“十三五”普通高等教育规划教材

大数据技术导论

程显毅　主编

提供电子课件

http://www.cmpedu.com

机械工业出版社
CHINA MACHINE PRESS

“十三五”普通高等教育规划教材

大数据技术导论

主　编　程显毅

参　编　陈伏兵　吴云霞　孙丽丽　温长吉

任越美　段先华　代冉冉　褚慧敏

机械工业出版社

本书以面向应用、面向实战为指导思想，紧扣企业技术人才培养的特点，在知识点讲解和实验中避免复杂的理论，使读者能快速上手体验、验证大数据处理的魅力，以激发读者的学习兴趣。

本书覆盖了大数据生命周期中的主要技术要点，全书共 8 章，第 1 章介绍大数据的产生和特点及思维的变革，第 2 章了解大数据生态系统，第 3～7 章按照大数据的生命周期，分别讨论大数据采集与预处理、大数据管理、大数据分析、大数据可视化、大数据应用的基本原理和方法，第 8 章讨论大数据安全面临的挑战。

本书可作为本科、高职院校大数据技术或数据科学课程的参考书或教材，也可供数据科学相关技术人员阅读。

本书配套授课电子课件，需要的教师可登录 www.cmpedu.com 免费注册，审核通过后下载，或联系编辑索取（QQ：308596956，电话：010-88379753）。

图书在版编目（CIP）数据

大数据技术导论 / 程显毅主编. —北京：机械工业出版社，2019.4
“十三五”普通高等教育规划教材
ISBN 978-7-111-62171-3

Ⅰ. ①大… Ⅱ. ①程… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 048545 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
策划编辑：汤　枫　　责任编辑：汤　枫
责任校对：李亚娟　　责任印制：张　博
三河市宏达印刷有限公司印刷
2019 年 4 月第 1 版 • 第 1 次印刷
184mm×260mm • 11.5 印张 • 276 千字
0001－3000 册
标准书号：ISBN 978-7-111-62171-3
定价：39.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
服务咨询热线：（010）88379833
读者购书热线：（010）68326294

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
教育服务网：www.cmpedu.com
金　书　网：www.golden-book.com

前　言

“大数据”已经成为近年来备受关注的热词，越来越多的人逐渐认识到，大数据将是新一轮产业革命的新动力、新引擎。相关报告预计未来 5 年，大数据或者数据工作者的岗位需求将激增，其中大数据分析师的缺口在 140 万～190 万。但大数据人才培养以及数据科学研究似乎远未做好准备。据教育部公布数据显示：继 2016 年北京大学、中南大学、对外经贸大学首批设立大数据相关学科后，2017 年中国人民大学、北京邮电大学、复旦大学等 32 所高校成为第二批成功申请“数据科学与大数据技术”本科新专业的高校。2018 年新增“数据科学与大数据技术”专业的高校达 248 所，2019 年又新增 406 所高校开设“数据科学与大数据技术”专业。

《大数据技术导论》是“数据科学与大数据技术”专业必修的第一门专业基础课，目前市场上大多数《大数据技术导论》教材只是作为专业技术课程的综述，专业术语过多，学生学起来很困难。本书的宗旨是导知识、导方法、导思维、导意识和导职业，而不是导技术。目的是把大数据思维、原理传递给想实践大数据的读者手中，不是让读者掌握大数据深奥的数学理论和复杂的环境搭建细节。因此，本书重点是在已搭建好的大数据平台下，实施大数据应用方案，注意力完全集中在能有效工作的大数据技术应用上，这样可以用最少的时间、最快的速度消化和部署大数据应用项目。

本书具有以下特点：

1）让读者从实践中学习大数据思维、原理和方法。书中给出了大量的故事和实验指导案例，指导读者一步一步迈向大数据世界。

2）学习大数据不需要很深的数学作为前提。无论你是谁，无论你来自哪里，无论你的受教育背景如何，都有能力使用书中提供的方法，解决大数据应用问题。

3）每一章都提供了一定数量的习题，用于检查学习效果。

4）为了减轻读者对编程基础的依赖，使文科专业也能学习大数据，本书采用 R 语言作为编程环境。

5）大数据生态环境 Hadoop 采用集群安装，实验更接近实际应用。

6）不仅理工科学生要掌握大数据技术，非理工科的学生也要掌握最基本的大数据技术，本书适合各类专业学习大数据技术。

带*号的章节为选学内容。

由于大数据领域发展迅猛，对许多问题编者并未做深入研究，一些有价值的新内容也来不及收入本书。加上编者知识水平和实践经验有限，书中难免存在不足之处，敬请读者批评指正。

