，一个元数据管理系统，HCatalog 现已集成到 Facebook 开源的 Hive 中。Hortonworks 的 Stinger 开创性的极大的优化了 Hive 项目。Hortonworks 为入门提供了一个非常好的，易于使用的沙盒。

（6）Hortonworks 开发了很多增强特性并提交至核心主干，这使得 Apache Hadoop 能够在包括 Window Server 和 Windows Azure 在内的 Microsoft Windows 平台上本地运行。定价以集群为基础，每 10 个节点每年为 12500 美元。

**五、思考和练习**

1. Hadoop 与传统数据处理软件有什么区别？
2. Hadoop 生态系统中的常用组件有哪些？主要功能有哪些？
3. Hadoop 的商用公司都有哪些？业务模式如何？

## 实训三 虚拟机环境准备

**一、实训目的和要求**

在 Hadoop 运行环境搭建之前必须做虚拟机环境准备、JDK 安装等。

**二、实训内容**

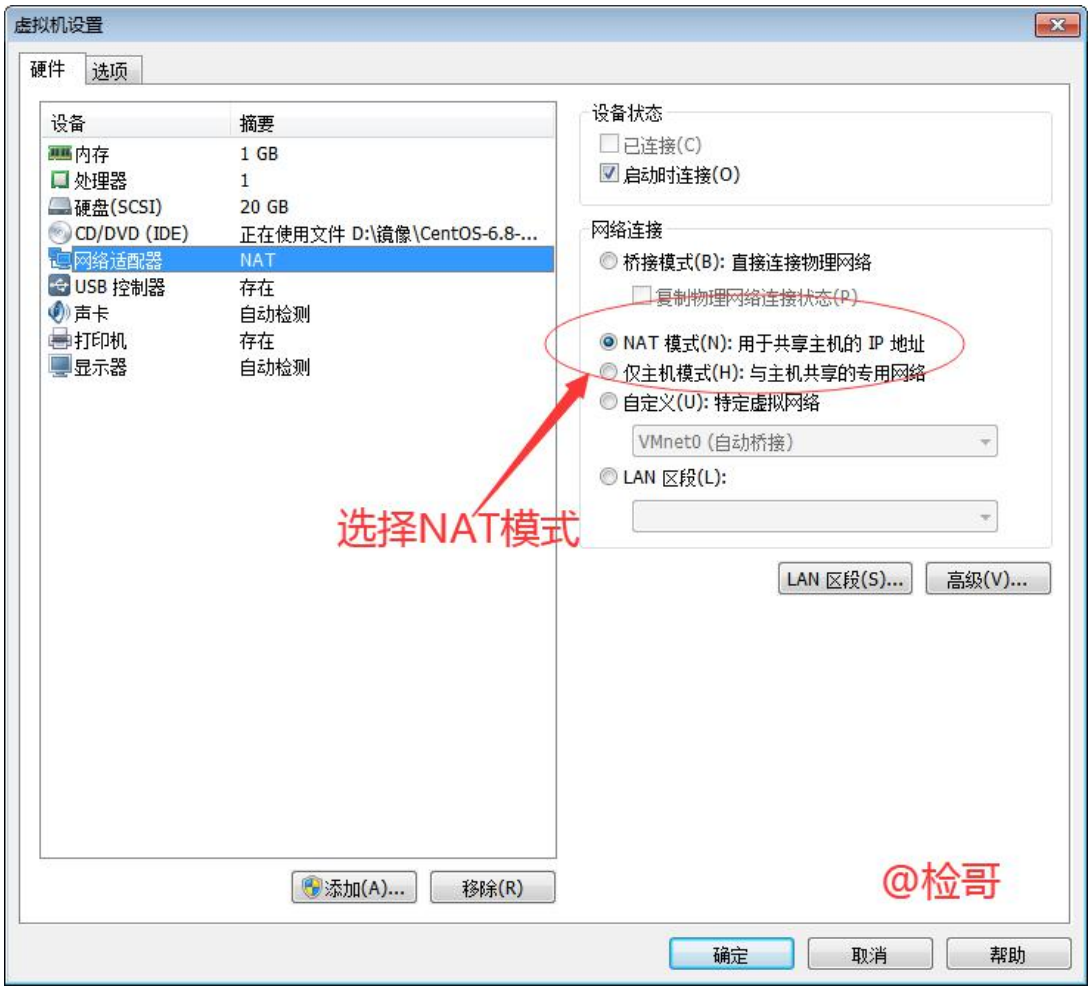
虚拟机环境准备、克隆虚拟机，修改静态 IP、主机名，准备 3 台虚拟机。

**三、实训准备**

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

前期准备：虚拟机网络模式设置为 NAT




1、克隆虚拟机

详情请参考检哥 Linux 第二次课

2、修改克隆虚拟机的静态 IP

2.1 使用命令：vim /etc/udev/rules.d/70-persistent-net.rules

进入如下页面，删除 eth0 该行；将 eth1 修改为 eth0，同时复制物理 ip 地址

2.2 修改 IP 地址

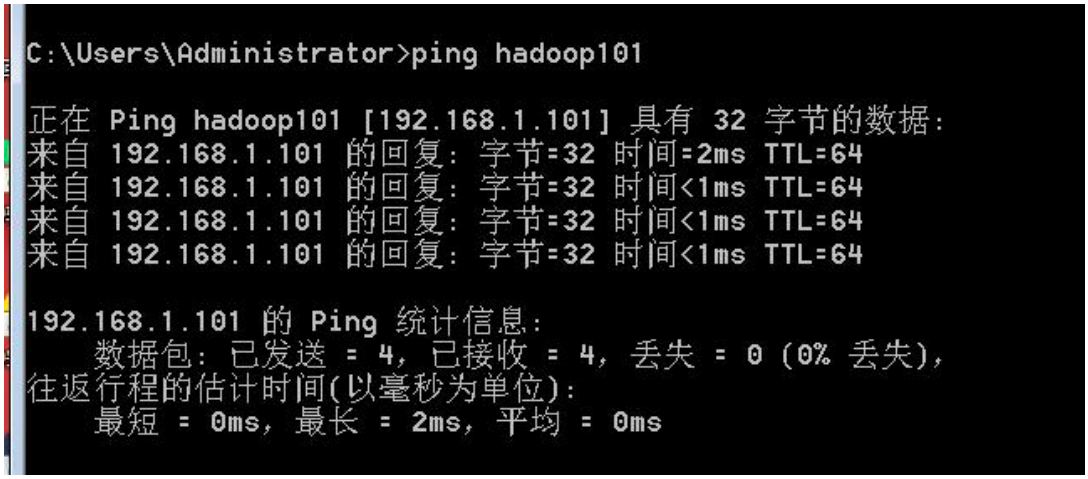
使用命令：vim /etc/sysconfig/network-scripts/ifcfg-eth0

2.3 执行命令：service network restart 重启网络服务

```
[root@localhost 桌面]# serv
servertool   service      serviceconf
[root@localhost 桌面]# service network restart
关闭环回接口：                                             [确定]
弹出环回接口：                                             [确定]
弹出界面 eth0：  错误：激活连接失败：The connection is not for this device.
                                                           [失败]
[root@localhost 桌面]# service network restart
