**实验一 基本操作实验**

**1.1 实验目的**

1． 熟悉大数据实验一体机并了解如何搭建集群；

2． 熟悉Linux基本命令；

3． 掌握vi编辑器的使用；

4． 了解SSH免密登录的原理以及为何需要配置SSH免密登录；

5． 掌握如何配置SSH免密登录；

6． 掌握Java基本命令；

7． 熟悉集成开发软件Eclipse的安装和使用。

**1.2 实验要求**

本次实验完成后，要求学生能够：

1． 使用大数据实验一体机搭建自己的集群；

2． 通过SSH工具登录集群服务器；

3． 实现每台服务器相互之间的免密登录；

4． 通过vi编辑器编写Java程序；

5． 通过Java命令编译和运行编写的Java程序；

6． 通过Jar命令打包编写的Java程序；

7． 安装Eclipse并在其中编写Java程序。

**1.3 实验原理**

**1.3.1 大数据实验一体机**

随着移动互联网、云计算、物联网的快速发展，特别是智能手机端博客、社交网络、位置服务（LBS）等信息发布方式的不断涌现，数据正以较快的速度在不断地增长和累积，大数据时代已经来到。

在海量数据面前，大数据人才无疑是其中关键环节之一，然而，不论国内外，大数据人才却紧缺相当稀缺，比如，当前我国数据人才缺口高达150万，而在未来5-10年，随着市场规模不断增加，这一缺口还将不断增加。

然而，在创新探索大数据教学面前，高校却碰到了一系列困难，如大部分高校大数据课程体系并不完善，在实验环节，由于缺乏实验设备，大数据实训案例匮乏、实验难开展。

针对大数据专业建设的三大难题，云创大数据为各大高校量身定制了大数据软硬件一体化的教学科研平台——大数据实验一体机。大数据实验一体机通过应用容器技术，以少量机器虚拟大量实验集群，可供大量学生同时拥有多套集群进行试验，而每个学生的实验环境不仅相互隔离，方便高效地完成实验，而且实验彼此不干扰，即使某个实验环境被破坏，对其他人也没有影响，一键重启就可以拥有一套新集群，大幅度节省了硬件和人员管理的投入成本。

此外，作为一个可供大量学生完成大数据与云计算实验的集成环境，该平台同步提供了配套的培训服务，对于教学组件的安装、配置，教材、实验手册等具体应用提供一站式服务，有助于高校更好地满足课程设计、课程上机实验、实习实训、科研训练等多方面需求，并在一定程度上缓解大数据师资不足的问题。对于各大高校而言，即使没有任何大数据实验基础，该平台也能助其轻松开展大数据与云计算的教学、实验与科研。

具体而言，大数据实验一体机从以下三个方面解决了高校大数据的教学科研难题：

（1）完整的大数据课程体系及配套资源，一步解决入门难内容充实的课程体系

在《实战Hadoop2.0——从云计算到大数据》和实验手册的指导之下，大数据实验一体机解决方案涵盖大数据算法、接口、工具、平台等多方面内容，从大数据监测与收集、大数据存储与处理、大数据分析与挖掘直至大数据创新，帮助高校构建完善的大数据课程体系。

综合36个大数据实验的实验手册及配套高清视频课程，涵盖原理验证、综合应用、自主设计及创新的多层次实验内容。每个实验呈现详细的实验目的、实验内容、实验原理和实验流程指导。配套相应的实验数据和高清视频课程，参照手册即可轻松完成每个实验。国内专业领域数一数二的网站——中国大数据、中国云计算、中国存储等提供全线支持，一网打尽各类优质资源。

（2）安全可靠的实验环境，大幅度提升大数据技能

基于Docker容器技术，大数据实验一体机可瞬间创建随时运行的实验环境。使用几台机器即可虚拟出大量实验集群，方便上百学生同时使用。采用Kubernetes+ZooKeeper架构管理集群，实验集群完全隔离。实验环境互不干扰，如果实验环境破坏，一键重启即可建立新集群。内置数据挖掘等教学实验数据，可导入高校各学科数据进行教学、科研，校外培训机构同样适用。

（3）热门实战项目贯穿始终，进一步提高教学效果与就业率

大数据实验一体机解决方案采用理论与实践相结合的人才培养模式，帮助教师提高教学水平，促使学生完善大数据知识体系。基于真实的企业基地实训经验，提供丰富的项目实训案例。结合高校各专业实际情况进行行业数据研究，培养实用型人才的专业项目能力。

（4）更多潜在效益，同步增强高校的硬实力和影响力

大数据上升为国家战略，发改委明确组建13个大数据实验室，大数据实验一体机有助于高校大数据实验室建设以及高层次大数据人才的深度培育。大数据实验一体机解决方案将理论与实践双管齐下，帮助提升了高校信息化管理水平和实验项目研究水平。大数据产业迎来发展黄金期，以大数据实验一体机提高大数据专业就业率，可进一步增强高校的硬实力和影响力。

在2016年暑期全国高校大数据培训中，云创大数据利用大数据实验一体机搭建了Docker容器云，为每个学员分配5套虚拟服务器集群，为其提供了简洁易用的上机操作环境，得到了学员的一致好评。在理论讲解的基础上，讲师通过这一实践平台，为学员提供每一步的操作指导，真正做到了学思结合、知行统一，所有学员的大数据应用能力均得以提升，并获得了相应的大数据能力等级证书。

大数据实验一体机基本操作主要包括账号管理、集群管理、集群登录和辅助功能四大部分，其中账号管理完成新建和销毁用户账号，集群管理完成新建和销毁集群，集群登录指的是通过SSH登录到集群各机器，辅助功能模板提供了部分软件下载等实用小功能。

