辽宁科协资助

·娄岩 编著

大数据技术应用导论

Introduction to Big Data Technology Application

辽宁科学技术出版社
LIAONING SCIENCE AND TECHNOLOGY PUBLISHING HOUSE

大数据技术应用导论

娄岩　编著

辽宁科学技术出版社
沈　阳

图书在版编目（CIP）数据

大数据技术应用导论/娄岩编著. —沈阳：辽宁科学技术出版社，2017.7

（辽宁省优秀自然科学著作）

ISBN 978-7-5591-0232-4

Ⅰ. ①大… Ⅱ. ①娄… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 092086 号

出版发行：辽宁科学技术出版社

（地址：沈阳市和平区十一纬路 25 号 邮编：110003）

印 刷 者：沈阳天正印刷厂

幅面尺寸：185 mm×260 mm

印 张：9.75

字 数：210 千字

印 数：1~1000

出版时间：2017 年 7 月第 1 版

印刷时间：2017 年 7 月第 1 次印刷

责任编辑：郑 红

策划编辑：陈广鹏

封面设计：嵘 嵘

责任校对：栗 勇

书 号：ISBN 978-7-5591-0232-4

定 价：30.00 元

联系电话：024-23284526

邮购热线：024-23284502

http://www.lnkj.com.cn

前 言

如今，大数据浪潮正在迅速向我们涌来，并将触及各个行业和生活的许多方面。大数据浪潮将比之前发生过的浪潮更大、触及面更广，给人们的工作和生活带来的变化和影响也更大。

毋庸置疑，大数据的应用激发了一场思想风暴，也悄然改变了我们的生活方式和思维习惯。大数据正以前所未有的速度颠覆人们探索世界的方法，引起工业、商业、医学、军事等领域的深刻变革。因此，在当前大数据浪潮的猛烈冲击下，人们迫切需要充实和完善自己原有的IT知识结构，掌握两种全新的技能：一是掌握大数据基本技术与应用，使大数据成为我们所用的技能；二是掌握数据之间隐藏的规律与关系以及可视化方法，使大数据更好地服务于社会发展的技能。

本书注重实用性，围绕大数据及其相关技术这一主题，采用深入浅出、图文并茂的叙述方式，简明扼要地阐述了大数据及其相关技术的基本理论和发展趋势，使广大读者通过阅读本书，深入了解和掌握大数据的理论和应用，从而更好地把握时代发展的脉搏和历史赋予的机遇。

本书的目标是给广大读者提供一本既通俗易懂又具有严谨、完整、结构化特征的图书。其独到之处是，既阐明了大数据技术的系统性和理论性，又对传统数据和大数据在来源、结构、特征、存储方式、使用方法等方面，通过大量的表格和图形方式进行了有针对性的对比和阐述，使读者对两者之间的区别一目了然，对理解和掌握大数据理论技术具有事半功倍的效果。另外，考虑到大数据技术涉及许多新名词和专业性极强的词汇，故在全书的每一章中均附有相关术语的注释，从而方便读者查阅和自学。

本书还力求将大数据技术晦涩难懂的理论知识以通俗易懂的语言和方式，由浅入深地展现在读者面前，便于读者理解和掌握。本书内容重点突出、语言精练易

懂，非常便于自学，可作为想了解、使用大数据技术的相关人员，如工程技术人员、IT工作者、企业策划和管理人员的参考书，也可作为相关学习班的培训教材。

本书在写作过程中参阅了大量的中外书籍和相关资料，在此对各位作者表示真诚的谢意。本书得到了中国医科大学沙宪政教授和东北大学杨广明教授的大力支持，特此感谢！由于水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评斧正。

