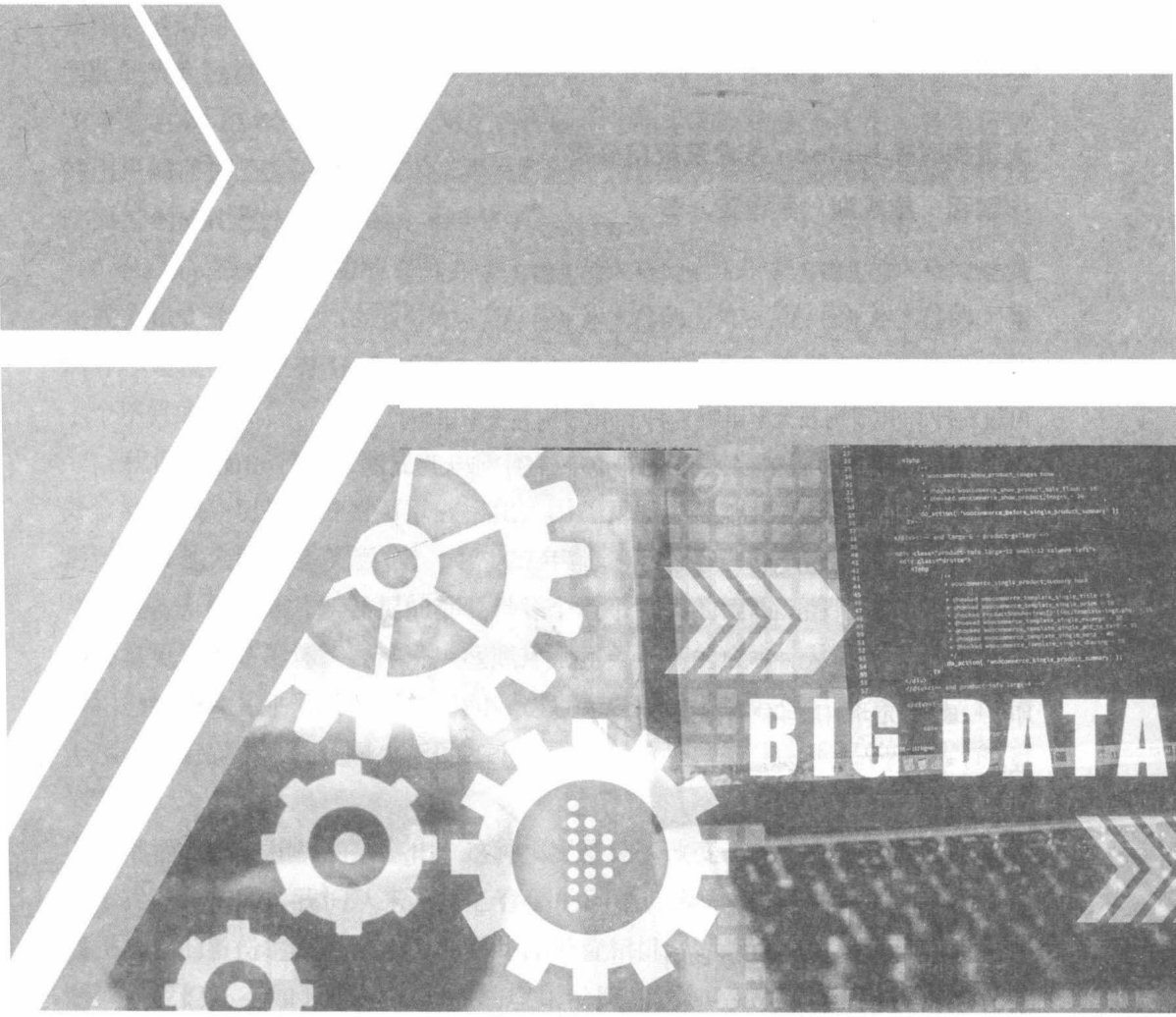
大数据时代
Hadoop 技术及应用分析

韦鹏程　施成湘　蔡银英　著

电子科技大学出版社
University of Electronic Science and Technology of China Press

大数据时代
Hadoop 技术及应用分析

韦鹏程　施成湘　蔡银英　著



电子科技大学出版社
University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

大数据时代hadoop技术及应用分析 / 韦鹏程, 施成湘, 蔡银英著. -- 成都 : 电子科技大学出版社, 2018.1
ISBN 978-7-5647-5570-6

