Broadview®
www.broadview.com.cn

CDA数据分析师系列丛书

Spark大数据分析
技术与实战

经管之家 主编　董轶群 曹正凤 赵仁乾 王安 编著

Spark大数据分析技术全覆盖：Spark GraphX、Spark SQL、Spark Streaming和Spark MLlib；
零基础学习Spark大数据分析技术：有数据、有案例、有分析。

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

CDA数据分析师系列丛书

Spark大数据分析技术与实战

经管之家 主编　董轶群 曹正凤 赵仁乾 王安 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

Spark 作为下一代大数据处理引擎，经过短短几年的飞跃式发展，正在以燎原之势席卷业界，现已成为大数据产业中的一股中坚力量。

本书着重讲解了 Spark 内核、Spark GraphX、Spark SQL、Spark Streaming 和 Spark MLlib 的核心概念与理论框架，并提供了相应的示例与解析。

全书共分为 8 章，其中前 4 章介绍 Spark 内核，主要包括 Spark 简介、集群部署、工作原理、核心概念与操作等；后 4 章分别介绍 Spark 内核的核心组件，每章系统地介绍 Spark 的一个组件，并附以相应的案例分析。

本书适合作为高等院校计算机相关专业的研究生学习参考资料，也适合大数据技术初学者阅读，还适合所有愿意对大数据技术有所了解并想要将大数据技术应用于本职工作的读者阅读。

图书在版编目（CIP）数据

Spark 大数据分析技术与实战 / 经管之家主编；董轶群等编著. —北京：电子工业出版社，2017.7
（CDA 数据分析师系列丛书）
ISBN 978-7-121-31903-7

Ⅰ. ①S… Ⅱ. ①经… ②董… Ⅲ. ①数据处理软件—技术培训—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 133619 号

策划编辑：张慧敏
责任编辑：徐津平
特约编辑：顾慧芳
印　　刷：北京中新伟业印刷有限公司
装　　订：北京中新伟业印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：14.5　字数：330 千字
版　　次：2017 年 7 月第 1 版
印　　次：2017 年 7 月第 1 次印刷
定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

董轶群，吉林大学计算机科学与技术学院博士毕业。曾在吉林大学“符号计算与知识工程”教育部重点实验室从事空间关系建模研究，参与了多个国家自然科学基金重点项目与面上项目的申报与研究工作，并在项目中主要负责空间方向关系建模、空间拓扑关系建模的研究工作。

目前作为经管之家（原人大经济论坛）大数据讲师，主讲Spark、Hbase、Scala等大数据核心课程，并从事大数据相关的理论与应用研究工作。重点关注海量数据背景下空间关系建模与智能交通的结合研究，并在国内期刊和国际会议上发表了一系列相关理论的研究成果。

曹正凤，统计学博士，经管之家（原人大经济论坛）大数据中心总工程师，经管之家CDA大数据分析师培训负责人，北京博宇通达科技有限公司技术总监。致力于大数据分析前沿领域研究，主持首发集团智慧交通大数据中心建设项目，基于大数据平台的互联网金融风险监控系统项目，参与国家社科基金项目《基于大数据整合的空气质量测度方法研究》。

赵仁乾，北京邮电大学管理科学与工程硕士，现就职于北京电信规划设计院任高级经济师，从事移动、联通集团及各省分公司市场、业务、财务规划，经济评价及运营咨询。重点研究方向包括离网用户挖掘、市场细分与精准营销、移动网络价值区域分析、潜在价值客户挖掘等。

王安，布本智能首席数据官，北京大学光华管理学院MBA，北京大学商务智能中心专家组成员。专注数据化决策，互联网金融风险管理与精准营销。在数据决策领域拥有十多年的实践经验，曾服务多家大中型银行、保险公司及互联网金融公司。同时也积极参与数据决策教育领域，为北京大学、人民大学、北京航空航天大学、北京理工大学等院校机构提供相关课程和数据教育辅导。

