HITE 7.0 软件开发与应用工程师

HDFS+MapReduce
分布式存储与计算实战

武汉厚溥数字科技有限公司　编著

清華大學出版社
北　京

内 容 简 介

本书按照高等院校计算机专业课程基本要求，注重理论和实践相结合，采用先实践再总结的方式，突出计算机课程的实践性特点。本书共包括 9 个单元：大数据概述，大数据必备 Linux 知识，Hadoop 伪分布式安装及其部署，HDFS 原理详解，MapReduce 计算框架详解，搭建 Hadoop 完全分布式环境，资源调度框架(YARN)与运用，Hive 初识，项目实战。

本书内容安排合理，结构清晰，通俗易懂，实例丰富，可作为各类高等院校、培训机构的教材，也可供大数据程序开发人员学习和参考。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn。

图书在版编目(CIP)数据

HDFS+MapReduce 分布式存储与计算实战 / 武汉厚溥数字科技有限公司编著. —北京：清华大学出版社，2023.2

(HITE 7.0 软件开发与应用工程师)

ISBN 978-7-302-62007-5

I. ①H… II. ①武… III. ①存储技术 IV. ①TP333

中国版本图书馆 CIP 数据核字(2022)第 187065 号

责任编辑：刘金喜
封面设计：王 晨
版式设计：孔祥峰
责任校对：成凤进
责任印制：沈 露

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 **邮 编：**100084
社 总 机：010-83470000 **邮 购：**010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市龙大印装有限公司
经 销：全国新华书店
开 本：185mm×260mm **印 张：**14 **彩 插：**2 **字 数：**288 千字
版 次：2023 年 3 月第 1 版 **印 次：**2023 年 3 月第 1 次印刷
定 价：69.00 元

产品编号：099255-01

前　言

Hadoop 是一个由 Apache 基金会开发的分布式系统基础架构。利用 Hadoop，用户可以在不了解分布式底层细节的情况下，开发分布式程序；充分利用集群的威力进行高速运算和存储。Hadoop 实现了分布式文件系统，其中一个组件是 HDFS。HDFS 有高容错性的特点，可部署在低廉的硬件上，提供高吞吐量来访问应用程序的数据，适合用在有着超大数据集的应用程序中。Hadoop 框架最核心的设计就是 HDFS 和 MapReduce，HDFS 为海量的数据提供了存储，而 MapReduce 则为海量的数据提供了计算。

本书是“工信部国家级计算机人才评定体系”中的一本专业教材。“工信部国家级计算机人才评定体系”由武汉厚溥数字科技有限公司开发，是以培养符合企业需求的软件工程师为目标的 IT 职业教育体系。在开发该体系之前，我们对 IT 行业的岗位序列做了充分的调研，包括研究从业人员在技术方向、项目经验和职业素养等方面的需求，通过对所面向学生的特点、行业需求的现状以及项目实施等方面的详细分析，结合我公司对软件人才培养模式的认知，按照软件专业总体定位要求，进行软件专业产品课程体系设计。该体系集应用软件知识和多领域的实践项目于一体，着重培养学生的熟练度、规范性、集成和项目实施能力，从而达到预定的培养目标。

本书共包括 9 个单元：大数据概述，大数据必备 Linux 知识，Hadoop 伪分布式安装及其部署，HDFS 原理详解，MapReduce 计算框架详解，搭建 Hadoop 完全分布式环境，资源调度框架(YARN)与运用，Hive 初识，项目实战。

我们对本书的编写体系做了精心的设计，按照“理论学习—知识总结—上机操作—课后习题”这一思路进行编排。“理论学习”部分描述通过案例要达到的学习目标与涉及的相关知识点，使学习目标更加明确；“知识总结”部分概括案例所涉及的知识点，使知识点得以完整系统地呈现；“上机操作”部分对案例进行了详尽分析，通过完整的步骤帮助读者快速掌握该案例的操作方法；“课后习题”部分帮助读者理解章节的知识点。本书在内容编写方面，力求细致全面；在文字叙述方面，注意言简意赅、重点突出；在案例选取方面，强调案例的针对性和实用性。

本书凝聚了编者多年来的教学经验和成果，可作为各类高等院校、培训机构的教材，也可供广大程序设计人员学习和参考。

本书由武汉厚溥数字科技有限公司编著，由王伟、寇立红、熊勇、余剑、王鹏、胡富文、杜同海等多名企业实战项目经理编写。本书编者长期从事项目开发和教学实施，并且对当前高校的教学情况非常熟悉，在编写过程中充分考虑不同学生的特点和需求，加强了项目实战方面的教学。在本书的编写过程中，得到了武汉厚溥数字科技有限公司各级领导的大力支持，在此对他们表示衷心的感谢。

参与本书编写的人员还有：宣化科技职业学院张兵，湖北国土资源职业学院管胜波，铜川职业技术学院张华、田宇，陕西国际商贸学院孙玮，江西机电职业技术学院吴赵盼，湖北科技职业学院田野、王智超，闽西职业技术学院陈坤定等。

限于编写时间和编者的水平，书中难免存在不足之处，希望广大读者批评指正。

服务邮箱：476371891@qq.com。

编　者

2022 年 11 月

目　录

单元一　大数据概述 ……1
1.1　大数据基本概念……2
1.1.1　大数据与生活 ……2
1.1.2　大数据的特征 ……4
1.1.3　大数据的发展史 ……4
1.1.4　云计算、大数据和人工智能…5
1.1.5　大数据平台——Hadoop……9
1.2　学习 Hadoop 的环境准备工作……12
单元小结……24
单元自测……24

单元二　大数据必备 Linux 知识……27
2.1　Linux 目录结构 ……28
2.2　Linux 运行级别 ……29
2.3　Linux 常用命令 ……30
2.3.1　帮助命令 ……30
2.3.2　显示当前目录绝对路径命令 ……32
2.3.3　列出目录命令 ……32
2.3.4　切换目录命令 ……33
2.3.5　创建目录命令 ……33
2.3.6　删除文件或目录命令 ……34
2.3.7　创建空文件 ……34
2.3.8　复制命令 ……35
2.3.9　移动/重命名命令 ……36
2.3.10　查看内容命令 ……36
2.3.11　分屏显示文件内容命令 ……37
2.3.12　输出重定向命令 ……37
2.3.13　输出内容到控制台命令 ……38
2.3.14　软链接命令 ……38
2.3.15　查看历史执行命令……39
2.3.16　显示当前时间命令……40
2.3.17　查看日历命令……40
2.3.18　tar 文件解压命令 ……41
2.3.19　在指定的目录下查找命令…41
2.3.20　全局查找命令……42
2.3.21　在文本中查找命令……42
2.4　Linux 用户管理 ……43
2.4.1　添加用户命令……43
2.4.2　创建用户组命令……44
2.4.3　添加用户并指定所属组命令 ……44
2.4.4　修改用户所属组命令……44
2.4.5　删除用户命令……45
2.4.6　删除用户组命令……45
2.4.7　设置用户密码命令……45
2.4.8　查看用户信息命令……45
2.4.9　切换用户命令……46
2.4.10　查看登录用户信息命令……46
2.4.11　用户、用户组的相关文件…47
2.5　Linux 组和权限管理 ……48
2.5.1　Linux 中的权限 ……48
2.5.2　修改文件/目录的所有者命令 ……49
2.5.3　修改文件/目录的所属组命令 ……50
2.5.4　修改文件所有者和所属组命令 ……51
2.5.5　修改权限命令……52

2.6 Linux 磁盘管理 …… 53
2.6.1 查看系统整体磁盘情况命令 …… 53
2.6.2 查看指定目录的磁盘占用情况命令 …… 54
2.7 Linux 网络 …… 54
2.7.1 修改 IP 地址 …… 55
2.7.2 修改主机名 …… 55
2.8 Linux 进程管理 …… 56
2.8.1 显示系统执行的进程命令 …… 56
2.8.2 显示子父进程的关系命令 …… 57
2.8.3 终止进程命令 …… 57
2.9 Linux 服务管理 …… 57
2.10 Linux RPM 和 YUM …… 59
2.10.1 RPM 相关命令 …… 59
2.10.2 YUM 相关命令 …… 60
2.11 Linux vim 编辑器 …… 61
2.11.1 vim 的普通模式 …… 61
2.11.2 vim 的编辑模式 …… 62
2.11.3 vim 的命令模式 …… 62
单元小结 …… 63
单元自测 …… 63