作　者

2016年9月

目 录

1 大数据概论

大数据（Big Data）不仅仅是一个新概念，也是一种技术，一种针对企业问题的解决方案，更是迄今为止知识界对数据形态最彻底的诠释。几千年来，人类一直利用数据，遗憾的是数据的真正价值，特别是大数据价值最近几年才成为人们关注的焦点。传统数据库技术，在数据处理的能力上有着难以逾越的局限性，超过100 T量级的数据，要么处理效率急剧下降，要么成本十分昂贵。所以大量的历史数据与过程数据，按传统的IT技术，既无法存储，也无法处理。大数据的出现引领我们从对数据简单的理解和一成不变的处理方法走向探索多极和无穷数据的奥秘，不再满足于用传统的方法认识和处理数据，即形成一种崭新的认识和处理数据信息的宇宙观。最终将我们身边无所不在、形态各异的隐秘数据信息可视化地呈现出来，并充分地加以利用。那么，什么是大数据，其相关技术、应用领域以及未来的发展趋势等将是本章重点介绍的内容。

1.1 大数据概述

就传统IT业来看，其结构化和非结构化的数据增长惊人，如非结构化数据，由2005年22 EB到2015年增至1 600 EB，年复合增长率约60%，远快于摩尔定律。大数据是指当传统的数据分析和处理技术对今天我们所认识的许多数据黔驴技穷时所使用的处理过程。如，数据是半结构化、非结构化、实时性强或信息量巨大，以至于无法通过关系数据库引擎进行处理，需要新的技术手段和具有分布式处理数据功能的并行硬件设备来实现。

1.1.1 大数据的发展简史

大数据的出现可以说是IT产业发展的链条之一和必然趋势。而IT产业发展改变了已有的社会秩序，重新定义了处理问题的规范，并为进入大数据时代的新纪元铺平了道路。

20世纪60年代和70年代的大型机阶段是以Burroughs，Univac，NCR，Control Data和Honeywell等公司为首的。在步入80年代后，小型机涌现出来，这时为首的公司包括DEC，IBM，Data General，Wang，Prime等。在90年代，IT产业进入了微处理器或个人计算机阶段，领先者为Microsoft（微软），Intel，IBM和Apple等公司。从90年代中期开始，IT产业进入了网络化阶段。如今，全球在线的人数已经

超过了 10 亿，这一阶段由 Cisco，Google，Oracle，EMC，Salesforce. com 等公司领导。IT 产业的下一个阶段人们更愿意称其为云计算/大数据阶段。

大数据快速增长的原因之一是智能设备的普及，如传感器、医疗设备及智能建筑（如楼宇和桥梁）。此外，非结构化信息，如文件、电子邮件和视频，将占未来 10 年新生数据的 90%。非结构化信息增长的另一个原因是高宽带数据的增长，如视频。用户的手机和移动设备是数据量爆炸的一个重要原因。

对于信息社会上的每一个人而言，只要看看周围正在变化的一切，就可以知道，大数据对每个人的重要性不亚于人类初期对火的使用。大数据让人类对一切事物的认识回归本源，其通过影响经济生活、政治博弈、社会管理、军事、文化教育科研、医疗、保健、休闲等行业，与每个人产生密切的联系。

大数据时代已悄然来到我们身边，并渗透到我们每个人的日常生活之中，谁都无法回避。它提供了光怪陆离的全媒体、自媒体、难以琢磨的云计算、无法抵御的虚拟仿真环境和随处可在的网络服务。它再也不仅仅是人们津津乐道的一种时尚，而是成为生活上的向导和助手。

早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。从技术层面上看，大数据是无法用单台计算机进行处理的，必须采用分布式计算架构。其特色在于对海量数据的挖掘，但它又必须依托一些现有的数据处理方法，如流式处理、分布式数据库、云存储与虚拟化技术（图 1-1）。

