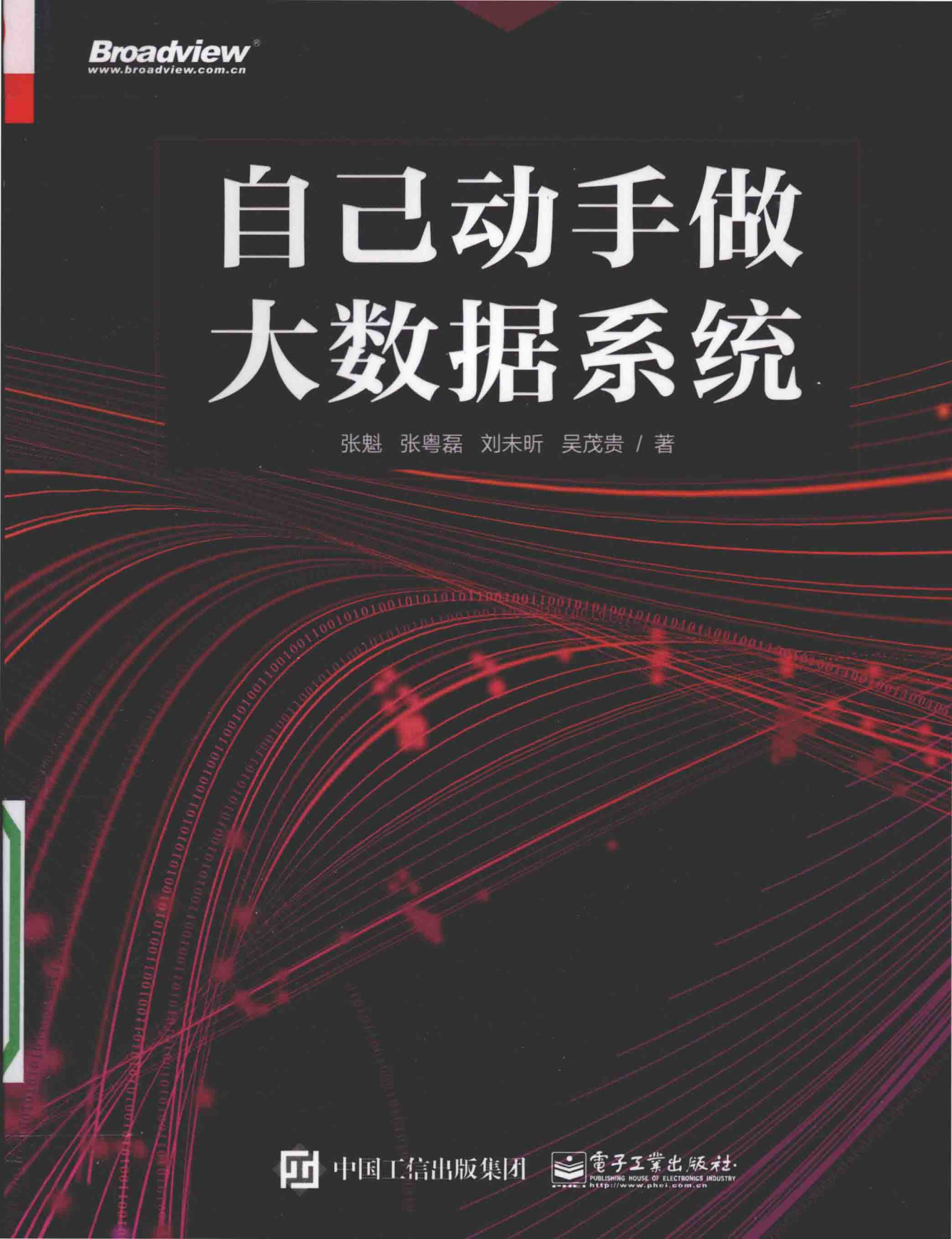
Broadview®
www.broadview.com.cn

自己动手做大数据系统

张魁 张粤磊 刘未昕 吴茂贵 / 著



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

自己动手做大数据系统

张魁 张粤磊 刘未昕 吴茂贵 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

如果你是一位在校大学生，对大数据感兴趣，也知道使用的企业越来越多，市场需求更是日新月异，但苦于自己基础不够，心有余而力不足；也看过不少大数据方面的书籍、博客、视频等，但感觉进步不大；如果你是一位在职人员，但目前主要使用传统技术，虽然对大数据很有兴趣，也深知其对未来的影响，但因时间不够，虽有一定的基础，常常也是打两天鱼、晒三天网，进展不是很理想。

