大数据分析：
R基础及应用

走进R，走进大数据时代数据分析的潮流尖端，掌握R语言，
熟悉大数据的基础概念和R与Hadoop结合进行大数据的处理分析。

深圳国泰安教育技术股份有限公司
中科院深圳先进技术研究院-国泰安金融大数据研究中心 | 编著

清華大學出版社

大数据分析：R基础及应用

深圳国泰安教育技术股份有限公司
中科院深圳先进技术研究院-国泰安金融大数据研究中心
编著

内 容 简 介

在大数据时代，R 以其强大的数据分析挖掘、可视化绘图等功能，越来越受到社会各个领域的青睐。现在，R 的计算引擎、性能、程序包都得到了提升，其中 R 与大数据分析平台 Hadoop 的结合，实现了 R 对大数据的分析式处理分析。这些不仅大大扩展了 R 的应用，也扩大了 R 在各行业的需求。

为了更好地适应新形势，掌握大数据分析处理的相关知识是很有必要的。本书从理论基础、方法、实证三方面详细地阐释了 R 和 RHadoop 的相关理论、技术以及应用，使读者了解大数据的基础概念，掌握 R 以及 Rhadoop 大数据分析技术。本书不仅适合高等院校的各相关专业的本专科生、研究生，也适合零编程基础的科研人员以及对大数据分析技术感兴趣的人士阅读。本书在内容的选择和结构的安排上进行了深入的思考，使得不论是 R 或 RHadoop 的初学者还是具备一定相关专业知识的人员都能从本书中得到一定的收获或启发。

图书在版编目(CIP)数据

大数据分析：R 基础及应用/深圳国泰安教育技术股份有限公司，中科院深圳先进技术研究院—国泰安金融大数据研究中心编著. --北京：清华大学出版社，2016

ISBN 978-7-302-42863-3

Ⅰ. ①大… Ⅱ. ①深… ②中… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 024162 号

责任编辑：彭　欣
封面设计：李文钰
责任校对：王荣静
责任印制：王静怡

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：010-62770175　　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：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：12.5　　**字　　数**：300 千字
版　　次：2016 年 3 月第 1 版　　**印　　次**：2016 年 3 月第1 次印刷
印　　数：1～4000
定　　价：49.00 元

产品编号：068005-01

CONTENTS 目录

第一部分 大数据简介

第二部分 R 语 言

第四部分 RHadoop案例分析

PART 1

第一部分

大数据简介

CHAPTER 1

第1章 大数据概述

大数据时代早已到来，《大数据时代》的作者维克托·迈尔·舍恩伯格说，世界的本质就是数据，大数据将开始一次重大的时代转型。其实早在1980年，美国著名未来学者托夫勒便在《第三次浪潮》一书中提出“数据就是财富”，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。作为云计算领域的重要延伸，大数据正在引领信息革命进入新的时代。2001年，全球最具权威的IT研究与顾问咨询公司Gartner提出大数据面临4个V的挑战；《自然》杂志(2008年)推出《大数据》专刊，全方面介绍大数据问题；美国总统奥巴马(2012年)将数据定义为“未来的新石油”。2013年，Gartner在一篇报告中指出，64%的受访企业都表示他们正在或是即将进行大数据工作。信息技术、计算机技术和互联网技术的迅速发展，使得人类社会各类数据呈现出爆炸性增长，对这些复杂大数据的有效管理，现已成为当前社会的热点问题。

1.1 大数据的概念

大数据(Big Data)，或称为巨量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的资讯。大数据一般指在10TB(1TB=1024GB)规模以上的数据量，其基本特征可以用4个V来总结：数据规模大(Volume)、数据类别多(Variety)、数据处理速度快(Velocity)、价值密度低(Value)[①]。

然而，“大数据”的概念远不止大量的数据(TB)和处理大量数据的技术，或者所谓的“4个V”之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，而这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力。

① http://www.cfern.org/wjgg/wjggDisplay.asp?Id=2353.