```

2.4 重启：reboot

3、修改主机名
3.1 查看主机名命令：hostname
3.2 修改主机名：vi /etc/sysconfig/network
3.3 配置主机名与 IP 同步
使用命令：vim /etc/hosts
3.4、重启 reboot
3.5、验证之前的操作是否成功
3.6 修改 window7 的 hosts 文件
（1）进入 C:\Windows\System32\drivers\etc 路径
（2）打开 hosts 文件并添加如下内容>
192.168.1.100 hadoop100
192.168.1.101 hadoop101
192.168.1.102 hadoop102
192.168.1.103 hadoop103
192.168.1.104 hadoop104
192.168.1.105 hadoop105
192.168.1.106 hadoop106
192.168.1.107 hadoop107
192.168.1.108 hadoop108
192.168.1.109 hadoop109
192.168.1.110 hadoop110
3.7 验证：打开 cmd 输入 ping hadoop101




4、关闭防火墙
1）查看防火墙开机启动状态
[root@hadoop101 ~]# chkconfig iptables --list
2）关闭防火墙
[root@hadoop101 ~]# chkconfig iptables off

5、创建用户
6、配置用户管理权限（暂时省略）

## 五、思考和练习

1. 为什么要修改静态 IP、主机名？
2. 克隆虚拟机能否保留用户和组？

# 实训四 安装 Hadoop

## 一、实训目的和要求

Hadoop 运行环境的搭建是 Hadoop 开发重点，要求学生必须掌握。

## 二、实训内容

Hadoop 运行环境搭建。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

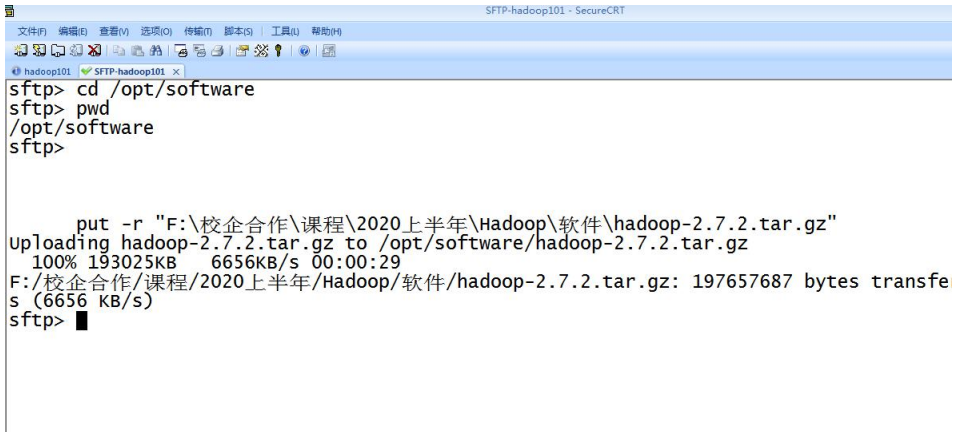
1 安装 Hadoop

1.1Hadoop 下载地址：

https://archive.apache.org/dist/hadoop/common/hadoop-2.7.2/

1.2 用 SecureCRT 工具将 hadoop-2.7.2.tar.gz 导入到 opt 目录下面的 software 文件夹下面

切换到 sftp 连接页面，选择 Linux 下编译的 hadoop jar 包拖入




1.3 进入到 Hadoop 安装包路径下

```
[sun@hadoop101 ~]$ cd /opt/software/
```

1.4 解压安装文件到/opt/module 下面

```
[sun@hadoop101 software]$ tar -zxvf hadoop-2.7.2.tar.gz -C /opt/module/
```

