企业级卓越人才培养解决方案“十三五”规划教材

DASHUJU GAOKEYONG HUANJING DAJIAN YU YUNWEI

大数据高可用环境搭建与运维

天津滨海迅腾科技集团有限公司◎编著

天津大学出版社
TIANJIN UNIVERSITY PRESS

企业级卓越人才培养解决方案“十三五”规划教材

大数据高可用环境搭建与运维

天津滨海迅腾科技集团有限公司　编著

天津大学出版社
TIANJIN UNIVERSITY PRESS

图书在版编目(CIP)数据

大数据高可用环境搭建与运维 / 天津滨海迅腾科技集团有限公司编著. —天津：天津大学出版社，2019.1
企业级卓越人才培养解决方案“十三五”规划教材
ISBN 978-7-5618-6371-8

Ⅰ.①大… Ⅱ.①天… Ⅲ.①数据处理软件－教材
Ⅳ.①TP274

中国版本图书馆CIP数据核字(2019)第027835号

主　编：刘晓丹　孙　峰
副主编：畅玉洁　冯德万　孟英杰　郭思延　罗芸芸
　　李思广　冯时昌　徐均笑　刘　健

出版发行　天津大学出版社
地　　址　天津市卫津路92号天津大学内(邮编:300072)
电　　话　发行部:022-27403647
网　　址　publish.tju.edu.cn
印　　刷　天津泰宇印务有限公司
经　　销　全国各地新华书店
开　　本　185mm×260mm
印　　张　17.75
字　　数　443千
版　　次　2019年1月第1版
印　　次　2019年1月第1次
定　　价　49.00元

企业级卓越人才培养解决方案“十三五”规划教材
编写委员会

范文涵　山西财贸职业技术学院
宋　军　山西煤炭职业技术学院
李庶泉　周口职业技术学院
孙　刚　南京信息职业技术学院
宋国庆　天津电子信息职业技术学院
王　英　天津滨海职业学院
刘　盛　天津城市职业学院
郭社军　河北交通职业技术学院
麻士琦　衡水职业技术学院
王　江　唐山职业技术学院
张　捷　唐山科技职业技术学院
张　晖　山东药品食品职业学院
赵红军　山东工业职业学院
杨　峰　山东胜利职业学院
王建国　烟台黄金职业学院
景悦林　威海职业学院
郑开阳　枣庄职业学院
常中华　青岛职业技术学院
赵　娟　山西旅游职业学院
陈怀玉　山西经贸职业学院
任利成　山西轻工职业技术学院
郭长庚　许昌职业技术学院
许国强　湖南有色金属职业技术学院
夏东盛　陕西工业职业技术学院
张雅珍　陕西工商职业学院
周仲文　四川广播电视大学
许　磊　重庆电子工程职业学院
夏先玉　重庆房地产职业学院
董新民　安徽国际商务职业学院
王国强　甘肃交通职业技术学院
杨志超　四川华新现代职业学院
王柱京　重庆电讯职业学院
谭维齐　安庆职业技术学院
李洪德　青海柴达木职业技术学院

企业级卓越人才培养解决方案简介

企业级卓越人才培养解决方案（以下简称“解决方案”）是面向我国职业教育量身定制的应用型、技术技能人才培养解决方案，是以教育部－滨海迅腾科技集团产学合作协同育人项目为依托，依靠集团研发实力，通过联合国内职业教育领域相关政策研究机构、行业、企业、职业院校共同研究与实践获得的科研成果。本解决方案坚持“创新校企融合协同育人，推进校企合作模式改革”的宗旨，消化吸收德国“双元制”应用型人才培养模式，深入践行基于工作过程“项目化”及“系统化”的教学方法，设立工程实践创新培养的企业化培养解决方案。在服务国家战略——京津冀教育协同发展、中国制造2025（工业信息化）等领域培养不同层次的技术技能人才，为推进我国实现教育现代化发挥了积极作用。