Ⅰ. ①大… Ⅱ. ①韦… ②施… ③蔡… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第010563号

大数据时代 hadoop 技术及应用分析

韦鹏程　施成湘　蔡银英　著

策划编辑　李述娜

责任编辑　李述娜

出版发行　电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼　邮编　610051

主　　页　www.uestcp.com.cn

服务电话　028-83203399

邮购电话　028-83201495

印　　刷　定州启航印刷有限公司

成品尺寸　170mm × 240mm

印　　张　19

字　　数　298千字

版　　次　2019年3月第一版

印　　次　2019年3月第一次印刷

书　　号　ISBN 978-7-5647-5570-6

定　　价　69.00元

前　言

2006 年 3 月份，Map/Reduce 和 Nutch DistributedFileSystem (NDFS) 分别被纳入称为 Hadoop 的项目中。Hadoop 是最受欢迎的在 Internet 上对搜索关键字进行内容分类的工具，但它也可以解决许多要求极大伸缩性的问题。例如，如果您要 grep 一个 10TB 的巨型文件，会出现什么情况？在传统的系统上，这将需要很长的时间。但是 Hadoop 在设计时就考虑到这些问题，采用并行执行机制，因此能大大提高效率。从 2006 年雅虎等团队开始研发 Hadoop 技术至今已整整 10 年。在这 10 年中技术发展迅速，Hadoop 上的生态系统逐渐扩大，各个行业的用户都在基于这一新的技术来开发各种应用，还有很多企业将原先基于传统 IT 系统的应用逐步向 Hadoop 上迁移。

Hadoop 技术能成功的最根本原因在于它是把传统的集中式运算有效地转化成分布式计算的一种有效手段。集中计算演变成分布式是一个必然趋势，当然并不是说一定只有 Hadoop 才是这个演进的唯一手段，不过它至少是可选的一个不错的手段。

我们处在一个由数据主导决策的时代。存储成本在降低，网络速度在提升，周围的一切都在变得可以数字化，因此我们会毫不犹疑地下载、存储或与周围的其他人分享各类数据。大约 20 年前，相机还是一个使用胶片来捕捉图片的设备，每张照片所捕捉的都要是一个近乎完美的镜头，且底片的存储也要小心翼翼，以防损坏。要冲洗这些照片则需要更高的成本。从你按动快门到看到拍摄的图片几乎需要一天的时间。这意味着捕捉下来的信息要少得多，因为上述因素阻碍了人们记录生活的各个瞬间，只有那些被认为重要的时刻才被记录下来。

然而，随着相机的数字化，这种情况得到了改变。我们几乎随时随地都会毫不犹疑地拍照；我们从来不担心存储的问题，因为 TB 级别的外部磁盘可以提供可靠的备份；我们也很少到哪儿都带着相机，因为可以使用移动设备拍摄照片；我们还有如 Instagram 这样的应用给照片添加特效并分享这些美图；我们收集关于图片的意见和信息，还会基于这些内容做出决策。

在商业上，大数据时代也带来了类似的变化。每项商业活动的方方面面都

被记录了下来：为提高服务质量，记录下用户在电子商务页面上的所有操作；为进行交叉销售或追加销售，记录下用户买下的所有商品。商家连客户的 DNA 恨不得都想掌握，因此只要是能得到的客户数据，他们都会想办法得到，并一个一个掐指研究。商家也不会受到数据格式的困扰，无论是语音、图像、自然语言文本，还是结构化数据，他们都会欣然接受。利用这些数据点，他们可以驱使用户做出购买决定，并且为用户提供个性化的体验。数据越多，越能为用户提供更好、更深入的个性化体验。