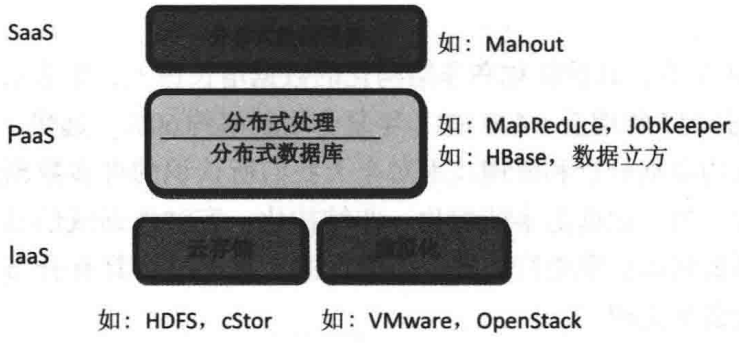

图 1-1 大数据与云技术

当前大数据的处理技术，特别是云存储与云计算技术的成熟应用，为大数据的存储和处理提供了技术可能性。企业可以利用生产系统以及管理系统中产生的大量数据，对海量的数据进行存储和挖掘分析。值得注意的是，大数据和云计算之间的区别在于：首先，大数据和云计算在概念上不同，云计算改变了 IT，而大数据改变了业务。其次，大数据和云计算的目标受众不同，如在一家公司中，那么云计算就是技术层，大数据就是业务层。但需要指出的是，大数据对云计算有一定的依赖性。

1.1.2 大数据的来源

大数据的来源非常多，如信息管理系统、网络信息系统、物联网系统、科学实

验系统等，其数据类型包括结构化数据、半结构化数据和非结构化数据。

1.1.2.1 信息管理系统

企业内部使用的信息系统，包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据，其产生的大数据大多数为结构化数据，通常存储在数据库中，一般为关系型数据。

1.1.2.2 网络信息系统

基于网络运行的信息系统即网络信息系统是大数据产生的重要方式，如电子商务系统、社交网络、社会媒体、搜索引擎等都是常见的网络信息系统。网络信息系统产生的大数据多为半结构化或非结构化的数据，在本质上，网络信息系统是信息管理系统的延伸，专属于某个领域的应用，具备某个特定的目的。因此，网络信息系统有着更独特的应用。

1.1.2.3 物联网系统

物联网是新一代信息技术，其核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络，其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信，而其具体实现是通过传感技术获取外界的物理、化学、生物等数据信息。

1.1.2.4 科学实验系统

主要用于科学技术研究，可以由真实的实验产生数据，也可以通过模拟方式获取仿真数据。

1.1.3 大数据的产生方式

从数据库技术诞生以来，产生大数据的方式主要有 3 种。

1.1.3.1 被动式生成数据

数据库技术使得数据的保存和管理变得简单，业务系统在运行时产生的数据可以直接保存到数据库中，由于数据是随业务系统运行而产生的，因此该阶段所产生的数据是被动的。

1.1.3.2 主动式生成数据

物联网的诞生，使得移动互联网的发展大大加速了数据的产生。例如，人们可以通过手机等移动终端，随时随地产生数据。用户数据不但大量增加，同时用户还主动提交了自己的行为，使之进入了社交、移动时代。大量移动终端设备的出现，使用户不仅主动提交自己的行为，还和自己的社交圈进行了实时互动，因此数据大量地产生出来，且具有极其强烈的传播性。显然如此生成的数据是主动的。

1.1.3.3 感知式生成数据

物联网的发展使得数据生成方式得以彻底的改变。例如遍布在城市各个角落的摄像头等数据采集设备源源不断地自动采集并生成数据。

大数据技术的核心在于为客户从数据中挖掘出蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式和数据产生方式的研究和探

索将是大数据产业健康发展的关键。

1.1.4 大数据与传统数据的区别

在大数据背景下，数据的采集、分析、处理较之传统方式有了颠覆性的改变，如表 1-1 所示。

表 1-1 传统数据与大数据的特点比较

项目	传统数据	大数据
数据产生方式	被动采集数据	主动生成数据
数据采集密度	采样密度较低，采样数据有限	利用大数据平台，可对需要分析事件的数据进行密度采样，精确获取事件全局数据
数据源	数据源获取较为孤立，不同数据之间添加的数据整合难度较大	利用大数据技术，通过分布式技术、分布式文件系统、分布式数据库等技术对多个数据源获取的数据进行整合处理
数据处理方式	大多采用离线处理方式，对生成的数据集中分析处理，不对实时产生的数据进行分析	较大的数据源、响应时间要求低的应用可以采取批处理方式集中计算；响应时间要求高的实时数据处理采用流处理的方式进行实时计算，并通过对历史数据的分析进行预测分析