前　言

随着电子信息、物联网等产业的高速发展，智能手机、平板电脑、可穿戴设备与物联网设备已经渗入到现代化生产与生活的各个方面，每时每刻产生着大量的数据，当今社会已经进入数据爆炸的时代。各领域中的相关数据不仅量大，而且种类繁多、变化速度快、价值密度低。这些日益凸显的大数据特征在全球范围内掀起了一场全新的思维、技术与商业变革，无论是产业界还是学术界都在持续加大在大数据技术和相关领域中的投入。“中国制造 2025”战略规划和“互联网+”概念的提出再次为国内大数据技术的发展注入了强劲的动力，大数据技术已被提升到了前所未有的高度，预示了其未来广阔的发展空间与应用前景。

在大数据背景下，各领域对数据相关服务的需求不断提升，迫切需要一种高效通用的大数据处理引擎。相对于第一代大数据生态系统 Hadoop 中的 MapReduce，Spark 是一种基于内存的、分布式的大数据处理引擎，其计算速度更快，更加适合处理具有较多迭代次数的问题；Spark 中还提供了丰富的 API，使其具有极强的易用性；与此同时，Spark 实现了“一栈式”的大数据解决方案，即在 Spark 内核基础上提出了 Spark GraphX、Spark Streaming、Spark MLlib、Spark SQL 等组件，使其不仅能够对海量数据进行批处理，同时还具备流式计算、海量数据交互式查询等功能，可以满足包括教育、电信、医疗、金融、电商、政府、智慧城市和安全等诸多领域中的大数据应用需求。

Spark 作为下一代大数据处理引擎，经过短短几年的飞跃式发展，正在以燎原之势席卷业界，现已成为大数据产业中的一股中坚力量。

本书主要针对大数据技术初学者，着重讲解了 Spark 内核、Spark GraphX、Spark SQL、Spark Streaming 和 Spark MLlib 的核心概念与理论框架，并提供了相应的示例与解析，以便读者能够尽快了解 Spark。

全书共分为 8 章，其中前 4 章介绍 Spark 内核，主要包括 Spark 简介、集群部署、工作原理、核心概念与操作等；后 4 章分别介绍 Spark 内核的核心组件，每章系统地介绍 Spark 的一个组件，并附以相应的案例分析。

- 第 1 章：Spark 导论。概述 Spark 的发展背景与起源，对比 MapReduce 介绍了 Spark 的特征、原理与应用场景等。
- 第 2 章：Spark 集群部署。该章详细介绍了 Ubuntu 下 Spark 集群的部署过程与注意事项，首先利用 VMware Workstation 搭建 Hadoop 分布式集群；然后在集群中安装 Scala；最后搭建 Standalone 模式的 Spark 集群。

- 第 3 章：RDD 编程。该章对 Spark 中的弹性分布式数据集（Resilient Distributed Dataset —RDD）这一核心概念进行了详细介绍，重点讲解了与之相关的定义、特征及其操作，并附以相应的示例与解析。
- 第 4 章：Spark 调度管理与应用程序开发。该章阐述了 Spark 底层的工作机制，介绍了 Spark 应用程序从产生作业到最终计算任务的整个流程；基于 IntelliJ IDEA 讲解了 Spark 应用程序的开发过程，并介绍了如何在本地与集群模式下提交运行 Spark 应用程序。
- 第 5 章：GraphX。该章介绍了 GraphX 的基本原理，着重讲解了 GraphX 中弹性分布式属性图的定义、表示模型、存储方式以及其上的丰富操作；以经典的 PageRank 与三角形计数等图计算算法为例，讲解了 GraphX 中相关接口的使用方法。
- 第 6 章：Spark SQL。该章包含了 Spark SQL 概述、SQL 语句的处理流程、DataFrame 数据模型的概念与相关操作等；并将 Spark SQL 与 Hive 相结合，给出了一个学生信息管理系统的设计与实现。
- 第 7 章：Spark Streaming。该章介绍了 Spark Streaming 的发展与应用场景以及批处理时间间隔、窗口间隔、滑动间隔等核心概念；着重讲解了 DStream 数据模型的概念与相关操作；针对不同应用场景下的流式计算需求，给出了有状态与无状态模式下的 Spark Streaming 应用案例与解析。
- 第 8 章：Spark MLlib。该章介绍了 Spark MLlib 中向量、LabeledPoint、矩阵等核心数据类型的定义与使用；详细介绍了机器学习中分类、回归、聚类、协同过滤等经典算法的 Spark 实现与应用，并附以相应的案例与解析。