单元三 Hadoop 伪分布式安装及其部署 …… 67
3.1 前期知识准备 …… 68
3.2 Linux 环境配置 …… 70
3.2.1 修改主机名和计算机名 …… 70
3.2.2 配置静态 IP 地址 …… 71
3.2.3 配置 SSH 无密码连接 …… 74
3.2.4 远程连接配置 …… 77
3.3 JDK 配置 …… 78
3.3.1 卸载 Open JDK …… 78
3.3.2 下载 Oracle JDK …… 79
3.3.3 安装 Oracle JDK(root 用户权限执行) …… 80
3.4 安装与部署 Hadoop …… 81
3.4.1 安装 CDH …… 82
3.4.2 修改 hadoop-env.sh …… 83
3.4.3 修改 core-site.xml …… 83
3.4.4 修改 hdfs-site.xml …… 83
3.4.5 修改 slaves 文件 …… 84
3.4.6 追加 HADOOP_HOME 到环境变量中 …… 84
3.4.7 格式化 HDFS …… 85
3.4.8 启动 Hadoop 并验证安装 …… 85
3.4.9 安装验证 …… 86
单元小结 …… 87
单元自测 …… 87

单元四 HDFS 原理详解 …… 89
4.1 HDFS 概述以及设计目标 …… 90
4.1.1 HDFS 概述 …… 90
4.1.2 HDFS 设计理念 …… 91
4.1.3 HDFS 目标 …… 92
4.1.4 HDFS 缺点 …… 93
4.2 HDFS 架构 …… 93
4.3 HDFS 副本机制 …… 97
4.3.1 数据复制 …… 97
4.3.2 副本存放机制 …… 98
4.4 HDFS 读取文件和写入文件 …… 99
4.4.1 通过 HDFS 读取文件 …… 99
4.4.2 通过 HDFS 写入文件 …… 100
4.5 HDFS 的基本文件操作 …… 105
4.5.1 -help [cmd] …… 105
4.5.2 -mkdir <path> …… 106
4.5.3 -ls(r) <path> …… 106
4.5.4 -put <localsrc> <dst> …… 106
4.5.5 -du(s) <path> …… 108
4.5.6 -count[-q] <path> …… 109
4.5.7 -mv <src> <dst> …… 109
4.5.8 -cp <src> <dst> …… 109
4.5.9 -rm(r) …… 110
4.5.10 -moveFromLocal<localsrc> <dest>/-moveToLocal<dest> <localscr> …… 110

4.5.11　-get [-ignorecrc] <src> <localdst> ····· 110
4.5.12　-cat <src> ····· 111
单元小结 ····· 111
单元自测 ····· 112

单元五　MapReduce 计算框架详解 ··· 115
5.1　认识 MapReduce ····· 116
5.1.1　什么是 MapReduce ····· 116
5.1.2　MapReduce 的特点 ····· 116
5.2　MapReduce 编程思想 ····· 117
5.3　MapReduce 执行流程 ····· 119
5.3.1　MapReduce 流程分解 ····· 119
5.3.2　MapReduce 详解 ····· 120
5.4　Java 版中 wordcount 功能的实现 ····· 121
5.5　Combiner 应用程序开发 ····· 128
5.5.1　MapReduce 中 Combiner 的作用 ····· 128
5.5.2　Combiner 的原理 ····· 128
5.5.3　代码实现 ····· 130
5.6　Partitioner 应用程序开发 ····· 131
5.6.1　MapReduce 中 Partitioner 的作用 ····· 131
5.6.2　代码实现 ····· 131
单元小结 ····· 134
单元自测 ····· 135

单元六　搭建 Hadoop 完全分布式环境 ····· 137
6.1　Hadoop 的集群规划 ····· 138
6.2　前置安装 ····· 141
6.3　安装 JDK ····· 142
6.4　Hadoop 集群的部署 ····· 143
6.5　作业提交到 Hadoop 集群上运行 ····· 145
单元小结 ····· 146
单元自测 ····· 146

单元七　资源调度框架(YARN)与运用 ····· 149
7.1　YARN 产生的背景 ····· 150
7.2　YARN 架构 ····· 152
7.3　YARN 的执行流程 ····· 154
7.4　YARN 的环境搭建 ····· 155
7.5　提交作业到 YARN 上执行 ····· 157
单元小结 ····· 158
单元自测 ····· 158

单元八　Hive 初识 ····· 161
8.1　认识 Hive ····· 162
8.2　Hive 的安装和配置 ····· 163
8.2.1　安装 MySQL ····· 163
8.2.2　安装 Hive ····· 168
8.2.3　验证安装 ····· 170
8.3　Hive 操作快速入门 ····· 171
单元小结 ····· 173
单元自测 ····· 173

单元九　电商用户行为分析项目实战 ··· 175
9.1　背景知识 ····· 176
9.2　项目基本介绍 ····· 179
9.2.1　用户日志分析 ····· 179
9.2.2　常用的电商术语 ····· 180
9.2.3　用户行为日志的意义 ····· 181
9.3　项目需求分析 ····· 182
9.3.1　需求分析 ····· 182
9.3.2　数据处理流程 ····· 183
9.4　实现项目功能 ····· 184
9.4.1　各省份浏览量统计功能实现 ····· 184
9.4.2　页面浏览统计功能实现 ····· 188
9.4.3　ETL 的介绍和实现 ····· 192
9.4.4　功能升级 ····· 195
9.4.5　打包上传服务器运行 ····· 202
9.5　项目功能优化 ····· 206

单元

一

大数据概述

课程目标

- 了解大数据的基本概念
- 理解 Hadoop 及其体系
- 掌握 Hadoop 的环境配置方法

简介

随着科技的进步和社会的发展，尤其是以计算机为代表的信息技术的飞速发展，各种信息呈爆炸式增长，数据渗透到各行各业。很多企业也越来越重视对数据的收集与分析，三大运营商(中国电信、中国移动、中国联通)、阿里云、景安网络等高新技术企业纷纷为客户建立起专业的大数据服务平台，对客户需要的数据进行收集与分析，提供基于大数据的运营指导。

大数据正在成为引领性的先进技术，它是信息技术领域的制高点。未来将会是大数据应用蓬勃发展的时代，海量的数据将会成为企业制定战略决策的重要依据。大数据时代究竟会给企业信息化带来怎样的影响？

1.1 大数据基本概念

未来的时代将不再是 IT 时代，而是 DT 时代。这里所说的 DT(data technology，数据科技)，是指以服务大众、激发生产力为主的技术。当前，万物互联下的数据流进入了人们的每个生活场景，尤其是当“数据资产是企业核心资产”的概念深入人心之后，企业对于数据管理便有了更清晰的界定，并逐渐将大数据管理作为企业的核心竞争力。

其实，“大数据”概念早在 20 世纪 80 年代就已经有人提出，不过彼时“大数据”只是形容数据集很大的一个词。直到计算机硬件升级和“云计算”技术出现之后，“大量数据和应用算法”才逐渐展现出其隐藏的秘密。从本质来说，大数据是指“无处不在的数据”，而随着人们对数据重要性认识的深入，人们对数据的挖掘和开发也在快速地发展。

1.1.1 大数据与生活

案例 1：医疗大数据，看病更高效

医疗行业是让大数据分析最先发挥作用的传统行业之一。医疗行业拥有大量的病例、病理报告、治愈方案、药物报告等，如果这些数据可以被整理和应用，将会极大地帮助医生和病人。我们面对的种类众多的病菌、病毒、肿瘤细胞，都处于不断进化的过程中，在发现或诊断疾病时，疾病的确诊和治疗方案的确定是最困难的。

在未来，借助于大数据平台，我们可以收集不同的病例和治疗方案，以及病人的基本特征，建立针对疾病特点的数据库。如果未来基因技术发展成熟，可以根据病人的基因序列特点进行分类，建立医疗行业的病人分类数据库。在医生诊断病人时可以参考病人的疾

病特征、化验报告和检测报告，参考疾病数据库，以快速帮助病人确诊，明确定位疾病。在制订治疗方案时，医生可以依据病人的基因特点，调取基因、年龄、人种、身体情况相同的其他病人的有效治疗方案，制订出适合该病人的治疗方案，使更多人及时得到治疗。同时，这些数据也有利于医药行业开发出更加有效的药物和医疗器械。

医疗行业的数据应用一直在进行，但是数据没有被打通，都是孤岛数据，没有办法进行大规模应用。未来需要将这些数据统一收集起来，纳入统一的大数据平台，为人类健康造福。

案例 2：《英雄联盟》，你为什么会越陷越深？

《英雄联盟》是风靡全球的网络游戏，每天夜深，当大多数玩家已经奋战一天，呼呼大睡的时候，数据服务器正紧张地忙碌着。世界各地的运营商会把当日的数据发送到北美的数据中心，随即一个巨大的数据分析引擎开始转动，执行上千个数据分析任务。当日所有的比赛都会被分析，数据分析师若发现某个英雄单位太强或太弱，在接下来的两三周内，会推出新补丁，及时调整所有的平衡性问题，并加入一个新单位。整个游戏被保持在一个快速更新，并且良好平衡的状态。正是靠大数据的魔力，《英雄联盟》才能成为这个时代最受欢迎的游戏之一。