如果你有上述疑惑或遇到相似问题，本书正好比较适合你。本书从 OpenStack 云平台搭建、软件部署、需求开发实现到结果展示，以纵向角度讲解了生产性大数据项目上线的整个流程；以完成一个实际项目需求贯穿各章节，讲述了 Hadoop 生态圈中互联网爬虫技术、Sqoop、Hive、HBase 组件协同工作流程，并展示了 Spark 计算框架、R 制图软件和 SparkRHive 组件的使用方法。本书的一大特色是提供了实际操作环境，用户可以在线登录云平台来动手操作书中的数据和代码，登录网址请参考 http://www.feiguyun.com/support。

图书在版编目（CIP）数据

自己动手做大数据系统 / 张魁等著. —北京：电子工业出版社，2016.10
ISBN 978-7-121-29586-7

Ⅰ. ①自… Ⅱ. ①张… Ⅲ. ①数据处理系统 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 205183 号

策划编辑：符隆美
责任编辑：葛　娜
印　　刷：北京京师印务有限公司
装　　订：北京京师印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：15.5　　字数：348 千字
版　　次：2016 年 10 月第 1 版
印　　次：2016 年 10 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

前　言

一个游泳爱好者，最大的烦恼是什么？没有好的教练？缺少好的教材？也许不是。如果哪天自己能拥有一个游泳池，可随时畅游，而且维护成本很低廉，甚至免费，同时还有教练的指导和一些游泳爱好者一起，那应该是一件很美的事。对于一个大数据爱好者，如果也能拥有一个属于自己的大数据实践环境，能够方便、快捷、随时随地使用真实环境，同时还有一些实战性、生产性的项目或课件，与一些志同道合的小伙伴一起攻坚克难，应该也是一件令人期待的事。

“纸上得来终觉浅，绝知此事要躬行”。要掌握一门技术，尤其像大数据相关技术，涉及的内容多，范围广，对环境的要求高，如果只是看看书、看看视频，很难深入理解，更不用说融会贯通了。一些有条件的学生，他们可以搭几个节点，组成一个微型大数据群，照着书中的一些实例练习，但这些练习往往支离破碎，缺乏系统性、生产性，更不用说包含生产性项目中的版本控制、质量管理和流程规范等。而这些对实施生产项目来说很重要，有时其重要性超过了对技术的要求。本书，就是为弥补这些内容而写的。

除了实战性、生产性的课件外，我们还提供了随时随地可操作、可实践的大数据云平台——飞谷云，这是我们自主开发的大数据平台，该平台用户可通过外网登录，与论坛及门户实现无缝连接。此外，还有很多志同道合的大数据爱好者一起学习、一起做项目。

本书主要内容

第 1 章，介绍我们为什么需要自己动手做大数据系统。

第 2 章，介绍动手做大数据系统的项目背景、项目架构及相关基础知识。

第 3 章，介绍大数据系统环境的搭建和配置，主要包括如何搭建和配置 Hadoop 集群、Sqoop、Hive、HBase、ZooKeeper、Spark、MySQL 等，图文并茂，内容翔实。

第 4 章，介绍大数据系统中数据获取相关技术，包括如何利用爬虫技术获取平面数据和使用

Sqoop 获取结构化数据。

第 5 章，介绍大数据系统中数据仓库工具 Hive 的使用方法及进行 ETL 的过程详解。

第 6 章，介绍大数据系统中数据库 HBase 的使用方法及和 Hive 之间的数据对接。

第 7 章，介绍如何使用数据展示利器 R 来展示 HDFS 中的数据。

第 8 章，介绍使用 Spark 计算模型来实时处理数据及 SparkRHive 组件的使用。

第 9 章，介绍如何搭建支撑大数据系统的云平台，以保证大数据系统的稳定性。

读者范围

- 对大数据感兴趣的院校师生。
- 对大数据有一定的基础，还想进一步熟悉整个生态系统的大数据爱好者。

勘误与支持

尽管我们仔细对待本书的写作，由于水平和能力有限，错误还是不可避免的。如果你在书中发现不妥或错误之处，请访问 http://www.feiguyun.com/support，留下宝贵意见，我们将非常感谢你的支持和帮助。

致谢

首先要感谢大数据实战团队，参与飞谷云大数据公益项目（www.feiguyun.com）的所有大数据爱好者，正是有了大家的支持和积极参与，才使得从飞谷一期的四个人，发展到目前飞谷七期的近四百人，短短一年多的时间，让我们真正感受到了共同坚持、诚信进取、协同分享的飞谷价值观所带来的收获和快乐，每期的项目线下启动会、交流会、项目结束总结会总能感受到大家积极参与的热情！同时也要感谢苏州大学计算机科学与技术学院何书萍老师、上海理工大学管理学院张帆老师、上海交通大学大数据分析俱乐部蒋军杰同学、中国社科院研究生院孙思栋同学、上海华师大数据分析俱乐部罗玉雪同学、上海大学黄文成同学等。