由于时间短，加之笔者水平有限，书中难免有疏漏之处，敬请读者朋友批评指正。

编　者

2017 年 5 月

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/31903

CDA数据分析师

CDA(Certified Data Analyst)，亦称“CDA数据分析师”，指在互联网、零售、金融、电信、医学、旅游等行业专门从事数据的采集、清洗、处理、分析并能制作业务报告、提供决策的新型数据分析人才。CDA秉承着总结凝练最先进的商业数据分析实践为使命，明晰各类数据分析从业者的知识体系为职责，旨在加强全球范围内正规化、科学化、专业化的大数据及数据分析人才队伍建设，进一步提升数据分析师的职业素养与能力水平，促进数据分析行业的高质量持续快速发展。

“CDA数据分析师认证”是一套专业化，科学化，国际化，系统化的人才考核标准，分为CDA LEVEL Ⅰ，LEVEL Ⅱ，LEVEL Ⅲ，涉及金融、电商、医疗、互联网、电信等行业大数据及数据分析从业者所需要具备的技能，符合当今全球大数据及数据分析技术潮流，为各界企业、机构提供数据分析人才参照标准。经管之家为中国区CDA数据分析师认证考试唯一主办机构，于每年6月与12月底在全国范围举办线下数据分析师考试，通过考试者可获得CDA数据分析师认证证书。

“CDA数据分析师培训”是根据CDA数据分析师认证体系标准而设立的一套专业化、科学化、系统化的学习方案。培训内容不仅包含认证标准中的技能知识要求，还有着企业环境中的真实项目和案例，能满足不同层次的学员需求，使学员能学到真本事技能并能够落地运用，实现商业价值。

品读经典，分享精华
我们期待您的加入

投稿邮箱：
zhanghm@phei.com.cn

交流学习：

目　录

第1章

Spark导论

半个多世纪以来，人类社会正由工业社会全面进入信息社会，其主要动力来自于以计算机技术、通信技术和控制技术为核心的现代化信息技术的飞速发展和扩展应用。信息继物质、能源之后，成为人类社会的第三大资源，可以说信息维系并推动着社会与经济的生存与发展。

近年来，互联网+概念的提出为信息技术的发展注入了更强劲的动力。所谓“互联网+”是互联网思维的进一步实践与拓展，它代表一种先进的生产力，推动经济形态不断演变，从而为实体经济的改革、发展和创新提供了广阔的网络平台。“互联网+”利用信息通信技术把互联网和包括传统行业在内的各行各业结合起来，例如互联网+金融衍生出互联网金融；互联网+零售衍生出电子商务；互联网+制造业衍生出工业4.0等。但这些并不是简单的两者相加，而是利用信息通信技术以及互联网平台让互联网与传统行业进行深度融合，创造新的发展生态。由此可见，互联网+、工业4.0等这些概念都与信息技术有着最密切和最直接的关系。信息技术的发展已经提升到国家战略层面这一前所未有的高度。

一般来说，信息既是对各种事物的特征和变化的反映，又是事物之间相互作用和联系的表征。人们通过信息来认识事物。信息以数据作为载体，数值、文字、语言、图形、图像等都是不同形式的数据。所以信息技术所处理的最主要对象就是数据。目前我们正处在一个数据爆炸的时代，大量涌现的智能手机、平板电脑、可穿戴设备及物联网设备每时每刻都在产生新的数据。据统计，有90%的数据是在过去短短两年内产生的，到2020年，将有500多亿台的互联设备产生Zeta字节级的数

据。在这一背景和趋势下，产业界将数据驱动的发展策略逐渐提升到前所未有的高度；在传媒、金融、电信等众多传统领域中，以往沉积的“垃圾”数据被重新审视，如何利用大数据分析技术“变废为宝”已经成为企业进行创新与盈利的一种重要模式；在学术界，国内外越来越多的高校和研究机构在云计算和大数据相关领域开展了深入的研究工作。然而在大数据时代，带来革命性改变的并非仅仅数据量本身，还有数据的种类、变化速度和价值密度。随着大数据特征的日益凸显，传统的信息技术已经难以准确高效地处理大数据背景下的数据采集、表示、存储、传输、处理等一系列问题。因此大数据相关技术是当前信息技术发展中亟需解决的关键技术。