从某些方面来讲，我们已经准备好接受大数据的挑战了。然而，分析这些数据的工具呢？它们能处理如此庞大、快速、多样化的新数据吗？理论上说，所有数据都可以放到一台机器上 . 何这样一台机器的成本要多少？它能满足不断变化的负载需求吗？我们知道超级计算机可以做到这一点，但是全世界的超级计算机也就那么几台，而且都不具有伸缩性。替代方案就是构建一组机器、一个集群或者串联的计算单元来完成一项任务。一组使用高速网络互相连接的机器可以提供更好的伸缩性和灵活性，但那还不够 3 这些集群还要可编程。大量的机器，就像一群人，需要更多的协调和同步。机器的数量越多，集群中出现故障的可能性就越大。如何使用一种简单的方法处理同步和容错，从而减轻程序员的负担呢？答案是使用类似于 Hadoop 的系统 :

Hadoop 可以认为是大数据处理的同义词 3 简单的编程模型，“一次编码，任意部署”，和日益增长的生态圈，使得 Hadoop 成为一个可供不同技能水平的程序员共同使用的平台。今天，它是数据科学领域首屈一指的求职技能。要去处理和分析大数据，Hadoop 成了理所当然的工具：Hadoop2.0 扩张了它的羽翼，使其能覆盖各种类型的应用模式，并解决更大范围的问题。它很快成为所有数据处理需求的一个通用平台，并将在不久的将来成为各个领域中每个工程师的必备技能。

由于时间的仓促，编者水平有限，本书难免存在不足之处，在此出版之际，我们真诚地希望读者对本书提出宝贵的意见和建议。

目 录

第1章　大数据的产生发展

从2011年开始，大数据作为一项技术进入人们的视野，至今已经超过5年，而Hadoop的诞生是10年前的事情了。Hadoop发展最快的就是过去的这5年，和大数据技术的快速发展是同步的。在过去的5年中，大数据技术被各个行业所使用，而出现在各个不同应用场景上实际应用的系统就是Hadoop。

1.1　互联网和物联网上的数据

传统企业也好，新兴的互联网企业也罢，凡是想要做精细化管理的企业，对数据都是非常关注的，因而在新出现的各种技术中，对“大数据”这一项有相当多的偏好和关注。今天的企业CIO和CTO，如果不说自己也在作一些“大数据研究和应用项目，都感觉自己好像落伍了。

在互联网上奉行的开放和透明的理念，应用到精细化管理和工业管理上也是一样的，而这里我们说的开放和透明，就是基于数字的。

1.1.1　互联网上越来越多的数据被存储

随着互联网和移动互联网的发展，越来越多的数据被存储和使用，这是毋庸置疑的。移动互联网上数据的特殊性首先在于它能够锁定一个特定用户，其次在于它能够获取用户的地理位置信息，再次在于移动互联网上的时空信息等多样化的数据种类。从而导致移动互联网上的数据数量会比传统互联网更大，形式也比传统互联网更加丰富，也有更高的价值。

在今天，数据的产生无论是数量、速度还是类型上都发生了很大的变化。下面我们看一个对比。

New York Times是世界上最老牌的报纸之一，他们把从创立之初的1851年到1980年的所有存档都扫描并转化成PDF格式，一共才有4TB的

数据。而今天一家普通的线上媒体每个月采集的包括高清照片、视频在内的素材，其数据量都可以轻松超过这个数字。

正如 Mary Meeker 在报告中所说，数据在今天越来越重要，下一波技术浪潮会是充分利用今天畅通的互联网渠道和存储来收集、整合、关联及翻译所有的这些数据，从而对人们的生活和企业的有效运作产生价值。

与传统互联网数据不同的是，在移动互联网数据中，文字以外的其他信息占到更多的比例。从数据的属性上来讲，移动互联网上的数据比传统互联网更加复杂，其中一个原因是这些数据包含了大量的时间和空间信息，也就是说我们需要把数据挖掘延伸到时空数据领域（spatio-temporal data mining）。因为多了一个维度，时空数据挖掘的复杂度比一般的数据挖掘又深了一层，虽然说研究方法和算法还是类似的。

在各种不同场景中产生的各种数据，其应用方式是不同的。有些数据会被存储起来，用作业务分析和流程管控，而有些数据则需要被实时或者准时监控、分析和处理。

那么各家公司的数据量究竟有多大呢?