1.5 查看是否解压成功

```
[sun@hadoop101 software]$ ls /opt/module/
hadoop-2.7.2
```

1.6 将 Hadoop 添加到环境变量

（1）获取 Hadoop 安装路径

```
[sun@hadoop101 hadoop-2.7.2]$ pwd
/opt/module/hadoop-2.7.2
```

（2）打开/etc/profile 文件

```
[sun@hadoop101 hadoop-2.7.2]$ sudo vi /etc/profile
```

在 profile 文件末尾添加 JDK 路径：（shitf+g）

```
##HADOOP_HOME
export HADOOP_HOME=/opt/module/hadoop-2.7.2
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
```

（3）保存后退出

```
:wq
```

（4）让修改后的文件生效

```
[sun@ hadoop101 hadoop-2.7.2]$ source /etc/profile
```

1.7 测试是否安装成功

```
[sun@hadoop101 hadoop-2.7.2]$ hadoop version
Hadoop 2.7.2
```

1.8 重启(如果 Hadoop 命令不能用再重启)

```
[sun@ hadoop101 hadoop-2.7.2]$ sync
[sun@ hadoop101 hadoop-2.7.2]$ sudo reboot
```

2 Hadoop 目录结构

1、查看 Hadoop 目录结构

```
[sun@hadoop101 hadoop-2.7.2]$ ll
总用量 52
drwxr-xr-x. 2 sun sun  4096 5月  22 2017 bin
drwxr-xr-x. 3 sun sun  4096 5月  22 2017 etc
drwxr-xr-x. 2 sun sun  4096 5月  22 2017 include
drwxr-xr-x. 3 sun sun  4096 5月  22 2017 lib
drwxr-xr-x. 2 sun sun  4096 5月  22 2017 libexec
-rw-r--r--. 1 sun sun 15429 5月  22 2017 LICENSE.txt
-rw-r--r--. 1 sun sun   101 5月  22 2017 NOTICE.txt
-rw-r--r--. 1 sun sun  1366 5月  22 2017 README.txt
```

```
drwxr-xr-x. 2 sun sun  4096 5月  22 2017 sbin
drwxr-xr-x. 4 sun sun  4096 5月  22 2017 share
```

2、重要目录

（1）bin 目录：存放对 Hadoop 相关服务（HDFS,YARN）进行操作的脚本

（2）etc 目录：Hadoop 的配置文件目录，存放 Hadoop 的配置文件

（3）lib 目录：存放 Hadoop 的本地库（对数据进行压缩解压缩功能）

（4）sbin 目录：存放启动或停止 Hadoop 相关服务的脚本

（5）share 目录：存放 Hadoop 的依赖 jar 包、文档、和官方案例

3 Hadoop 官方文档

官网：https://hadoop.apache.org/

## 五、思考和练习

1. 在安装 Hadoop 之前需要准备哪些组件？
2. 验证 Hadoop 安装是否成功，并运行测试程序。

# 实训五 Hadoop 本地运行模式

## 一、实训目的和要求

学习 Hadoop 本地运行模式是为搭建 Hadoop 分布式集群做准备，本地模式即单机模式，无需节点，一台虚拟机即可。

## 二、实训内容

搭建 Hadoop 本地运行模式，执行官方 Grep、WordCount 案例。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

1 Hadoop 运行模式

Hadoop 运行模式包括：本地模式、伪分布式模式以及完全分布式模式。

Hadoop 官方网站：http://hadoop.apache.org/

本地模式（默认模式）：不需要启用单独进程，直接可以运行，测试和开发时使用。

伪分布式模式：等同于完全分布式，只有一个节点。

完全分布式模式：多个节点一起运行。

1.1 本地运行模式

1.1.1 官方 Grep 案例

1. 创建在 hadoop-2.7.2 文件下面创建一个 input 文件夹

[sun@hadoop101 hadoop-2.7.2]$ mkdir input

2. 将 Hadoop 的 xml 配置文件复制到 input

[sun@hadoop101 hadoop-2.7.2]$ cp etc/hadoop/*.xml input

3. 执行 share 目录下的 MapReduce 程序

[sun@hadoop101 hadoop-2.7.2]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar grep input output 'dfs[a-z.]+'