综上所述，信息技术是推动未来社会发展的一种主流技术，信息技术处理的对象是数据，而当前的数据具有大数据特征，所以大数据技术成为了信息技术的重中之重。本书介绍的 Spark 以其卓越的性能，是目前大数据领域中最具潜力和影响力的处理引擎。

1.1 Spark 的发展

第一代大数据生态系统 Hadoop 已经非常成功，其采用 HDFS 实现分布式存储，使用 MapReduce 进行分布式计算。MapReduce 是一个简单通用和自动容错的批处理计算模型，不仅极大地简化了并行程序的开发过程，而且提高了程序执行的效率。然而在 MapReduce 任务内部，为了防止 Reduce 任务的失败，Map 通常会把结果存储在磁盘上。由于 MapReduce 每次都需要将中间数据写回磁盘，导致网络通信、磁盘 I/O 等消耗了大量的系统资源，尤其是在处理具有较多迭代次数的计算任务时，这一缺点尤为突出。因此 MapReduce 的运算性能仍难以满足面向大数据的交互式查询、迭代计算、流式计算等方面的需求。为了弥补 MapReduce 这一不足，涌现出了一系列专用的数据处理模型，例如 Storm、Impala、GraphLab 等。随着新模型的不断出现，不同类型的作业需要一系列不同的处理框架才能很好地完成，然而这无形中又增加了系统在部署、测试、运维等方面的成本。

针对 MapReduce 及各种专用数据处理模型在计算性能、集成性、部署运维等方面的问题，2009 年美国加洲大学伯克利分校开始研发全新的大数据处理框架，即 Spark。2010 年，Spark 实现开源。自从 2013 年 Spark 进入 Apache 的孵化器项目后，发生了翻天覆地的变化。2014 年初，Spark 成为了 Apache 排名第三的顶级项目，其发展势头更加迅猛，一个多月左右就会发布一个小版本，两三个月左右会发布一个大版本，2015 年 6 月份发布了 1.4.0，2015 年 9 月份发布了 1.5.0，至本书作者执笔时已经发布到 2.0。

Spark 的迅猛发展和其突出的计算性能引起了大数据处理相关领域的广泛关注，使得其迅速成为了业内一门具有强劲竞争力的热门技术。在短短的三年时间里，世界各地开设了多次 Spark 主题峰会，反映出 Spark 技术的前沿性与火爆的发展势头。在 2014 年的 Spark 峰会上，Hadoop 三大发行商均声称未来将会把精力投入到 Spark 的研究中；Yahoo 有全世界最强大的 Hadoop 集群（Hadoop 的 70%是由 Yahoo 贡献的），但早在几年前 Yahoo 便已经开启了 Spark 的研发工作；Intel、IBM 等大公司纷纷宣布其产品支持 Spark；亚马逊完全基于 Spark 搭建了云服务平台；Google、Facebook 也陆续开展了转向 Spark 框架的研发工作。

同样也是在 2014 年 Spark 在中国的发展达到了一个前所未有的火爆状态。国内许多重量级的数据企业纷纷搭建了自己的 Spark 集群。例如 2015 年百度搭建了一个大规模的 Spark 集群，其中最大单集群规模达上千台节点，包含了数万个核心和上百 TB 的内存，与此同时公司内部还运行着大量的小型 Spark 集群。淘宝的推荐系统已经用 Spark 取代了部分的 MapReduce。腾讯通过 Spark 对数据实时采集、算法实时训练、系统实时预测，进而实现了大数据背景下的精准推荐。目前国内最大规模的 Spark 集群来自于腾讯，其中包含了 8000 个节点，而最大的单个 Job 则来自于阿里巴巴。虽然上述集群中节点数目看似不多，但是 1000 个节点的 Spark 集群，其性能相当于包含 5000 个节点的 Hadoop 集群。

1.2 什么是 Spark

Apache 官网对 Spark 定义如下：

“Apache Spark is a fast and general engine for large-scale data processing”