大数据的意义在于可以通过人类日益普及的网络行为附带生成，并被相关部门、企业所采集，蕴含着数据生产者的真实意图、喜好，其中包括传统结构和非传统结构的数据。

1.1.5 大数据处理流程

大数据的处理流程可以定义为在适合工具的辅助下，对不同结构的数据源进行抽取和集成，结果按照一定的标准统一存储，利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展示给终端用户。大数据处理的基本流程如图 1-2 所示。

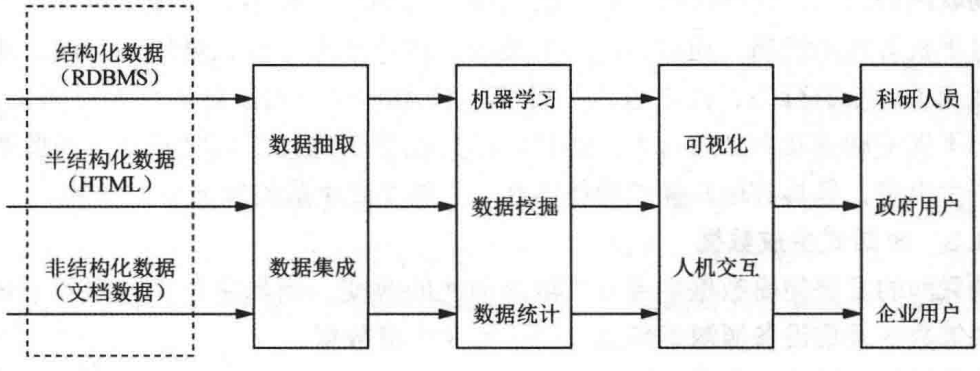

图 1-2 大数据处理的基本流程

1.1.5.1 数据抽取与集成

这是一个数据准备阶段。其中包括 3 个部分：数据集成、数据选择和预处理。由于大数据处理的数据来源类型广泛，而其第一步是对数据进行抽取和集成，从中找出关系和实体，经过关联、聚合等操作，再按照统一的格式对数据进行存储，现有的数据抽取和集成引擎有 3 种：基于物化或 ETL 方法的引擎、基于中间件的引擎、基于数据流方法的引擎。

1.1.5.2 大数据分析

大数据分析是指对规模巨大的数据进行分析。大数据分析是大数据处理流程的核心步骤。通过抽取、集成和预处理环节，从不同结构的数据源中获得用于大数据处理的原始数据，用户根据需求对数据进行分析处理，如数据挖掘、机器学习、数据统计，数据分析可以用于决策支持、商业智能、推荐系统、预测系统等。

1.1.5.3 数据可视化

用户最关心的是数据处理的结果及以何种方式在终端上显示结果。因此采用什么方式展示处理结果非常重要。就目前来看，可视化和人机交互是数据解释的主要技术。

数据可视化主要是借助于图形化手段，清晰有效地传达与沟通信息。数据可视化技术的基本思想，是将数据库中每一个数据项作为单个图元元素表示，大量的数据集合构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入地观察和分析。而使用可视化技术可以将处理结果通过图形方式直观地呈现给用户，如标签云、历史流、空间信息等；人机交互技术可以引导用户对数据进行逐步分析，参与并理解数据分析结果。

1.1.6 大数据的信息形态

从 IT 角度来看，信息结构类型大致经历了 3 个阶段。必须注意的是，旧的阶段仍在不断发展，如关系数据库的使用。因此 3 种数据结构类型一直存在，只是其中一种结构类型往往主导其他结构。