4. 查看输出结果

[sun@hadoop101 hadoop-2.7.2]$ cat output/*

1.1.2 官方 WordCount 案例

1. 创建在 hadoop-2.7.2 文件下面创建一个 wcinput 文件夹

[sun@hadoop101 hadoop-2.7.2]$ mkdir wcinput

2. 在 wcinput 文件下创建一个 wc.input 文件

[sun@hadoop101 hadoop-2.7.2]$ cd wcinput

[sun@hadoop101 wcinput]$ touch wc.input

3. 编辑 wc.input 文件

[sun@hadoop101 wcinput]$ vi wc.input

在文件中输入如下内容

hadoop yarn

hadoop mapreduce

sun

sun

保存退出:: wq

4. 回到 Hadoop 目录/opt/module/hadoop-2.7.2

5. 执行程序

[sun@hadoop101 hadoop-2.7.2]$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount wcinput wcoutput

6. 查看结果

[sun@hadoop101 hadoop-2.7.2]$ cat wcoutput/part-r-00000

sun 2

hadoop 2

mapreduce 1

yarn 1

## 五、思考和练习

1. Hadoop 本地运行模式应用场景。

2. Hadoop 本地运行模式注意事项。

# 实训六 Hadoop 伪分布式运行模式

## 一、实训目的和要求

Hadoop 伪分布式运行模式等同于完全分布式，只有一个节点。

## 二、实训内容

搭建 Hadoop 伪分布式运行模式，准备 3 台虚拟机。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

1.1 伪分布式运行模式

1.1.1 启动 HDFS 并运行 MapReduce 程序

1. 分析

（1）配置集群

（2）启动、测试集群增、删、查

（3）执行 WordCount 案例

2. 执行步骤

（1）配置集群

（a）配置：hadoop-env.sh（在 etc/hadoop 目录下）

Linux 系统中获取 JDK 的安装路径：

```
[sun@ hadoop101 ~]# echo $JAVA_HOME
/opt/module/jdk1.8.0_144
```

修改 JAVA_HOME 路径：

```
export JAVA_HOME=/opt/module/jdk1.8.0_144
```

（b）配置：core-site.xml

```
<!-- 指定 HDFS 中 NameNode 的地址 -->
<property>
<name>fs.defaultFS</name>
    <value>hdfs://hadoop101:9000</value>
</property>

<!-- 指定 Hadoop 运行时产生文件的存储目录 -->
<property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/module/hadoop-2.7.2/data/tmp</value>
</property>
```

（c）配置：hdfs-site.xml

```
<!-- 指定 HDFS 副本的数量 -->
<property>
    <name>dfs.replication</name>
    <value>1</value>
```

</property>

（2）启动集群

（a）格式化 NameNode（第一次启动时格式化，以后就不要总格式化）

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs namenode -format

（b）启动 NameNode

[sun@hadoop101 hadoop-2.7.2]$ sbin/hadoop-daemon.sh start namenode

（c）启动 DataNode

[sun@hadoop101 hadoop-2.7.2]$ sbin/hadoop-daemon.sh start datanode

（3）查看集群

（a）查看是否启动成功

[sun@hadoop101 hadoop-2.7.2]$ jps

13586 NameNode

13668 DataNode

13786 Jps

注意：jps 是 JDK 中的命令，不是 Linux 命令。不安装 JDK 不能使用 jps

（b）web 端查看 HDFS 文件系统

http://hadoop101:50070/dfshealth.html#tab-overview

注意：如果不能查看，看如下帖子处理

http://www.cnblogs.com/zlslch/p/6604189.html

（4）操作集群

（a）在 HDFS 文件系统上创建一个 input 文件夹

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -mkdir -p /user/sun/input

（b）将测试文件内容上传到文件系统上

[sun@hadoop101 hadoop-2.7.2]$bin/hdfs dfs -put wcinput/wc.input /user/sun/input/

（c）查看上传的文件是否正确

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -ls /user/sun/input/

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -cat /user/sun/input/wc.input

（d）运行 MapReduce 程序

[sun@hadoop101 hadoop-2.7.2]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount /user/sun/input/ /user/sun/output

