首部全面讲解R语言与Hadoop技术结合应用于大数据分析的优秀著作

系统阐释R与Hadoop集成的各种实用方法、工具和最佳实践，深入剖析各种常见问题，包含大量实例，能为用户高效利用R语言与Hadoop技术进行大数据分析提供翔实指导

HZ BOOKS 华章科技

[PACKT PUBLISHING]

大数据技术丛书

Big Data Analytics with R and Hadoop

R与Hadoop大数据分析实战

（印） Vignesh Prajapati◎著

李明 王威扬 孙思栋◎等译

机械工业出版社
China Machine Press

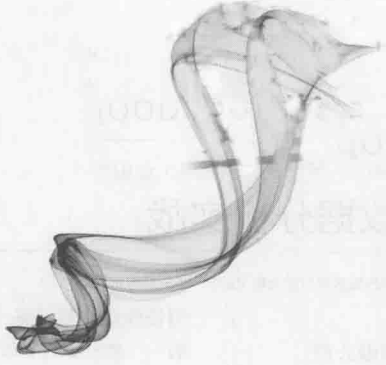
技术丛书

Big Data Analytics with R and Hadoop

R与Hadoop大数据分析实战

（印） Vignesh Prajapati◎著

李明 王威扬 孙思栋◎等译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 与 Hadoop 大数据分析实战 /（印）普贾帕提（Prajapati，V.）著；李明等译 . —北京：机械工业出版社，2014.11
（大数据技术丛书）
书名原文：Big Data Analytics with R and Hadoop

ISBN 978-7-111-48352-6

I. R… II. ① 普… ② 李… III. ① 程序语言 – 程序设计 ② 数据处理软件 IV. ① TP312 ② TP274

中国版本图书馆 CIP 数据核字（2014）第 246828 号

本书版权登记号：图字：01-2014-4757

R 与 Hadoop 大数据分析实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：秦 健　　责任校对：董纪丽
印　刷：北京市荣盛彩色印刷有限公司　　版　次：2014 年 11 月第 1 版第 1 次印刷
开　本：186mm × 240mm 1/16　　印　张：11.25
书　号：ISBN 978-7-111-48352-6　　定　价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

The Translator's Words 译 者 序

本书出版后在 Amazon 上获得了极高的关注度，因为它是当时全球仅有的一本讲述 R 语言同 Hadoop 技术结合的权威书籍。我于当年年末拿到此书并仔细研读，在此过程中便萌发出把此书翻译为中文版本的想法。而机械工业出版社以极快的速度同我敲定此事，可见他们独到的眼光和敏锐的市场洞察力。

本书由 10 余位小伙伴共同翻译而成，整个翻译过程充分体现了当下互联网的合作精神。首先我在个人博客以及豆瓣同城上发布了想翻译此书的想法，并迅速得到了几十名同学的报名响应。之后我制定并发布了整套书籍的术语以及译稿的样式。在接下来的 1 个月中小伙伴们陆续完成翻译工作。最后由我挑选出较好的译稿，并进行后期整合以及校验工作。

当下互联网甚至整个社会都在谈论大数据概念，而大数据之所以可以为互联网企业所推崇，其中一个重要原因是提出了 Hadoop 技术。它使可处理的数据量不再局限于某台单机的性能，而是通过计算机集群的方式极大地提高了可处理的数据量。而 R 语言则是另一款算法全面、易学易用的数据统计开源语言。它使得数据分析师以及数据挖掘人员可以把精力更多地放在算法本身，而非程序语言的繁琐语法上。但是 R 语言较大的缺点就是它只能在单机上运行，这就使其数据处理能力受限于本机的内存。所以如何使 R 语言处理大数据就成了当下新的研究热点。而现阶段较好的方式就是把 Hadoop 同 R 语言结合，实现在集群上运行 R 语言。