1.1.6.1 结构化信息

这种信息可以在关系数据库中找到，多年来一直主导着 IT 应用，是关键任务 OLTP 系统业务所依赖的信息。另外，这种信息还可对结构数据库信息进行排序和查询。

1.1.6.2 半结构化信息

包括电子邮件、文字处理文件及大量保存和发布在网络上的信息。半结构化信息是以内容为基础的，可以用于搜索，这也是 Google（谷歌）等搜索引擎存在的理由。

1.1.6.3 非结构化信息

该信息在本质形式上可认为主要是位映射数据。数据必须处于一种可感知的形

式中（如可在音频、视频和多媒体文件中被听或看到）。许多大数据都是非结构化的，其庞大的规模和复杂性需要高级分析工具来创建或利用一种更易于人们感知和交互的结构。

总而言之，大数据的结构就体现了它最突出的特征。结构化数据，一般包括预定义的数据类型、格式和结构的数据，可进行事务性数据和联机分析处理。半结构化数据具有可识别的模式并可以解析的文本数据文件，如自描述和具有定义模式的XML数据文件。非结构化数据是那些数据结构不固定，按不同类型文档保存的数据信息，如TXT文本文档、PDF文档、图像和视频等。

1.1.7 大数据的基本特征

大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像MapReduce那样的框架来向数十、数百或甚至数千的计算机分配工作。简言之，从各种各样类型的数据中，快速获得有价值信息的能力，就是大数据技术。大数据呈现出“4V1O”的特征。

1.1.7.1 数据量大（Volume）

是大数据的首要特征，包括采集、存储和计算的数据量非常大。大数据的起始计量单位至少是100TB。通过各种设备产生的海量数据，其数据规模极为庞大，远大于目前互联网上的信息流量，PB级别将是常态。

1.1.7.2 多样化（Variety）

表示大数据种类和来源多样化，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，多样化对数据的处理能力提出了更高的要求，编码方式、数据格式、应用特征等多个方面都存在差异性，多信息源并发形成大量的异构数据。

1.1.7.3 数据价值密度化（Value）

表示大数据价值密度相对较低，需要很多的过程才能挖掘出来。随着互联网和物联网的广泛应用，信息感知无处不在，信息量大，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。

1.1.7.4 速度快，时效高（Velocity）

随着互联网的发展，数据的增长速度非常快，处理速度也较快，时效性要求也更高。例如，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实时完成推荐，这些都是大数据区别于传统数据挖掘的显著特征。

1.1.7.5 数据是在线的（On-Line）

表示数据必须随时能调用和计算。这是大数据区别于传统数据的最大特征。现在谈到的大数据不仅大，更重要的是数据是在线的，这是互联网高速发展的特点和趋势。例如好大夫在线，患者的数据和医生的数据都是实时在线的，这样的数据才有意义。如果把他们放在磁盘中或者是离线的，显然这些数据远远不及在线的商业

价值大。

总之，无所遁形的大数据时代已经到来，了解大数据的基本特征，并快速渗透到每个职能领域，如何借助大数据持续创新发展，使企业成功转型，具有非凡的意义。

1.1.8 大数据的应用领域

大数据在社会生活的各个领域得到了广泛的应用，如科学计算、金融、社交网络、移动数据、物联网、医疗、网页数据、多媒体、网络日志、RFID 传感器、社会数据、互联网文本和文件，互联网搜索索引，呼叫详细记录，天文学、大气科学、基因组学、生物和其他复杂或跨学科的科研、军事侦察、医疗记录，摄影档案馆视频档案，大规模的电子商务等。不同领域的大数据应用具有不同特点，其响应时间、稳定性、精确性的要求各不相同，解决方案也层出不穷，其中最具代表性的有 Informatica Cloud 解决方案、IBM 战略、Microsoft 战略、京东框架结构等，对此我们将在后续章节中讨论。

