第 1 篇　大数据技术原理与应用

通常来说，数据本身是没有什么直接意义的，但是借助于统计、分类、萃取、特征抽取等一系列技术手段，便能够从数据中产生信息与知识。鉴于数据目前已被视为重要的战略资源，蕴藏着巨大的经济价值，由此也引起了科学界和企业界的高度重视。在有效地组织和使用数据之后，将对经济发展产生巨大的推动作用。在大数据的背景下，会诞生许多前所未有的机遇，通过对大数据的交换、整合和分析，可以发现新的知识，创造新的价值。

越来越多的企业和机构意识到数据正在成为最重要的资产，数据分析能力正在成为核心竞争力。之前历经了由 PC 转向软件和服务的成功期，现在则更多地专注于因数据分析而带来的全新业务增长点，而将远离服务与咨询。在未来，数据会变为各行各业决定胜负的根本因素，最终数据将成为人类极其重要的自然资源。如今已经到了数据化运营的黄金时期，如何利用好并整合这些数据成为未来发展的关键性任务。

大数据既代表着信息技术发展的高度抽象和概括，同时也体现了信息技术服务于数据蕴藏的巨大价值。大数据的出现给数据的采集、存储、共享、维护带来了具有研究意义的现象和挑战。对于人类来说，处理、分析、使用大量数据，通过对这些数据的处理、整合和分析，可以发现新知识，创造新价值，带来大知识、大科学和大发展，从而逐渐走向创新社会化的新信息时代。

大数据全生命周期可以划分为“数据产生—数据采集—数据传输—数据存储—数据处理—数据分析—数据发布、展示和应用—产生新数据”等阶段。目前已经形成了大数据的“生产与集聚层—组织与管理层—分析与发现层—应用与服务层”的产业链，而 IT 基础设施则为各环节提供基础支撑。

第 1 章　大数据的来源、采集与基本概念

谈起当下最流行的技术，你第一时间能想到的一定是人工智能技术和大数据，基本上每位关注科技和发展最新动态的年轻人都会对此有所涉猎。但一旦谈起大数据的历史发展，你是否真的完全了解呢？本章主要叙述的就是大数据的来源与发展，同时还会介绍一些大数据的基本概念，方便之后深入学习。

1.1　大数据的来源与发展

通常认为，Google 的“三驾马车”是大数据概念的起源，包括“The Google file system”“MapReduce: simplified data processing on large clusters”“Bigtable: a distributed storage system for structured data”。在 2007 年，Amazon 也同时发表了一篇关于 Dynamo 系统的论文。这几篇论文奠定了大数据的基础[1]。

为何在 Google 发表了三篇论文之后，“大数据”的概念就应运而生了呢？在互联网经济泡沫破灭后，Google 作为全世界第一个上市的大型 IT 公司，上市之后借助于其优秀的发展策略，市值飞速增长。究其原因，在于 Google 在广告方面投入很多，大数据技术则在其中扮演着至关重要的角色。当时，大部分相关的互联网企业都认为大数据是改变自己命运的机会，借此纷纷加入大数据圈子，同期加入的有 Facebook、LinkedIn、Twitter、Microsoft、阿里巴巴、Yahoo 等公司。

那么大数据是怎样让互联网产业再一次蓬勃发展的呢？这主要归功于上面提到的“三驾马车”。众所周知，搜索引擎是 Google 发家的主力军，而搜索引擎主要完成两个工作，一个是索引构建，另一个是网页抓取。在这个流程中，会存在大量需要存储和计算的数据。这“三驾马车”就是用来解决这个问题的：一个数据库系统、一个文件系统、一个计算框架。现在分布式、大数据之类的词对大众来说并不陌生，但要追溯到 2004 年左右，当时整个互联网还处于懵懂时代，Google 发布的论文则让整个业界为之一振，如梦初醒，突然领悟到还可以这么利用大数据。

当时，大多数公司的关注点聚焦在单机上，他们都在探索如何提升单机的性能，以及寻找更贵更好的服务器。而 Google 的思路则是部署一个大规模的服务器集群，通过分布式的方式将海量数据存储在这个集群上，然后利用集群上的所有机器进行数据计算。这样，Google 不需要买很多很贵的服务器，而只要把这些普通的机器组织到一起就非常厉害了。