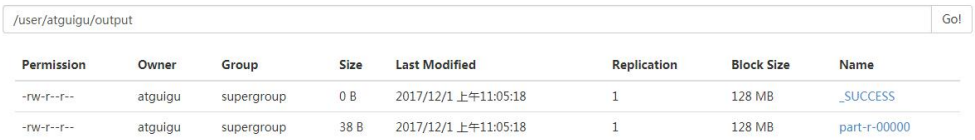
（e）查看输出结果

命令行查看：

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -cat /user/sun/output/*

浏览器查看，如下图所示

Browse Directory




（f）将测试文件内容下载到本地

[sun@hadoop101 hadoop-2.7.2]$ hdfs dfs -get

/user/sun/output/part-r-00000 ./wcoutput/

（g）删除输出结果

[sun@hadoop101 hadoop-2.7.2]$ hdfs dfs -rm -r /user/sun/output

### 1.1.2 启动 YARN 并运行 MapReduce 程序

1. 分析

（1）配置集群在 YARN 上运行 MR

（2）启动、测试集群增、删、查

（3）在 YARN 上执行 WordCount 案例

2. 执行步骤

（1）配置集群

（a）配置 yarn-env.sh

（b）配置 yarn-site.xml

（c）配置：mapred-env.sh

（d）配置：mapred-site.xml

（2）启动集群

（a）启动前必须保证 NameNode 和 DataNode 已经启动

（b）启动 ResourceManager

[sun@hadoop101 hadoop-2.7.2]$ sbin/yarn-daemon.sh start resourcemanager

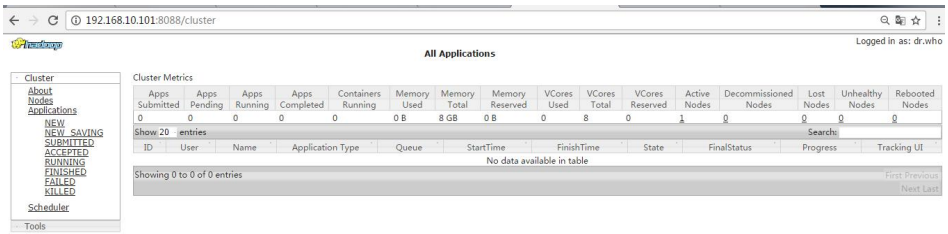
（c）启动 NodeManager

[sun@hadoop101 hadoop-2.7.2]$ sbin/yarn-daemon.sh start nodemanager

（3）集群操作

（a）YARN 的浏览器页面查看，如图 2-35 所示

http://hadoop101:8088/cluster




（b）删除文件系统上的 output 文件

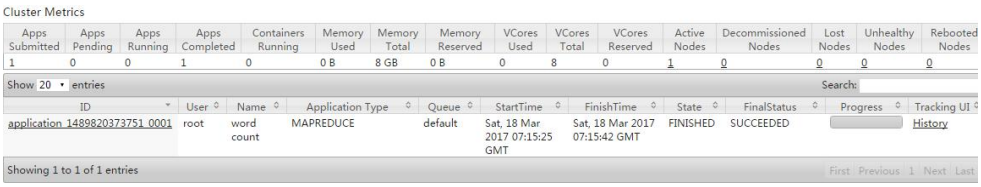
[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -rm -r /user/sun/output

（c）执行 MapReduce 程序

[sun@hadoop101 hadoop-2.7.2]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount /user/sun/input /user/sun/output

（d）查看运行结果，如图 2-36 所示

[sun@hadoop101 hadoop-2.7.2]$ bin/hdfs dfs -cat /user/sun/output/*

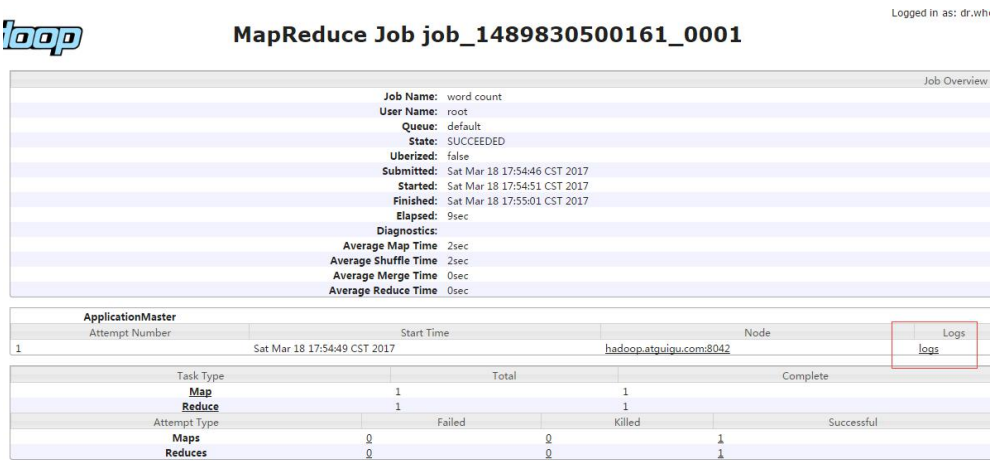



### 1.1.3 配置历史服务器

为了查看程序的历史运行情况，需要配置一下历史服务器。具体配置步骤如下：

1. 配置 mapred-site.xml
[sun@hadoop101 hadoop]$ vi mapred-site.xml
在该文件里面增加如下配置。
```
<!-- 历史服务器端地址 -->
<property>
<name>mapreduce.jobhistory.address</name>
<value>hadoop101:10020</value>
</property>
<!-- 历史服务器 web 端地址 -->
<property>
    <name>mapreduce.jobhistory.webapp.address</name>
    <value>hadoop101:19888</value>
</property>
```
2. 启动历史服务器
[sun@hadoop101 hadoop-2.7.2]$ sbin/mr-jobhistory-daemon.sh start historyserver
3. 查看历史服务器是否启动
[sun@hadoop101 hadoop-2.7.2]$ jps
4. 查看 JobHistory
http://hadoop101:19888/jobhistory