本书的阐述主要围绕如何实现 Hadoop 与 R 语言的结合，主要分为 4 个部分。

第一部分（基础概念），包括第 1 ~ 2 章，主要讲解 R 语言以及 Hadoop 的计算原理以及概念。

第二部分（初级应用），包括第 3 ~ 4 章，主要讲解 RHIPE、RHadoop 以及 streaming 三种实现方案。

第三部分（高级实例），包括第 5 ~ 6 章，主要以 RHadoop 为技术背景，讲解多个实际应用案例。

第四部分（数据库连接），包括第 7 章，主要讲解在 RHadoop 下如何同各类数据库进行连接。

相信书中大量的实际案例以及作者的精妙阐述可以帮助各位读者把 RHadoop 这项技术真正应用到实际工作中。

最后我要感谢我的老婆刘慧，如果没有她，我将失去做任何事情的动力。并把此书送给我刚出生的侄女李沐瑶，愿她健康快乐成长。

除封面署名译者外，参与本书翻译的还有以下译者：

张粤磊：从事过各行业（制造、金融、互联网）业务及大数据技术实践工作，关注大数据架构及分析，目前在平安付担任大数据平台架构师。

扶至钦：SuccezBI 商业智能研发工程师，对 Hadoop 平台下的 BI 数据分析挖掘有浓厚兴趣。

李学沧：致力于将 Hadoop 大数据技术应用于跨组织癌症的致癌机理、基于电子病历的医疗欺诈行为识别等医疗大数据研究工作。

游皓麟：在互联网、电信、电力领域拥有丰富建模经验，精通 Clementine、R 语言等数据挖掘工具，对 Anomaly Detection、广告反作弊、推荐系统、客户及营销建模有一定研究。

龚君泰：毕业于中国人民大学统计学院，研究方向为数据挖掘，现任中电广通科技有限公司数据分析师，从事数据分析及数据挖掘在政府统计及企业中的应用产品研发工作。

张春强：毕业于哈尔滨工业大学机械电子工程专业。曾就职于中兴通讯，从事软件开发以及大数据相关工作，R 语言爱好者。

齐舰：一直围绕着数据和数据库工作。精通 Oracle、MySQL、PostgreSQL、MongoDB 等各种数据库，在数据库设计、开发和管理上拥有丰富的经验，同时对数据分析和数据挖掘亦有心得。

志洪新：毕业于北京邮电大学，研究方向是基于 ERP 业务支持智能营销的高可用大电商平台。从事过大型电商平台（当当网等多个电商平台）的快速开发以及多个邮政物流和银行的大数据项目。

刘奔：长期关注 R 语言和 Hadoop 架构以及数据报表项目的搭建工作。

李岚凤：武汉大学信息学院情报学在读研究生。研究方向为竞争情报与信息分析，对 R 语言有较为浓厚的兴趣，并对其有所研究。

彭震：毕业于华南师范大学计算机学院，对 R 语言以及统计学有较深入研究。

谈申申：中南财经政法大学 2014 届金融系研究生，在多个科研项目中承担了数据挖掘与分析工作。

李明

Preface 前　言

组织获得的数据量每一天都在成倍增加。现在组织可以存储大量信息在低成本平台上，例如 Hadoop。

如今这些组织面临的难题是如何处理这些数据以及如何从这些数据中获取关键的见解。于是 R 应运而生。R 是一个令人惊讶的工具，它是一个在数据中能够运行先进统计模型的单元，它将原始模型翻译成丰富多彩的图形和可视化视图，而且它有很多和数据科学相关的功能。

然而，R 的一个主要缺点是它的可扩展性较差。R 的核心技术引擎可以加工和处理非常有限的数据量。正因为在大数据处理中 Hadoop 十分流行，所以为了可扩展性，下一步符合逻辑的方法将是把 R 和 Hadoop 结合起来。

本书介绍了 R 和 Hadoop，以及如何通过使用一个平台（如 Hadoop）进行 R 的数据分析操作以实现其可扩展性。

出于这样一个目标，本书将适合广大范围的读者，包括数据统计者、数据科学家、数据架构师和任何正在寻找使用 R 和 Hadoop 来处理和分析大量信息的解决方案工程师。