编　者

目　录

第1章 概　论

大数据作为继云计算、物联网之后 IT 领域又一次颠覆性的理念，备受人们的关注。大数据已经渗透到各行各业众多领域，对人类的社会生产和生活产生重大而深远的影响。那么大数据是如何产生的？什么是大数据？大数据能做什么？本章将回答这些问题。

1.1 揭秘大数据

1.1.1 大数据产生历史必然

（1）数据产生方式的变革促成大数据时代的来临

由于物联网技术的成熟，数据产生方式经历了被动产生→主动产生→自动产生三个阶段（见图 1.1）。

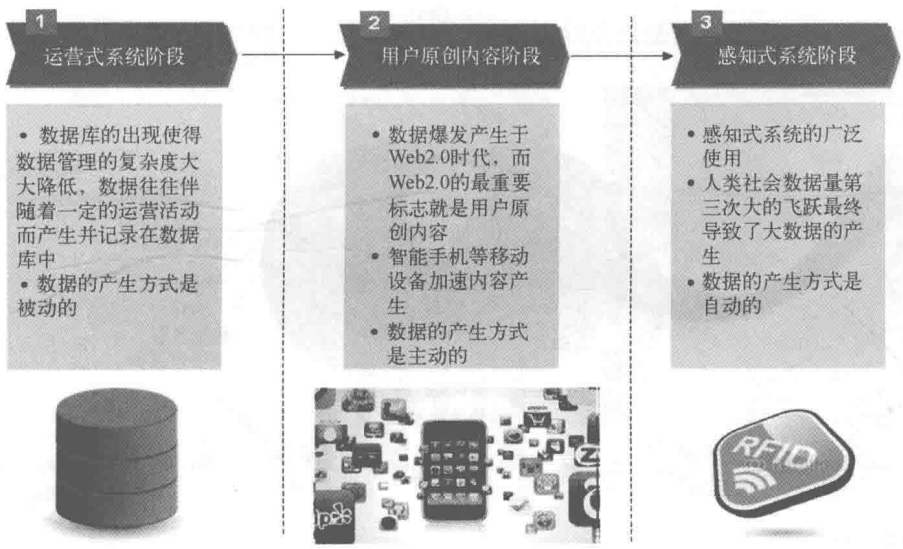

图 1.1　数据产生方式的三个阶段

（2）云计算是大数据诞生的前提和必要条件

在云计算出现之前，传统的计算机是无法处理如此量大的“非结构数据”。以云计算为基础的信息存储、分享和挖掘手段，可以便宜、有效地将这些大量、高速、多变化的终端数据存储下来，并随时进行分析与计算。图 1.2 给出了云的发展历程。