1.2 大数据的特征

大数据具有以下 4 个基本特征：数据规模大、数据类别多、数据处理速度快、价值密度低。

1. 数据规模大

大数据的基本属性是数据量巨大。目前，各个行业中的各个企业每天都会产生大量的数据，数据呈爆炸式的增长，数据量已从 TB 级别跃升到 PB 级别，甚至到了 EB 数量级。面对海量数据，传统的数据库系统处理能力已经难以应对，而且数据量仍在大规模增长，产生数据的来源也变得更加多样化。

2. 数据类别多

大数据除了传统的商业活动产生的数据外，还包括互联网上社交媒体产生的文本数据及时刻产生的传感器数据等。数据类型除了结构化数据外，还有半结构化和非结构化数据，如图片、网页、视频等，数据种类繁多。

3. 数据处理速度快

大数据和传统数据挖掘最显著的一个区别就是大数据要求处理速度快。面对如此大规模的数据，有效处理数据的效率也就牵系着企业的命运。对数据的实时处理、分析及反馈变得十分重要，创建实时数据已经成为一种趋势。

4. 价值密度低

价值密度往往与数据量成反比，在大量数据中有用的信息可能是非常少的，而且要有效地获取这些有用的信息也是比较困难的。比如，连续的监控产生大量的视频信息，而我们需要的数据可能就只有一两秒。针对大数据价值密度低这一特征，如何有效地挖掘出其中有用信息变得尤为重要。

1.3 大数据的产生

大数据的产生是计算机和网络通信技术被广泛运用的必然结果。互联网、移动互联网、物联网、云计算、社交网络等新一代信息技术的发展对大数据的产生起到了促进的作用。数据产生方式的变化表现为以下 4 个方面。

(1) 数据产生由企业内部向企业外部扩展。

(2) 数据产生由 Web1.0 向 Web2.0 扩展。

(3) 数据产生由互联网向移动互联网扩展。

(4) 数据产生由计算机或互联网(IT)向物联网(IOT)扩展。

这 4 个方面的变化让数据产生的源头成几何数增长，数据量也呈现出大幅度地快速增加。

1.4 大数据应用案例

大数据在各行业中有着大量的应用案例，比如金融行业中的信贷分析、银行风险分析及公司的交易分析等，医疗行业中的流行病学研究、病房的实时监控等，以及在亚马逊、淘宝

网、Facebook等互联网企业中的应用等。下面给出一个典型的大数据应用案例——余额宝。

余额宝的问世改变了天弘基金由原来国内排名中下并且连年亏损的状态，使得它位居国内基金管理公司之首，世界排名14。该公司将天弘增利宝货币基金从零开始发展到用户数量超过1亿元、资金规模达到5742亿元，超出了预计的10倍，成为世界第四大货币基金。

余额宝产生的背景是天弘基金欲借助最大电商阿里平台，在支付宝上向用户推销基金。阿里负责余额宝在支付宝端的建设，天弘基金负责与支付宝对接的直销和清算系统的建设。面对大规模的数据量，余额宝之前的系统已经不能满足需求，需要重建。余额宝的系统建设分为两期，然而随着数据量和交易量暴增，使得第一期系统仍无法负载日益增长的海量数据。于是进行了第二期系统的建设，阿里金融云提供了云计算服务，使得该系统的性能得到了相当大的提高，在很大程度上缩短了清算时间。在2013年11月11日的“双11”活动中，余额宝完成了1679万笔赎回，1288万笔申购的清算工作，成功为639万用户正确分配收益，当天处理了61.25亿元的消费赎回，119.97亿元的转入申购，而系统只用了46分钟就将全部清算工作完成。

实际上，二期系统现已不是简单的直销和清算系统，它每天面对着50个数据库里海量用户和交易数据的暴涨。那么，这些数据的使用、价值最大化吸引了企业机构的眼球。对此，天弘基金选择了阿里云提供的ODPS(开放数据处理服务)作为大数据平台，其中ODPS是阿里集团进行离线数据处理的平台，支撑了阿里金融、淘宝等多家BU的大数据业务。天弘基金将目标锁定在余额宝产生的海量数据分析上，以求把握上亿用户的理财需求及不同的风险接受能力，创造出更多更丰富的理财产品[①]。

① http://www.csdn.net/article/2014-05-26/2819939.