在 Hadoop 上使用 R 将提供一个弹性的数据分析平台，其规模取决于所需分析的数据集大小。富有经验的程序员可以用 R 语言编写 Map/Reduce 模块，并用 Hadoop 的 Map/Reduce 并行处理机制运行它以识别数据集的模式。

R 简介

R 是一个可以对数据进行统计分析的开源软件包。R 是一种编程语言，它受到数据科学统计师以及其他需要进行数据统计分析和从数据使用机制中寻找关键因素的人所青睐，这些机制包括回归、聚类、分类和文本分析。R 采用 GNU（General Public License，通用公共许可证）。它是由新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman 开发，他们目前领

导一个 R 语言开发核心团队。它可以看做是 S 语言的另一种实现，S 语言由贝尔实验室的 Johan Chambers 开发。它们有一些重要的差异，但是大部分用 S 语言编写的代码可以直接在 R 编译器下使用。

R 提供广泛的统计分析、机器学习（线性和非线性建模、经典的统计检验、时间序列分析、分类、聚类）和图形技术，并且高度可扩展。针对统计、机器学习和可视化 R 有多种内置的可扩展的功能，例如：

- ❑ 数据提取
- ❑ 数据清洗
- ❑ 数据加载
- ❑ 数据转换
- ❑ 统计分析
- ❑ 预测建模
- ❑ 数据可视化

它是一种当今市场上所提供的最流行的开放源代码统计分析软件包。它是跨平台的并具有广泛的社区支持，这意味着每天都会有数量庞大并不断增长的用户群体添加新的程序包。随着程序包数量的增加，R 现在可以与其他数据存储，如 MySQL、SQLite、MongoDB 和 Hadoop，相连接以进行数据存储。

R 的特点

R 不同的实用特点如下：

- ❑ 高效的编程语言
- ❑ 支持关系型数据库
- ❑ 数据分析
- ❑ 数据可视化
- ❑ 通过庞大的 R 程序包库文件进行扩展

R 的受欢迎程度

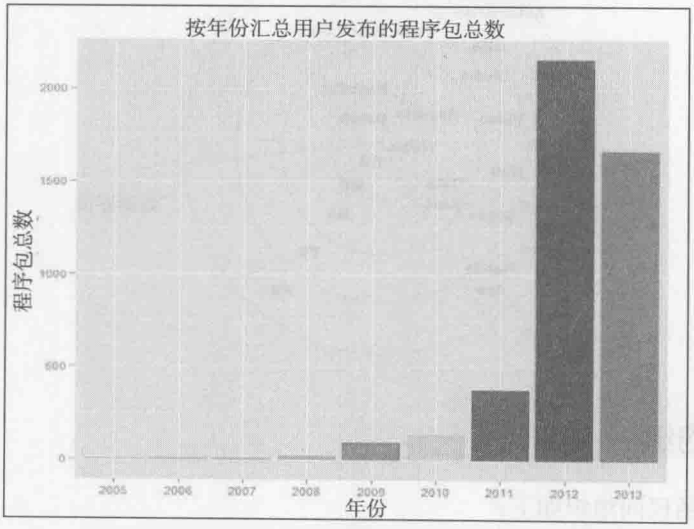
本图来源于 KD，图中表明 R 是用于数据分析和数据挖掘最流行的语言。

What programming languages you used for data mining / data analysis in the past 12 months? [570 voters]	
R (257)	45%
SQL (184)	32%
Python (140)	25%
Java (139)	24%
SAS (121)	21%
MATLAB (83)	15%
C/C++ (73)	13%
Unix shell/awk/gawk/sed (59)	10%
Perl (45)	7.9%
Hadoop/Pig/Hive (35)	6.1%
Lisp (4)	0.7%
Other (70)	12.0%
None (7)	1.2%

下图提供从 2005 年到 2013 年由 R 用户发布的 R 程序包总数的详细信息。这是我们如何探究 R 用户的依据。2012 年该数据呈指数增长，并且 2013 年似乎要超过前一年。