由此可见，Spark 是一个快速通用的大规模数据处理引擎。Spark 的功能看似与 Hadoop 相同，但是两者却有着明显的区别。Hadoop 是一个开源分布式计算平台，是第一代大数据生态系统，也是目前应用最为广泛的大数据生态系统。Hadoop 以分布式文件系统 HDFS 和分布式计算 MapReduce 为核心，为用户提供了系统底层细节透明的分布式基础框架。HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上；MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。因此用户利用 Hadoop 能够轻松地组织计算机资源，从而搭建自己的分布式计算平台，并且可以充分利用集群的计算和存储能力完成海量数据的处理。Spark 作为大数据处理引擎与 Hadoop 有着密切的联系，但是两者并非是一个层面的概念。Hadoop 不仅有数据的处理还有数据的存储，而 Spark 仅仅是大数据处理框架。因此 Spark 其实与 Hadoop 上的 MapReduce 是一个层面的概念。目前的 Hadoop 已经趋于完善，其中的 Yarn 与 HDFS 非常经典，已经成为了业内大数据存储和分布式资源管理的标准，在未来短时间内难以被轻易取代。很多资料均预言 Spark 将取代 Hadoop，其实是指取代 Hadoop 中的 MapReduce。

1.3 Spark 主要特征

Spark 是一种基于内存的、分布式的、大数据处理框架，它与 Hadoop 上的 MapReduce 是一个层面上的概念，这意味着两者在诸多方面存在着竞争与可比性。本节将通过与 MapReduce 的对比分析来介绍 Spark 的主要特征。

1.3.1 快速

面向磁盘的 MapReduce 受限于磁盘读/写性能和网络 I/O 性能的约束，在处理迭代计算、实时计

算、交互式数据查询等方面并不高效，但是这些却在图计算、数据挖掘和机器学习等相关应用领域中非常常见。针对这一不足，将数据存储在内存中并基于内存进行计算是一个有效的解决途径。Spark 是面向内存的大数据处理引擎，这使得 Spark 能够为多个不同数据源的数据提供近乎实时的处理性能，适用于需要多次操作特定数据集的应用场景。

在相同的实验环境下处理相同的数据，若在内存中运行，那么 Spark 要比 MapReduce 快 100 倍，如图 1.1 所示；在磁盘中运行时 Spark 要比 MapReduce 快 10 倍，如图 1.2 所示。综合各种实验表明，处理迭代计算问题 Spark 要比 MapReduce 快 20 多倍，计算数据分析类报表的速度可提高 40 多倍，能够在 5~7 秒的延时内交互式扫描 1TB 数据集。

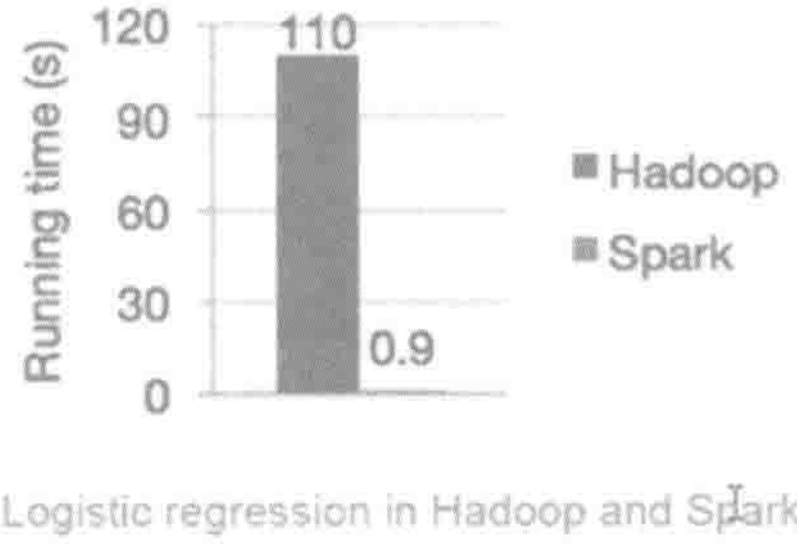


图 1.1　基于内存 Spark 与 MapReduce 执行逻辑回归的性能对比

排序问题是最考验系统性能的问题之一。图 1.2 是 Spark 与 MapReduce 对相同的 100TB 数据样本排序的性能对比。在实验中，MapReduce 用了 2100 台节点，用时 72 分钟；而 Spark 仅用 207 台节点，是前者的 1/10，用时 23 分钟，是前者的 1/3。