该解决方案由“初、中、高”3个培养阶段构成，包含技术技能培养体系（人才培养方案、专业教程、课程标准、标准课程包、企业项目包、考评体系、认证体系、社会服务及师资培训）、教学管理体系、就业管理体系、创新创业体系等，采用校企融合、产学融合、师资融合“三融合”的模式在高校内共建大数据（AI）学院、互联网学院、软件学院、电子商务学院、设计学院、智慧物流学院、智能制造学院等，并以“卓越工程师培养计划”项目的形式推行，将企业人才需求标准、工作流程、研发规范、考评体系、企业管理体系引进课堂，充分发挥校企双方优势，推动校企、校际合作，促进区域优质资源共建共享，实现卓越人才培养目标，达到企业人才招录的标准。本解决方案已在全国几十所高校开始实施，目前已形成企业、高校、学生三方共赢的格局。

天津滨海迅腾科技集团有限公司创建于2004年，是以IT产业为主导的高科技企业集团。集团业务范围已覆盖信息化集成、软件研发、职业教育、电子商务、互联网服务、生物科技、健康产业、日化产业等。集团以科技产业为背景，与高校共同开展“三融合”的校企合作混合所有制项目。多年来，集团打造了以博士研究生、硕士研究生、企业一线工程师为主导的科研及教学团队，培养了大批互联网行业应用型技术人才。集团先后荣获：天津市“五一”劳动奖状先进集体、天津市政府授予“AAA”级劳动关系和谐企业、天津市“文明单位”、天津市“工人先锋号”、天津市“青年文明号”、天津市“功勋企业”、天津市“科技小巨人企业”、天津市“高科技型领军企业”等近百项荣誉。集团将以“中国梦，腾之梦”为指导思想，在2020年实现100所以上高校合作，形成教育科技生态圈格局，成为产学协同育人的领军企业。2025年形成教育、科技、现代服务业等多领域100%生态链，实现教育科技行业“中国龙”目标。

前　言

在大数据时代，数据的存储与分析至关重要，为保证存储可靠、分析精准，对大数据环境部署与运维的要求日益提高。医疗、交通、金融等多个行业在追求大数据处理平台的高可靠性、高扩展性及高容错性的同时，还希望能够降低成本，本书为实现这些需求提供了解决案例。

本书主要以 Hadoop 生态体系环境部署为主线，讲解其各组件的功能与基础应用以及大数据生态系统的维护和安全解决方案等方面的知识。全书知识点的讲解由浅入深，能使每一位读者都有所收获，也保持了整本书的知识深度。

本书主要涉及 11 个项目，即大数据分布式集群、分布式集群基础配置、ZooKeeper 分布式协调系统、Hadoop 高可用、Hive 分布式数据仓库工具、HBase 分布式数据库、大数据协作框架、Linux 自动化部署、Ambari 大数据环境搭建利器、企业级 Hadoop 调优方案、企业级 Hadoop 安全方案，严格按照生产环境中的操作流程对知识体系进行编排。书中介绍的环境搭建并不限于虚拟机，对于有条件的公司和学校，参照书中介绍的搭建过程，同样可以将大数据平台搭建在多台实体计算机上，以便更加接近于大数据真实的运行环境。与此同时，为了方便读者学习，本书还配有教材中所用到的全部软件以及安装包和工具，可以节省读者查找下载相关工具的时间。

本书中每个模块都设有学习目标、学习路径、任务描述、任务技能、任务实施和任务总结。结构条理清晰、内容详细，任务实现可以将所学的理论知识充分地应用到实际操作中。

本书由刘晓丹、孙峰共同担任主编，畅玉洁、冯德万、孟英杰、郭思延、罗芸芸、李思广、冯时昌、徐均笑、刘健担任副主编，孙峰和刘晓丹负责整书编排，项目一和项目二由畅玉洁、冯德万负责编写，项目三和项目四由冯德万、郭思延负责编写，项目五由孟英杰负责编写，项目六由郭思延、冯时昌负责编写，项目七由罗芸芸负责编写，项目八和项目九由李思广和冯时昌共同负责编写，项目十和项目十一由徐均笑和刘健负责编写。