1）云计算转变了数据的服务方式。

2）虚拟化为进入大数据时代铺平了道路。

基于以上两点，大数据的出现是历史的必然。科技发展到今天，正处于大数据时代的前

夜，不管你接受与否、承认与否，大数据必将对全人类的生产生活方式带来一次深刻的变革。

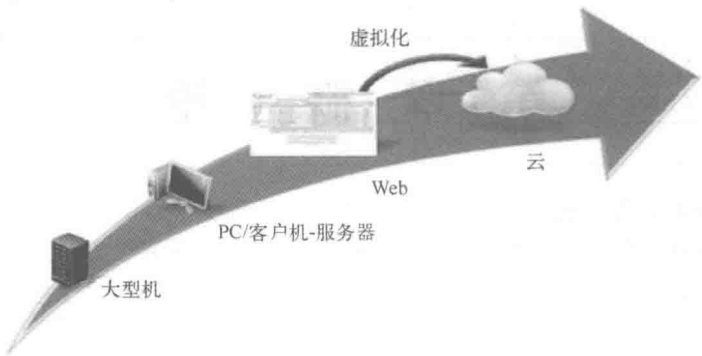

图 1.2　云的发展历程

1.1.2　大数据概念和特征

大数据（Big Data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

大数据具有 4V 特征，如图 1.3 所示。

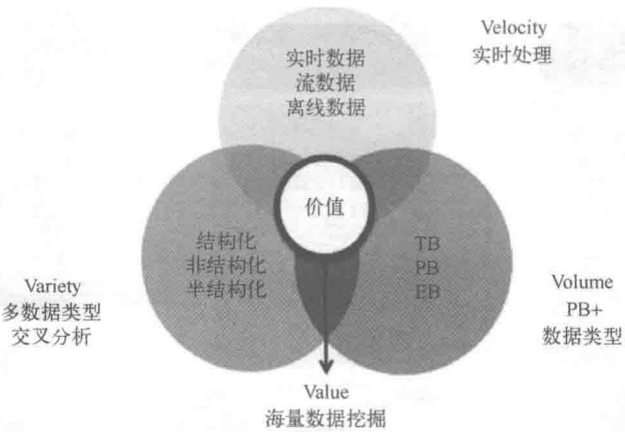

图 1.3　大数据 4V 特征

几点说明如下：

1）容量度量最小单位是 bit，1B=8bit，1KB=1024B，按从小到大顺序给出常用度量单位：bit、B、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。从 KB 开始它们按照进率 1024 来计算。

为了让读者理解数据量有多大，图 1.4 给出了一个示例，2014 年美国国会图书馆藏书的数据量约 235TB，而百度每天的数据处理量约为其 5000 倍。

2）数据的种类如图 1.5 所示。

3）速度快包括两个方面：产生速度快、处理速度快。图 1.6 给出了数据产生的增长速度示意图。

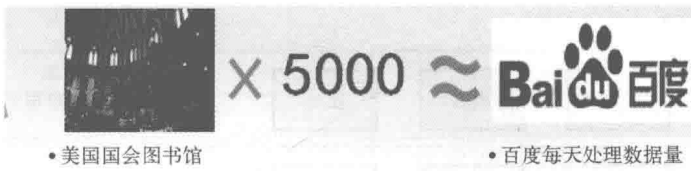

图 1.4　数据量示例

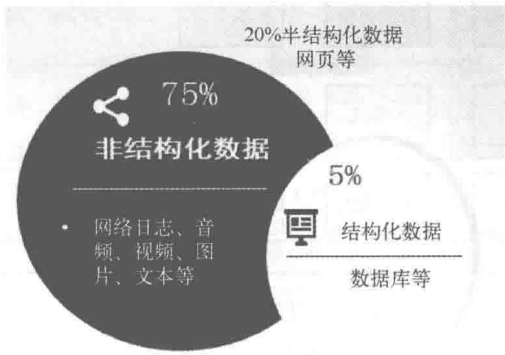

图 1.5　数据种类

4）价值是相对的、有时效的，隐藏较深，人们看到的只是冰山一角（见图 1.7）。

图 1.6　数据产生的增长速度

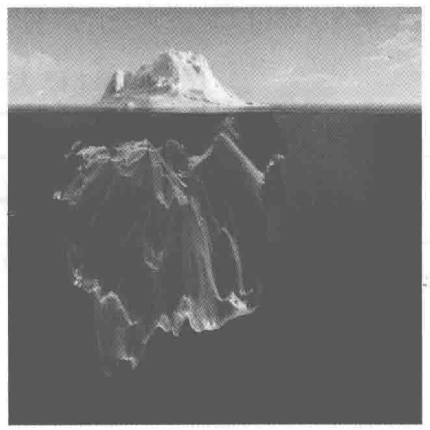

图 1.7　大数据价值

1.1.3　大数据生命周期

大数据采集、存储、处理、解释和应用，形成了大数据生命周期（见图 1.8）。

1）数据采集：ETL（Extract-Transform-Load）负责将分布的、异构数据源中的数据，如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