R 允许通过各种统计和机器学习进行数据分析，具体如下：

- ❑ 回归
- ❑ 分类
- ❑ 聚类
- ❑ 推荐
- ❑ 文本挖掘



大数据简介

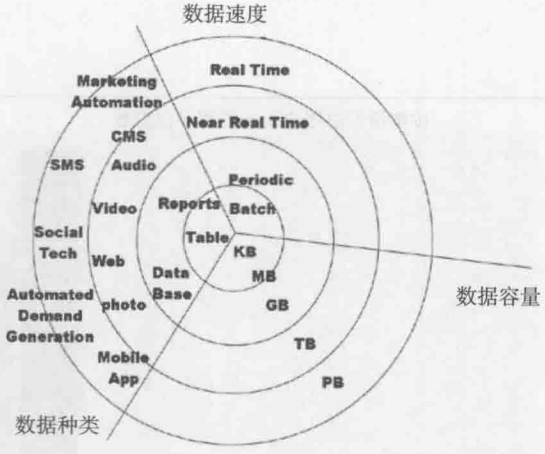
大数据意味着必须要处理结构化的、半结构化的和无结构的大型复杂数据集，并且通常会造成内存外溢。它们必须被适当执行，这是指计算必须发生在存在数据驻留等待被处理的过程中。当提到开发者，也就是那些实际建立大数据系统和应用的人时，我们能够更加清楚地理解他们所说的 3V 的含义。他们通常会提到大数据的 3V 模型，也就是速度、容量和种类。

速度是指在应用分析功能时的低延时和实时速度。这方面的一个典型例子是对一系列从一个社交网站上获得的连续数据流进行分析或者对不同来源的数据进行整合。

容量是指数据集的大小，其大小可能是 KB、MB、GB、TB 或 PB，这取决于那些生成或接受数据的应用程序的类型。

种类是指可以存储各种数据的类型，例如，文本、音频、视频和照片。

大数据通常将数据集的大小考虑在内。对于这样的系统，在企业规定的时限内处理此数据量不太可能。大数据的容量是一个持续性变化的指标，例如，2012 年一个数据集可以从几十 TB 变化到很多 PB。面对这个看似不可逾越的困难，全新的处理平台被称为大数据平台。



采用大数据的组织

采用大数据的民间组织如下：

- Facebook：它有 40PB 的数据并且日获取数据量为 100TB。
- Yahoo!：它有 60PB 的数据。

❑ Twitter：日获取数据量为 8TB。

❑ eBay：它有 40PB 的数据并且日获取数据量为 50TB。

有多大的数据量才能被视为大数据？这个问题对于所有公司各不相同。尽管现实往往是：一个公司的大数据对于另一个公司来说只是小数据，但是大数据也有一些共同点：内存溢出、磁盘溢出、有大量数据的快速涌入需要加以处理和将受益于分布式软件栈。对一些公司来说，10TB 的数据可能就被视为大数据，但对于其他公司来说，1PB 的数据可能才算大数据。所以只有你可以判断数据是否是真正的大数据，这足以说明，它将从 TB 级范围内开始。

此外，一个问题也值得一问，当你不能获取并存储足够的数据时，你是否认为你没有一个大数据的问题呢？在某些情况下，公司随便丢弃数据，因为没有一个符合成本效益的方式来存储和处理它。如果有 Hadoop 这样的平台，就可以开始获取和存储所有这样的数据。

Hadoop 简介

Apache Hadoop 是一个开源的 Java 框架，用于处理和查询存放在大型商用硬件集群上的大量数据。由雅虎和 Doug Cutting 发起和领导的 Hadoop 是一个最高级别的 Apache 项目。它的成功依赖于来自世界各地的社区志愿者的贡献。