本书理论内容简明、扼要，实例操作讲解细致、步骤清晰，实现了理实结合，操作步骤后有相对应的效果图，便于读者直观、清晰地看到操作效果，从而牢记书中的操作步骤，使读者在学习 Hadoop 生态体系相关知识的过程中能够更加顺利掌握。

天津滨海迅腾科技集团有限公司

技术研发部

目　录

项目一　大数据分布式集群 …… 1

学习目标 …… 1
学习路径 …… 1
任务描述 …… 2
任务技能 …… 3
任务实施 …… 16
任务总结 …… 19
英语角 …… 19
任务习题 …… 19

项目二　分布式集群基础配置 …… 21

学习目标 …… 21
学习路径 …… 21
任务描述 …… 22
任务技能 …… 23
任务实施 …… 39
任务总结 …… 45
英语角 …… 45
任务习题 …… 45

项目三　ZooKeeper 分布式协调系统 …… 47

学习目标 …… 47
学习路径 …… 47
任务描述 …… 48
任务技能 …… 49
任务实施 …… 66
任务总结 …… 69
英语角 …… 70
任务习题 …… 70

项目四　Hadoop 高可用 …… 72

学习目标 …… 72
学习路径 …… 72

任务描述 ……73
任务技能 ……74
任务实施 ……90
任务总结 …… 100
英语角 …… 100
任务习题 …… 101

项目五 Hive 分布式数据仓库工具 …… 102

学习目标 …… 102
学习路径 …… 102
任务描述 …… 103
任务技能 …… 104
任务实施 …… 117
任务总结 …… 131
英语角 …… 131
任务习题 …… 132

项目六 HBase 分布式数据库 …… 133

学习目标 …… 133
学习路径 …… 133
任务描述 …… 134
任务技能 …… 135
任务实施 …… 148
任务总结 …… 154
英语角 …… 154
任务习题 …… 155

项目七 大数据协作框架…… 156

学习目标 …… 156
学习路径 …… 156
任务描述 …… 157
任务技能 …… 158
任务实施 …… 168
任务总结 …… 173
英语角 …… 174
任务习题 …… 174

项目八 Linux 自动化部署 …… 176

学习目标 …… 176
学习路径 …… 176

任务描述 …… 177
任务技能 …… 178
任务实施 …… 187
任务总结 …… 202
英语角 …… 202
任务习题 …… 203

项目九 Ambari 大数据环境搭建利器 …… 204

学习目标 …… 204
学习路径 …… 204
任务描述 …… 205
任务技能 …… 206
任务实施 …… 217
任务总结 …… 229
英语角 …… 229
任务习题 …… 229

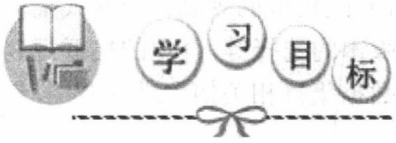
项目十 企业级 Hadoop 调优方案 …… 231

学习目标 …… 231
学习路径 …… 231
任务描述 …… 232
任务技能 …… 233
任务实施 …… 241
任务总结 …… 244
英语角 …… 244
任务习题 …… 244

项目十一 企业级 Hadoop 安全方案 …… 246

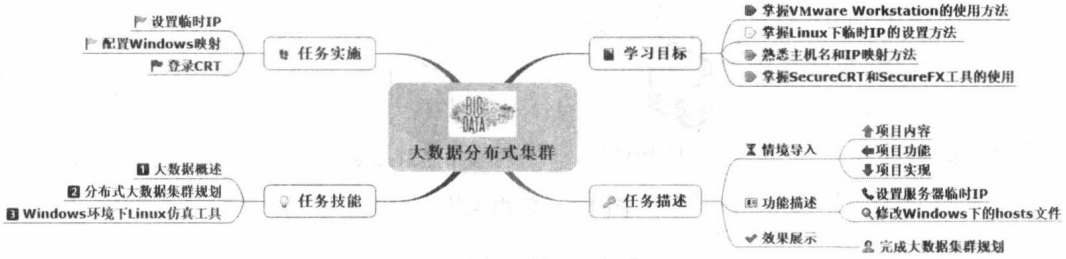
学习目标 …… 246
学习路径 …… 246
任务描述 …… 247
任务技能 …… 247
任务实施 …… 262
任务总结 …… 268
英语角 …… 268
任务习题 …… 269