**1.3.1.1 界面管理**

输入本校大数据实验一体机网址后，请输入相应账号与密码，点击登录即可。如图1-1所示：



**图1-1 登录界面**

**1.3.1.2 账号管理**

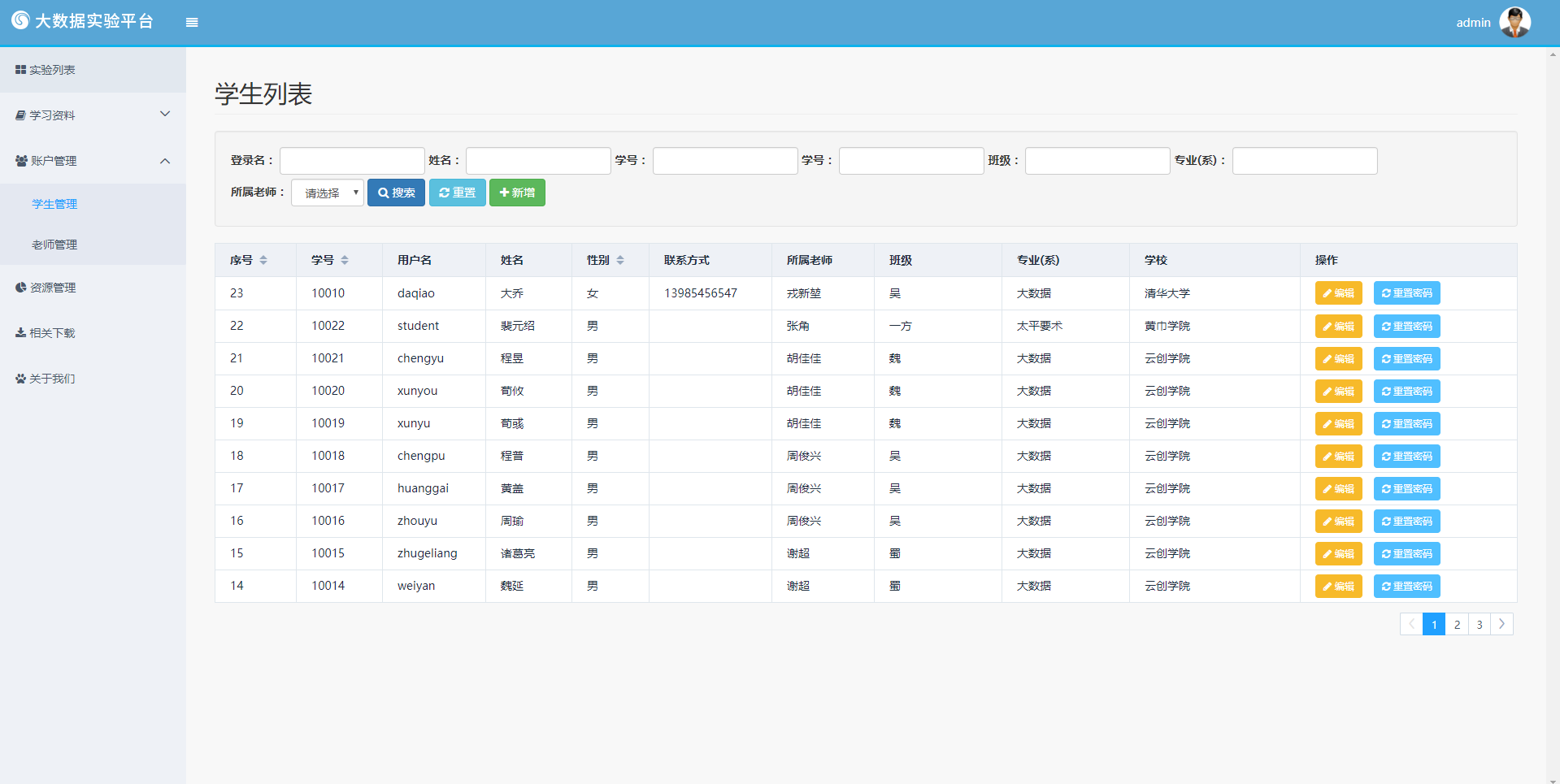
系统管理员和教师角色登录后，可以看到用户账户管理界面。

其中：

系统管理员用户可以在该界面中查看或修改所有的教师和学生用户信息，并可以注册或销毁教师或学生用户账户；

教师用户可以在该界面中查看或修改自己建立的所有学生用户信息，并可以注册或销毁自己的学生用户账户。

如图1-2所示：



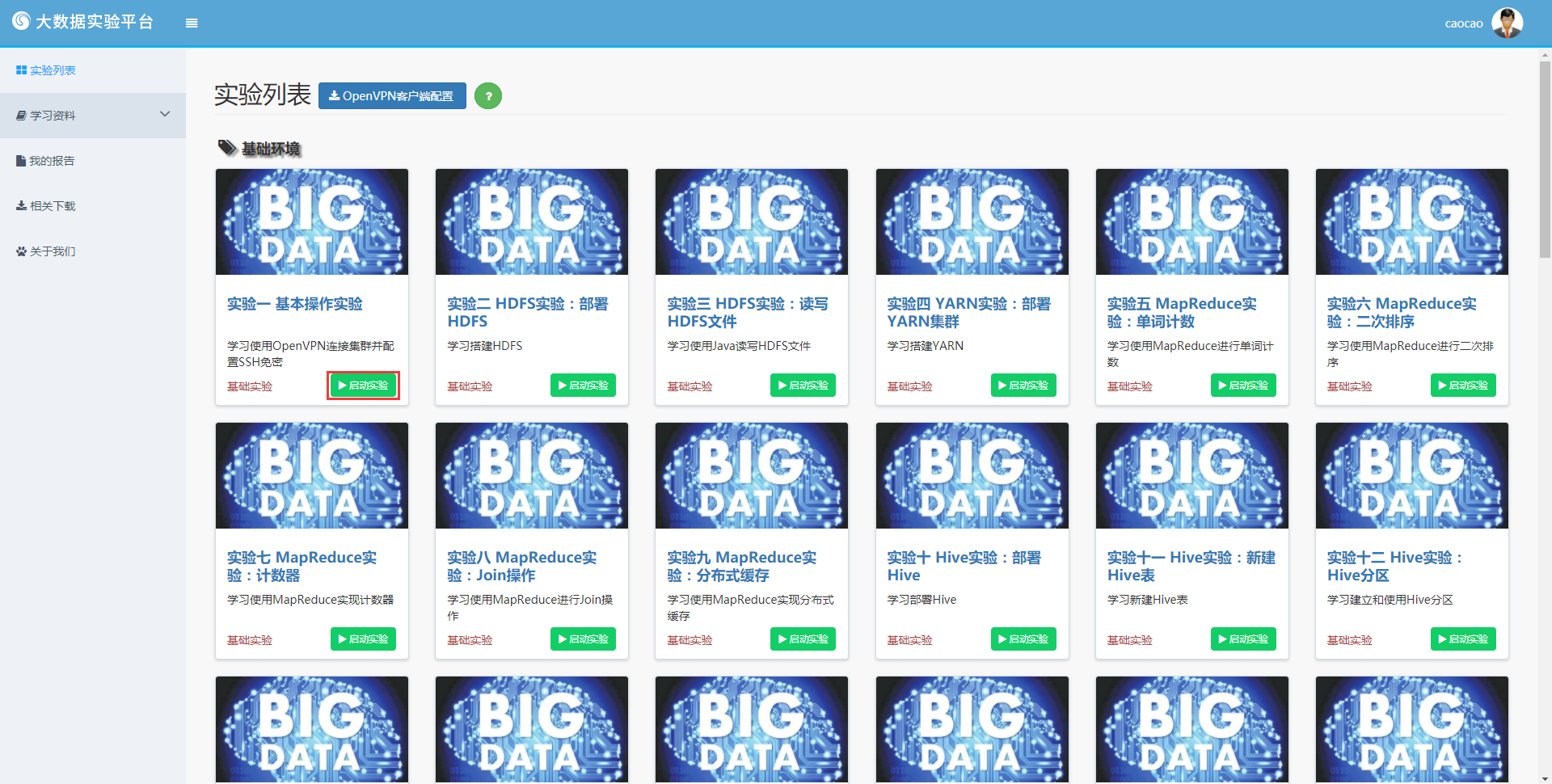
**图1-2 账号管理**

**1.3.1.3 集群管理**

此处的集群管理包含“开始实验”和“终止实验”，由于云创大数据实验一体机采用“Docker”技术，因此能够在几乎不占用系统资源情况下，实现大量机器快速创建与销毁，不必担心资源消耗高、启动销毁慢、管理维护难等问题。