1.1.1　抱团取暖的 Hadoop 圈

当时有一位天才程序员，即 Lucene 开源项目的创始人 Doug Cutting，正在开发开源搜索

引擎 Nutch。他阅读了 Google 发表的论文后，根据论文原理初步实现了类似 GFS 和 MapReduce 的功能。

2006 年，Doug Cutting 将与大数据相关的功能从 Nutch 中分离出来，启动了一个独立的项目专门开发维护大数据技术，这就是后来著名的 Hadoop。它主要包括 Hadoop 分布式文件系统 HDFS 和大数据计算引擎 MapReduce。Hadoop 的主要贡献者是 Yahoo，Facebook、LinkedIn、Twitter 等公司也都贡献了一些影响深远的项目。

在 Hadoop 被发布后不久，Yahoo 很快就用了起来。2007 年后，百度和阿里巴巴也逐渐开始运用 Hadoop 来进行大数据的存储与计算。Hadoop 在 2008 年正式成为 Apache 软件基金会的顶级项目，后来 Apache 软件基金会的主席也由 Doug Cutting 本人担任。同年，作为专门运营 Hadoop 的一家企业，商业公司 Cloudera 成立。借助于公司强大的宣传和雄厚的资产，Hadoop 得到了更快的发展。

考虑到用 MapReduce 进行大数据编程太麻烦，Yahoo 随即发布了 Pig。Pig 是一种脚本语言，使用类 SQL 的语法。利用 Pig 脚本，开发者可以描述要对大数据集进行的操作，Pig 经过编译后会生成 MapReduce 程序，然后在 Hadoop 上运行。

当时，直接使用 MapReduce 编程要比编写 Pig 脚本稍难一些。但若想深入 Pig 脚本的应用，仍然还要学习新的脚本语法，为此 Facebook 公司发布了 Hive。Hive 中支持使用 SQL 语法来进行大数据计算，比如可以写个 Select 语句进行数据查询，然后 Hive 会把 SQL 语句转化成 MapReduce 的计算程序。

通过以上方法，对数据库应用十分熟悉的工程师和数据分析师可以很快地利用大数据来解决数据分析和处理方面的问题。Hive 的出现，很快得到企业和开发者们的拥簇，因为其在很大程度上降低了 Hadoop 的使用难度。据研究，在 2011 年左右，利用 Hive 使 Facebook 大数据平台上 90%的作业得以运行。

自此，大量 Hadoop 周边产品开始涌现，逐渐形成了以大数据为中心的生态体系，其中包括：专门将关系数据库中的数据导入导出到 Hadoop 平台的 Sqoop、针对大规模日志进行分布式收集、聚合和传输的 Flume、MapReduce 工作流调度引擎 Oozie 等。

回望众多软件开发的前世今生，你会惊讶地发现，有些软件在被开发出来之后无人问津或者只被极个别人在使用，而这样的现象不在少数。有些软件的出现则会缔造一个辉煌的行业，它被数以亿计的人在使用着，同时也给行业带来富可敌国的利润和价值，如 Windows、Linux、Java。而如今，这个名单上多了一个叫 Hadoop 的软件。

当下来评价 Yahoo，可能认为在中国，雅虎时代已彻底落幕。但若把时钟拨到 2008 年，Yahoo 可以说是风靡一时的公司，曾经被称为“互联网第一股”。在那个门户网站飞速发展的时代，Yahoo 主导着其他互联网公司一起“造一个轮子”，它成为一些互联网初创公司的中心，并试图打造一套可以和 Google 的“三驾马车”相提并论的系统。

在大数据领域，除了 Hadoop 以外的系统主要有两个：一个是 Microsoft 研发的 Cosmos，中文名叫作“宇宙”；另一个是阿里巴巴的 ODPS（open data processing service），现在已经更名为 MaxCompute。

1.1.2 一场大论战

2008 年，大数据行业内爆发了一件引人注目的争论事件。对立的一方是数据领域的元老级人物迈克尔·斯通布雷克和大卫·德威特，另一方是 Google Brain（谷歌大脑）团队的负责人杰夫·迪恩。两大战队争论的焦点是 MapReduce 究竟是创新还是倒退，双方各执一词。