项目一　大数据分布式集群



通过完成集群的创建，了解大数据的定义及其发展趋势，熟悉集群规划、主机规划、软件规划以及数据目录规划，掌握相关工具的使用，在任务实施过程中：

- 掌握 VMware Workstation 的使用方法；
- 掌握 Linux 下临时 IP 的设置方法；
- 熟悉主机名和 IP 映射方法；
- 掌握 SecureCRT 和 SecureFX 工具的使用。



【情境导入】

由于互联网技术的兴起，大数据技术被越来越多的人所熟知，更多的企业和个人开始学习大数据技术。但在系统学习大数据技术前，需要进行分布式集群的规划。分布式集群的规划阶段包括确认集群部署方式、节点 IP 的配置和软件版本的选择等，合理的集群规划有助于集群的高效运行和维护的便利性。本项目主要对大数据集群相关知识进行介绍，最终完成集群的创建。

【功能描述】

- 设置服务器临时 IP。
- 修改 Windows 下的 hosts 文件。

【效果展示】

通过对本项目的学习，完成大数据集群中服务器规划部署，最终效果如图 1-1 和表 1-1 所示。

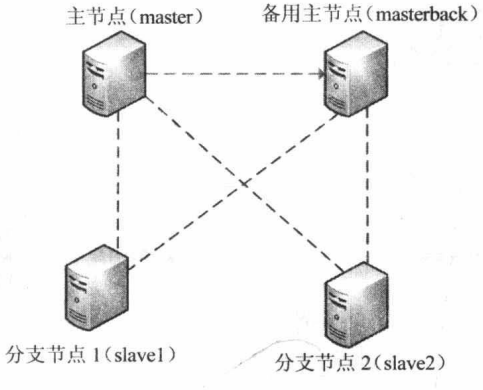


图 1-1 集群架构

表 1-1 服务器规划

主机名	节点类型	IP	系统	内存
master	主节点	192.168.10.110	CentOS7	4 GB
masterback	备用主节点	192.168.10.111	CentOS7	4 GB
slave1	分支节点 1	192.168.10.112	CentOS7	2 GB
slave2	分支节点 2	192.168.10.113	CentOS7	2 GB

技能点一　大数据概述

大数据的概念最早由维克托·迈尔-舍恩伯格和肯尼斯·库克耶提出。在二人编写的《大数据时代》中，大数据被定义为不用抽样调查这样的捷径，而采用对“所有数据”进行分析处理的数据。

随着时代的发展，大数据的定义也在不断变化。现如今，大数据（Big Data）是指：无法在一定时间内使用常规软件工具或方法进行收集、管理和处理的数据集合，需要全新的处理模式才能具有更强的决策力、洞察力和流程优化能力的信息资产。大数据同时也是高增长、海量和多样化的信息资产，可应用在计算机、信息科学、人工智能等领域。

1. 大数据应用

现代社会正处在信息时代，以“大数据”为代表的新兴技术，在迅猛发展的同时也引领着互联网的发展。

党的十八届五中全会审议通过了《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》（以下简称《建议》）。《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，超前布局下一代互联网”，可见大数据已经成为国家发展战略。