（1）开始实验

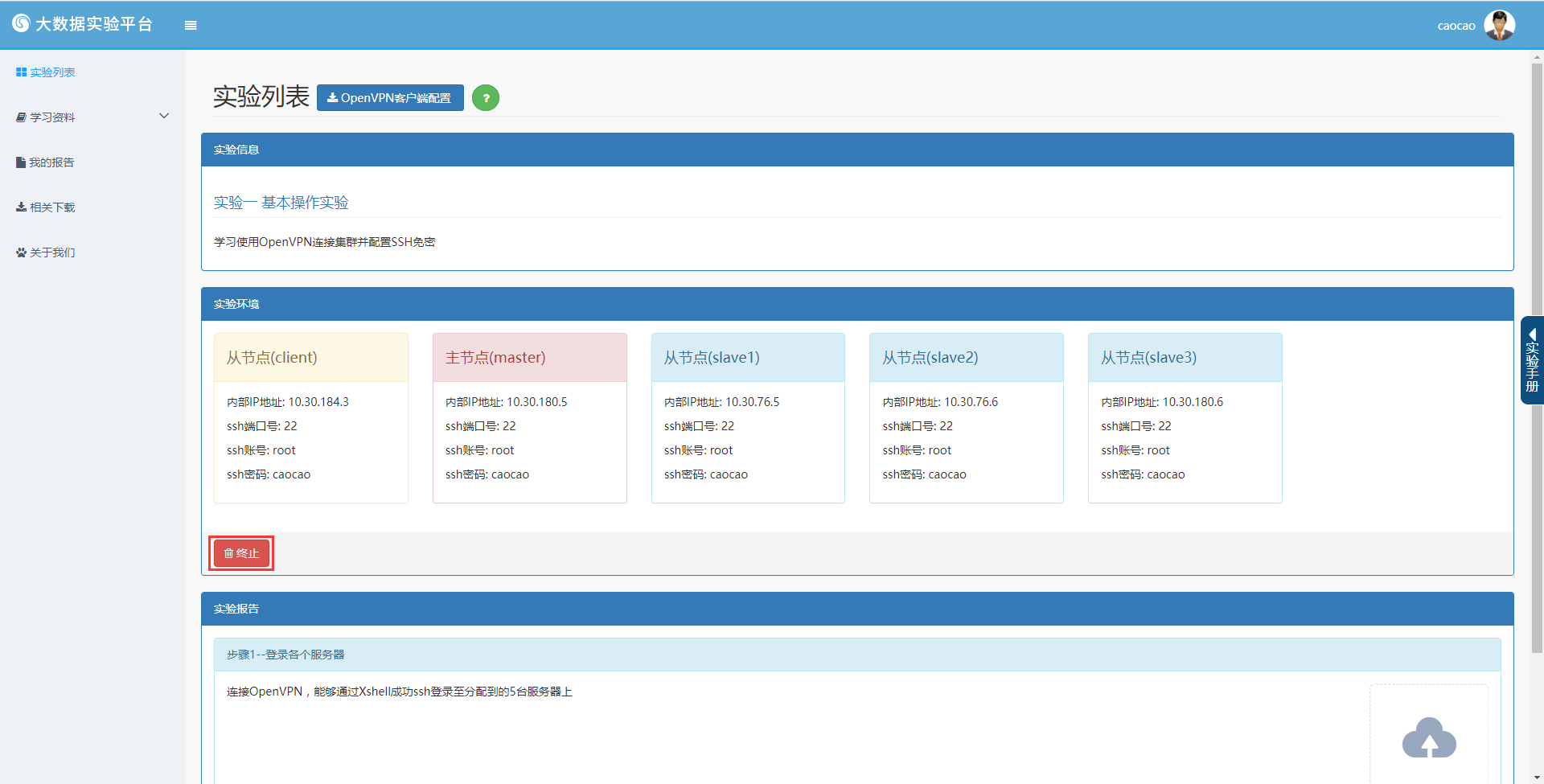
当需要开始实验时，直接点击实验界面对应实验面板上的启动按钮即可，后台会快速为用户新建五台能够支撑该实验的预装CentOS 7操作系统的机器，并配置好各自的主机名和IP地址等。如图1-3所示：