案例 3：电商大数据，精准营销的法宝

电商是最早利用大数据进行精准营销的行业，除了精准营销，电商还可以依据客户消费习惯提前为客户备货，并利用便利店作为货物中转点，在客户下单 15 分钟内将货物送上门，提高客户体验。菜鸟网络宣称的 24 小时完成在中国境内的送货，以及京东宣称未来将在 15 分钟内完成送货上门都是基于客户消费习惯的大数据分析和预测。

电商可以利用其交易数据和现金流数据，为其生态圈内的商户提供基于现金流的小额贷款，电商业也可以将此数据提供给银行，同银行合作为中小企业提供信贷支持。由于电商的数据较为集中，数据量足够大，数据种类较多，因此未来电商数据应用将会有更多的想象空间，包括预测流行趋势、消费趋势、地域消费特点、客户消费习惯、各种消费行为的相关度、消费热点、影响消费的重要因素等。依托大数据分析，电商的消费报告将有利于品牌公司的产品设计，生产企业的库存管理和生产计划制订，物流企业的资源配置，生产资料提供方的产能安排等，有利于精细化和社会化大生产。

随着 Web 2.0 和移动互联网的兴起，大数据时代已经到来，因为我们每时每刻创造数据的速度在不断加快，各行各业数据的规模也越来越大，从原来的吉字节(GB)发展为太字节(TB)，再到拍字节(PB)，以后一定会往泽字节(ZB)发展，而且数据的类型也越来越多，有日志、视频、音频、地理位置信息等。随着云计算的普及，数据也越来越集中到云端服务器之上，在这样的前提下，我们就可以对数据集中进行挖掘、分析。典型场景有科学数

据、物联网数据、交通数据、社交数据、零售数据、金融数据等，这些场景下无时无刻不存在着大数据，大数据已经和人们的生活息息相关。

1.1.2 大数据的特征

在日常生活中，我们经常会听到“大数据”这个词。那么，充满“神秘感”的大数据到底是什么？我们先来看看麦肯锡全球研究院给出的定义。

大数据是一种规模大到在获取、存储、管理、分析方面大大超出传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。可以看出，用以区分大数据的核心是上述四大特征。由于这四大特征的英文单词都以 V 开头，我们称之为 4V 特征。

(1) Volume

Volume 指数据量大，包括采集、存储和计算的量都非常大。大数据的起始计量单位至少是 PB(1PB＝1024TB，1TB＝1024GB)，甚至是 EB(1EB=1024PB)及 ZB(1ZB＝1024EB)。

(2) Variety

Variety 指种类和来源多样化，包括结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等，多类型的数据对数据的处理能力提出了更高的要求。

(3) Value

Value 指数据价值密度相对较低，或者说是浪里淘沙却又弥足珍贵。随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低。如何结合业务逻辑并通过强大的机器算法挖掘数据价值，是大数据时代最需要解决的问题。

(4) Velocity

Velocity 指数据增长速度快，处理速度也快，时效性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法要求尽可能实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

我们可以把具有以上特征的数据统称为“大数据”。例如，每年的“双十一”购物节产生的巨量交易数据就属于大数据的范畴。

1.1.3 大数据的发展史

大数据在当今是炙手可热的技术，但是它的发展也不是一蹴而就的。早在 20 世纪 90 年代，美国国家航空航天局的大卫·埃尔斯沃思和迈克尔·考克斯在他们研究数据可视化

时首次使用了“大数据”的概念。1998 年，*Science* 杂志发表了一篇题为《大数据科学的可视化》的文章，大数据作为一个专用名词正式出现在公共期刊上。不过在当时大数据只是作为一个概念或假设，少数学者对其进行了研究和讨论，其意义仅限于数据量的巨大，对数据的收集、处理和存储没有进一步的探索。

今天我们常说的大数据技术，其实起源于 Google 在 2004 年前后发表的 3 篇论文，分别是《分布式文件系统 GFS》《大数据分布式计算框架 MapReduce》《NoSQL 数据库系统 BigTable》。

2003 年，Google 发表“The Google File System”论文。

2004 年，Google 发表“MapReduce”论文。

2006 年，Google 发表“BigTable”论文。受论文启发 Doug Cutting 启动相关项目，同年 Hadoop 从 Nuth 中分离出来，为了便于 MapReduce 开发，Yahoo 研发了 Pig。

2007 年，Hadoop 成为 Apache 顶级项目，专门运营 Hadoop 的商业公司 Cloudera 成立。

2008 年，Facebook 将 Hive 贡献到开源社区。

2010 年，HBase 独立成为 Apache 顶级项目。

2012 年，Yarn 作为独立项目，负责资源调度。Sqoop 成为 Apache 顶级项目(将关系数据库的数据导入到 Hadoop 平台)。

2014 年，Spark 成为 Apache 顶级项目，Strom 成为 Apache 顶级项目，Flink 成为 Apache 顶级项目。

至此，大数据的技术发展已基本成型，已经出现众多被各互联网企业使用的成熟大数据平台技术，包括 Hadoop、Spark、Strom、Flink 等。大数据技术已经进入到百花齐放的时期。

1.1.4　云计算、大数据和人工智能

当今社会最热门的技术词汇有 3 个，分别是云计算、大数据和人工智能。其实这些所谓的前沿技术之间是存在相关性的，弄清楚它们之间的关系，有利于我们理解这些技术，并理清我们的学习方向和路径，以便合理规划职业发展。

1.1.4.1　云计算概述

云计算是一种可以通过网络方便地接入共享资源池，按需获取计算资源(这些资源包括网络、服务器、存储、应用、服务等)的服务模型。共享池中的资源可以通过较少的管理代价和简单业务交互过程而快速部署和发布。

1. 云计算的 5 个特点

(1) 按需提供服务

消费者不需要或很少需要云服务提供商的协助，就可以单方面按需获取云端的计算资源。

(2) 广泛的网络访问

消费者可以随时随地使用任何云终端设备接入网络并使用云端的计算资源。常见的云终端设备有手机、平板电脑、笔记本电脑、PDA 掌上电脑和台式机等。

(3) 资源池化

云端计算资源需要被池化，以便通过多租户形式共享给多个消费者，也只有池化才能根据消费者的需求，动态分配或再分配各种物理的和虚拟的资源。消费者通常不知道自己正在使用的计算源的确切位置，但是在自助申请时允许指定大概的区域范围(比如在哪个国家、哪个省或者哪个数据中心)。

(4) 高可伸缩性

消费者能方便、快捷地按需获取和释放计算资源，也就是说，需要时能快速获取资源从而扩展计算能力，不需要时能迅速释放资源以便降低计算能力，从而减少资源的使用费用。对于消费者来说，云端的计算资源是无限的，可以随时申请并获取任何数量的计算资源。但是即使是投资巨大的工程，也不一定具备超大规模的运算能力。其实一台计算机就可以组建一个最小的云端，云端建设方案务必采用可伸缩性策略，刚开始是采用几台计算机，然后根据用户数量规模来增减计算资源。

(5) 可量化服务

消费者使用云端计算资源是要付费的，付费的计量方法有很多，比如根据某类资源(如存储、CPU、内存、网络带宽等)的使用量和时间长短计费，也可以按照每使用一次来计费。但不管如何计费，对消费者来说，价码要清楚，计量方法要明确，而运营服务提供商需要监视和控制资源的使用情况，并及时输出各种资源的使用报表，做到供/需双方费用结算清楚、明白。

由于“云”的特殊容错措施，可以采用极其廉价的节点来构成云，“云”的自动化集中式管理使大量企业无须负担日益高昂的数据中心管理成本，“云”的通用性使资源的利用率较之传统系统大幅提升，因此用户可以充分享受“云”的低成本优势，经常只要花费几百美元、几天时间就能完成以前需要数万美元、数月时间才能完成的任务。