2）数据存储：数据存储方式主要有关系型数据库 SQL、非关系型数据库 NoSQL、分布

式数据库 NewSQL 等。

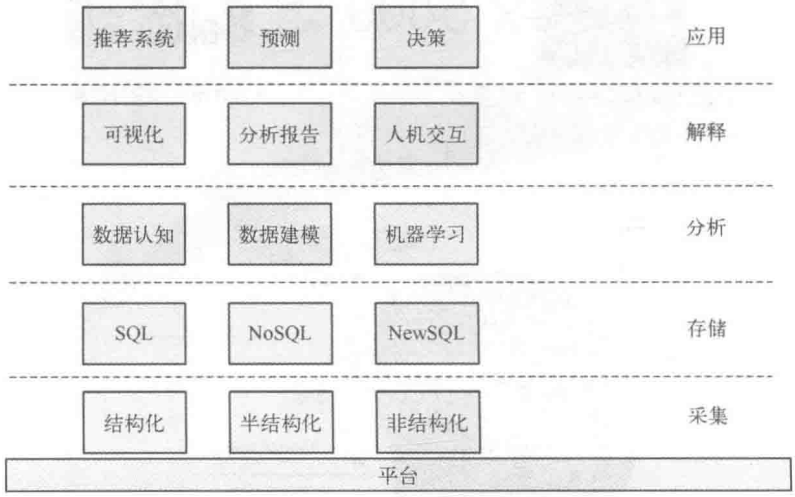

图 1.8　大数据生命周期

3）数据分析：包括假设检验、显著性检验、差异分析、相关分析、T 检验、方差分析、卡方分析、回归分析、因子分析、聚类分析、主成分分析、因子分析、关联分析等，这是生命周期最重要的阶段。

4）结果解释：包括可视化、数据分析报告等。

表 1.1 给出了大数据生命周期各个阶段相关技术产品。

表 1.1　大数据生命周期各个阶段相关技术产品

类　别		产　品
平台	本地	Hadoop、MapR、Hortonworks
	云端	Cloudera、AWS、Google Compute Engine
数据存储	关系型数据库 SQL	Greenphum、Aster Data、Vertica
	非关系型数据库 NoSQL	云数据库：Datastore
		键值对数据库：Redis
		文档数据库：MongoDB
		图数据库：Neo4j、GraghDB
		列表式数据库：HBase
	分布式数据库 NewSQL	AmazonDB、Azure、Smanner、VoltDB
数据分析	数据仓库	Hive
	批模式	MapReduce、Spark
	流模式	Storm、Kafka、Spark
	图模式	GraphX、Pregel
	查询分析模式	Hive
	机器学习	Mahout、Weka、R、Python

（续）

类　别		产　品
数据解释	日志处理	Flume
	可视化	Echarts、Excel、SPSS、R、Python、Tableau
	数据分析报告	RMarkDown

1.1.4　大数据与物联网、云计算、人工智能

物联网、云计算和大数据三者互为基础，物联网产生大数据，大数据需要云计算。物联网将物品和互联网连接起来，进行信息交换和通信，以实现智能化识别、定位、跟踪、监控和管理，云计算解决万物互联带来的巨大数据量，所以三者互为基础，又相互促进，可以将它们看作一个整体，相互发展、相互促进（见图 1.9）。