大数据不仅是国家战略，也在不知不觉中影响着人们的生活甚至日常行为。

很多人可能有这样的经历，使用某浏览器或手机客户端在淘宝、京东等购物网站上购买过一部手机，在之后的很长一段时间内，浏览器两侧的广告栏里或者手机客户端的订阅栏会出现关于此款手机的相关产品的情况，如：该型号的手机膜、手机保护壳、手机的充电线、耳机等。对于经常浏览的商品类型，系统会默认用户有购买这些产品的意向，在推送广告时，也会把此类的商品推送给用户。大多数人并不反感这些广告，因为这些往往是系统经过对用户的行为分析而“精准化”推送的服务，甚至很多用户达到了乐此不疲的程度，就算没有购买的欲望，也会登录购物网站进行浏览。这就是“大数据”最为简单的应用之一。

大数据的应用主要分为精准化定制和预测两个方面。

1）精准化定制

精准化定制即通过对用户的需求进行数据分析，然后提供相对指定化的服务。具体可分为以下三类。

➢ 个性化产品：使用智能化搜索引擎搜索相同内容，不同的用户将得到不同的结果。

➢ 精准营销：常见的互联网营销如“百度推广”“淘宝推广”等，或是基于地理位置的推送如“美团”“大众点评”，当用户到达某个地理位置时会收到周边的消费信息。

➢ 选址定位：零售店选址、公共基础设施选址。

大数据通过获取需求方的个性化需求，帮助供应方精准定位目标，然后依据需求提供定制服务，最终实现供需双方的最佳匹配。

精准化定制服务如图 1-2 所示。

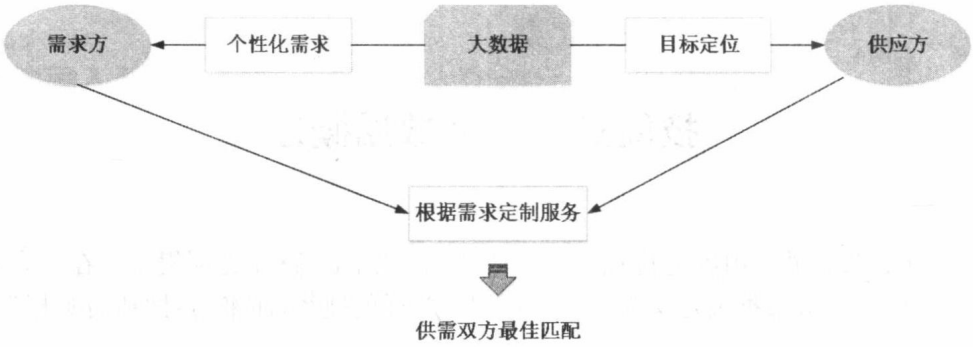


图 1-2 精准化定制

2）预测

预测目标数据，通过相关的数据分析作出预警，或是实时动态优化。具体可分为以下三类。

➢ 决策支持类：企业运营决策、证券投资决策、医疗行业临床诊疗支持以及电子政务等。

➢ 风险预警类：疫情预警、健康管理疾病预警、公共安全预警、设备设施运营维护等。

➢ 实时优化类：导航路线主能规划、实时计价等。

大数据收集大量数据并对相关因素进行分析后得到对未来发展的精准预测，而做到预警和动态优化。

预测服务如图 1-3 所示。

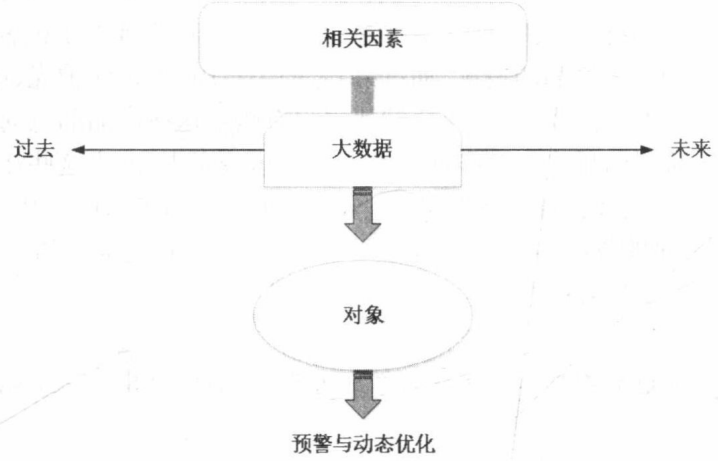


图 1-3 预测

2. 大数据的特征

虽然对于大数据的概念人们往往有不同的见解，但是无论作何描述，大数据的本质是由

如下 4 点组成的：Volume（海量 / 规模性）、Velocity（高速性）、Variety（多样性）以及 Value（价值性），见表 1-2。

表 1-2　大数据特性

特性	解释
Volume（海量 / 规模性）	数据的价值和隐藏价值由数据的大小决定
Velocity（高速性）	获得数据的速度
Variety（多样性）	数据的多样性
Value（价值性）	合理运用大数据，以低成本创造高价值