2. 云计算的服务模式

云计算主要分为三种服务模式(图 1-1)，而且这个三层的分法主要是从用户体验的角度出发的，这三种服务模式之间及其与传统 IT 的区别如图 1-2 所示。

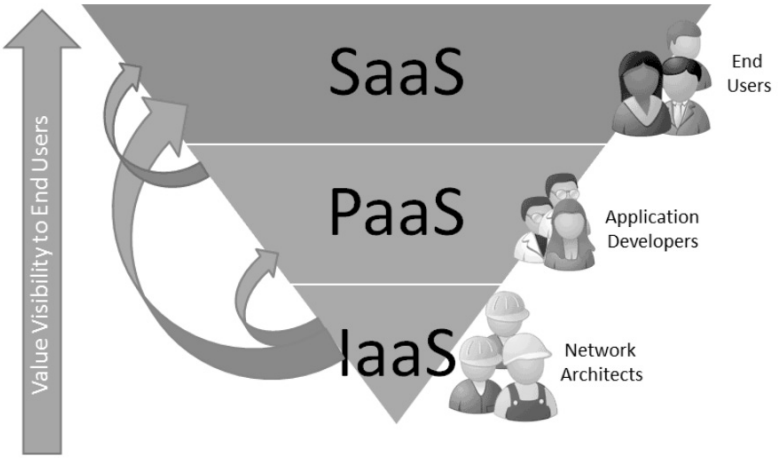

图 1-1

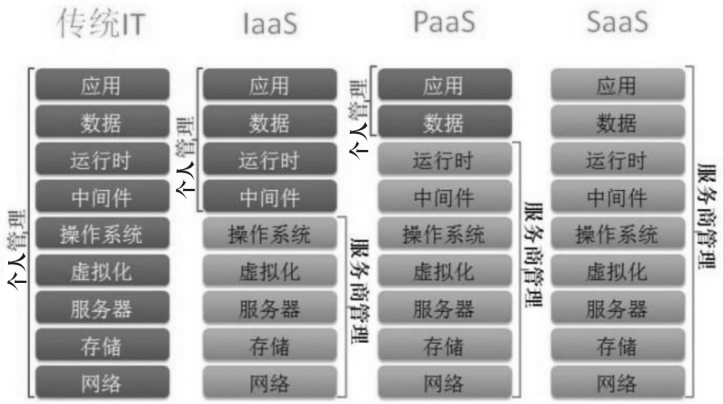

图 1-2

(1) 软件即服务

软件即服务(Software as a Service，SaaS)，该层的作用是将应用作为服务提供给客户。SaaS 的针对性更强，它将某些特定的应用软件封装成服务，而只提供某些专门用途的服务供应用调用。这一层应该是和我们接触最频繁的一层。现在的很多软件系统基础都会使用 SaaS 产品，例如企业中使用非常普遍的企业 IM(即时通信)软件“钉钉”“企业微信”等。

(2) 平台即服务

平台即服务(Platform as a Service，PaaS)，该层的作用是将一个开发平台作为服务提供给用户。PaaS 是对资源进行更进一步的抽象，它提供了用户应用程序的应用环境。PaaS 自身负责资源的动态扩容、容错、灾备，用户的应用程序不需要过多考虑节点间的配合问题，但与此同时，用户的自主权降低，必须使用特定的编程环境并遵照特定的编程模型。例如谷歌的 Google App Engine 服务，它提供的是包括 SDK、文档和测试环境等在内的开发平台，开发人员能够非常方便地开发应用，而且在部署或者运行的时候，用户都无须为

服务器、操作系统、网络和存储等资源的管理操心。一台运行 Google App Engine 的服务器能够支撑成千上万的应用，所以 PaaS 服务是非常高效并且经济的，而 PaaS 的主要用户是开发人员。

(3) 基础设施即服务

基础设施即服务(Infrastructure as a Service，IaaS)，该层的作用是提供虚拟机或者其他资源作为服务提供给用户。IaaS 作为云技术的架构的最底层，利用虚拟化技术将硬件设备等基础资源封装成服务供用户使用，用户相当于在使用裸机，既可以让它运行 Windows 系统，也可以让它运行 Linux 系统；既可以作为 Web 服务器，也可以作为数据库服务器。IaaS 最大的优势在于它运行用户动态申请或释放节点，按使用量和使用时间计费。例如大家现在普遍使用的阿里云、腾讯云等云平台，其中就提供虚拟主机的定制功能，并根据定制的规格提供相关收费的服务，如图 1-3 所示。

图 1-3

接下来通过一个场景来讲解三者的区别。假设一个程序开发员，想开发一个基于 Web 的互联网应用程序，想使用云端服务，如果选择的是 IaaS，就不用自己搭建或者购买服务器，可以在对应的云平台(例如前面提到的阿里云或者腾讯云)上购买虚拟主机，然后自己在虚拟主机中搭建开发环境进行开发、运行、部署等相关工作。如果选择采用 PaaS，既不需要购买服务器，也不需要自己安装相关软件和开发环境，只需要在对应的云服务器应用(例如前面提到的 Google App Engine)上进行程序的开发即可。再进一步，如果采用的是 SaaS，就可以直接购买具有相关业务的程序，不需要自己开发，而且服务提供商会负责程序的升级、维护、增加服务器等，客户只需要专心运营即可。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

1.1.4.2　人工智能概述

其实数据本身并没有用，必须经过一定的处理。例如，用户每天下载的电影、收听的音乐、浏览的网页等，我们都称之为“数据”。数据本身没有什么用处，但数据里面包含一个很重要的东西，叫作信息(Information)。

经过梳理和筛选后的数据才能叫作信息。信息会包含很多规律，我们需要从信息中将规律总结出来，称为知识(Knowledge)。

有了知识，然后通过运用并实践，就产生了智慧(Intelligence)。数据、信息、知识和智慧之间的关系如图 1-4 所示。

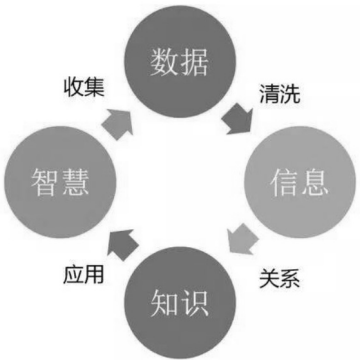

图 1-4

数据的处理分以下几个步骤，完成了才会产生智慧：数据收集→数据传输→数据存储→数据处理和分析→数据检索和挖掘→数据收集。这其实就是人工智能的基础，所谓的人工智能，指的是计算机科学软件的一个分支，它模拟人类的决策和活动。它通过先进的算法和机器学习通过“数据”产生“智慧”，模仿“学习”和“解决问题”等活动。

可以看出，一方面人工智能需要大量的数据作为“思考”和“决策”的基础，另一方面大数据也需要人工智能技术进行数据价值化操作。在大数据价值的两个主要体现当中，数据应用的主要渠道之一就是智能体(人工智能产品)，为智能体提供的数据量越大，智能体运行的效果就会越好，因为智能体通常需要大量的数据进行“训练”和“验证”，从而保障运行的可靠性和稳定性。

1.1.5　大数据平台——Hadoop

前面的章节已经大致对大数据的概念、场景做了介绍，接下来我们进入到对大数据更加具体的认知。先来了解大数据平台——Hadoop。

(1) Hadoop 概况

Hadoop 可以以一种可靠、高效、可扩展的方式存储、管理“大数据”，为管理、挖掘

大数据提供一整套成熟可靠的解决方案，Hadoop 可以称作是一个“大数据管理和分析平台”。Hadoop 和大数据之间的关系如图 1-5 所示。

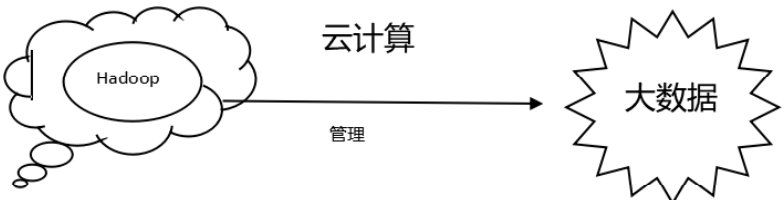

图 1-5

同时 Hadoop 也是 Apache 下的一个开源框架，它允许在整个集群使用简单编程模型计算机的分布式环境存储并处理大数据。它的目的是从单一的服务器扩展到上千台机器，每台机器都可以提供本地计算和存储。

(2) Hadoop 的由来

2003—2004 年，Google 公司公布了部分 GFS 和 MapReduce 思想的细节，受此启发的 Doug Cutting 等用两年的业余时间实现了 DFS 和 MapReduce 机制，使 Nutch 性能得到了很大的提升。然后 Yahoo 招募了 Doug Gutting 及其项目。

2005 年，Hadoop 作为 Lucene 的子项目 Nutch 的一部分正式引入 Apache 基金会。2006 年 2 月被分离出来，成为一套完整独立的软件，起名为 Hadoop。

Hadoop 名字不是一个缩写，而是一个生造出来的词，是 Hadoop 之父 Doug Cutting 以儿子的毛绒玩具象命名的，如图 1-6 所示。

图 1-6

(3) Hadoop 的特点

① 高可靠性

Hadoop 按位存储和处理数据的能力值得人们信赖。

② 高扩展性

Hadoop 是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

③ 高效性

Hadoop 能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

④ 高容错性

Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新做分配。

⑤ 低成本

Hadoop 是开源的，使用的是廉价的硬件设备，项目的软件成本因此会大大降低。

(4) Hadoop 及体系介绍

随着 Hadoop 的不断发展，围绕 Hadoop 的许多软件蓬勃出现，构成了一个生机勃勃的生态圈。

图 1-7 用一只大象生动地描述了 Hadoop 生态的组成部分。当然，这么多组件中核心的组件只有 HDFS 和 MapReduce 两个，这两个组件是整个 Hadoop 平台的基础，也是 Hadoop 中最早出现的组件。在 Hadoop 慢慢演化的过程中不断产生了各种各样提供不同服务的组件。