传统数据库方发布了一篇以“MapReduce：a step backward”为标题的文章，由此挑起了纷争，文中主张 MapReduce 是数据库领域早就被淘汰的、令人不屑一顾的技术。而 Google 一方的专家则认为 MapReduce 是个伟大的发明。双方都有自己的立场和观点，却有失客观性。经过辩证地看待两方的观点后，加州大学伯克利分校的 AMP 实验室开发了新的系统——Spark。在那时，AMP 实验室的马铁博士意识到利用 MapReduce 进行机器学习计算的时候性能十分不理想，因为机器学习算法通常需要进行很多次的迭代计算，而 MapReduce 每执行一次 Map 和 Reduce 计算都需要重新启动一次作业，带来大量的无谓消耗。还有一点就是 MapReduce 主要使用磁盘作为存储介质，而 2012 年的时候，内存已经突破容量和成本限制，成为数据运行过程中主要的存储介质。由此 Spark 市场化以后，立即受到业界的强烈关注，并在接下来的时间里逐渐取代了 MapReduce 在企业应用中的地位。

通常，像 MapReduce、Spark 这类计算框架处理的业务场景都被称作批处理计算。它们通常针对以“天”为单位产生的数据进行一次计算，然后得到需要的结果，这中间计算需要花费的时间大概是几十分钟甚至更长的时间。因为计算的数据是非在线得到的实时数据，即历史数据，这类计算也被称为大数据离线计算。

在典型的大数据的业务场景下，数据业务最通用的做法是采用批处理计算处理历史全量数据，采用流式计算处理实时新增数据。而像 Flink 这样的计算引擎可以同时支持流式计算和批处理计算。

除了大数据批处理和流处理，NoSQL 系统处理的主要也是大规模海量数据的存储与访问，所以也被归为大数据技术。NoSQL 曾经在 2011 年左右非常火爆，涌现出 HBase、Cassandra 等许多优秀的产品，其中 HBase 是从 Hadoop 中分离出来的、基于 HDFS 的 NoSQL 系统。

1.2 何为大数据

“大数据”（big data）这个词是 2008 年在维克托·迈尔–舍恩伯格及肯尼斯·库克耶编写的《大数据时代》这本书中被首次提出的。

麦肯锡（美国首屈一指的咨询公司）是研究大数据的先驱，其在报告“Big data：the next frontier for innovation，competition，and productivity”中给出的大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB（太字节）值的数据集才能算是大数据。

Amazon（全球最大的电子商务公司）的大数据科学家 John Rauser 给出了一个简单的定义：大数据是任何超过了一台计算机处理能力的庞大数据量。

1.2.1　不断增生的数据巨浪

维克托 • 迈尔–舍恩伯格在自己出版的《大数据时代》中提到，数据是世界的本质。直到今天，现实也能很好地印证这些话。当下，点击一个小程序，微信上点个赞或发个评论，或者是拨打一通电话，都会伴随着海量数据资料的产生。人们生活中的每一笔移动支付，上网时发送和接收到的消息，全都是数据资料。每天每时每秒，由人类产生的数据信息越来越密集。假设数据位肉眼可见，那么身处其中的我们都像是一个个采蜂人，全身上下裹着一层又一层、厚厚的“位”数据。

视线转到另一侧，某同学在学校的自习室里喝着刚买的咖啡。瑞幸的视频监控显示，他半小时前购买了一杯拿铁；微信的消费记录表明，他一小时前从宿舍旁的小卖部购买了一支铅笔；之前他还往校园卡里充值了一笔钱。从消费记录、转账记录、定位这些数据中，可以很明显地看到该同学的每日动向及资金去向。短短一个多小时，又有好几亿的数据产生。

如今人们所生活的世界，不仅仅包含能制造大量数据的个人，更是一个不断被大数据淹没的新世界。每一秒，一家大型医院会产生 12 万笔生理健康数据；每一分钟，YouTube 网站上传影片总时长达 72 小时；每一天，一家银行要处理 500 万笔信用卡交易，一个 Twitter 网站上上传 2.3 亿条推文。如果再加上全世界同一时间约 5 亿部智能手机、10 亿台计算机和数万亿个传感器同时运作所产生的各种文字、声音和图片数据，每一天的数据量高达 25 亿 GB（吉字节），需要用 4 000 万台 64 GB 的 iPad 才能装载。而且，单单过去 20 年间，人们制造的数据就占了当今全球数据总量的 90%左右。