海量 / 规模性，是指数据的规模巨大。现如今，数据的规模已经超出了历史上任何一个时代。在互联网飞速发展的今天，数据量的增速已经达到了前所未有的程度。根据 IDC（国际数据公司）的统计，到 2020 年全球将会有超过 35 ZB（1 ZB ≈ 1 万亿 GB）的数据量，如果需要存储这些数据大约需要 6.2 亿块容量为 60 TB 的硬盘。海量规模性是大数据的首要特性。

高速性，是指大数据的传播速度快。与过去的纸质传播和磁盘相比，在线（online）使数据的传播速度变得比以往的任何时期都快，将来数据的传播速度只会更快，所以在线也是高速的本质。高速性是大数据的关键特性。

多样性，是指数据的来源方式与形态众多。数据来源的方式，随着网络和科技的发展而改变。在现代社会，无数的传感器与互联网是数据来源的主要方式，社交网络、传感器网络、工业生产过程等无时无刻不在产生数据，数据也从之前单一的文字和符号变为图片、视频、音频等非结构化与半结构化数据。多样性是大数据的自然属性。

价值性，是指数据是“可用”的且“有用”的。例如，在海量数据中，有价值的可能就是一个文本中的一句话或者是一个视频中的两三秒的镜头，这就需要对数据进行清洗和筛选，清洗掉不需要的或者是有干扰信息的数据，筛选出有用的、有价值的数据。价值性是大数据的基本属性。

3. 大数据发展趋势

1）数据分析

数据分析是大数据技术的核心，在数据处理过程中占据很重要的位置。大规模数据集合的处理是大数据价值的体现，要想从大规模的数据中获取有用的信息，对数据进行挖掘分析是必需的。而数据的采集、存储、管理均是数据分析的基础步骤，对数据分析后得到的结果会应用到各个领域。大数据的发展与数据分析是密切相关的。

2）实时数据处理

如今用户获取信息的速度越来越快，为满足用户需求，大数据系统也同样需要不断升级。目前大数据主要采用具有一定局限性的批处理方式，主要应用于数据报告频率低的场合，因为数据要求频率较高的场合，批处理方式达不到要求，如实时个性化推荐、实时路况等数据处理就要求在短时间内完成。在大数据的发展趋势中数据实时处理将会成为主流，推动大数据发展。

3）基于云平台的数据分析

随着云计算技术的飞速发展，其应用范围也越来越广。云计算的发展为大数据技术提

供了能够进一步发展的技术平台。此外，随着云计算技术的日趋完善，大数据技术和数据处理水平也会得到显著提升。

4. 大数据技术

大数据技术和大数据是容易被初学者混淆的两个概念。前文已经对大数据的概念进行了详细的解释，大数据技术可以理解为对大数据的处理方法。

大数据技术通常包括如下方面。

（1）数据采集：使用数据采集工具将分布的、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，而后加载到数据仓库或数据集之中，使之成为联机分析处理、数据挖掘的基础。