此外，要感谢飞谷管理团队的各位老师：陈健、刘军、吴嘉瑜、张勤池、王继红、张海峰、许小平、陶方震和刘李涛。诸君对飞谷大数据项目的热心参与及全力配合，是此公益项目得以持续推进的不懈动力。特别感谢为飞谷云提供实战项目的企业数据负责人；飞谷七期电商比价项目提供者——张晓雷先生及飞谷八期汽车推荐模型需求提供者——章水鑫先生，正是有了你们提供

的需求、数据和业务指导，才使得飞谷大数据小伙伴们有了学习大数据的真实场景，在实践中体会大数据分析价值和魅力。

飞谷云在全国一些大学还建立了交流群，作为每个群的组织者：中国科技大学张海洋同学、河南工程学院孟祥杰同学、南京农业大学邬家栋同学、西安电子科技大学刘东航同学等，为飞谷公益项目在院校中的推广，亦发挥了积极作用，在此一并表示感谢。

目　录

第 1 章
为什么要自己动手做大数据系统

1.1 大数据时代

“双 11”购物的狂欢、滴滴的便捷出行、阿尔法狗（AlphaGo）的进步、聊 QQ、刷微信、旅游、看病、健身等，我们已不由自主地置身于大数据时代，我们不但是大数据的创造者，同时也是大数据的体验者、受益者。随着 4G 的普及、移动互联网及“互联网+”的不断发展、5G 的临近、物联网的不断发展，大数据与我们的关系也越来越密切，取得的成果越来越鼓舞人心，对我们的影响深远。

2015 年 11 月初我国发布的《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》提出，拓展网络经济空间，推进数据资源开放共享，实施国家大数据战略，提前布局下一代互联网，可以说这是我国推行的国家大数据战略。各招聘网的大数据人才需求日新月异，需要的大数据人数在不断增加，涉及的行业范围在不断扩展，开出的薪资也很诱人。总之，不管从国家层面还是企业、个人，大数据已经跟我们息息相关。

大数据的需求增长迅猛，但是大数据方面的人才供给却相对滞后。其原因有二：一是人才培养不够。据我们了解，现在开设大数据相关课程的院校屈指可数，有好消息是复旦大学于 2015 年 8 月成立了大数据学院，2016 年 9 月开始招生；二是目前大数据技术的应用主要集中在“BAT”类型的互联网公司，整个技术团队的维护成本很高，这也抬高了其他科技公司使用大数据技术的门槛。

现在很多大学生或在职人员学习大数据技术，大都通过网络、博客、购买书籍等方式学习，这种学习往往效率不高，虽然书中或网上有这些内容的介绍，但很多都是理论性或概要性的，缺少具体的、详细的内容或步骤。大数据相关技术要求的面较广，包括操作系统、开发语言、数据库、网络，甚至需要一定的英语基础，虽然大数据相关技术都是开源的，如 Hadoop、Sqoop、HBase、Redis、MongoDB、GraphDB、Hive、Spark、Storm、OpenStack 等，但即使是工作多年且有一定经验的在职人员，自己搭建大数据环境挑战也不少，其中不少因问题多半途而废，就更不用说在

校大学生了。

大数据技术实践性很强，仅仅看一些书或视频，没有实际操作，很难有深刻的理解，更不用说融会贯通了。缺乏实际操作，在具体实施项目时一遇到问题，往往会不知所措，不知如何分析、定位、处理问题，感叹书到用时方恨少。针对诸如此类的问题，我们希望通过本书对一个实战性项目的详细介绍，给正在或将学习大数据的同学或在职人员提供一些思路或帮助。

1.2 实战大数据项目

目前很多想学习或正在学习大数据的人，大都面临一些问题或困惑，本书的第一个特点就是系统性，覆盖了如何利用爬虫、Sqoop 等获取各种数据，如何利用 HDFS、HBase 等存储大数据，如何利用 MapReduce、Hive、Pig、Python、Spark 等技术来处理大数据，如何利用 Spark 及 R 分析展示大数据整个过程，而且这些过程我们都可以以实战项目的方式在云平台上完成，这又体现出本书的第二个特点，即操作的便捷性。同时，本书既有对大数据主要原理的概述，又不乏对一些关键细节的详细描述，如 SSH、NTP 等详细配置等，重点、难点部分还贴有源码，这体现了本书的第三个特点，即实战性。因此，这些内容降低了在校大学生学习、搭建、实践大数据的难度，对有一定工作经验的在职人员也有参考价值。