**图1-3 启动实验**

（2）终止实验

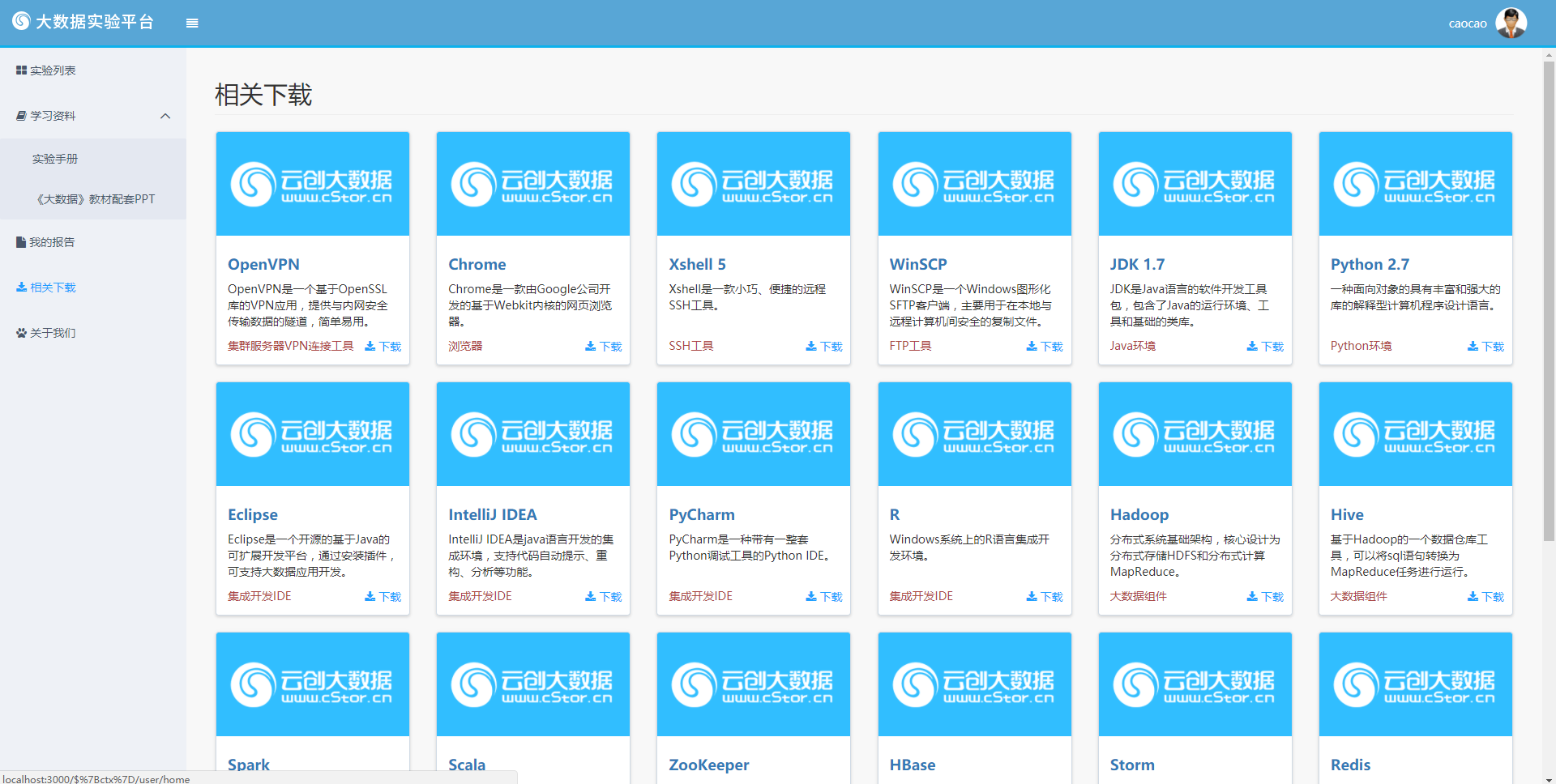
若实验过程中，由于命令敲错等各种原因导致集群无法使用，可在实验页面中随时终止您的实验进程，销毁失效的集群，之后再重新启动该实验，创建新的集群。如图1-4所示：



**图1-4 终止实验**

**1.3.1.4 相关下载**

大数据实验一体机的相关下载界面提供了实验所需的软件及插件的下载，为避免软件版本不同导致实验环境配置错误，请尽量下载和使用此处指定的软件版本与插件。如图1-5所示：



**图1-5 相关下载**

**1.3.2 Linux基本命令**

云创大数据实验平台搭建的集群服务器均为预装Linux操作系统的服务器。

Linux是一套免费使用和自由传播的类Unix操作系统，是一个基于POSIX和UNIX的多用户、多任务、支持多线程和多CPU的操作系统。它能运行主要的UNIX工具软件、应用程序和网络协议。它支持32位和64位硬件。Linux继承了Unix以网络为核心的设计思想，是一个性能稳定的多用户网络操作系统。

Linux操作系统诞生于1991年10月5日。Linux存在着许多不同的Linux版本，但它们都使用了Linux内核。Linux可安装在各种计算机硬件设备中，比如手机、平板电脑、路由器、视频游戏控制台、台式计算机、大型机和超级计算机。

严格来讲，Linux这个词本身只表示Linux内核，但实际上人们已经习惯了用Linux来形容整个基于Linux内核，并且使用GNU工程各种工具和数据库的操作系统。

本小节将介绍实验中涉及到的Linux操作系统命令：

(1)查看当前目录

pwd命令用于显示当前目录：

[root@master ~]# pwd

/root

(2)目录切换

cd命令用来切换目录：

[root@master ~]# cd /usr/cstor

[root@master cstor]# pwd

/usr/cstor

[root@master cstor]#

(3)文件罗列

ls命令用于查看文件与目录：

[root@master cstor]# ls

(4)文件或目录拷贝

cp命令用于拷贝文件，若拷贝的对象为目录，则需要使用-r参数：

[root@master cstor]# cp -r hadoop /root/hadoop

(5)文件或目录移动或重命名

mv命令用于移动文件，在实际使用中，也常用于重命名文件或目录：

[root@master ~]# mv hadoop hadoop2 #当前位于/root，不是/usr/cstor

(6)文件或目录删除

rm命令用于删除文件，若删除的对象为目录，则需要使用-r参数：

[root@master ~]# rm -rf hadoop2 #当前位于/root，不是/usr/cstor

(7)进程查看

ps命令用于查看系统的所有进程：

[root@master ~]# ps # 查看当前进程

(8)文件压缩与解压

tar命令用于文件压缩与解压，参数中的c表示压缩，x表示解压缩：

[root@master ~]# tar -zcvf /root/hadoop.tar.gz /usr/cstor/hadoop

[root@master ~]# tar -zxvf /root/hadoop.tar.gz

(9)查看文件内容

cat命令用于查看文件内容：

[root@master ~]# cat /usr/cstor/hadoop/etc/hadoop/core-site.xml

(10)查看服务器IP配置

ip addr命令用于查看服务器IP配置：

[root@master ~]# ip addr

1: lo: <LOOPBACK,UP,LOWER\_UP> mtu 65536 qdisc noqueue state UNKNOWN

link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00

inet 127.0.0.1/8 scope host lo

valid\_lft forever preferred\_lft forever

inet6 ::1/128 scope host

valid\_lft forever preferred\_lft forever

125: eth0@if126: <BROADCAST,MULTICAST,UP,LOWER\_UP> mtu 1500 qdisc noqueue state UP

link/ether 02:42:ac:11:00:0c brd ff:ff:ff:ff:ff:ff link-netnsid 0

inet 172.17.0.12/16 scope global eth0

valid\_lft forever preferred\_lft forever

inet6 fe80::42:acff:fe11:c/64 scope link

valid\_lft forever preferred\_lft forever

[root@master ~]#

**1.3.3 Vi编辑器**

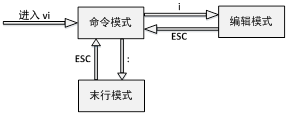
vi编辑器通常被简称为vi，而vi又是visual editor的简称。它在Linux上的地位就像Edit程序在DOS上一样。它可以执行输出、删除、查找、替换、块操作等众多文本操作，而且用户可以根据自己的需要对其进行定制，这是其他编辑程序所没有的。

vi 编辑器并不是一个排版程序，它不像Word或WPS那样可以对字体、格式、段落等其他属性进行编排，它只是一个文本编辑程序。没有菜单，只有命令，且命令繁多。vi有3种基本工作模式：命令行模式、文本输入模式和末行模式。