CHAPTER 2

第2章 大数据相关技术

大数据处理流程主要是指从海量数据中获取需要的信息并进行加工分析得到有用知识的输出过程。大数据处理流程的关键技术包括大数据存储和管理及大数据检索使用(包括数据挖掘和智能分析)。围绕大数据,一批新兴的数据存储、数据挖掘、数据处理与分析技术不断涌现,使得对海量数据的处理变得更加简便快速。大数据处理流程一般包括以下几个步骤:数据采集/清洗、数据存储、数据挖掘及数据呈现,如图 2.1 所示。

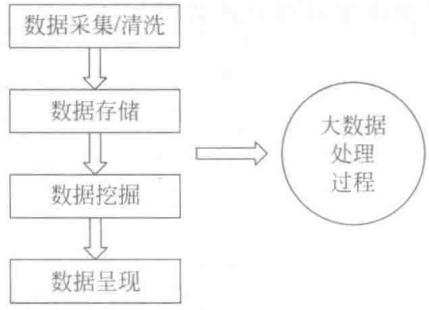

图 2.1 大数据处理流程

2.1 数据采集和准备

数据采集,即数据获取,是指从传感器或其他待测设备中获取信息的过程[①]。大数据采集包括对实时数据、非实时数据的采集,数据类型包括结构化、半结构化及非结构化数据。

大数据采集的方法有系统日志采集、数据库采集、网络数据采集等,采集的工具包括传感器、网络爬虫、移动基站及使用者自身产生的信息。

---

① http://baike.baidu.com/link?url=lnD8lmwKE4vGOneQhSBhNfFPNt7MfXl-sSyubVzcdYMN2Xsf9ylWBOLSLZt0YpVWInArgZunuSpSgv6G2bGrI_.

1. 传感器

传感器是一种检测装置，它采集数据的过程为：首先传感器感受被测量的信息，然后将其按一定规律变换成为电信号或其他形式的信息并输出。传感器是大规模数据的来源，比如，监控大型强子对撞机或四发动机大型喷气式客机需要成千上万的传感器通道，从而产生数百 TB 的数据。

2. 网络爬虫

网络爬虫是一种按照一定的规则，自动提取互联网网页信息的程序或脚本。互联网的数据形式多样，包括结构化的数据及图片、音频、视频等非结构化数据，对于这些海量数据，传统的获取方法已经不能满足需求，所以网络爬虫技术应运而生。网络爬虫可以定向地抓取用户所需的与某一特定主题相关的网页内容。

3. PON

日常通信过程中产生的海量信息。

4. 使用者自身产生的信息

随着微信、微博及邮件等的普及，使得它们拥有庞大的用户群。在人们使用这些软件的同时会产生巨大的信息，这些信息也是海量数据的重要来源。

在进行数据挖掘与分析前需要对数据进行一定的处理，即数据的准备。数据的准备是数据分析整个过程中的一个重要阶段，可以为后续的挖掘分析提供高质量的数据，从而保证了分析结果的有效性。数据准备包括数据的导入、数据的抽取、转换和装载等。数据导入指的是将外部数据导入到数据库或数据仓库中，关键是针对数据库的存储方式及具体的应用场景定义数据合适的模式。数据的抽取(Extract)是指将所需数据从源数据中抽取出来；数据的转换(Transform)是将获取的源数据按照一定的业务需求转换成所需要的形式，包括对数据的清洗和加工等操作；数据的装载(Load)指的是将经过转换后的数据装载到目的数据数源中。ETL 过程包括对数据空值的处理、数据格式的规范化处理、数据的替换及正确性验证的处理等，是数据挖掘分析的基础。