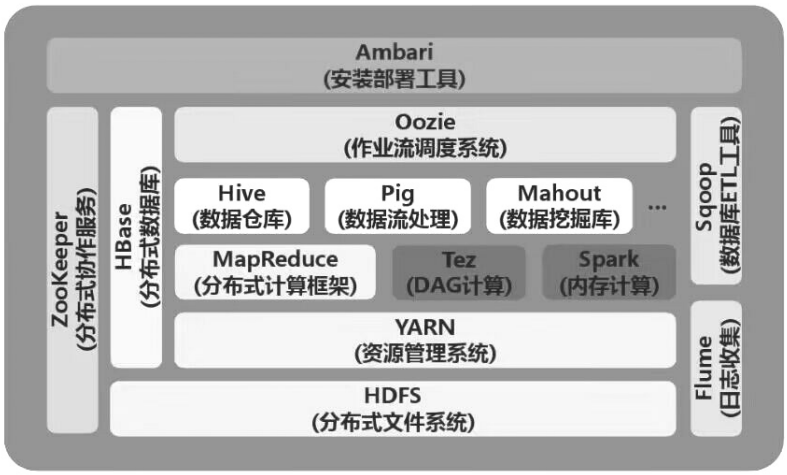

图 1-7

① HDFS

目前大量采用的分布式文件系统，是整个大数据应用场景的基础通用文件存储组。

② MapReduce

分布式计算的基本计算框架。

③ YARN

分布式资源调度，可以接收计算的任务并把它分配到集群各节点处理，相当于大数据操作系统，通用性好，生态支持好。

④ Spark

Spark 提供了一个更快、更通用的数据处理平台。和 Hadoop 相比，Spark 可以让程序在内存中运行的速度提升 100 倍，或者在磁盘上运行的速度提升 10 倍。

⑤ HBase

一种 NoSQL 列簇数据库，支持数十亿行、数百万列大型数据存储和访问，尤其是写数据的性能非常好，数据读取实时性较好，提供一套 API(应用程序接口)，不支持 SQL 操作，数据存储采用 HDFS(Hadoop 分布式文件系统)。

⑥ Hive

通过 HQL(类似 SQL)来统计、分析、生成查询结果，通过解析 HQL 生成可以分布式执行的任务，典型的应用场景是与 HBase 集成。

⑦ Zookeeper

用于协调分布式系统上的各种服务。例如，确认消息是否准确到达，防止单点失效，处理负载均衡等。

⑧ Flume

分布式的海量日志采集、聚合和传输的系统，主要作用是收集和传输数据，也支持非常多的输入输出数据源。

⑨ Sqoop

主要用于在 Hadoop(Hive)与传统的数据库(MySQL、PostgreSQL 等)间进行数据的传递，可以将一个关系数据库(如 MySQL、Oracle、PostgreSQL 等)中的数据导入 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入关系数据库中。

1.2 学习 Hadoop 的环境准备工作

首选，由于 Hadoop 本身是一个分布式软件框架，而在学习的过程中使用真实的分布式环境不太现实，所以采用虚拟技术来模拟分布式环境(Oracle VM VirtualBox 或者 Vmware WorkStation)；其次，由于 Hadoop 的推荐运行环境和大部分的生成环境都是基于 Linux 的，所以学习 Hadoop 需要有 Linux 操作系统(CentOS/Ubuntu)和 Linux 操作的基础(了解基本操作命令)；再者，Hadoop 是基于 Java 环境的，所以需要在 Linux 中配置 Java 环境。

(1) 下载 Oracle VM VirtualBox

Oracle VM VirtualBox 虚拟机(简称 VM 虚拟机)软件是一款桌面计算机虚拟软件，让用户能够在单一主机上同时运行多个不同的操作系统。每个虚拟操作系统的磁盘分区、数据配置都是独立的，不用担心会影响到自己计算机中原本的数据。而且 VM 还支持实时快照、

虚拟网络、文件拖曳传输以及网络安装等方便实用的功能。此外，还可以把多台虚拟机构建成一个专用局域网，使用起来很方便。

① 打开网页 https://www.virtualbox.org/wiki/Downloads，如图 1-8 所示。

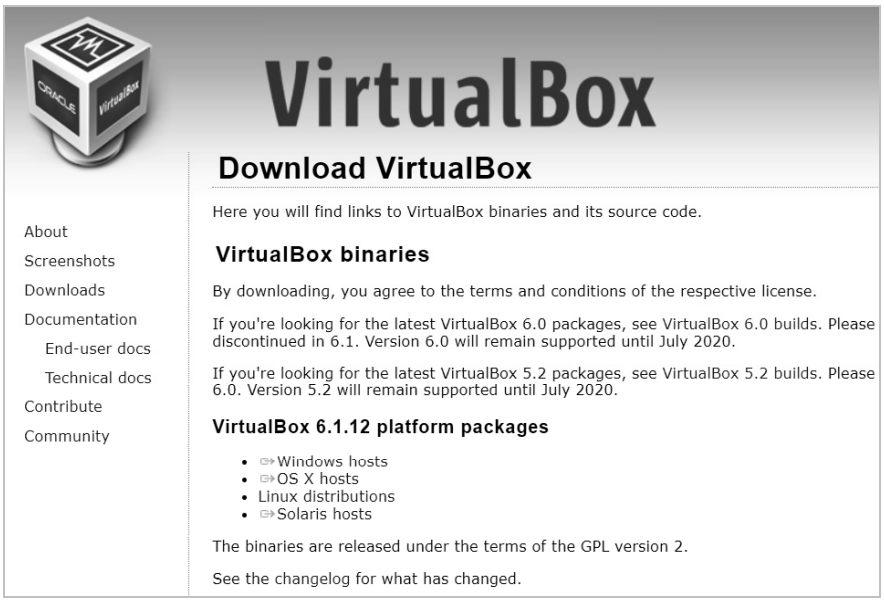

图 1-8

② 选择 Windows hosts 选项，单击“下载”按钮，下载 VirtualBox。

③ 按照提示，不断单击“下一步”按钮进行安装。

(2) 下载 CentOS 7.x

CentOS 是 Community Enterprise Operating System 的缩写，也叫作社区企业操作系统。它是企业 Linux 发行版领头羊 Red Hat Enterprise Linux(以下称之为 RHEL)的再编译版本(也就是所谓的再发行版本)，而且在 RHEL 的基础上修正了不少已知的缺陷(Bug)，相对于其他 Linux 发行版，其稳定性值得信赖。

CentOS 是免费的，用户可以像使用 RHEL 一样使用它构筑企业级的 Linux 系统环境，而不需要向 Red Hat 付任何的费用。CentOS 的技术主要通过社区的官方邮件列表、论坛和聊天室来进行更新维护。

由于 CentOS 是一个免费的且开源的操作系统，所以有很多知名的 IT 公司或学校都会对其进行修改或重新编译而生成自己使用的系统，如阿里云、腾讯云、网易等都有自己的编译系统。由于目前 CentOS 8 版本已经停止更新，为了能够在后面的项目部署中使用一些网络的软件仓库，所以本次项目使用了一个较新的 CentOS 版本，即 CentOS 7。

目前市面上有很多的 CentOS 系统可供下载，比如 CentOS 官方网站、阿里云、腾讯云、清华大学开源软件镜像站等。为了能够更好地体验 Linux 系统的魅力，我们在这里选择使用国内的阿里云作为此次项目的服务器系统。