依照这种每年约 50%的增长速度计算，国际数据公司 2008 年估测，到 2021 年，全球数据总量将增长 44 倍，达到 35.2 ZB（泽字节，相当于 1 万亿 GB）。如果把这些数据全都装在 64 GB 的 iPad 里，这些 iPad 叠起来的高度超过 13 万座玉山总高度之和。

ZB 到底有多大？有人形容 1 ZB 的数据量相当于全世界海滩上的沙子数量之和。更惊人的是，全球数据量还在不断增长！如果以每一分钟在网络世界流动的信息量来看，当你打上关键词，按下“Google 搜索”的这一刻，你只是 200 万人中的 1 人；当你写好电子邮件，按下“发送”的那一刹那，这封电子邮件也只是 2 亿封中的 1 封。

其他更惊人的一分钟网络数据资料包括：Facebook 上产生超过 68 万条内容；超过 27 万美元的网络购物交易；苹果商店里的 App 被下载 47 000 次；Flickr 用户分享了 3 125 张照片；有 217 名移动网络新用户诞生。

而上述的数据只是现状，未来呢？思科视觉网络指数（Cisco VNI）报告显示，以前人们只用计算机上网，但现在通过计算机、手机、平板电脑等多种设备随时上网的生活状态已逐渐成为文明世界的常态。2011 年，全球网络联机设备为 103 亿台，以地球上 70 亿人口计算，每人分配到 1.5 台；预计到 2023 年，全球网络联机设备将达 489 亿台。

在中国，对不少人来说，拥有一台台式计算机或一台笔记本电脑、一部智能手机是再普通不过的了，现在可能还要加上一台平板电脑。而人们“随时在线”也让全球网络流量正以每年 1 ZB 的速度在增加，2016 年前已达到 1.3 ZB，平均流量已达到 245 TB/s，相当于有 200 万人在同时观看高清影片。

“大数据”已经渗透到人们生活中的方方面面。比如打开手机淘宝，呈现在每位用户面前的界面是不一样的，而它推送的商品往往却能够抓住用户的需求和心理，这其实就是大数据分析出的结论。

淘宝平台对每一个浏览及购买过商品的用户都进行了全数据分析。它可以轻松获取用户的很多信息，如性别、年龄、家庭成员、喜好、是否结婚、是否有孩子、孩子的性别，甚至细致到用户对服饰的喜好类型等。平台经过分析和处理，能进一步推测出用户可能会订购的商品并进行推送，让用户花更少的时间进行检索却花更多的钱进行消费。假如用户购买了一些孕妇用品，可能在不久之后，平台就会推送一些相关联的婴儿用品。而用户消费后的评价与反馈，又使得平台不断改进自己。例如调整不同卖家的钻石星级，或者清退一些不合格的卖家等这些行为，就是淘宝对自身的改进。这种互利互惠的双回路的运转模式，可以看作是卖家与买家间的一种良性的互动方式。这种互动方式在传统的卖场里面是不可想象的，也是难以实现的。

当今世界产生的庞大数据量只是大数据的基石，如果不加以应用，它将毫无价值。大数据研究的目的是让这些海量的数据“活”起来。

1.2.2 让数据发声

“大数据”的存在是为了让人们理解信息内容和发现信息与信息之间的内在关系。曾有人提出，要让数据“说话”。但直到现在，尽管人们使用数据已经过了相当长的一段时间了，人们依旧对此难以把握。无论是平时进行的海量的非正式历程，还是过去几百年在专业层面上用各种算法及高级测试程序进行的量化研究，均与数据相关。

在纷繁复杂的数字化时代，各种工具和算法的设计能够使数据处理更加快速和简易化，短时间内处理成千上万的数据已成为可能。然而，探究能“说话”的数据则远远不止以上内容。

实际上，大数据与三个重大的思维转变有关，这三个转变是相互联系和相互作用的。首先，要分析与某事物相关的所有数据，而不是分析少量的数据样本；其次，要乐于接收数据的纷繁复杂，而不再一味追求精确性；最后，要转变思想，不再探求难以捉摸的因果关系，转而关注事物的相互关系。