1.1.4 配置日志的聚集
日志聚集概念：应用运行完成以后，将程序运行日志信息上传到 HDFS 系统上。
日志聚集功能好处：可以方便的查看到程序运行详情，方便开发调试。
查看日志，如图 2-37，2-38，2-39 所示
http://hadoop101:19888/jobhistory




1.1.5 配置文件说明
Hadoop 配置文件分两类：默认配置文件和自定义配置文件，只有用户想修改某一默认配置值时，才需要修改自定义配置文件，更改相应属性值。
（1）默认配置文件：
表 2-1

| 要获取的默认文件 | 文件存放在 Hadoop 的 jar 包中的位置 |
| --- | --- |

| | |
|---|---|
| [core-default.xml] | hadoop-common-2.7.2.jar/ core-default.xml |
| [hdfs-default.xml] | hadoop-hdfs-2.7.2.jar/ hdfs-default.xml |
| [yarn-default.xml] | hadoop-yarn-common-2.7.2.jar/ yarn-default.xml |
| [mapred-default.xml] | hadoop-mapreduce-client-core-2.7.2.jar/ mapred-default.xml |

（2）自定义配置文件：

core-site.xml、hdfs-site.xml、yarn-site.xml、mapred-site.xml 四个配置文件存放在$HADOOP_HOME/etc/hadoop 这个路径上，用户可以根据项目需求重新进行修改配置。

1.1.6 关闭安全模式

[sun@hadoop101 hadoop-2.7.2]$ bin/hadoop dfsadmin -safemode leave

## 五、思考和练习

1. 熟悉分布式 Hadoop 的安装流程，搭建 Hadoop 环境。

# 实训七 完全分布式运行模式

## 一、实训目的和要求

完全分布式模式：多个节点一起运行。

## 二、实训内容

搭建 Hadoop 伪分布式运行模式，准备至少 3 台虚拟机，3 个节点。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

1. 集群部署规划

表 2-3

| | hadoop102 | hadoop103 | hadoop104 |
|---|---|---|---|
| HDFS | NameNode<br>DataNode | DataNode | SecondaryNameNode<br>DataNode |
| YARN | | ResourceManage | |

| | | | |
|---|---|---|---|
| | NodeManager | r<br><br>NodeManager | NodeManager |

2. 配置集群

（1）核心配置文件

配置 core-site.xml

```
[sun@hadoop102 hadoop]$ vi core-site.xml
```

在该文件中编写如下配置

```
<!-- 指定 HDFS 中 NameNode 的地址 -->
<property>
      <name>fs.defaultFS</name>
       <value>hdfs://hadoop102:9000</value>
</property>

<!-- 指定 Hadoop 运行时产生文件的存储目录 -->
<property>
      <name>hadoop.tmp.dir</name>
      <value>/opt/module/hadoop-2.7.2/data/tmp</value>
</property>
```

（2）HDFS 配置文件

配置 hadoop-env.sh--JDK 路径

```
[sun@hadoop102 hadoop]$ vi hadoop-env.sh
export JAVA_HOME=/opt/module/jdk1.8.0_144
```

配置 hdfs-site.xml

```
[sun@hadoop102 hadoop]$ vi hdfs-site.xml
```

在该文件中编写如下配置

```
<property>
      <name>dfs.replication</name>
      <value>3</value>
</property>

<!-- 指定 Hadoop 辅助名称节点主机配置 -->
<property>
      <name>dfs.namenode.secondary.http-address</name>
      <value>hadoop104:50090</value>
</property>
```

（3）YARN 配置文件

配置 yarn-env.sh

```
[sun@hadoop102 hadoop]$ vi yarn-env.sh
```

```
export JAVA_HOME=/opt/module/jdk1.8.0_144
```

配置 yarn-site.xml

```
[sun@hadoop102 hadoop]$ vi yarn-site.xml
```

在该文件中增加如下配置

```
<!-- Reducer 获取数据的方式 -->
<property>
     <name>yarn.nodemanager.aux-services</name>
     <value>mapreduce_shuffle</value>
</property>

<!-- 指定 YARN 的 ResourceManager 的地址 -->
<property>
     <name>yarn.resourcemanager.hostname</name>
     <value>hadoop103</value>
</property>
```

（4）MapReduce 配置文件

配置 mapred-env.sh

```
[sun@hadoop102 hadoop]$ vi mapred-env.sh
export JAVA_HOME=/opt/module/jdk1.8.0_144
```

配置 mapred-site.xml

```
[sun@hadoop102 hadoop]$ cp mapred-site.xml.template mapred-site.xml

[sun@hadoop102 hadoop]$ vi mapred-site.xml
```

在该文件中增加如下配置

```
<!-- 指定 MR 运行在 Yarn 上 -->
<property>
     <name>mapreduce.framework.name</name>
     <value>yarn</value>
</property>
```

3. 在集群上分发配置好的 Hadoop 配置文件

```
[sun@hadoop102 hadoop]$ xsync /opt/module/hadoop-2.7.2/
```

4. 查看文件分发情况

```
[sun@hadoop103 hadoop]$ cat /opt/module/hadoop-2.7.2/etc/hadoop/core-site.xml
```