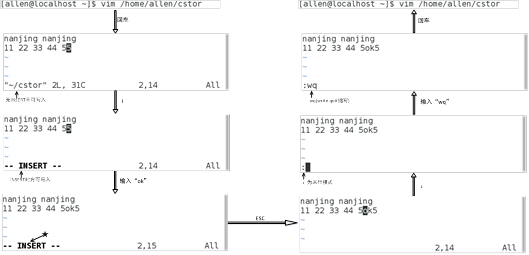
Vim是vi的加强版，比vi更容易使用。vi的命令几乎全部都可以在vim上使用。

vi编辑器是Linux和Unix上广泛使用的文本编辑器，工作在字符模式下。由于不需要图形界面，vi是效率很高的文本编辑器。尽管在Linux上也有很多图形界面的编辑器可用，但vi在系统和服务器管理中的功能是那些图形编辑器所无法比拟的。

Vi或vim是实验中用常用的文件编辑命令，命令行嵌入“vi/vim 文件名”后，默认进入“命令模式”，不可编辑文档，需键盘点击“i”键，方可编辑文档，编辑结束后，需按“ESC”键，先退回命令模式，再按“：”进入末行模式，接着嵌入“wq”方可保存退出。下述的图1-6为vi/vim三种模式转换示意图，图1-7为vi/vim操作实例。



**图1-6 vi/vim命令模式**

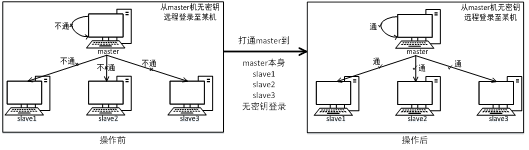


**图1-7 vim实例**

**1.3.4 SSH免密认证**

实验中，我们需要从实验室机器登录到集群中的Linux服务器上，而绝大多数Linux服务器采用的是SSH（Secure Shell）登录方式，因此，我们需要在实验室机器上安装一个SSH登录工具。常用的SSH工具包括XShell、Secure CRT、putty等，大数据实验一体机的相关下载界面中提供了XShell工具的下载。

Hadoop的基础是分布式文件系统HDFS，HDFS集群有两类节点以管理者-工作者的模式运行，即一个namenode（管理者）和多个datanode（工作者）。在Hadoop启动以后，namenode通过SSH来启动和停止各个节点上的各种守护进程，这就需要在这些节点之间执行指令时采用无需输入密码的认证方式，因此，我们需要将SSH配置成使用无需输入root密码的密钥文件认证方式。如图1-8所示：



**图1-8 实验集群master服务器SSH免密登录示意图**

**1.3.5 Java基本命令**

在安装Java环境后，可以使用Java命令来编译、运行或者打包Java程序。

(1)查看Java版本

[root@client ~]# java -version

java version "1.7.0\_79"

Java(TM) SE Runtime Environment (build 1.7.0\_79-b15)

Java HotSpot(TM) 64-Bit Server VM (build 24.79-b02, mixed mode)

(2)编译Java程序

[root@client ~]# javac Helloworld.java

(3)运行Java程序

[root@client ~]# java Helloworld

Hello World!

(4)打包Java程序

[root@client ~]# jar -cvf Helloworld.jar Helloworld.class

added manifest

adding: Helloworld.class(in = 426) (out= 289)(deflated 32%)

由于打包时并没有指定manifest文件，因此该jar包无法直接运行：

[root@client ~]# java -jar Helloworld.jar

no main manifest attribute, in Helloworld.jar

(5)打包携带manifest文件的Java程序

manifest文件用于描述整个Java项目，常用功能是指定项目的入口类：

[root@client ~]# cat manifest.mf

Main-Class: Helloworld

打包时，加入-m参数，并指定manifest文件名：

[root@client ~]# jar -cvfm Helloworld.jar manifest.mf Helloworld.class

added manifest

adding: Helloworld.class(in = 426) (out= 289)(deflated 32%)

之后，即可使用java命令直接运行该jar包：

[root@client ~]# java -jar Helloworld.jar

Hello World!

**1.3.6 Eclipse集成开发环境**

Eclipse 是一个开放源代码的、基于Java的可扩展开发平台。就其本身而言，它只是一个框架和一组服务，用于通过插件组件构建开发环境。幸运的是，Eclipse 附带了一个标准的插件集，包括Java开发工具（Java Development Kit，JDK）。

Eclipse是跨平台的自由集成开发环境（IDE）。最初主要用来Java语言开发，通过安装不同的插件Eclipse可以支持不同的计算机语言，比如C++和Python等开发工具。Eclipse的本身只是一个框架平台，但是众多插件的支持使得Eclipse拥有其他功能相对固定的IDE软件很难具有的灵活性。许多软件开发商以Eclipse为框架开发自己的IDE。