1. 首先是更多

在信息处理较为困难的时代，生活发展需要数据分析，却苦于没有收集数据的相关工具。由此，出现了随机采样方法，它可以作为当时的代表性产物之一。现在，计算和制表变得越来越容易，传感器、GPS、网络点击量及 Facebook 等可以收集大量数据，而计算机能够十分轻松地对这些数据进行处理。

众所周知，采样是为了从最少的数据中获得最多的信息量。而当可以获取大量数据时，采样就没有什么意义了。如今，日新月异的数据处理技术被不断应用，但人们的方法和思维却没有跟上这种改变。

采样存在一个不容忽视的缺点：缺少对细节方面的观测。尽管有时候只能采用采样分析来进行观测，然而在其他很多应用领域方面，从采集部分数据到采集越来越多的数据的变化已经

在产生了。如有可能，人们会采取手段收集所有数据，即“样本等于总体”。在这个程度上，大数据概念中的“大”显然不是绝对意义上的大。

2. 其次是更杂

在绝大多数情况下，可获取的数据被我们使用的频率越来越高，但由于数据量剧增，同时人为或者计算失误导致的错误数据也会掺入数据库中，导致结果会变得越来越不准确。但人们能够努力规避这些问题，或许大家从不认为这些问题是无法解决的，而且也似乎开始正视它们。这是由“小数据”到“大数据”非常关键的一步。

就“小数据”来说，它最应该做到，或者说它最基础的功能要求就是减少错误，确保质量。由于采集到的信息量相对较少，必须保证精确的数据信息内容。科学家们都善于使用精确率较高的仪器或工具来观测遥远的天体物质或者是微小的单细胞生物，而放到采样中，对精确度的要求便更高且更严格了。由于收集的信息有限，这意味着细微的错误会被放大，甚至有可能影响整个结果的准确性。

但是，在多数情况下，容许精确度不是那么高已经变为了一个较为前沿的亮点，而不是缺点。缘由是我们降低了容错的标准，因而掌握的数据也越来越多样化，同时还能够利用这些数据做更多更有效的事情。由此，不是大量数据比少量数据优化这么简单，而是大量数据已经展现了更优质的结果。

与此同时，我们需要与各种混乱作斗争。随着数据量的增加，错误率也相应增加。在整合来自不同来源的各种信息时，混淆的程度会增加，因为它们通常不完全一致。例如，与服务器处理投诉时的数据相比，使用语音识别系统识别呼叫中心收到的投诉虽更易产生不准确的结果，但有助于我们掌握整个事情的总体情况。

混淆也可以指格式上的不一致，实现一致性需要在数据处理之前仔细清理数据，这在大数据环境下很难做到。大数据专家 D. J. Patil 指出，沃森实验室和国际商用机器公司都可以用来指代 IBM，或许还有数千种表示方式。当我们做数据转换的时候，我们把它变成了别的东西。所以在提取或处理数据时会出现混淆。比如我们做 Twitter 消息的情感分析来预测好莱坞票房的时候，就有一定的困惑。

“大数据”通常用概率而不是确定性来说话。整个社会需要很长时间才能习惯这种思维，虽会出现一些问题，但就目前而言，当我们试图扩大数据规模时，我们学会了接受混乱。

3. 最后是更好

在大数据时代，知道“是什么”就够了，不一定要知道“为什么”。我们不必知道现象背后的原因，而是让数据自己说话。

谈到这里，就不得不提起一个叫作 Amazon 的公司。

早在 1997 年，热爱计算机和人工智能的 Greg Linden 在网上卖书。他的网店仅仅营业了两年就赚得盆满钵满，那时候他年仅 24 岁。他曾在回忆录里写道：“那时候的我喜欢卖书和一些自己认为用得到的知识，以此帮助人们找到下一个他们可能会感兴趣的认知点。”而这家网店，就是以后名满世界的 Amazon。

Amazon 的技术先进性不仅依赖于软件技术，在内容生产的方式上也有其独到之处。当时，Amazon 雇用了一个由 20 多名审稿人和编辑组成的团队对图书进行评论，推荐新书，并挑选非常独特的图书放在 Amazon 网站上售卖。这个团队创造了《亚马逊之声》，它成了