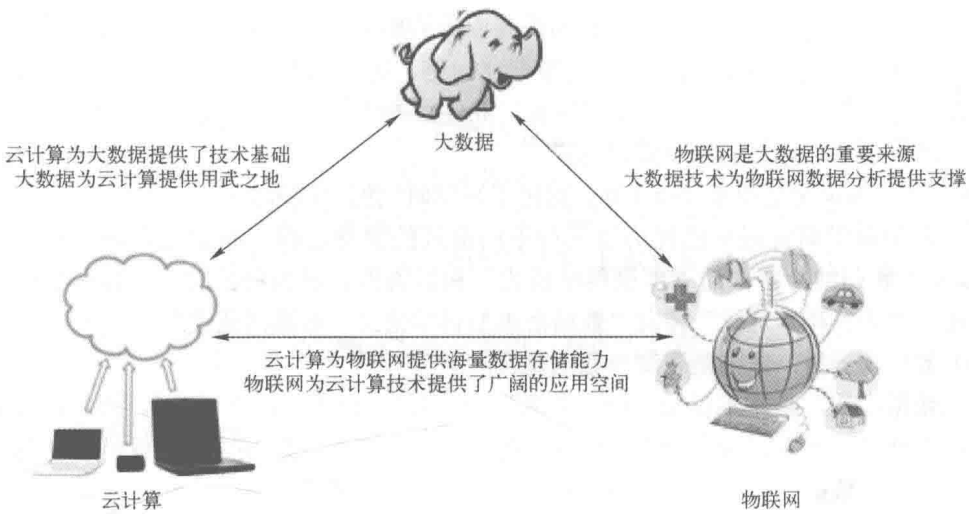

图 1.9　云计算、大数据与物联网之间关系

1.1.5　大数据时代的八个重大变革

（1）决策方式：目标驱动型→数据驱动

传统科学思维中，决策制定往往是“目标”或“模型”驱动的——根据目标（或模型）进行决策。然而，大数据时代出现了另一种思维模式，即数据驱动型决策，数据成为决策制定的主要“触发条件”和“重要依据”。例如，近年来，很多高新企业中的部门和岗位设置不再是“固化的”，而是根据所做项目与所处的数据环境，随时动态调整其部门和岗位设置。然而，部门和岗位设置的敏捷性往往是基于数据驱动的，根据数据分析的结果灵活调整企业内部结构。

（2）方法论：基于知识的方法→基于数据的方法

传统的方法论往往是“基于知识”的，即从“大量实践（数据）”中总结和提炼出一般性知识（定理、模式、模型、函数等）之后，用知识去解决（或解释）问题。因此，传统的问题解决思路是“问题→知识→问题”，即根据问题找“知识”，并用“知识”解决“问

题”。然而，数据科学中兴起了另一种方法论——“问题→数据→问题”，即根据问题找“数据”，并直接用数据（不需要把“数据”转换成“知识”的前提下）解决问题。

（3）计算方式：复杂算法→简单分析

“只要拥有足够多的数据，我们可以变得更聪明”是大数据时代的一个新认识。因此，在大数据时代，原本复杂的“智能问题”变成简单的“数据问题”——只要对大数据进行简单查询就可以达到“基于复杂算法的智能计算的效果”。为此，很多学者曾讨论过一个重要话题——“大数据时代需要的是更多数据还是更好的模型?”。机器翻译是传统自然语言技术领域的难点，虽曾提出过很多种“算法”，但应用效果并不理想。近年来，Google 翻译等工具改变了“实现策略”，不再仅靠复杂算法进行翻译，而对他们之前收集的跨语言语料库进行简单查询的方式，提升了机器翻译的效果和效率。

（4）管理方式：业务数据化→数据业务化

在大数据时代，企业需要重视一个新的课题——数据业务化，即如何“基于数据”动态地定义、优化和重组业务及其流程，进而提升业务的敏捷性，降低风险和成本。但是，在传统数据管理中人们更加关注的是业务的数据化问题，即如何将业务活动以数据方式记录下来，以便进行业务审计、分析与挖掘。可见，业务数据化是前提，而数据业务化是目标。

（5）研究范式：第三范式→第四范式

2007 年，图灵奖获得者 Jim Gray 提出了科学研究的第四范式——数据密集型科学。在他看来，人类科学研究活动已经历过三种不同范式的演变过程（原始社会的“实验科学范式”、以模型和归纳为特征的“理论科学范式”和以模拟仿真为特征的“计算科学范式”），目前正在从“计算科学范式”转向“数据密集型科学范式，即第四范式”。

（6）数据的属性：数据是资源→数据是资产

在大数据时代，数据不仅是一种“资源”，而更是一种重要的“资产”。因此，数据科学应把数据当作“一种资产来管理”，而不能仅仅当作“资源”来对待。也就是说，与其他类型的资产一样，数据也具有财务价值，且需要作为独立实体进行组织与管理。