2.2　分布式数据库

大数据包括结构化数据、半结构化数据及非结构化数据，大数据的存储与普通数据存储的差别主要表现在数量级别和能否存储索引非结构化数据上。对于声音、图片、视频等非结构化数据，传统的关系型数据库无法满足存储需求，因此非关系型数据库变得尤为重要。大数据处理系统将通过 NoSQL 来存储这些非结构化数据并对这些数据进行相关的检索。

NoSQL 数据库指的是非关系型的数据库。NoSQL 数据库主要面向 Web 应用，支持分布式存储，能够满足对数据库高并发读写需求、海量数据的高效存储需求、数据库高扩展性和高可用性的需求等。NoSQL 数据库可以分为以下三类：面向高性能读写的数据库、面向文档的数据库及面向分布式计算的数据库(比如 Cassandra 数据库)。NoSQL 具有自由灵活的数据模型，典型的 NoSQL 数据库是以键值(Key-Values)的形式存储数据的。

NoSQL 满足 CAP 理论、BASE 原则。CAP 指的是对于以下三个特性：一致性、可用性及分区容错性，分布式系统不能同时满足，最多只能满足三个特性中的两个。BASE 指的是 Basically Available、Soft state、Eventually consistent。Basically Available(基本可用)指的

是对于系统短时间内的不可用是可容忍的；Soft state(柔性状态)指的是系统有异步的情况存在，即在某个时期可以不同步；Eventually consistent(最终一致性)指的是只要最终的数据满足一致性即可，不要求时刻满足一致性。NoSQL 数据库的设计一般针对具体的应用，遵循以上两个原则，比较注重数据的读写效率、数据的容量和系统的可扩展性等。

目前普遍使用的关系型数据库采用的是关系型数据模型，对数据存储增加及一些需要满足的数据范式，有时需要强行修改对象数据，以满足关系型数据库管理系统的需要，而 NoSQL 数据库完全改变了传统的观念，通过改变某些数据范式的严格要求，获得灵活的扩展性、灵活的数据模型、能够有效处理大数据、降低管理和维护成本等众多优点。表 2.1 对 NoSQL 数据库与关系型数据库的原理、规模、模式等进行了一个对比分析。

表 2.1 NoSQL 和关系型数据库的简单比较

比较标准	RDBMS	NoSQL	备　注
数据库原理	完全支持	部分支持	RDBMS 有数学模型支持，NoSQL 则没有
数据规模	大	超大	RDBMS 的性能会随着数据规模的增大而降低；NoSQL 可以通过添加更多设备以支持更大规模的数据
数据库模式	固定	灵活	使用 RDBMS 需要定义数据库模式，NoSQL 则不用
查询效率	快	简单查询非常高效、较复杂的查询性能有所下降	RDBMS 可以通过索引，能快速地响应记录查询(point query)和范围查询(range query)；NoSQL 没有索引，虽然 NoSQL 可以使用 MapReduce 加速查询速度，但仍然不如 RDBMS
一致性	强一致性	弱一致性	RDBMS 遵守 ACID 模型；NoSQL 遵守 BASE (Basically Available、Soft State、Eventually Consistent)模型
扩展性	一般	好	RDBMS 扩展困难；NoSQL 扩展简单
可用性	好	很好	随着数据规模的增大，RDBMS 为了保证严格的一致性，只能提供相对较弱的可用性；NoSQL 任何时候都能提供较高的可用性
标准化	是	否	RDBMS 已经标准化(SQL)；NoSQL 还没有行业标准
技术支持	高	低	RDBMS 经过几十年的发展，有很好的技术支持；NoSQL 在技术支持方面不如 RDBMS
可维护性	复杂	复杂	RDBMS 需要专门的数据库管理员(DBA)维护；NoSQL 数据库虽然没有 DBMS 复杂，但是也难以维护