Amazon 王冠上的一颗明珠。《华尔街日报》的一篇文章热情地写道，他们是美国最有影响力的书评家，因为他们推动了图书销量的增长。

Amazon 创始人兼首席执行官 Jeff Bezos 决定尝试一个创意：根据客户此前的购物偏好，向他们推荐特定的书籍。从一开始，Amazon 就从每个客户那里获取了大量数据。例如，他们买了什么书？他们关注哪些书？他们关注但不买东西有多久了？

由于拥有海量的客户信息数据，Amazon 必须先用传统的方式处理，通过样本分析找到客户之间的相似性。但这些分析结果给人的感觉就像在波兰买一本书，却被东欧其他地方的价格搞得不知所措；或者像在买一种婴儿用品时，却被一大堆类似的婴儿用品搞得不知所措，显得毫无道理。1996 年至 2001 年担任 Amazon 书评人的詹姆斯•马库斯在他的回忆录 Amazonia 中回忆道：“平台往往会推荐给你一些与你之前购买的产品略有不同的产品，而且这些产品会来来回回地被推荐给你。”

Greg Linden 很快找到了解决办法。他意识到，在推荐系统中没有必要将客户与其他客户进行比较，这在技术上很烦琐，而需要做的是找到产品之间的相关性。1998 年，他和他的同事申请了著名的“逐项”协同过滤技术的专利。方法的改变带来了技术上翻天覆地的变化。

由于可以提前进行估计，推荐系统具有闪电般的速度，并可以应用于各种各样的产品。因此，当 Amazon 跨境销售书籍以外的商品时，它也可以推荐电影或烤面包机等产品。因为系统使用了所有数据，所以推荐是最理想的。Greg Linden 回忆道：“小组中有个笑话说，如果系统运行良好，Amazon 应该只向你推荐一本书，而那就是你下一本要买的书。”

现在，据说 Amazon 三分之一的销售额来自其个性化推荐系统。Amazon 不仅关闭了许多大型书店和音乐商店，而且数百家本地书商也会根据推荐系统及时调整售卖书籍。在一定程度上，Greg Linden 的工作已经彻底改变了电子商务，现在几乎每个人都在使用电子商务。

也许了解人们为什么对这些信息感兴趣是有用的，但现在这个问题并不那么重要。知道“是什么”可以创造点击量，这种洞察力足以重塑许多行业，而不仅仅是电子商务。所有行业的销售人员长期以来都被告知，他们需要了解是什么让客户做出选择，并掌握客户做出决定背后的真正原因，因此专业知识和多年的经验是非常重要的。然而，大数据提出了另一种方法，在某些方面更有用。Amazon 的推荐系统梳理出了有趣的相关性，但原因尚不清楚。知道它是什么就足够了，没必要知道原因。

1.2.3 如何让数据“发声”？

在这个数据时代，谁拥有数据资源，谁就拥有“金库”。随着计算机处理能力的不断增强，所获得的数据量越大，可挖掘的价值也就越大。然而，要想访问“金库”中的财富，你需要一把钥匙来打开“金库”。大数据技术就是打开数据“金库”的一项新技术。

法国著名雕刻家罗丹曾经说过：“世界不是缺少美，而是缺少发现美的眼睛”。同样，世界上并不缺乏数据，而是缺乏开发数据的工具。大数据的价值在近几年才得到认可，至此分析数据的能力才有了革命性的突破。

无论是在工业、金融研究、办公、媒体还是日常生活中，产生的数据都可以成为大数据的一部分。但吸引我们的不仅仅是这些数据的庞大规模，而是我们可以利用它们做些什么。由于数据本身并不产生价值，如何分析和利用大数据来帮助企业才是关键。

从具体应用的角度来看，大数据的真正价值并不是特别明显，即没有“产品化”。企业目前热衷于做的不是从大数据中提取价值，而是把他们能收集到的数据先储存起来，必要的时候再从中提取价值。诚然，目前大数据的处理和分析还不成熟，特别是在采用 IT 方法将之转化为商业效益的过程中。可以说，大数据暂未真正找到“数据实现”的方法。

从“大数据的生成”到“价值的生成”，需要一个非常复杂的过程：首先，需要大数据的收集；其次，对采集到的数据进行预处理和存储；最后，对存储的数据进行分析和挖掘，并将其呈现给用户，使之成为真正有用的资产。