（7）数据处理模式：小众参与→大众协同

传统科学中，数据的分析和挖掘都是具有很高专业素养的“企业核心员工”的事情，企业管理的重要目的是如何激励和绩效考核这些“核心员工”。但是，在大数据时代，基于“核心员工”的创新工作成本和风险越来越大，而协同日益受到重视（见图 1.10）。

（8）思维方式：抽样思维→整体思维+相关思维+容错思维

1）整体思维。整体思维是根据全部样本得到结论，即“样本=总体”。因为大数据是建立在掌握所有数据，至少是尽可能多的数据基础上，所以整体思维可以正确地考查细节并进行新的分析。

如果数据足够多，则会让人们觉得有足够的能力把握未来，从而做出自己的决定。

结论：从抽样中得到的结论总是有水分的，而根据全部样本得到的结论水分就很少，数据越大，真实性也就越大。

2）相关思维。相关思维要求人们只需要知道是什么，而不需要知道为什么。在这个不确定的时代，等找到准确的因果关系，再去办事的时候，这个事情早已经不值得办了。所以，社会需要放弃它对因果关系的渴求，而仅需关注相关关系。

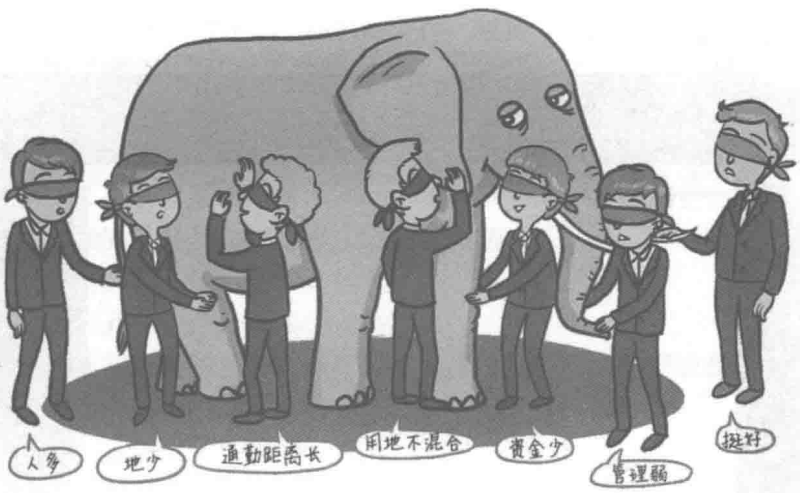

图 1.10　大数据需要协同

结论：为了得到即时信息，实时预测，寻找到相关性信息，比寻找因果关系信息更重要。

3）容错思维。实践表明，只有 5%的数据是结构化且能适用于传统数据库的。如果不接受容错思维，剩下 95%的非结构化数据都无法被利用。

对小数据而言，因为收集的信息量比较少，必须确保记下来的数据尽量精确。然而，在大数据时代，放松了容错的标准，人们可以利用这 95%数据做更多新的事情，当然，数据不可能完全错误。

结论：容错思维可以利用 95%的非结构化数据，帮助人们进一步接近事实的真相。

1.2　Linux 系统概述

1.2.1　Linux 版本

在 Linux 系统各个发行版本中，CentOS 系统和 Ubuntu 系统在服务端和桌面端使用占比最高，网络上资料最齐全，所以建议使用 CentOS 6.4 或 Ubuntu LTS 14.04 系统。

一般来说，如果要做服务器，可以选择 CentOS 或者 Ubuntu Server；如果做桌面系统，则选择 Ubuntu Desktop。但是在学习 Hadoop 方面，虽然这两个系统没有多大区别，但还是推荐新手使用 CentOS 操作系统。

虚拟机安装地址如下：

http://dlsw.baidu.com/sw-search-sp/soft/08/15321/VirtualBox_5.0.10.4061_104061_Win.1448355141.exe