随着互联网 Web 2.0 网站的兴起，传统的关系数据库在应付 Web 2.0 网站，特别是超大规模和高并发的 SNS 类型的 Web 2.0 纯动态网站已经显得力不从心，暴露了很多难以克服的问题，非关系型的数据库则由于其本身的特点得到了非常迅速的发展。

在信息技术融合应用的新时代，大数据就是像黄金一样的新型经济资产、像石油一样的重要战略资源。为满足大数据对处理和存储能力的无限需求，现今的计算机体系结构在数据存储方面要求具备庞大的水平扩展性(Horizontal Scalability，即要求满足能够连接多个软硬件的特性，这样可以将多个服务器从逻辑上看成一个实体)，而 NoSQL 致力于改变这一现状。目前 Google 的 BigTable 和 Amazon 的 Dynamo 使用的就是 NoSQL 数据库。NoSQL 数据库根据数据的存储模型和特点分为很多种类，如列存储、文档存储、Key-Value 存储、图存储、对象存储、xml 存储等数据库。表 2.2 给出了几种典型的 NoSQL 数据库及

其性能优缺点。

表 2.2　典型的 NoSQL 数据库分类

NoSQL 数据库类型	代表性产品	性能	扩展性	灵活性	复杂性	优点	缺点
键/值数据库	Redis Riak	高	高	高	无	查询效率高	不能存储结构化信息
列式数据库	HBase Cassandra	高	高	一般	低	查询效率高	功能较少
文档数据库	CouchDB MongoDB	高	可变	高	低	数据结构灵活	查询效率较低
图形数据库	Neo4J OrientDB	可变	可变	高	高	支持复杂的图算法	只支持一定的数据规模

在过去的 10 年里，正如交易率发生了翻天覆地的增长一样，需要存储的数据量也发生了急剧的膨胀，这种现象被称为“数据的工业革命”。为了满足数据量增长的需要，RDBMS（关系型数据库管理系统）的容量也在日益增加，但是对于一些企业来说，随着交易率的增加，单一数据库需要管理的数据约束的数量也变得越来越让人无法忍受了。现在，大量的“大数据”可以通过 NoSQL 系统来处理，它们能够处理的数据量远远超出了最大型的 RDBMS 所能处理的极限，很好地弥补了关系数据在某些方面的不足。

2.3　分布式数据分析框架

对于海量数据处理，一般可以分成离线数据处理和流式数据处理两大类。在海量数据的计算中，Hadoop 无疑是开源分布式离线处理技术的一大主力，而 Storm 则提供了分布式流处理框架，让实时大数据处理得以实现。

2.3.1　Hadoop

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，并且是以一种可靠、高效、可伸缩的方式进行处理的。Hadoop 的核心框架为 HDFS（Hadoop Distributed File System）、MapReduce 和 HBase，最底部是 HDFS，HDFS 的上一层是 MapReduce 引擎，如图 2.2 所示。其中 HDFS 实现对分布式存储的底层支持，用于存储 Hadoop 集群中所有存储节点上的文件，HBase 则为大量非结构化数据存储和索引提供了条件，MapReduce 则实现对分布式并行任务处理的程序支持，能够让用户编写的 Hadoop 并行应用程序运行更加简化。

Hadoop 作为开源的云计算平台已经在互联网领域得到了广泛的应用，互联网公司往往需要存储海量的数据并对其进行处理，而这正是 Hadoop 的强项。如 Facebook 使用 Hadoop 存储内部的日志拷贝以及数据挖掘和日志统计；Yahoo 利用 Hadoop 支持广告系统并处理网页搜索；Twitter 则使用 Hadoop 存储微博数据、日志文件和其他中间数据等。在国内，Hadoop 同样也得到了许多公司的青睐，如百度主要将 Hadoop 应用于日志分析和