图 1-1 中列出的是 2016 年 Tintri 公司走访了数百家有数据中心的公司做出的数据统计。从图中我们可以看到，已经有 24.4% 的公司的数据量在 IPB 以上，而只有 32% 的公司的数据量在 100TB 以下。

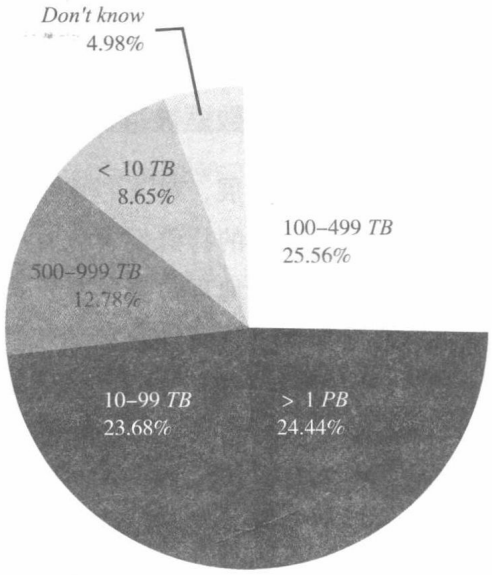


图 1-1　数据中心存储数据量的对比

注：请读者注意这里调查的是“有数据中心的公司”，所以数据量比较大是显然的。不过没有数据中心的公司也一样需要存储和处理和自己相关的数据。

如图 1-2 所示，各家公司存储的数据量基本上在 1TB~1EB（1000000TB），而这个量级恰恰是 Hadoop 系统最能发挥优势的量级。

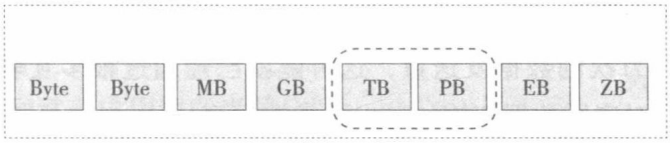


图 1-2　公司数据量级示意图

1.1.2　物联网带来更多的数据

“互联网 +”和“工业 4.0”的概念也为我们添加了更多的数据。工业机床、工业控制设备、RFID 阅读器、传感器网络、GPS 跟踪设备等这些设备每天、每小时甚至每分每秒都在产生新的数据。

我们可以认为互联网其实是一个连接人的网络，采集的数据大部分都是人的行为的数据，如人的交易数据、人的上网记录，而物联网（Internet of Things，IOT）采集的数据更多来自机器和设备。

物联网为我们提供了感知物理世界的接口和手段。遍布于各处的传感器，就如同人的眼耳鼻舌，是大数据系统的输入端。

在过去的 10 年中很多新科技的发展对物联网的发展起到很大的作用，比如：

（1）PV6；

（2）传感器技术；

（3）带宽价格；

（4）全面免费的 WiFi 覆盖；

（5）性价比更好的 CPU。

不过，对于物联网来说最重要的技术还是大数据。

根据 Gartner 的数据，在 2016 年已经有 64 亿个设备连接到互联网上，而且每天还在新增 550 万个设备，或者说每秒增加 63 个新设备。

和互联网数据相比，物联网数据的第一个差异是数据量更大。如果比

较这两个数据源，我们发现它们的数据量会差一个量级。全世界人口可能是 60 亿，但已经有上百亿的设备，如果我们将这些设备产生的数据都采集到的话，其数量会比来自互联网的数据更大，所以这会对数据系统架构产生一个新的、大的挑战。