图 1.11 所示为选择安装 CentOS 后 Desktop 使用界面。

1.2.2　Linux 系统目录结构

登录系统后，在当前命令窗口输入命令：

```
#ls /
```

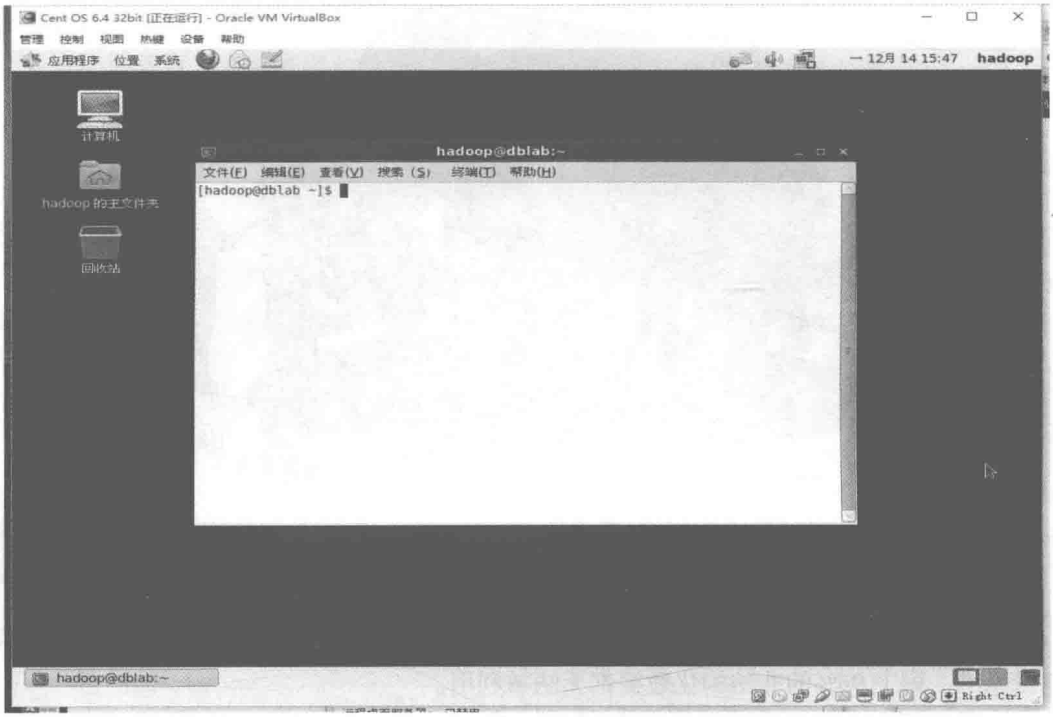

图 1.11　CentOS 下 Desktop 使用界面

执行结果如图 1.12 所示。

```
[root@localhost ~]# ls /
bin   dev  home  lost+found  mnt  proc  sbin     srv  tmp  var
boot  etc  lib   media       opt  root  selinux  sys  usr
```

图 1.12　执行 ls 命令截图

由图 1.12 可知，Linux 系统目录结构如图 1.13 所示。

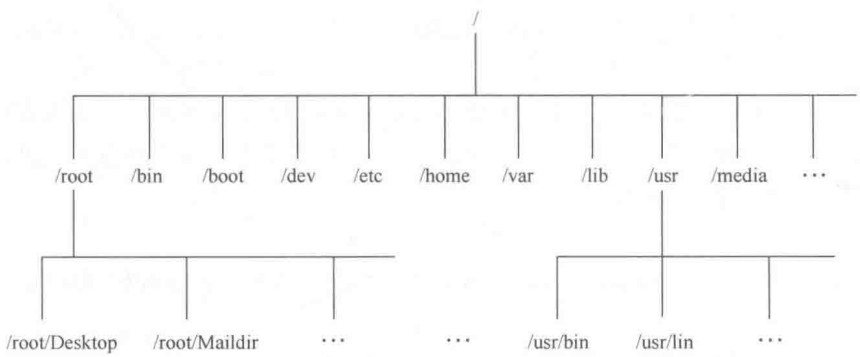

图 1.13　Linux 系统目录结构

/root：该目录为系统管理员，也称作超级权限者的用户主目录。
/sbin：这里存放的是系统管理员使用的系统管理程序。
/bin：bin 是 Binary 的缩写，这个目录存放着最常用的命令。