基于 Yahoo! 提供的重要技术支持，Apache 的 Hadoop 已经成为一个准企业级的云计算技术。它正在成为行业对于大数据处理的事实框架。

Hadoop 改变了经济学和大规模计算的动态变化，其影响可以归结为四个显著的特点。Hadoop 是可扩展性、高性价比、灵活性和容错的解决方案。

Hadoop 特征探讨

Apache Hadoop 有两个显著特征：

❑ HDFS（Hadoop 分布式文件系统）

❑ MapReduce

学习 Hadoop 组件

Hadoop 包含一个构建在核心 HDFS 和 MapReduce 平台上的其他产品的生态系统，以启动各种类型的操作。一些流行的 Hadoop 组件如下：

❑ Mahout：这是一个广泛的机器学习算法库。

- Pig：Pig 是一种用来分析大型数据集的高级语言（如 Perl）。它的语法用于描述数据分析程序以及用于评估这些程序的组件。
- Hive：Hive 是一个服务于 Hadoop 的数据仓库系统，有利于数据统计、特殊查询和分析存储在 HDFS 上的大型数据集。它自己有类似 SQL 的查询语言——Hive Query Language（HQL），用于向 Hadoop 发出查询命令。
- HBase：HBase（Hadoop 数据库）是一个分布式、面向数据列的数据库。HBase 使用 HDFS 作为底层存储。它同时支持使用 MapReduce 和原子查询（随机读取）的批量式计算。
- Sqoop：Apache Sqoop 是一个专为 Hadoop 和结构化关系数据库间有效传输大量数据而设计的工具。Sqoop 是一个 (SQ)L 到 Had(oop) 的缩写。
- ZooKeeper：ZooKeeper 是一种留存配置信息、命名、提供分布式同步和集团服务的集中服务，这对各种分布式系统来说非常有用。
- Ambari：Ambari 是一个基于网络的工具，用于配置、管理、监控 Apache Hadoop 集群，其中包括为 Hadoop HDFS、Hadoop MapReduce、Hive、HCatalog、HBase、ZooKeeper、Oozie、Pig 和 Sqoop 提供支持。

同时使用 R 和 Hadoop 的原因

我还要指出，有时数据驻留在 HDFS 上（以多种格式）。由于用 R 所做的大量数据分析都是极有成效的，因此自然而然使用 R 来计算这些通过 Hadoop 相关工具存储的数据。

正如前面所提到的，R 的优势在于它具有丰富的封装库分析数据的能力，但当处理非常大的数据集时会功亏一篑。另外，Hadoop 的优势在于其能在 TB 甚至 PB 范围上存储和处理非常庞大的数据。如此庞大的数据集不能在内存中进行处理，因为每台机器的 RAM 都无法容纳这么庞大的数据集。可以采用的方法是在有限的块上进行分析，也称为采样或将 R 的分析能力与 Hadoop 的存储和处理能力对应起来，然后你就得到了一个理想的解决方案。这种方案也可以利用像 Amazon EMR 一样的平台在云端实现。

本书涵盖的内容

第 1 章介绍 R 和 Hadoop 的安装过程。

第 2 章包括 Hadoop MapReduce 的基础知识，用 Hadoop 执行 MapReduce 的方式。

第 3 章通过各种数据处理过程，展示 RHadoop 与 RHIPE 的开发和简单 MapReduce 程序的运行。

第 4 章展示如何用 R 运行 Hadoop 流。

第 5 章通过展示真实世界的数据分析问题来介绍数据分析项目的生命周期。

第 6 章包括通过 RHadoop 机器学习技术来执行大数据分析。

第 7 章包括如何与流行的关系数据库衔接，从而通过 R 导入和导出数据。

附录介绍了涉及所有章节内容的资源链接。

阅读本书的准备工作

当打算用 R 和 Hadoop 进行大数据分析时，你就应该有关于 R 和 Hadoop 以及如何上机实践的基本知识，并且需要把 R 和 Hadoop 安装、配置好。如果你已经有了一个大数据集和可以用数据驱动技术解决的问题，比如 R 和 Hadoop 的功能，那真是太好了。