第二个差异是，物联网数据并发度非常高，而且数据一旦产生需要立刻处理。比如我们有一个真实的客户案例，客户目前有一千万个传感器，每秒有一千万次的数据发送量，这可能就已经超过很多互联网公司的数据量，所以它对底层数据架构的并发要求非常高。

第三个差异在于互联网数据可能是人的行为数据，主要用来分析，可以作一些营销；但是物联网数据更多的是用于发现一些自然规律，因为这里面使用到了大量的技术运算，也会用到大量复杂的物理和数学的方法。

1.2 大数据的使用

IBM 的研究报告说明，表现比较优异的公司和表现相对没那么好的公司之间最大的差距往往在于对数据的使用。对数据应用比较好的公司主要做对了下面的几件事。

（1）用大数据分析来吸引、发展和维护用户；

（2）用数据来优化流程；

（3）把所有可能产生的数据都收集起来；

（4）在可以应用数据作决策的地方都不作主观判断；

（5）快速获取信息；

（6）快速做出决策。

总而言之，这些表现比较优异的公司是基于数据作管理（data-driven decision making），而且往往会引领他们所在的领域和行业。

数据具体可以有哪些应用呢?

（1）数据长期的保存。因为有些数据需要实时分析，有些需要线下分析，还有一些目前可能用不到，不过在未来可能会有用。

（2）欺诈分析和预防。这不仅仅是在金融领域，还可以在任何与用户交互的地方。

（3）社交网络和人、企业等的关系分析。

（4）产品和市场的分析、设计和优化。

（5）根据物联网上采集的数据作数据分析，并作实时响应。

1.2.1 用户画像和任何企业都需要关注的数据

当我们在和任何一家企业讨论基于数据的管理时，首要的基础就是数据。我们来看看企业中都有哪些数据。

（1）网站和移动应用程序流量分析；

（2）产品和服务销售分析；

（3）市场调查分析；

（4）设备和机器监控和数据分析；

（5）人力资源员工数据分析；

（6）（潜在）竞争对手市场分析；

（7）互联网口碑分析。

我们在这里只是简单地列举了一些数据点，而实际情况是企业任何一个部门、任何一个员工、任何一台设备或者任何一个市场活动都会持续不断地产生各种各样的数据。对这些数据进行分析和挖掘，是企业转向基于数据管理的关键。

我们经常听到的一个词是“用户画像”，如图 1-3 所示，那么“用户画像”究竟是什么呢？

图 1-3　用户画像

我们认为用户画像其实就是关于这个用户的各种数据的整合。当我们获取了用户的各种数据，而且这些数据还能确保真实的时候，我们可能会比用户本人更加了解他自己。

（1）社交网络上的各种信息；

（2）游戏中的各种数据；

（3）用户所关注的娱乐内容；

（4）用户的信用、借贷和消费记录；

（5）用户在电商网站上的购买和浏览记录；

（6）用户在原有传统数据库中的数据等。

1.2.2 大数据的 3V、4V 和 N 个 V

最早是 IBM 提出了大数据领域的“3V”概念，即大量化（Volume）、多样化（Variety）和快速化（Velocity）。3V 是大数据时代的显著特征，正是这些特征给今天的企业带来了巨大的挑战。

业内也有学者和从业者提出了其他关于大数据的 V，比如：

（1）数据的价值（Value）；

（2）数据的可验证性（Verification）；

（3）数据的可变性（Variability）；

（4）数据的真实性（Veracity）；

（5）数据的邻近性（Vicinity）。

可验证性（Verification）指的是数据需要经过验证，因为数据量大了之后，带来的一个后果必然是数据质量的良莠不齐，以及因不同级别用户介入而产生的数据安全问题。可变性（Variability）主要指的是数据格式的可变性，着重于非关系型数据。真实性（Veracity）指的是因为数据来自不同的源头，而有些来源的数据（比如 Facebook 上的评论和 Twitter 上的跟帖）其本身的可信度是需要考虑的。邻近性（Vicinity）和大数据的存储相关，处理数据的程序和服务器需要能够就近获取资源，不然会造成大量的浪费和效率的降低。