## 五、思考和练习

1. Hadoop 三种运行模式。
2. Hadoop 三种运行模式的区别？

# 实训八 HDFS 客户端操作

## 一、实训目的和要求

通过学习 HDFS 文件系统常规操作，使用 Maven 项目操作 HDFS 文件系统。

## 二、实训内容

HDFS 客户端操作。

## 三、实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、实训步骤

1.1 HDFS 客户端环境准备

1.根据自己电脑的操作系统拷贝对应的编译后的 hadoop jar 包到非中文路径(例如：D:\softwareAn\HadoopAn\hadoop-2.7.2)，如图 3-4 所示。

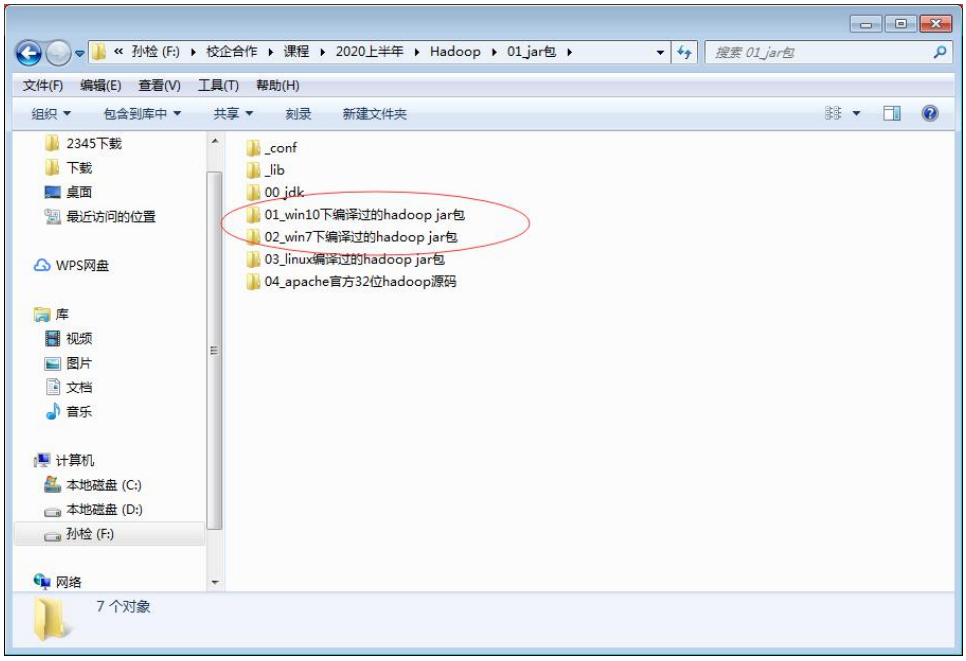



图 3-4　编译后的 hadoop　jar 包

2．配置 HADOOP_HOME 环境变量，如图 3-5 所示。

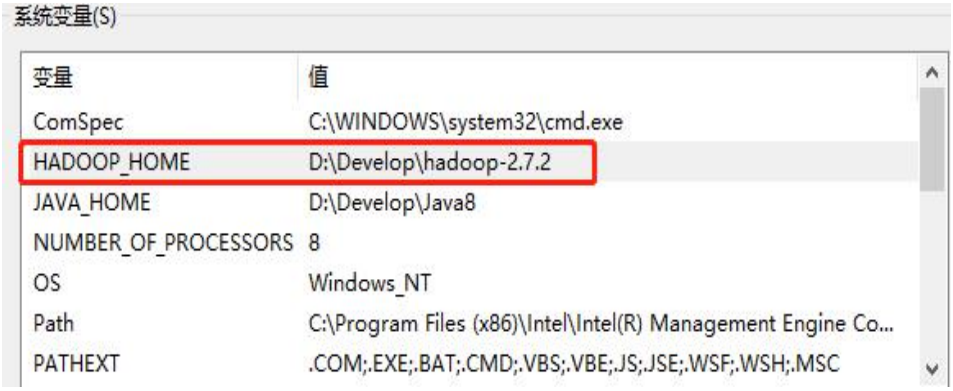


图 3-5　配置 HADOOP_HOME 环境变量

3. 配置 Path 环境变量，如图 3-6 所示。

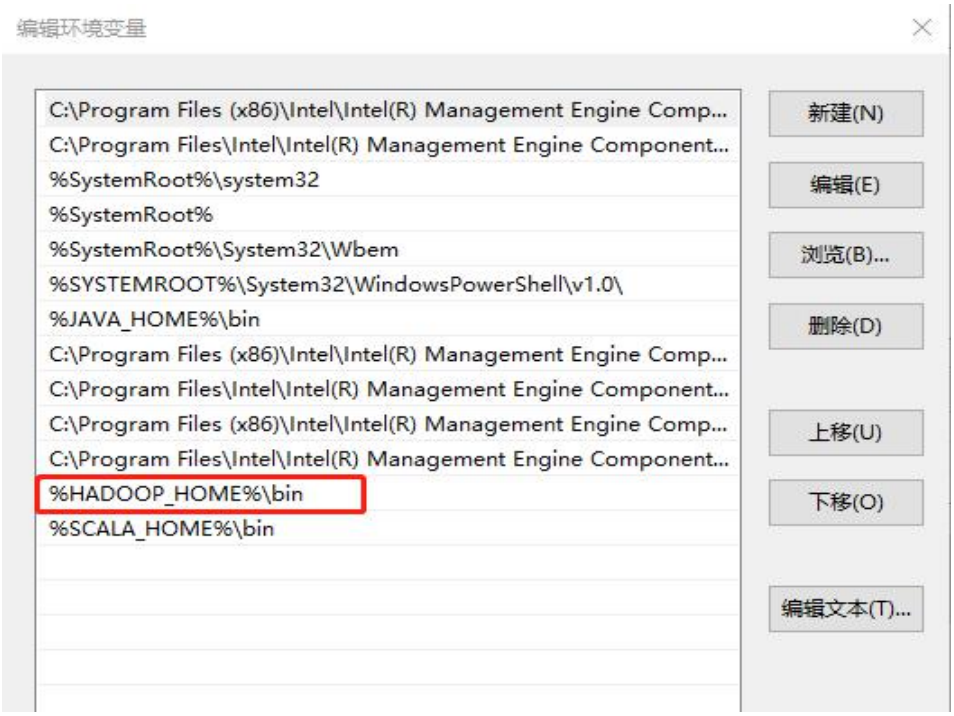



图 3-6　配置 Path 环境变量

4. 创建一个 Maven 工程 HdfsClientDemo

5. 导入相应的依赖坐标+日志添加

6. 创建包名：com.sun.hdfs

7. 创建 HdfsClient 类

```
public class HdfsClient{
@Test
public void testMkdirs() throws IOException, InterruptedException,
URISyntaxException{

        // 1 获取文件系统
        Configuration configuration = new Configuration();
        // 配置在集群上运行