目前 IT 领域最流行的两个词是“大数据”和“云计算”。几乎所有的 IT 公司都在进行改革，努力挤进“大数据”和“云计算”这两条轨道。事实上，大数据和云计算是相辅相成的。随着云计算服务器的出现，大数据有了运行的轨迹，其真正价值得以实现。如果把“大数据”比作汽车，支撑这些汽车的高速公路则是“云计算”。最著名的例子是 Google 搜索引擎，面对海量的 Web 数据，Google 在 2006 年首次提出了“云计算”的概念。Google 自己的云计算服务器支持公司的各种“大数据”应用。

近年来，大数据分析技术无处不在。IT 公司正在争先恐后地为不同的场景和需求提供各种解决方案和技术。本书将探讨大数据的处理技术，分析各种大数据处理技术的特点和适用场景，并针对不同的场景提供可行的处理方案，即如何让大数据“活起来”。

1.3　大数据的特点

以上介绍了什么是大数据，大数据时代人们的思维发生了怎样的变化。本节将讨论大数据的特点。

国际数据公司定义了大数据的四个特征，即大（volume）、快（velocity）、杂（variety）、疑（veracity）。此即大数据的“4V”。

1.3.1　大数据的“4V”之大（volume）

如前所述，人类产生的数据量呈爆炸式增长。根据国际数据公司的统计，仅 2011 年一年就产生了 1.8 ZB 的数据量：这相当于每个美国人每分钟发 3 条 Twitter 并持续 266 976 年产生的数据量；或者是一个人在 4 700 万年里每天观看 4 小时高清电影所消耗的数据量。

Twitter 每天生成 7 TB 的数据，Facebook 则高达 100 TB，一些公司正在以每小时数 TB 的速度累积 PB 级的数据，还有很多例子表明内部存储系统存储了 PB 级的数据。根据麦肯锡 2011 年的调查，2010 年，在美国 15 个行业中的任何一家公司存储的数据都比国会图书馆还要多。想象一下，当你阅读这本书时，实际的数据量则又远远超过了 2010 年的数据量。

1.3.2　大数据的“4V”之快（velocity）

数据的数量和种类都发生了变化，数据产生的速度也与以前大不相同。具体来说，实时移动的动态数据已经成为大数据时代面临的另一个挑战。以爱尔兰的戈尔韦湾为例，像世界

上许多其他水域一样，它正面临着水污染、鱼群减少和气候变化的威胁。由于地理环境、石油泄漏或其他污染事件，戈尔韦湾的污染扩散速度远快于公海，因此爱尔兰海洋研究所与IBM 合作，在戈尔韦湾安装了数百个浮标。浮标上的传感器通过无线电和网络链接，可以实时测量海洋和气候环境的变化。

通过多次取样和跟踪，任何细微的波高、盐度、水温变化及氧含量变化等信息将被实时记录，科学家们会不断更新数据以掌握海洋生态的变化，使用波高传感器在不同时间找到波浪的不同位置，进而较早地采取应对措施。

连续生成移动数据的另一个应用示例是医疗保健行业，这是一秒钟也不能延迟的。例如，早产儿出生时免疫系统不发达，经常需要插管、注射或测试。在疾病或感染的情况下，病情的变化可能来得很快，这是特别危险的。因此，加拿大安大略理工大学建立了早产儿健康监测系统。传感器安装在早产儿身上及其周围，该系统收集由传感器和其他监测设备产生的心跳、呼吸和其他数据，每秒可生成 512 个监测值。通过不断更新移动数据，系统可帮助医护人员提前 24 小时预防早产儿败血症。

过去，企业数据库通常每周或每月进行批量数据集成，然后对结构化的静态数据进行分析。用于处理静态数据的数据库管理模式已经不能处理频繁、庞大、需要实时响应的流数据。IBM 提出的“流计算”已经成为解决移动数据问题的一种新方法。流计算系统利用多节点 PC 服务器的内存进行大量处理，无须等待数据存储或从企业数据库中提取数据，即可自动收集动态数据，并直接处理流动的多结构数据，其分析和响应速度可控制在微秒之内。