专家和学者们会将上述的某一个或者几个 V 与 Volume、Variety、Velocity 合在一起，并称为 4V 或者 5V、6V，至于选用的是哪一个 V，则要看他们想要推送的理念、产品和服务与哪一个或者哪几个 V 最接近。

在这 N 个 V 中，我们认为最值得关注的当然是数据的价值（value）。所有的大数据应用如果不落到价值体现上，是没有意义的。以商业应用为核心，这是我们在本书中从头到尾都在讲述的概念。

1.2.3　从数据分析到数据挖掘

什么是数据挖掘呢？古人云“物以类聚，人以群分”，这句话其实描述的就是数据挖掘中的一类算法——聚类算法。

要看一个人是怎样的，只需要看他周围都有什么样的朋友；而从数据挖掘的角度来说，聚类算法要预测一个对象的特征，只需要看它周围对象的特征。

大数据挖掘在本书中的定义是在海量数据的基础上进行数据挖掘的过程，也就是对数据进行处理和研究，并从数据中提取有用信息和发现知识的过程。

对数据进行分析和处理，那么数据分析和数据挖掘之间有什么区别呢？

从本质上来说，数据分析和数据挖掘都是为了从收集来的数据中提取有用信息，发现知识，而对数据加以详细研究和概括总结的过程。在不少场景中，数据分析和数据挖掘这两个概念是可以互换的，而它们之间最大的区别是数据本身的不同，这主要表现在以下两个方面。

（1）数据量的不同。数据分析的数据对象通常是存储在数据库或者文件中，而数据挖掘对应的数据对象一般是在分布式数据库或者数据仓库中。在今天，一个数据分析应用的对象数量级会是在 MB 或是 GB，而数据挖掘的应用数据动辄 TB，甚至 PB。

（2）数据类型的不同。数据分析处理的对象一般是文本或者纯数字，而数据挖掘的对象不仅仅是文本，还有音频、视频和图片数据；数据挖掘面对的不仅仅是规范化数据，还有半规范化数据和不规范数据。

从某种意义上来讲，数据分析和数据挖掘之间的区别就像淘金客和矿山主，不同点在于淘金客只在一条小溪上工作，甚至几十个人共享一条小溪，通常只能通过手工作业用沙漏从沙里淘金；而矿山主则占有整座巨大的矿山，由于矿山拥有成分复杂的矿石和数量繁多的伴生矿物，这时候矿山主就不能仅仅依靠手工作业，而需要建立一个以机器为劳动力的现代化

工业企业，才能做到最大限度和效率的产出。

数据挖掘与传统的数据分析（如查询、报表、联机应用分析等）的本质区别在于数据挖掘往往是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得出的信息通常具有先前未知性、有效性和可实用性三个特征。而从本质上讲，数据分析主要是一个假设检验的过程，是一个严重依赖于数据分析师手工作业的过程。数据分析就像是我们在淘金，如果有高水平的淘金客，我们就能淘出金子。

数据挖掘或者大数据挖掘，是传统手工业式的数据分析的现代大工业形式。数据挖掘建立在拥有大量数据，并且能够让机器方便读取的数据仓库之上，采用机器学习的算法，是自动挖掘知识的过程。

当然这并不意味着数据分析会完全被数据挖掘所取代。就像现代大工业只是取代了手工生产的组织形式，而手工生产中的方法、技能等都被现代大工业吸收进来，重新赋予了意义。同样地，大数据挖掘也需要数据分析的算法和思路，只是用新的方法组织施行。而如今这一过程也才刚刚开始。