1.2 大数据的技术架构

大数据基础架构必须具有分布式计算能力，如此在接近用户的位置进行数据分析，从而减少跨越网络引起的延迟。云计算模式为大数据成功提供了很好的条件，以实现大数据分析所需要的效率、可扩展性、数据便携性和经济性。另外还可以用来跨越不相干的数据源比较不同类型的数据和进行模式匹配。正因为如此，大数据分析能够以全新的视角挖掘企业传统数据和前所未有的数据洞察力。基于上面考量，大数据可以采用 4 层堆栈技术架构。

1.2.1 基础层

第一层，为大数据技术架构的最底层，其特点是虚拟化、网络化，分布式可横向扩展的体系结构。其作用是可将过去的存储孤岛发展为具有共享能力的高容量存储池。

1.2.2 管理层

本层主要负责数据的存储、管理和计算。可处理结构化数据和非结构化数据，具有并行处理和线性可扩展性。大数据架构中需要一个管理平台，使得结构化数据和非结构化数据能够一体化管理，具备实时传递、查询和计算功能。

1.2.3 分析层

主要用于大数据分析，其特点是可提供自助服务和实时协作。分析层提供基于

统计学的数据挖掘和机器学习算法，用于分析和解释数据集，帮助企业获得对数据价值深入的认知度。

1.2.4 应用层

大数据的价值在于帮助企业决策，以及为终端用户提供服务。因此该层主要是提供实时决策，内置预测能力，利用数据驱动经济，使数据信息等同货币流通。大数据应用对其技术不断推出新的要求，而大数据技术也在不断发展变化中成熟起来。

总之，拥有数据并非坐拥金矿，数据的产生和存储必然要付出相应的成本代价。大数据只有通过数据分析和挖掘，发现知识和解决办法才能创造价值。大数据挖掘的应用将总结事物的发展规律，提升生产与管理活动的科学性，减少传统方式下的研究和探索成本，提高社会发展的总生产效率。

1.3 大数据分析的5种常见工具

大数据分析是在研究大量的数据的过程中寻找模式、相关性和其他有用的信息，以帮助企业更好地适应变化，并做出更明智的决策。

1.3.1 Hadoop

Hadoop是一个能够对大量数据进行分布式处理的软件框架，是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的应用程序。

1.3.1.1 高可靠性

Hadoop按位存储和处理数据的能力值得人们信赖。

1.3.1.2 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。

1.3.1.3 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。

1.3.1.4 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分配。

Hadoop带有用Java语言编写的框架，因此运行在Linux平台上是非常理想的。Hadoop上的应用程序也可以使用其他语言编写，如C++。

1.3.2 Spark

Spark是一个基于内存计算的开源集群计算系统，目的是更快速地进行数据分

析。Spark 由加州伯克利大学 AMP 实验室 Matei 为主的小团队使用 Scala 开发，其核心部分的代码只有 63 个 Scala 文件，非常轻量级。Spark 提供了与 Hadoop 相似的开源集群计算环境，但基于内存和迭代优化的设计，Spark 在某些工作负载表现更优秀。在 2014 年上半年，Spark 开源生态系统得到了大幅增长，已成为大数据领域最活跃的开源项目之一。那么 Spark 究竟以什么吸引了如此多的关注？

1.3.2.1 轻量级快速处理

着眼大数据处理，速度往往被置于第一位。Spark 允许 Hadoop 集群中的应用程序在内存中以 100 倍的速度运行，即使在磁盘上运行也能快 10 倍。Spark 通过减少磁盘 IO 来达到性能提升，它们将中间处理数据全部放到了内存中。

Spark 使用了 RDD（Resilient Distributed Dataset）的理念，这允许它可以在透明的内存中存储数据，只在需要时才持久化到磁盘。这种做法大大地减少了数据处理过程中磁盘的读写，大幅度地降低了所需时间。