（2）数据存取：将数据采集过程中得到的数据导出到关系型或非关系型数据库中。

（3）存储架构：云存储、分布式文件系统（HDFS）等。

（4）数据处理：把采集到的数据针对关键指标进行数据的处理和清洗等。

（5）统计分析：方差分析、回归分析、简单回归分析技术等。

（6）数据挖掘：分类、估计、模型预测、结果呈现等方式。

5. Hadoop 相关项目

目前 Hadoop 是最为流行的大数据处理平台。Apache 相关项目有很多，常见的 Hadoop 相关项目如下。

（1）Ambari：Ambari 是一种支持 Hadoop 集群供应、管理和监控的 Web 工具，主要用来解决大数据集群搭建复杂和对 CPU、HDFS 等相关指标监控难等问题。Ambari 除了提供仪表式的指标监控界面以外，还内嵌了报警系统，方便管理和及时发现并解决问题。

（2）Hive：Hive 是基于 Hadoop 的一个类似于关系型数据库的数据仓库工具，它的作用是能够将结构化的数据文件映射成为数据库表，并提供完整的查询功能。Hive 定义了类似于 SQL 的查询语言 HiveQL。HiveQL 可将 SQL 语句转换为 MapReduce 任务进行执行。

（3）HBase：HBase 是一个分布式的、面向列的数据库。HBase 利用 Hadoop HDFS 作为其文件存储系统，通过使用 Hadoop MapReduce 来处理 HBase 中的海量数据，使用 ZooKeeper 作为协同服务，提供对数据的随机随时读写与访问。

（4）ZooKeeper：ZooKeeper 是一个开源的分布式应用程序协调服务，是 Hadoop 和 HBase 的重要组件。它能够为分布式应用提供配置维护、域名服务、分布式同步、组服务等，其目的是为用户提供简单易用的接口和性能高效、功能稳定的服务。

技能点二　分布式大数据集群规划

分布式大数据集群需要由多台 Linux 主机组成，一个主备集群中只能有一个主节点、一个备用主节点和多个分支节点，在高可用集群中主节点和分支节点需要安装的软件会有所不同，导致所要完成的功能也会有所不同，所以在正式开始搭建集群前做好集群规划是很有必要的。

1. 集群拓扑

本书集群主要使用 master/slave（服务器主从方式）作为基础拓扑。服务器主从方式由一台节点正常运行并提供对外服务，另一台节点作为备用机，备用机在正常运行状态下不接受外部请求，但会实时对主服务器进行检测，当主服务器宕机时才会接管应用服务。因此设备利用率最高可达 50%。主从方式集群如图 1-4 所示。

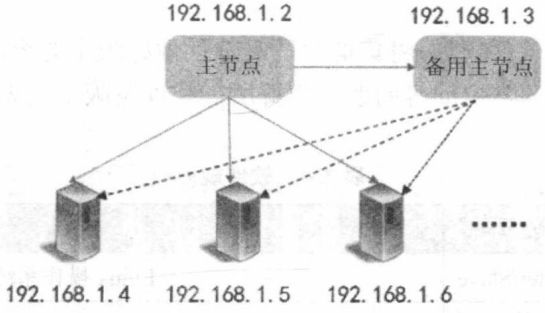


图 1-4　主从方式集群

更多大数据集群部署方式通过扫描下方二维码即可了解。

2. 主机规划

使用 4 台主机搭建 Hadoop2.0 高可用（目的是为了减少服务中断时间）分布式集群，选用两台机器分别作为主节点和备用节点，其余两台作为数据节点，这样规划的目的是为了简单实现高可用，采用两个数据节点是为了增加数据备份数量从而实现数据的安全性。虽然节点较少，但足以完成分布式集群的搭建，Hadoop 主机规划进程信息见表 1-3。

表 1-3　Hadoop 主机规划进程信息

进程	master	masterback	slave1	slave2
NameNode	是	是	否	否
DFSZKFailoverController	是	是	否	否
QuorumPeerMain	是	是	是	是
ResourceManager	是	是	否	否
JournalNode	否	否	是	是
NodeManager	否	否	是	是
DateNode	否	否	是	是