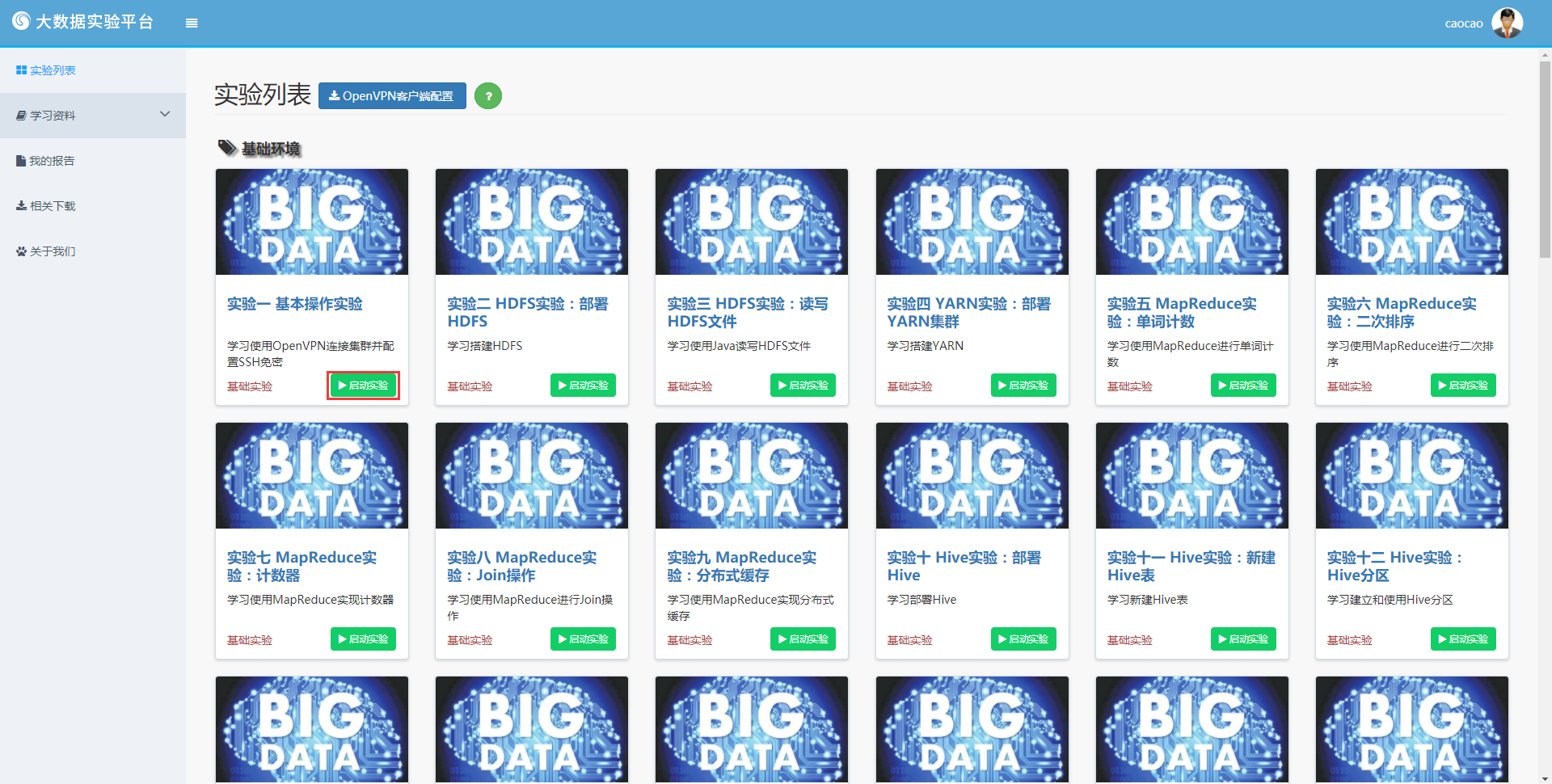
使用Eclipse可以帮助程序开发人员自动补全语义、方法名、方法参数、语句块等等，并且能够实时检查程序语法，提供错误和警告说明等等，能够极大地提高开发效率。