1.3.2.2 易于使用，支持多语言

自带 80 多个高等级操作符，允许在 Shell 中进行交互式查询。

1.3.2.3 支持复杂查询

在简单的“Map”及“Reduce”操作之外，Spark 还支持 SQL 查询、流式查询及复杂查询。同时，用户可以在同一个工作流中无缝地搭配这些能力。

1.3.2.4 实时的流处理

对比 MapReduce 只能处理离线数据，Spark 支持实时的流计算。

1.3.2.5 可以与 Hadoop 和已存 Hadoop 数据整合

Spark 可以独立地运行，除了可以运行在当下的 Yarn 集群管理之外，它还可以读取已有的任何 Hadoop 数据。

1.3.2.6 活跃和无限壮大的社区

Spark 起源于 2009 年，当下已有超过 50 个机构 250 个工程师贡献过代码，和 2008 年 6 月相比，代码行数几乎扩大 3 倍，这是个令人艳羡的增长。

1.3.3 HPCC

HPCC（高性能计算与通信）是美国实施信息高速公路而实施的计划，该计划的实施将耗资百亿美元，其主要目标是开发可扩展的计算系统及相关软件，以支持太位级网络传输性能；开发千兆比特网络技术，扩展研究和教育机构及网络连接能力。

1.3.4 Storm

Storm 是一种开源软件，一个分布式、容错的实时计算系统。Storm 可以非常可靠地处理庞大的数据流，用于处理 Hadoop 的批量数据。Storm 很简单，支持许多种编程语言，使用起来非常有趣。Storm 由 Twitter 开源而来。Storm 有许多应用领域。

包括实时分析、在线机器学习、不停顿的计算、分布式 RPC（远过程调用协议）等。

1.3.5 Apache Drill

Drill 作为 Apache 孵化器项目来运作，Drill 项目是从 Google 的 Dremel 项目中获得灵感的，该项目帮助 Google 实现海量数据集的分析处理，包括分析抓取 Web 文档、跟踪安装在 Android Market 上的应用程序数据、分析垃圾邮件、分析 Google 分布式构建系统上的测试结果等。通过开发 Apache Drill 开源项目，组织机构将有望建立 Drill 所属的 API 接口和灵活强大的体系架构，从而帮助支持广泛的数据源、数据格式和查询语言。

1.4 大数据的未来走向

大数据逐渐成为我们生活的一部分，它既是一种资源，又是一种工具，让我们更好地探索世界和认识世界。大数据提供的并不是最终答案，只是参考答案，它为我们提供的是暂时帮助，以便等待更好的方法和答案出现。

1.4.1 数据资源化

资源化是指大数据成为企业和社会关注的重要战略资源，并已成为大家争抢的新焦点，数据将逐渐成为最有价值的资产。

随着大数据应用的发展，大数据资源成为重要的战略资源，数据成为新的战略制高点。资源不仅仅只是指看得见、摸得着的实体，如煤、石油、矿产等，大数据已经演变成不可或缺的资源。《华尔街日报》在题为“大数据，大影响”的报告中提到，数据就像货币或者黄金一样，已经成为一种新的资产类别。

大数据作为一种新的资源，具有其他资源所不具备的优点，如数据的再利用、开放性、可扩展性和潜在价值。数据的价值不会随着它的使用而减少，而是可以不断地被处理和利用。

1.4.2 数据科学

1.4.2.1 催生新的学科和行业

数据科学将成为一门专门的学科，被越来越多的人所认知。越来越多的高校开设了与大数据相关的学科课程，为市场和企业培养人才。

一个新行业的出现，必将会增加工作职位的需求，大数据催生了一批与之相关的新的就业岗位。例如，大数据分析师、大数据算法工程师、数据产品经理、数据管理专家等。因此，具有丰富经验的大数据相关人才将成为稀缺资源。

1.4.2.2 数据共享

大数据相关技术的发展，将会创造出一些新的细分市场。针对不同的行业将会