此外，关于大数据生态系统的每个组成部分，比如 Hadoop、Hive、HBase 等，专门论述的书籍有很多，本书没有对某个部分的使用方法做大而全的罗列，而是着重于组件之间如何配合工作，并以同一个案例贯穿其中，旨在让读者从宏观上对大数据实现技术有一个全面的了解。

大数据技术实践性很强，相关技术也都是开源免费的，其中包含很多组件，各个组件更新很快，相互间的关系也较紧密，大数据开发涉及知识面广、数据复杂度高。纸上得来终觉浅，如果仅仅看看书、视频、博客等，则只能了解一些表面现象，很难做到融会贯通，更不用说深刻理解了；绝知此事要躬行，要想对大数据技术有一个较全面和深刻的掌握，亲自实践必不可少，通过自己亲手搭建、开发和实践，将大大加深对各组件功能、原理、架构、关系等的理解，为系统后续维护、升级、二次开发、问题定位、问题排查、问题解决等打下良好的基础。

1.3 大数据演练平台

我记得 10 年前大学毕业找工作时，发现企业需要的各项计算机实操技术自己都没实际做过，还是依靠自己在实验室的 SQL 实践进入一家外企从事 DBA。进入企业后，确实感觉到企业的实际需求环境与大学学习环境有很多不同。后来依靠自己买书来学习并和互联网论坛上共同的技术交流群配合，实践了数据库、大数据开发，先后担任了某大型咨询公司的大数据开发经理，以及某互联网金融公司的大数据架构师的职务。

反过来思考，如果我们在刚毕业二、三年或在校时，就能接触到真正的“准生产”项目实践，有二、三年的实际工作经验，那么接下来的机会选择会好很多。这样的经历感受，我身边几位工作十多年的技术朋友都有同感，于是便有了飞谷云大数据公益项目，我们几个大数据爱好者秉持对 10 年前的自己负责的态度，来对待目前像 10 年前现状的在校学生，以及刚毕业不久的同学们，让大家有机会来实践在学校、实验室、公司接触不到的大数据实践技术。

据笔者了解，因为学生在校一般两年多，所以研究生导师一般接的项目偏理论和咨询，不愿涉足具体实战开发的项目，担心学生毕业后难以维护。而实验室一般是基于一个问题点来的，很难有实际商业需求的连贯性，且一般公司，也只是一个岗位对应一种技术，很难有机会接触全貌。但飞谷云环境正好能解决这个问题。

通过飞谷云大数据公益项目来真正参与进来，自己可以动手做起来。在我们的公益项目中有实际的企业大数据负责人做需求文档，飞谷云的导师作为技术负责人来做需求分析设计，参与的同学来做各岗位的工程师（数据获取工程师、数据模型工程师、数据分析工程师、数据可视化工程师等），项目每期都有启动会说明，中间有交流分享会，项目结束后有总结，以对应大家的技术实践全过程，真正让自己动手与团队动手达到协同分享的目的。

同时，大数据依赖的各项基础开发，比如 shell、Python、MySQL、爬虫技术、机器学习等，都会以基础班的形式同步支持项目技术储备，真正让即使完全什么都不懂的同学，只要感兴趣想学，就可以跟着本书和飞谷云大数据公益项目动手做起来。所以不管你是在校学生、大公司某一岗位的开发人员，还是中小公司的开发者，都可以通过本书和飞谷项目来实现在个人环境中所无法企及的收获与快乐。

第 2 章
项目背景及准备

2.1 项目背景

急速增长的数据使传统的关系型数据库无所适从，如何存储大数据、如何处理大数据，如何挖掘大数据，这是大数据时代面临的一些挑战。毫无疑问，在大数据处理中，Hadoop、Spark 已成为当下的王者技术，经过开源社区无数贡献者的强大推动，Hadoop 以其显著的低成本、高可靠性、高性能等特性，成功征服了众多大数据处理需求的商业机构和科研团体，既有像 Google、Facebook、Yahoo、阿里、百度、腾讯、华为等这样的知名企业，也有数以万计的中小企业。

大数据是一个大趋势，Hadoop 又是处理大数据的大趋势，大数据实战项目就是为学习、传播 Hadoop 技术而设立的。

2.2 项目简介

项目数据的来源主要包括大数据网站、网络金融、网上招聘、网购等。我们通过爬虫技术把这些信息抓取到 HDFS 或 MySQL 中，其他数据源通过 Sqoop 导入 HDFS 或 MySQL 中。然后使用 MapReduce、Spark 等技术对数据进行处理，处理后的数据导入 Hive、HBase 等，最后用 Java、HiveQL、R 及 Spark 等进行数据分析与展示，其间包括实施生产性项目的主要流程，如设计、开发、调试、问题定位及处理、版本管理等。一键实现环境的搭建，系统调度管理实现自动化，同时对系统的性能等实现实时监控。