然而，使用Eclipse会占用较大的系统内存，因此，对于配置不高（32位操作系统或内存不足4G）的实验机器，不推荐安装Eclipse。

**1.4 实验步骤**

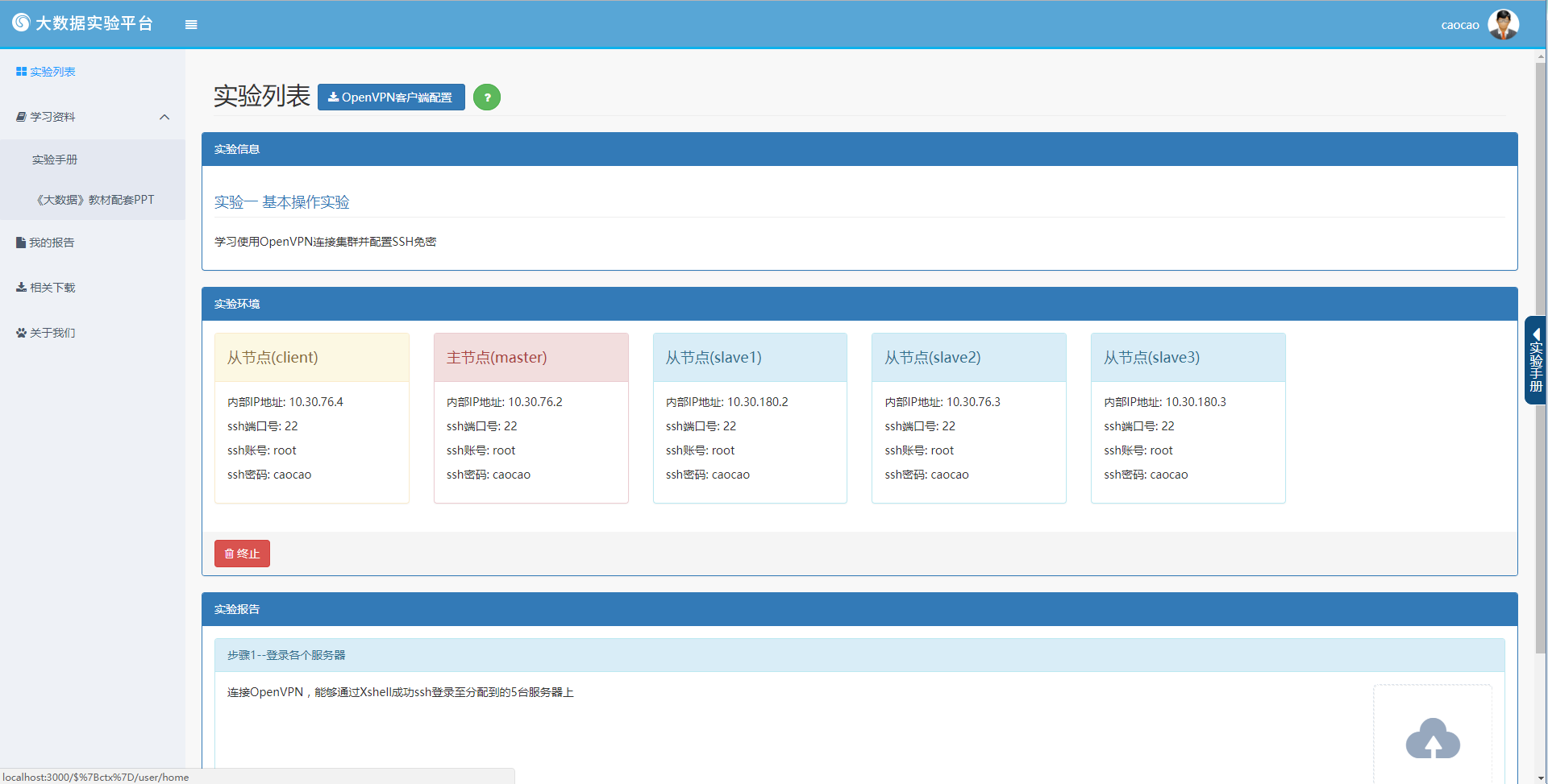
**1.4.1 搭建集群服务器**

使用自己的账号密码登录大数据实验一体机（默认密码为123456，登录后会自动跳转至密码修改界面，建议修改为自己的密码），进入实验界面。如图1-9所示：



**图1-9 登录大数据实验一体机**

选择实验一，点击启动，等待集群建立完成。如图1-10所示：



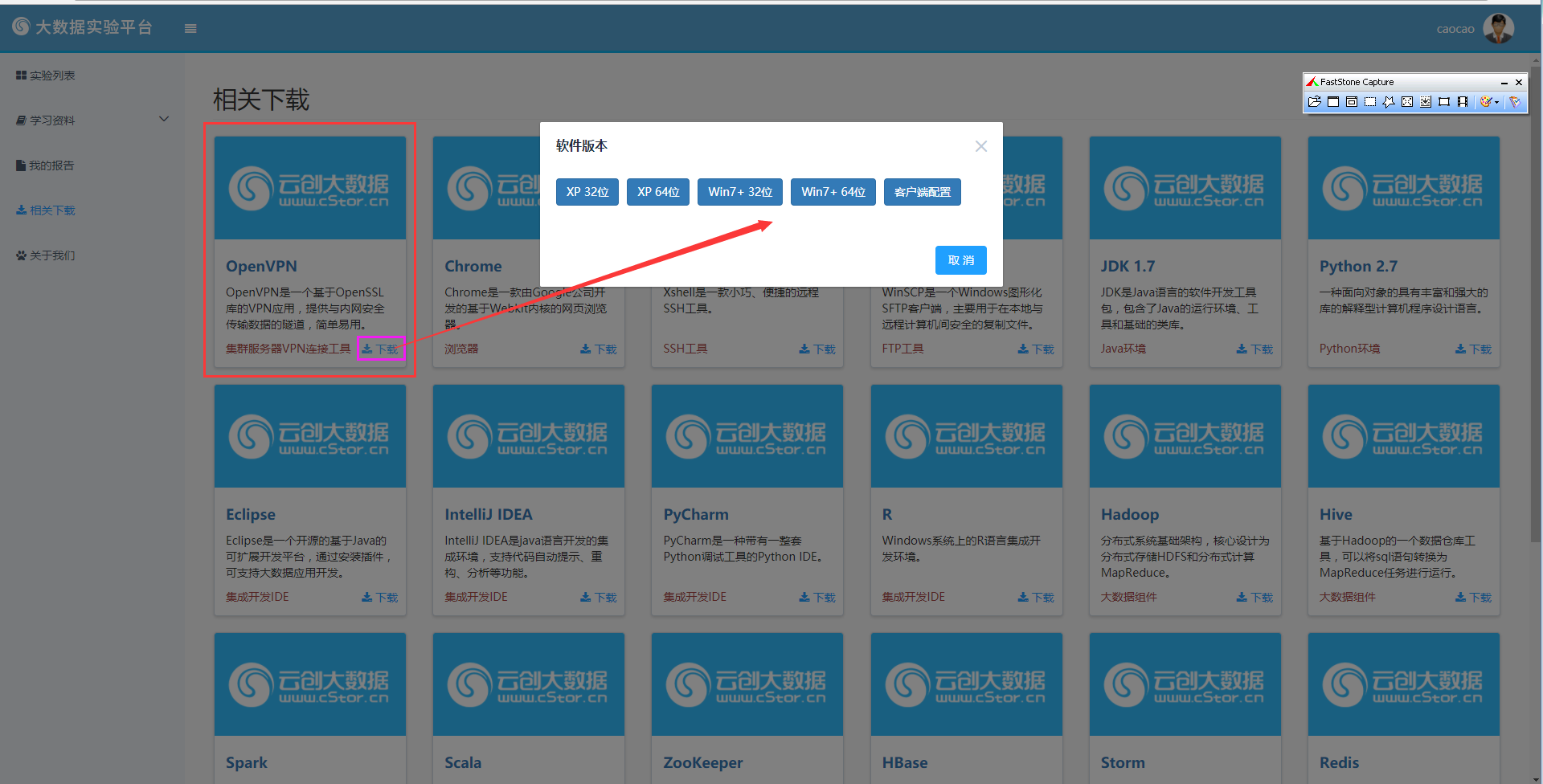
**图1-10 实验启动完成**

**1.4.2 使用SSH工具登录每台服务器**

在搭建好的Hadoop集群中，已经给出了所有五台服务器的内部ip地址、ssh端口号、ssh登录名以及ssh登录密码。

要想登录这些服务器，我们需要先下载OpenVPN客户端软件。

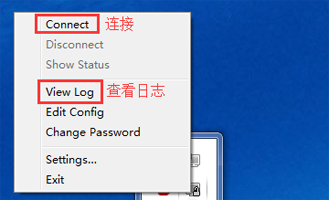
在相关下载中，根据PC的操作系统版本下载对应版本的OpenVPN客户端安装包，并下载客户端配置文件。如图1-11所示：



**图1-11 下载OpenVPN**

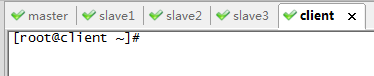
安装完成后，将下载的客户端配置压缩包解压，将其中的client.ovpn放于OpenVPN安装目录的config文件夹下。

以管理员身份运行OpenVPN GUI，任务栏将出现OpenVPN GUI图标，右键单击任务栏内OpenVPN GUI图标，点击「Connect」。如图1-12所示：



**图1-12 登录集群服务器**

当提示连接成功后，即可使用SSH工具登录大数据试验一体机分配的内网IP连接你的集群服务器。如图1-13所示：



**图1-13 登录集群服务器**

**1.4.3 添加域名映射**

系统搭建好的集群服务器已经完成修改主机名、关闭防火墙、安装JDK、同步时钟四步操作，为了可以安装大数据组件，还需为所有机器添加域名映射。如图1-14所示：



**图1-14 汇总各机机器名及其对应IP**

使用ssh工具登录到master服务器，使用vi命令编辑/etc/hosts文件：

[root@master ~]# vi /etc/hosts #root权限,编辑master的域名映射文件

在文件的末尾追加写入如下五行（具体的IP地址请替换为实际集群服务器内部IP）：

172.17.0.7 master

172.17.0.10 slave1

172.17.0.33 slave2

172.17.0.8 slave3

172.17.0.34 client

保存退出后，master服务器的域名映射即添加完成，使用cat命令查看/etc/hosts文件。如图1-15所示：



**图1-15 配置域名映射文件**

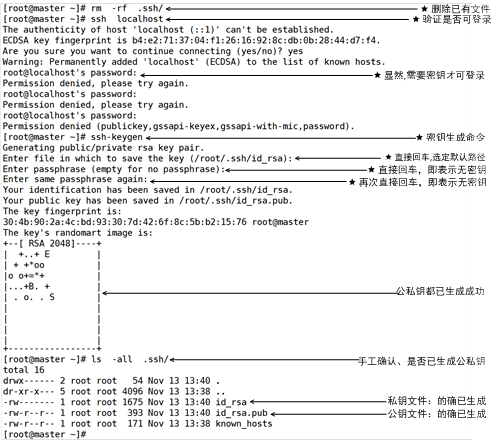
依次登录slave1~3和client服务器，重复该操作。

**1.4.4 配置SSH免密登录**

**1.4.4.1 生成master服务器密钥**

执行命令ssh-keygen，生成master服务器密钥。如图1-16所示：

[root@master ~]# ssh-keygen #root用户,master机,生成公私钥



**图1-16 master机密钥生成**

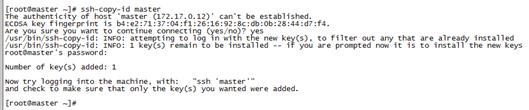
如图1-14所示，在master上执行“ssh-keygen”命令生成公私钥。第一个提示是询问将公私钥文件存放在哪，直接回车，选择默认位置。