打开阿里云官方网站，在“开发者”中找到“镜像站”，如图 1-9 所示。

图 1-9

单击“镜像站”，在打开的镜像站中，选择 OS 镜像，打开“下载 OS 镜像”对话框，选择下载的发行版本，然后单击“下载”按钮即可，如图 1-10 所示。

图 1-10

除单击“下载”按钮直接下载系统之外，还可以选择复制下载地址的内容，然后用其他下载工具下载。

接下来介绍在 Oracle VM 中安装 CentOS 7.x[h2]的操作过程。

双击打开 Oracle VM 软件，单击“新建”按钮，在打开的对话框中输入名称“CentOS 7.0”，如图 1-11 所示。

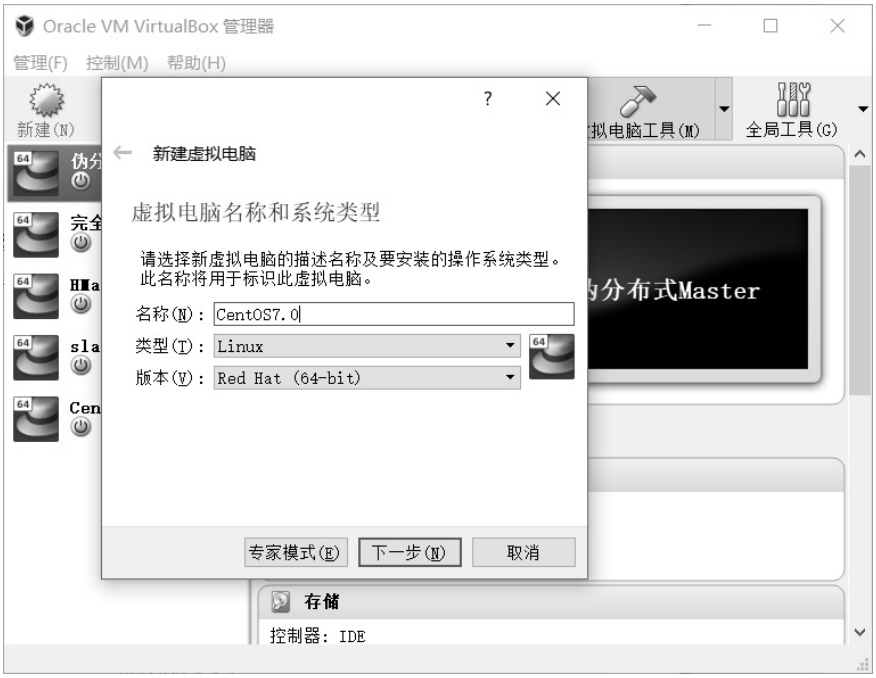

图 1-11

单击“下一步”按钮，设置内存大小，如图 1-12 所示。调整内存大小至少为 2048MB。

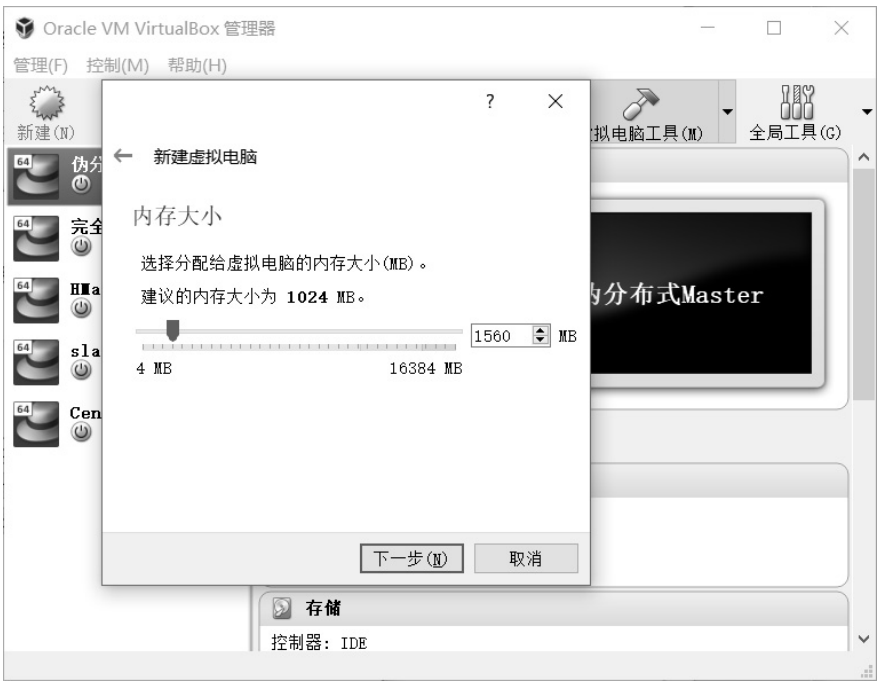

图 1-12

单击“下一步”按钮，设置硬盘类型，选择“现在创建虚拟硬盘”，如图 1-13 所示。

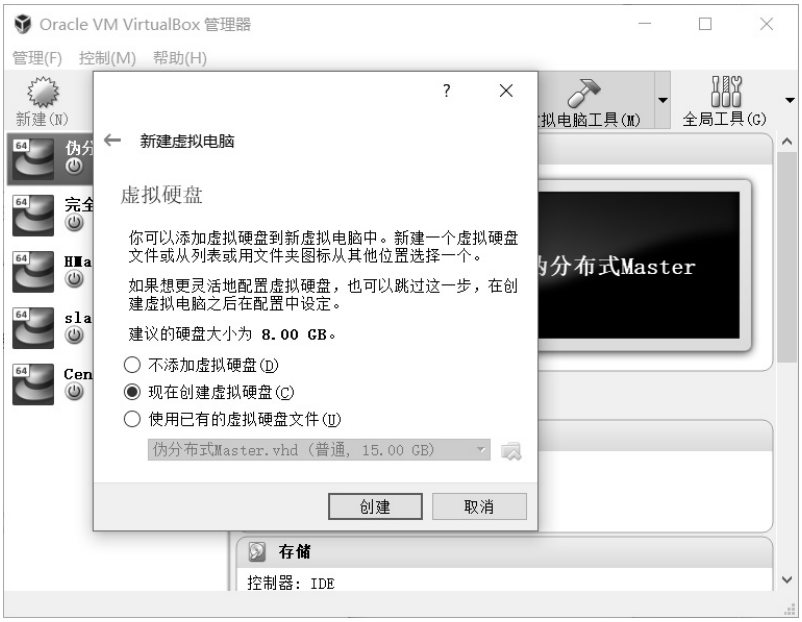

图 1-13

单击“创建”按钮进入“创建虚拟硬盘”界面，选择“VHD(虚拟硬盘)”，如图 1-14 所示。

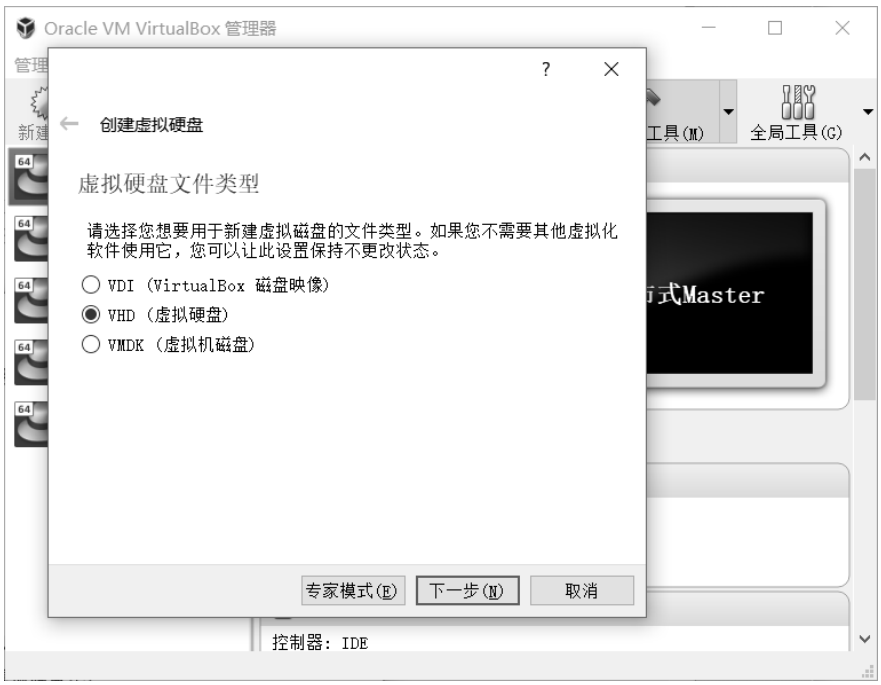

图 1-14

单击“下一步”按钮，选择“固定大小”，以固定容量的方式创建虚拟硬盘，如图 1-15 所示。

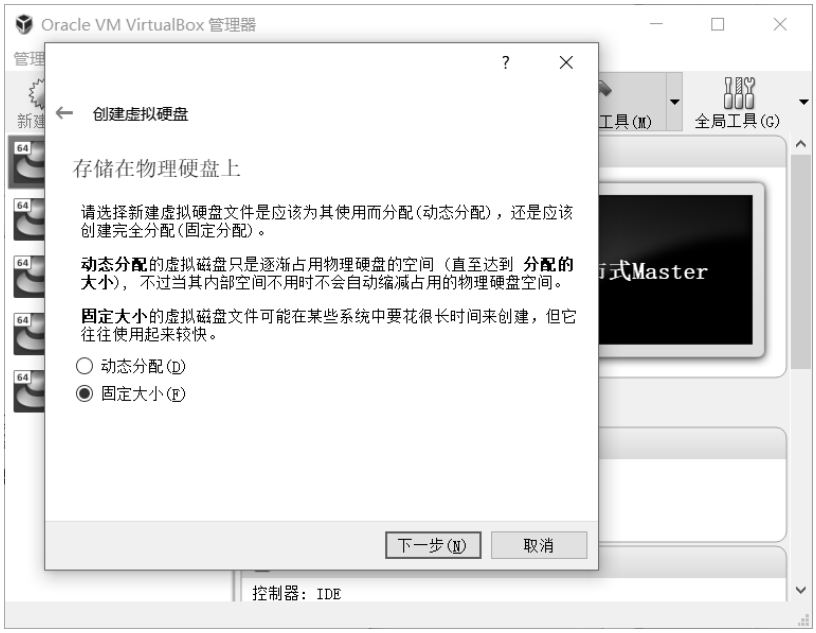

图 1-15

单击“下一步”按钮，选择虚拟硬盘文件的存放路径，如图 1-16 所示。因为虚拟机文件较大，因此建议不要存在系统盘。

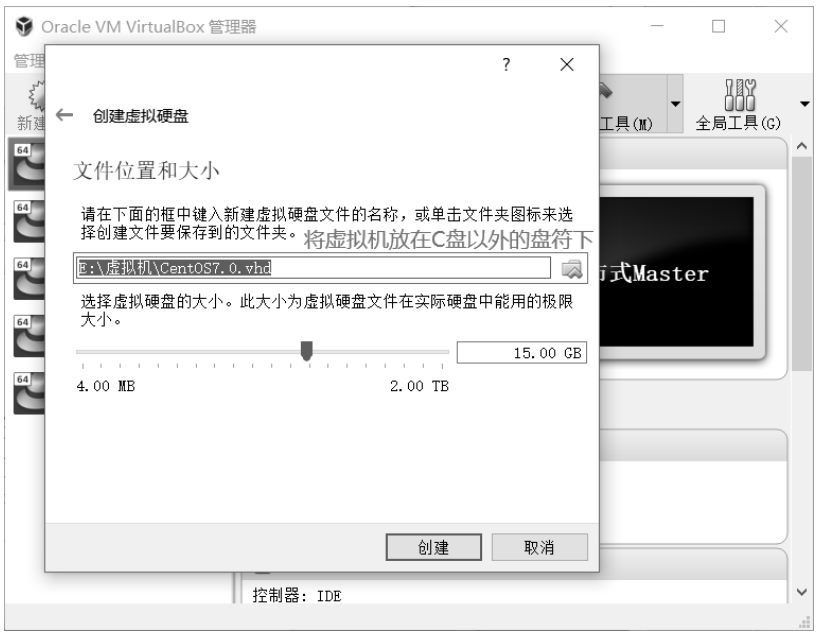

图 1-16