2.3 项目架构

项目使用的主要技术是开源的，如 Hadoop、Hive、HBase、Spark、Ganglia、R、OpenStack、MySQL 等，系统的安装调度等用 Java、shell、Python 等工具自主开发。项目的总体架构如图 2.1 所示。

图 2.1　项目的总体架构图

2.4　操作系统

有关 Linux 和 UNIX 的知识网上有很多，这里主要介绍一些简单、实用的内容及本项目使用的操作系统，使初学者有一个大体的认识和了解。已熟悉 Linux 或 UNIX 的读者可跳过本节。

Linux 与 UNIX 的主要区别是：Linux 是开源、免费的，UNIX 是商业软件；Linux 能够运行在多种平台上，而 UNIX 大多与硬件配套。

常用的 Linux 及 Unix 版本信息如表 2.1 所示。

表 2.1　操作系统版本说明

类别	发行版本	说明	安装
Linux	Debain	一种流行的非商业性质的版本	apt-get install wget
	Ubuntu	Debain 的精炼版本 帮助文档 URL：help.ubuntu.com	apt-get install wget
	RedHat Enterprise	Red Hat Linux 商业化版本 帮助文档 URL：redhat.com/docs	yum install wget
	CentOS	模仿 Red Hat Enterprise Linux 的免费版本	yum install wget
	Fedora	从 Red Hat Linux 分出的非商业化版本	yum install wget
UNIX	Solaris	始于 Sun，但目前属于 Cracle 公司	安装软件包
	HP-UX	用于 HP 的硬件平台	安装软件包
	AIX	用于 IBM 的硬件平台 帮助文档 URL：www.redbooks.ibm.com	安装软件包

编写脚本的语言有 shell、Perl、Python 等。在大多数系统上，默认的 shell 都是 bash（即 Bourne Again shell），但在几种 UNIX 上也用 sh（Bourne shell）和 ksh（Korn shell）。各种操作系统上都

有 shell，用 shell 编写的脚步可移植性好，除了其调用的命令外，需要依赖的东西不多，但如果要实现一些较复杂或高端的脚本，建议采用 Perl 或 Python 等。

分析和理解 Linux 或 UNIX 系统中的一些常见日志文件，对于调试和排查问题大有帮助。如表 2.1 所示，列出了几种主要日志文件的用途，Linux 系统日志文件一般在/var/log 目录下，HP-UX 及 AIX 的一般在/var/adm 目录下。

表 2.2　日志文件列表

日志文件	涉及程序	操作系统	说明
mail*	与 mail 有关	所有系统	记录有关 mail 的信息
messages	很多	UR	系统日志文件
syslog*	很多	UH	系统日志主文件
cron	cron	RAH	记录 cron 的执行情况及出错信息
debug	很多	U	调试输出日志文件
daemon.log	很多	U	记录所有守护进程的功能消息
dpkg.log	dpkg	U	软件包管理日志
/etc/httpd*	httpd	R	记录 Apache HTTP 服务器日志
wtmp	login	所有系统	用户登录记录

系统说明：
U=Ubuntu，R=Red Hat 和 CentOS，H=HP-UX，A=AIX

本项目使用的操作系统为 Ubuntu 14.0.4。

登录 Ubuntu 服务器，可通过 GNU 工具 PuTTY 客户端，配置如图 2.2 所示。

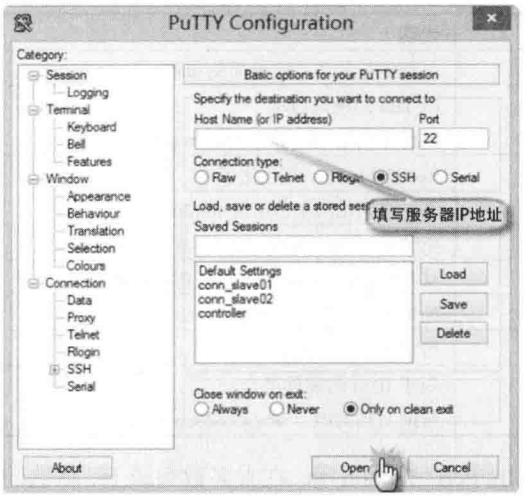


图 2.2　PuTTY 配置