第二个提示是请求用户输入密钥，既然操作的目的就是实现SSH无密钥登录，故此处必须使用空密钥，所谓的空密钥指的是直接回车，不是空格，更不是其他字符。此处请读者务必直接回车，使用空密钥。第三个提示是要求用户确认刚才输入的密钥，既然刚才是空密钥（直接回车即空），那现在也应为空，直接回车即可。

最后，可通过命令“ls -all /root/.ssh”查看到，SSH密钥文件夹.ssh目录下的确生成了两个文件id\_rsa和id\_rsa\_pub，这两个文件都有用，其中公钥用于加密，私钥用于解密。中间的rsa表示算法为RSA算法。

**1.4.4.2 拷贝master服务器公钥至本机**

执行命令ssh-copy-id master，将master服务器公钥拷贝至master服务器本身。如图1-17所示：



**图1-17 拷贝master公钥至master**

第一次连接master时，需要输入yes来确认建立授权的主机名访问，并需要输入root用户密码来完成公钥文件传输。

**1.4.4.3 验证master服务器ssh免密登录master本身**

公钥拷贝完成后，可以在master服务器上直接执行命令ssh master，查看是否可以免密登录master服务器：

[root@master ~]# ssh master #root用户,登录本机网络地址

[root@master ~]# exit #退出本次登录

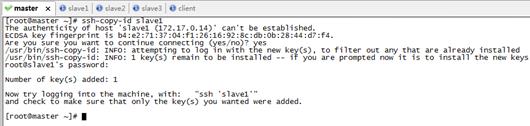
logout

Connection to master closed.

[root@master ~]#

**1.4.4.4 拷贝master服务器公钥至其余服务器**

执行命令ssh-copy-id slave1，将master服务器公钥拷贝至slave1服务器。如图1-18所示：



**图1-18 拷贝master公钥至slave1**

第一次连接slave1时，需要输入yes来确认建立授权的主机名访问，并需要输入root用户密码来完成公钥文件传输。

依照同样的方式将公钥拷贝至slave2、slave3和client服务器。

**1.4.4.5 验证master服务器ssh免密登录其余服务器**

公钥拷贝完成后，可以在master服务器上直接执行命令ssh master，查看是否可以免密登录slave1~3和client服务器：

[root@master ~]# ssh localhost #root用户,登录本机环回地址

[root@master ~]# ssh master #root用户,登录本机网络地址

[root@master ~]# ssh slave1 #root用户,从master远程登录slave1

[root@master ~]# ssh slave2 #root用户,从master远程登录slave2

[root@master ~]# ssh slave3 #root用户,从master远程登录slave3

**1.4.4.6 其余服务器配置ssh免密登录**

其余服务器按照同样的方式配置ssh免密登录，完成后验证是否可以互相之间实现SSH免密登录。

**1.4.5 在client服务器开发Java Helloworld程序**

使用ssh工具登录client服务器，使用vi编辑器编写Helloworld.java：

public class Helloworld {

public static void main(String args[]) {

System.out.println("Hello World!");

}

}

使用javac命令编译该程序，生成Helloworld.class文件：

[root@client ~]# javac Helloworld.java

[root@client ~]# ls

anaconda-ks.cfg data envSource Helloworld.class Helloworld.java

使用java命令执行该程序，输出Hello World!

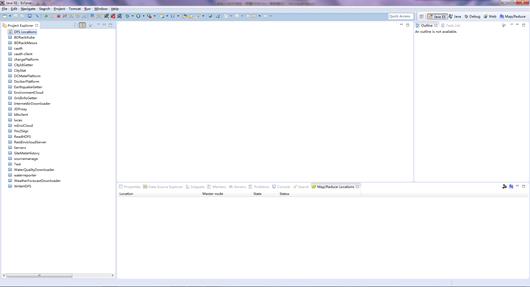
[root@client ~]# java Helloworld

Hello World!

**1.4.6 使用Eclipse开发Java Helloworld程序**

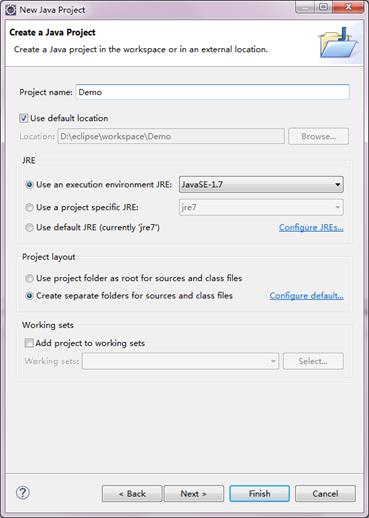
根据实验室机器的环境，下载并安装对应版本的JDK和Eclipse软件（若已安装则跳过该步骤）。

安装完成后，双击Eclipse图标，打开该软件。如图1-19所示：



**图1-19 Eclipse界面**

依次点击File->New->Java Project或File->New->Other->Java Project，项目名为Demo。如图1-20所示：



**图1-20 创建Java项目**

点击Finish，新建项目完成。在左侧导航栏中选中该项目，右键点击，选择New->Class，新建Helloworld类，并在该文件入如下内容：

public class Helloworld {

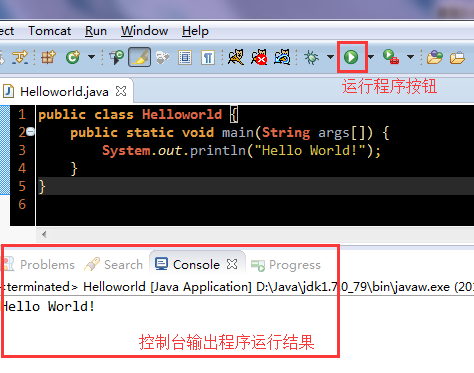
public static void main(String args[]) {

System.out.println("Hello World!");

}

}

完成后，点击上方的Run按钮，即可执行该程序。如图1-21所示：



**图1-21 Eclipse执行Java程序**