        // configuration.set("fs.defaultFS", "hdfs://hadoop102:9000");
        // FileSystem fs = FileSystem.get(configuration);

        FileSystem fs = FileSystem.get(new URI("hdfs://hadoop102:9000"),
configuration, "sun");

        // 2 创建目录
        fs.mkdirs(new Path("/1108/daxian/banzhang"));

        // 3 关闭资源
        fs.close();
    }
```

}

8．执行程序

运行时需要配置用户名称，如图 3-7 所示

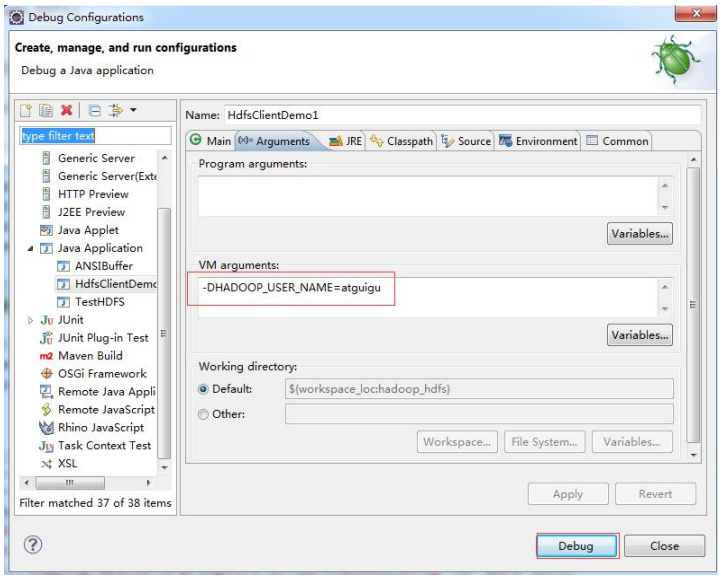



图 3-7　配置用户名称

客户端去操作 HDFS 时，是有一个用户身份的。默认情况下，HDFS 客户端 API 会从 JVM 中获取一个参数来作为自己的用户身份：-DHADOOP_USER_NAME=sun，sun 为用户名称。

## 五、思考和练习

1. Hadoop 流处理的原理是什么？
2. HDFS 文件系统组成。

# 总结

## 一、 实训目的和要求

将学生制作的作品进行综合的考核，并进行总结。

## 二、 实训内容

1. 对学生作品进行考核。
2. 选择典型的（优秀的和劣质的）作品分别进行总结。

## 三、 实训准备

Linux 操作系统、Centos6.5 以上版本、VMware 虚拟机软件、SecureCRT 连接工具。

## 四、 实训步骤

1. 对学生的作品依次进行综合考核。

2. 抽取典型（优秀和劣质）的作品进行全面的解析。

## 五、 实训方法

在机房利用 Linux 虚拟机完成。

## 六、 考核办法

此部分实训以考核为主，对学生组品进行整体的考核。采用全体考察的方法，以百分制为满分，具体评分标准如下：

系统文档　20 分

项目功能　40 分

Bug 调试　10 分

实训出勤　20 分

效果展示　10 分