数据挖掘并不是一门崭新的科学，而是综合了统计分析、机器学习、人工智能、数据库等诸多方面的研究成果的边缘学科。其与专家系统、知识管理等研究方向的不同之处在于，数据挖掘更侧重于企业应用。

在 2015 年年初，PWC 普华永道发布了一份针对 77 国逾 1300 位 CEO 的调查。结果显示，在推动数字技术发展、提高组织能力方面，提高客户参与度的移动技术排在第一位，而数据挖掘分析占有第二重要的战略地位。同时，这些 CEO 还认为，提供更好的客户体验并提高业务效率也是数据分析最为重要的一项能力。

笔者认为，数据逐渐成为最大的一类交易商品。在互联网上，继“入口为王”“流量为王”和“应用为王”之后，下一个概念理所当然应该是“数据为王”。在今天，大数据已经像公用设施一样，有数据提供方、管理方、运营商、第三方服务商和监管方，而且数据交易的流程也在被完善。

数据的供应、交易和处理将会形成一个新的大产业链，而 Hadoop 将是一把利器。

1.2.4 大数据处理的三个维度

当我们在讨论大数据的时候，需要更多关注的是对大数据的处理。如

果我们只是把数据存储在那里，而没有充分使用它们，那么这是没有意义的。

面对大数据，NetApp 公司作过一个值得借鉴的分析，如图 1-4 所示。

从图 1-4 中我们可以看到，大数据处理要分成三个维度。

（1）Content，在内容上，我们要有安全的无限数据存储；

（2）Brandwidth，在速度上，我们要能做快速的数据密集性处理；

（3）Analytics，在分析层面，我们要能处理超大的数据集。

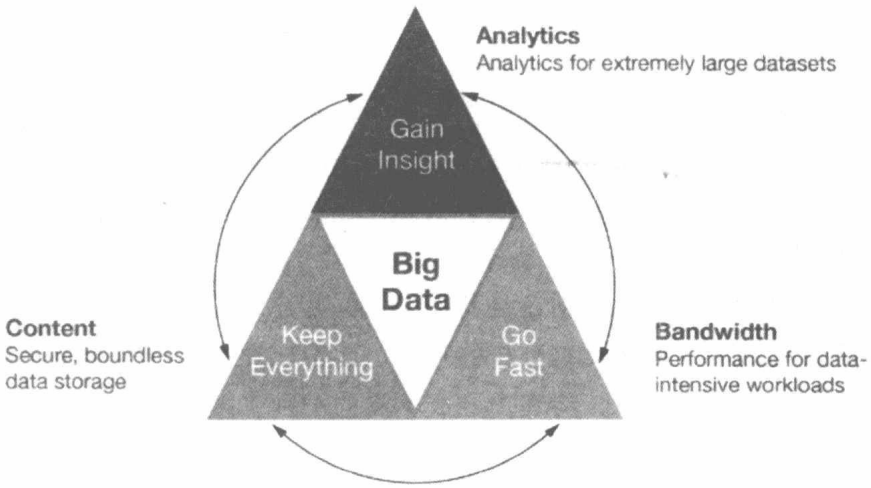


图 1-4　大数据处理的三个维度

其实，就前文所讲的 3V 来说，Content 对应的是 Volume，Bandwidth 对应的是 Velocity，而 Analytics 对应的是其中两项：Variety 和 Volume。

简而言之，数据挖掘（Data Milling）是有组织、有目的地收集数据，通过分析数据，使之成为信息，从而从大量数据中寻找潜在规律以形成规则或知识的技术。

（1）用户和市场行为产生大量的数据。

（2）数据在经过解析之后产生洞察和分析。

（3）数据要产生价值就需要把洞察应用到用户和市场行为上。

（4）优化了的用户和市场行为又产生了大量的数据，循环再一次开始。

我们经常听到的“大数据挖掘”其实包含了“大数据”和“数据挖掘”两个不同的概念，前者说的是数据的规模，而后者说的是数据的使用。

基于大数据的服务创新有很大的想象空间。