同时设置虚拟硬盘的大小。硬盘的大小建议设置为 15GB。单击“创建”按钮，开始创建虚拟硬盘，如图 1-17 所示。

图 1-17

创建成功后，会在 Oracle VM 界面左侧硬盘列表中看到创建好的虚拟硬盘。此时选择此虚拟硬盘，单击“启动”按钮，启动虚拟硬盘，如图 1-18 所示。

图 1-18

此时 Oracle VM 的右侧启动窗口会打开“选择启动盘”界面。选择下载好的 CentOS 7.x 镜像文件，如图 1-19 所示。

图 1-19

单击“启动”按钮，进入 CentOS 系统图形化安装界面，首先在语言选择栏选择“简体中文”，如图 1-20 所示。

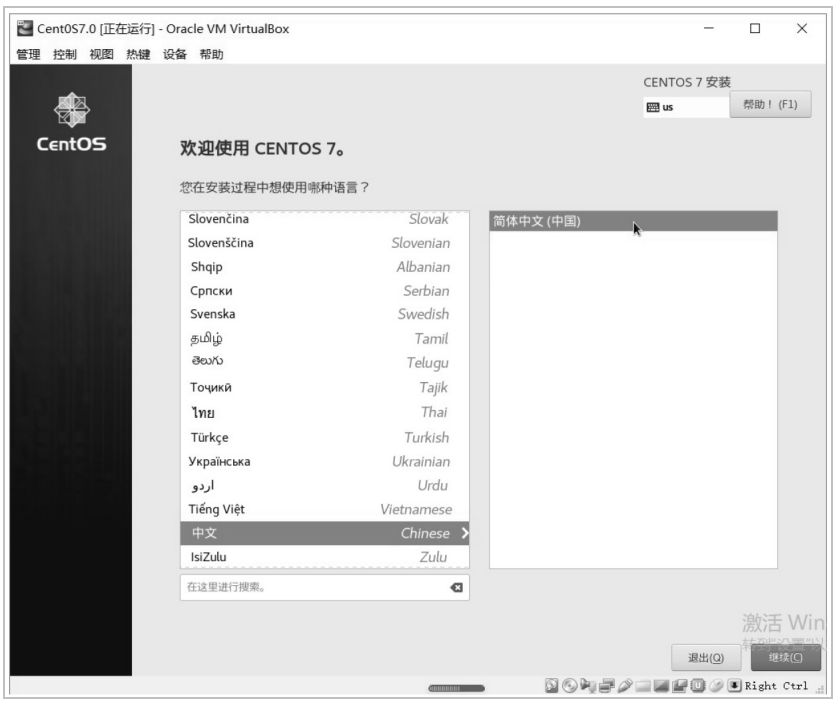

图 1-20

单击“继续”按钮，进入“安装信息摘要”界面。在该界面的系统栏中，单击“安装位置”，选择 CentOS 在硬盘中的安装位置，如图 1-21 所示。

图 1-21

选择好安装位置后，单击“开始安装”按钮，开始安装 CentOS。在安装 CentOS 系统的同时，可以配置 root(管理员)用户的密码，如图 1-22 所示。

图 1-22

此处为了方便，设置 root 账户的密码为“root12”，如图 1-23 所示。

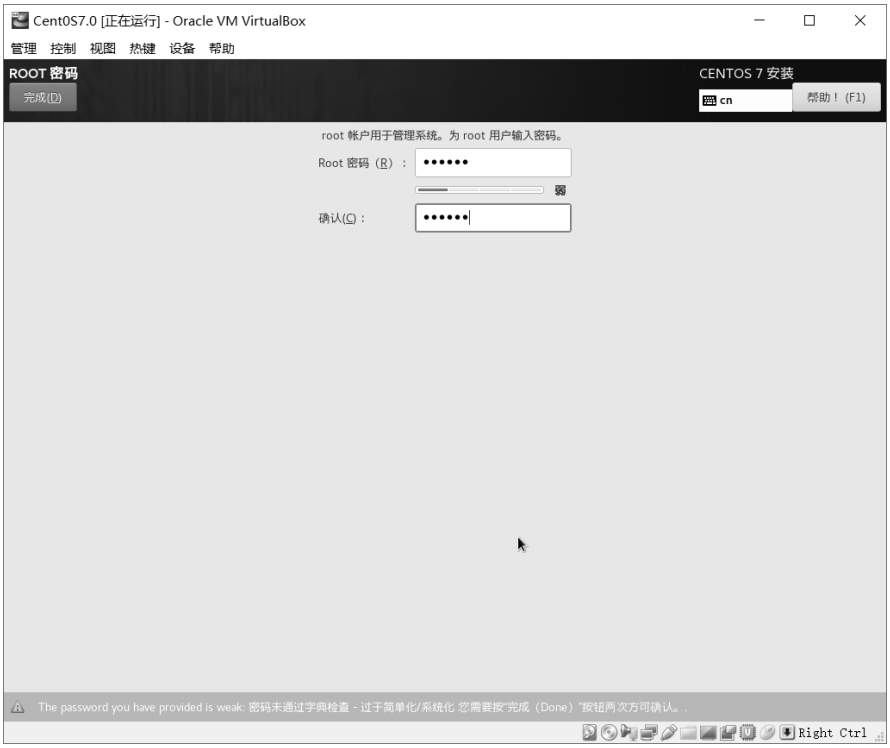

图 1-23

等待进度条到 100%，表示 CentOS 系统已经安装完成，如图 1-24 所示。单击“完成配置”，完成 CentOS 系统的安装操作，如图 1-25 所示。

图 1-24

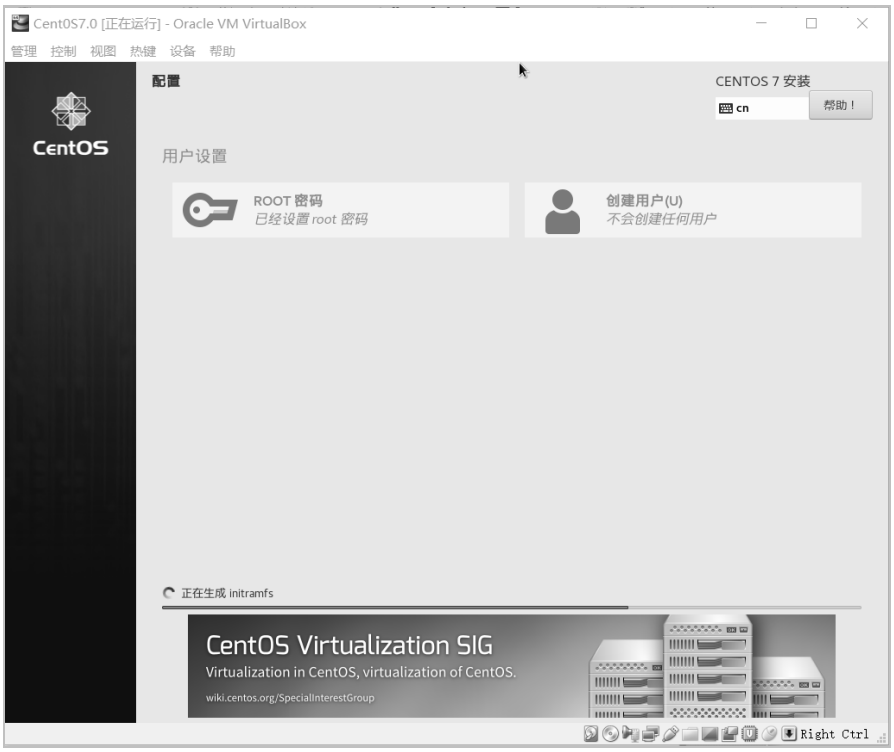

图 1-25

单击“完成配置”后，系统进入初始化阶段。完成初始化后，单击“重启”按钮，系统会自动重新启动，如图 1-26 所示。

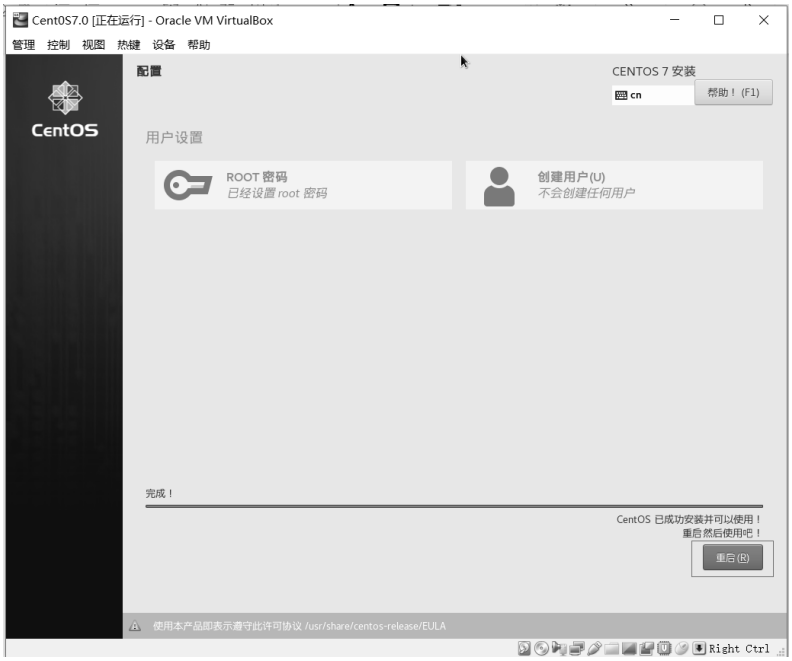