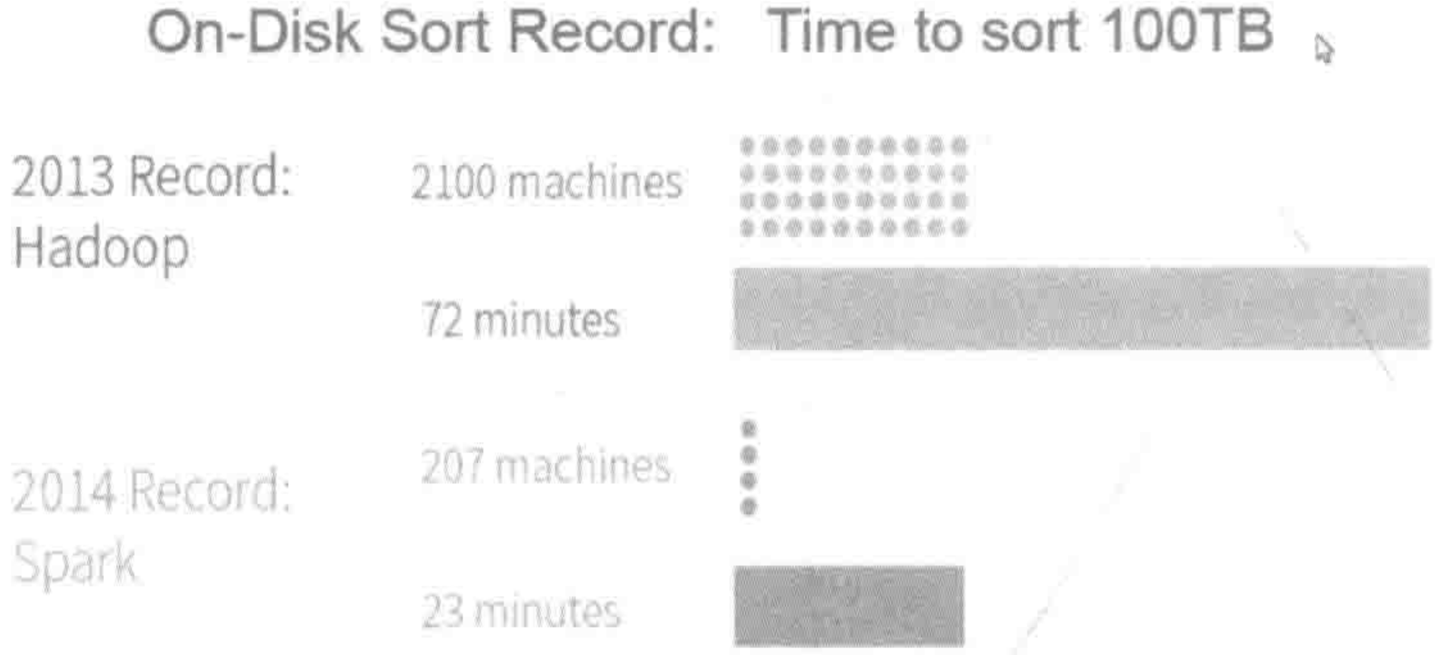


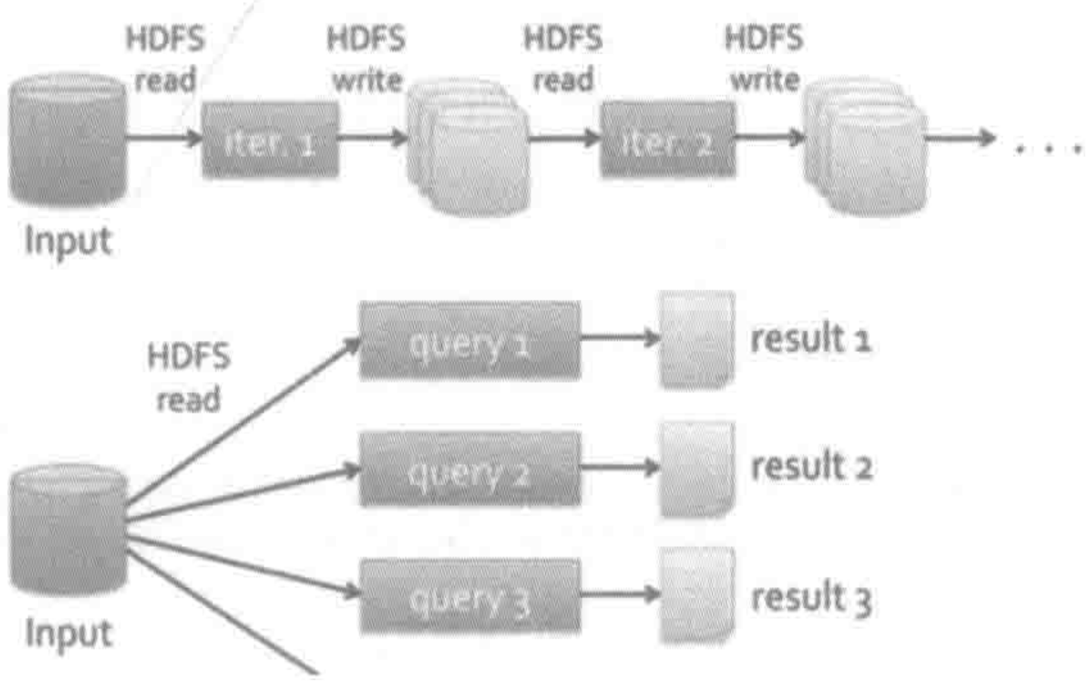
图 1.2　基于磁盘 Spark 与 MapReduce 对 100T 数据排序的性能对比

Spark 与 MapReduce 相比在计算性能上有如此显著的提升，主要得益于以下两方面。

1．Spark 是基于内存的大数据处理框架

Spark 既可以在内存中处理一切数据，也可以使用磁盘来处理未全部装入到内存中的数据。由于

内存与磁盘在读/写性能上存在巨大的差距，因此 CPU 基于内存对数据进行处理的速度要快于磁盘数倍。然而 MapReduce 对数据的处理是基于磁盘展开的。一方面，MapReduce 对数据进行 Map 操作后的结果要写入磁盘中，而且 Reduce 操作也是在磁盘中读取数据，另一方面，分布式环境下不同物理节点间的数据通过网络进行传输，网络性能使得该缺点进一步被放大。因此磁盘的读/写性能、网络传输性能成为了基于 MapReduce 大数据处理框架的瓶颈。图 1.3 为 MapReduce 数据处理流程示意图。



（a） MapReduce 处理流程

图 1.3　MapReduce 数据处理流程示意图

2. Spark 具有优秀的作业调度策略

Spark 中使用了有向无环图（Directed Acyclic Graph，DAG）这一概念。