本书读者对象

本书对于正在寻求用 Hadoop 这种方法执行大数据分析的 R 开发者是很有用的。他们希望了解整合 R 和 Hadoop 的所有技术、如何编写 Hadoop MapReduce、教程开发和在 R 上运行 Hadoop MapReduce。本书针对那些知道 Hadoop 并且想要用 R 程序包针对大数据建立一些智能应用的人。这本书对于那些拥有关于 R 基础知识的读者很有帮助。

下载示例代码

你可以从在 http://www.packtpub.com 上的账户下载你购买的所有 Packt 书籍的示例代码文件。如果你在其他地方购买这本书，你可以访问 http://www.packtpub.com/support，注册后，文件会直接通过电子邮件发送给你。

审校者简介 *About the Reviewers*

Krishnanand Khambadkone 有超过20年的综合经验。他最近在美国交通管制局大数据及Hadoop运作部门里担任解决方案高级设计师。他为世界500强客户设计并执行Hadoop解决方案，主要是大型银行机构。在此之前他主要的工作是利用Oracle中间件堆栈开发中间件和SOA解决方案，利用J2EE产品堆栈来开发软件。

他热衷于传播与大数据及Hadoop相关的知识。他已经在这个主题上撰写了数篇文章和白皮书，并在一些会议上发表演讲。

Muthusamy Manigandan 是Ozone Media公司的设计与架构主管。他在大规模软件系统设计的虚拟化、分布式版本控制系统、企业资源规划、供应链管理、机器学习和推荐引擎、基于行为的重新定位以及行为目标创造领域内有超过15年的经验。在加入Ozone Media公司前，他在VMware、Oracle、AOL和曼哈特联合软件公司中担任过重要职务。他在Ozone Media公司中负责产品、技术和研究计划。相关信息可以在mmaniga@yahoo.co.uk和http://in.linkedin.com/in/mmanigandan/上找到。

Vidyasagar N V 早年就对计算机技术很感兴趣。他在计算机和网络方面的尝试始于高中时代。之后他来到著名的贝拿勒斯印度教大学应用技术学院攻读技术学学士。他是一位软件开发人员、数据专家，主要开发和构建可升级系统。他曾使用各式第二、三、四代计算机语言，也曾从事与平面文件、索引文件、层次数据库、网络数据库、关系型数据库相关的工作，例如NoSQL数据库、Hadoop和相关技术。最近，他任职于Collective公司，作为高级开发人员利用网络与本地信息开发基于大数据的结构性数据析取技术。他热衷于开发高质量软件、基于网络的解决方案，以及设计安全、可升级的数据系统。

Siddharth Tiwari 在过去三年中一直从事机器学习、文本分析、大数据管理以及信息搜寻和管理方面的工作，从那时起他就已经加入了这一行业。最近，他任职于EMC公司的大数据管理、研究计划和产品工程部门，从事Hadoop配置工作。

在一家大型金融服务公司工作时，他参与完成了TeraSort和MinuteSort的世界纪录。

他在印度北方邦科技大学获得了技术学学士学位，平均绩点为8。

Acknowledgment 致　　谢

首先，我想感谢我挚爱的父母和弟弟 Vaibhav，感谢他们一直以来支持我的事业尤其是在我创作这本书时。如果没有他们的支持，这一份知识的共享就无法实现。当我开始写这本书时，我总会收到来自父亲（Prahlad Prajapati）的不断激励以及来自母亲（Dharmistha Prajapati）的定期关心与询问。同时，我也感谢我的朋友们，感谢他们鼓励我着手开始关于大技术（如 Hadoop 与 R）的写作。

在创作期间，我经历了人生中对我构成挑战的一些关键时期。我很感谢 Tatvic 的总裁和创始人 Ravi Pathak，他将我带进了机器学习和大数据这个宽广的领域，并帮助我意识到了自己的潜力。我也不能忘记 Packt 出版社的 James、Wendell 以及 Mandar，不能忘记他们宝贵的支持、激励以及指导，是他们帮助我抵达顶峰。我特别感谢他们填补了这本书在技术与图解部分之间的缺陷。

让我们一起步入大数据分析的未来吧！

目　录 *Contents*