图 1-26

重启完成后，进入 CentOS 系统，选择第一项，如图 1-27 所示。

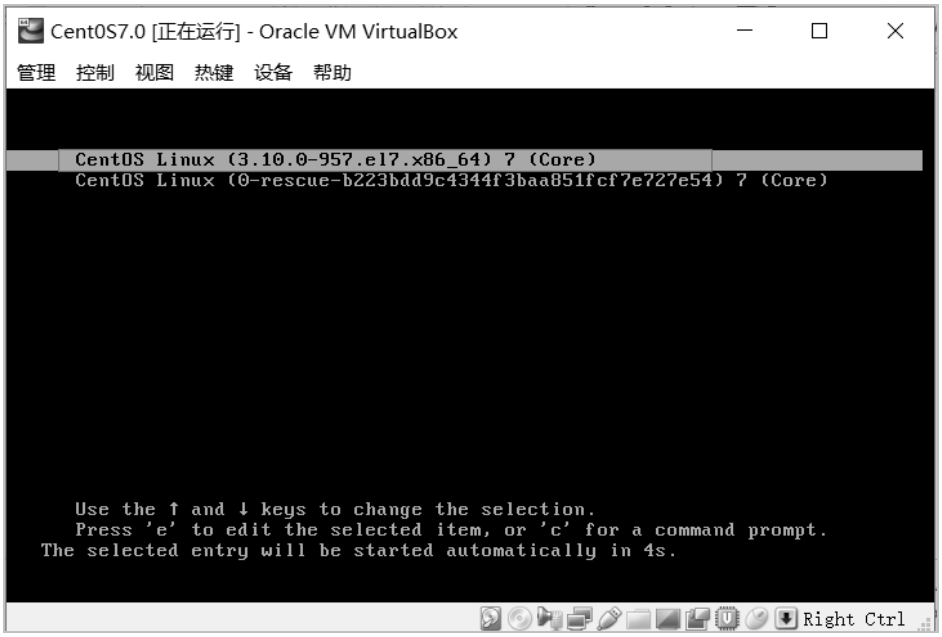

图 1-27

进入系统登录界面，在此界面中根据系统提示，输入账号 root 以及设置的密码 root12 登录系统，如图 1-28 所示。

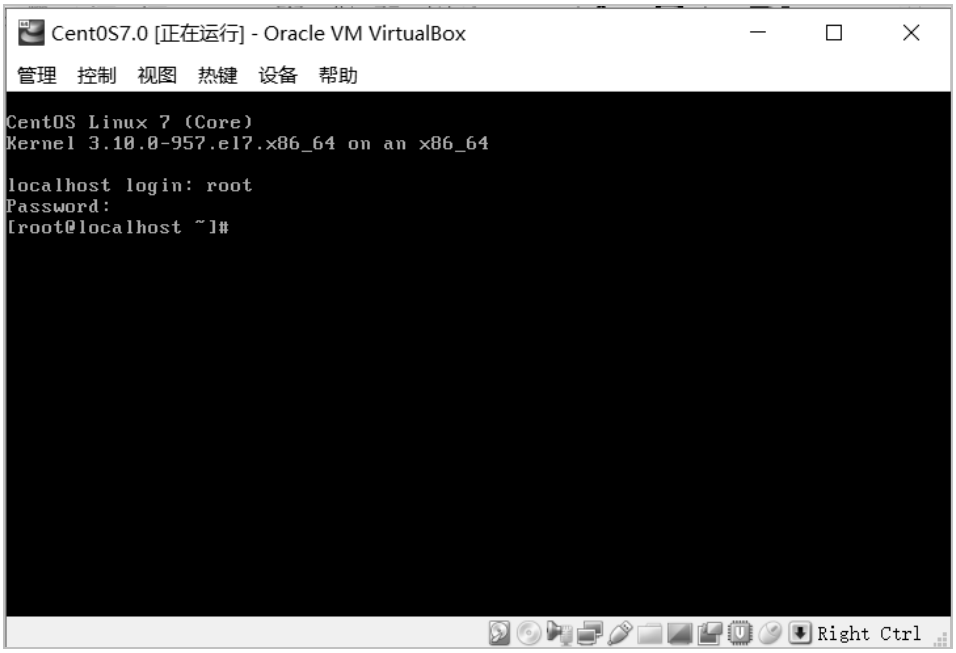

图 1-28

至此，我们已经在虚拟机中安装好了 CentOS 7 系统，在接下来的章节中我们会在此环境中进行各种相关操作和配置，以构建和使用 Hadoop 平台。

单元小结

- 大数据基本概述
- Hadoop 及体系介绍
- Hadoop 的环境准备工作

单元自测

选择题

1. Hadoop 的作者是(　　)。

A. Martin Fowler　　B. Kent Beck

C. Doug cutting

2. 下面选项(　　)不属于 Google 的三驾马车。

A. HDFS　　B. MapReduce

C. BigTable　　D. GFS

3. 下面选项(　　)是大数据的基本特征。

A. 数据体量大　　B. 数据类型多

C. 处理速度快　　D. 价值密度低

4. Hadoop 能够使用户轻松开发和运行处理大数据的应用程序，它具有的特点是(　　)。(多选题)

A. 高可靠性　　B. 可扩展性

C. 高效性　　D. 高容错性

5. Hadoop 能够使用户轻松开发和运行处理大数据的应用程序，下面(　　)不属于 Hadoop 的特性。

A. 高可靠性、高容错性　　B. 高扩展性

C. 高实时性　　D. 高效性

问答题

1. 理解什么是云计算、大数据和人工智能，梳理三者之间的关系。
2. 请解释什么是大数据的 4V 特性。

上机题

完成 CentOS 7 的环境安装。