一个 Spark 应用由若干个作业构成，首先 Spark 将每个作业抽象成一个图，图中的节点是数据集，图中的边是数据集之间的转换关系；然后 Spark 基于相应的策略将 DAG 划分出若干个子图，每个子图称为一个阶段，而每个阶段对应一组任务；最后每个任务交由集群中的执行器进行计算。借助于 DAG，Spark 可以对应用程序的执行进行优化，能够很好地实现循环数据流和内存计算。

1.3.2　简洁易用

Spark 不仅计算性能突出，在易用性方面也是其他同类产品难以比拟的。一方面，Spark 提供了支持多种语言的 API，如 Scala、Java、Python、R 等，使得用户开发 Spark 程序十分方便。另一方面，Spark 是基于 Scala 语言开发的，由于 Scala 是一种面向对象的、函数式的静态编程语言，其强大的类型推断、模式匹配、隐式转换等一系列功能结合丰富的描述能力使得 Spark 应用程序代码非常简洁。Spark 的易用性还体现在其针对数据处理提供了丰富的操作。在使用 MapReduce 开发应用程序时，通常用户关注的重点与难点是如何将一个需求 Job（作业）拆分成 Map 和 Reduce。由于 MapReduce 中仅为数据处理提供了两个操作，即 Map 和 Reduce，因此系统开发人员需要解决的一个难题是如何把数据处理的业务逻辑合理有效地封装在对应的两个类中。与之相对比，Spark 提供了 80 多个针对数据处理的基本操作，如 map、flatMap、reduceByKey、filter、cache、collect、textFile 等，这使得用户基于 Spark 进行应用程序开发非常简洁高效。以分词统计为例，虽然 MapReduce 固