数据存储具有流动性，计算机会对首次进入系统的数据进行分析，比如银行想要导入一个欺诈风险控制机制，只要设计好逻辑后，系统将根据这个逻辑来决定是否含有欺诈行为，经过比较之后，交易数据才会被写入数据库。

智能电表生成的数据、网络应用程序等都可以先经过分析然后再存储在数据库中。在实际的企业竞争中，先对数据进行分析有可能在混乱多变的商业环境中比竞争对手早一点发现新趋势、新问题和新机遇。

1.3.3 大数据的“4V”之杂（variety）

除了庞大的数据量，数据处理中心还面临着另一个新的挑战，即数据的多样性和杂乱性。

传感器、智能设备已渗透到人们的日常生活和工作环境中，企业必须处理的数据变得越来越复杂，从 Web 内容、网络日志文件（点击流）、搜索索引到电子邮件、文档、系统传感数据等，它们都会成为企业处理的对象。而且，大多数数据无法以传统的数据库技术进行管理。既有的系统根本难以储存和分析这些数据，更别想从中解读出什么意义了。尽管有些企业已经很积极地想要驾驭大数据分析，但是绝大多数企业现在才真正开始了解大数据分析所蕴含的商机，或体会到不懂大数据分析将付出多高的代价。

然而，目前，一个组织的成功与否取决于其从各种各样的数据（不仅仅是结构化数据）中过滤出有价值的策略以助其业务发展的能力。例如，当收集 Twitter 消息时，你会看到一个 JSON（JavaScript 对象表示法）格式的结构化程序，但是消息本身的文本是包含各种表单的非结构化数据。以 Facebook 为例，平均每个月用户发布的内容多达 300 亿条，包括文本、图片、视频和音频等各种数据。仅发布的照片每天就超过 30 亿张，点击次数最多的“赞”就

达 27 亿次，更不用说无数的小通知、生日祝福、请柬等了。这些音频、文本和图像很难以编程的方式存储在传统的关系数据库中。

过去，数据库管理人员常常花费大量时间处理世界上仅占 15%的数据，这些数据是经过整齐格式化以适应现有格式的结构化数据。但事实上，世界上超过 85%的数据存在于社交媒体、电子商务等非结构化数据中，或者最多是半结构化数据。旧的数据库已经不能满足如此多种类的数据格式，现在迫切需要新的解决方案。

1.3.4　大数据的“4V”之疑（veracity）

在处理、分析和应用数据之前，有一个非常关键的问题：这些数据可靠吗？随着数据真实性和可靠性问题的日益突出，第四个“V”即数据的疑（veracity）也开始受到重视。这里的“疑”表示不确定性。

过去，不同的组织包括企业通常会对数据来源进行仔细的检查，所以数据的可靠性比较高。但在当今社会，随着社交网站和传感器技术蓬勃发展，不完整、不可靠的数据越来越多。2015 年，在美国某研究中心随机收集到的信息中每 100 个就有 8 个以上的数据难以确定其可靠性。

数据的可靠性不够高，必然会影响信息分析的价值。例如，实时集成和分析数以百万计的病人的医疗记录，可以帮助研究人员在传染病暴发的初期尽早作出回应，但如果同样的疾病或治疗的医疗记录被标示成几个不同的名字，则会导致分析软件无法解析，进而导致估计和预测结果不够准确。

造成大数据不确定性的因素有很多，主要包括以下几个方面。

1. 制造过程不可靠

即使许多事情的处理被设计得更加精确，其结果也总是难以预测或掌握。例如，先进的半导体晶片工艺不能保证 100%的产品合格率；配送路线无论计划得多好，都不可能在高峰时段准确预测从 A 到 B 的行程；传感器感知到的数据可能会因为环境变化或使用寿命的原因而出现错误，甚至在数据传输过程可能被恶意破坏，从而导致数据制造过程中的数值不准确。

2. 数据内容不可靠

特别是“人”产生的数据，尤其不可靠，主要原因有以下 3 个方面。

（1）故意欺骗。互联网上的信息有多少是真实的？一些人在 Facebook 上发布假新闻，一些人在微博上注册多个假 ID，还有一些公司在网络论坛上利用“刷新”来淹没负面评论……这样的例子有很多。任何个人或组织只要有足够的资金和时间，都可以制造虚假的舆论和声势。

（2）无心错误。无心的错误也可能导致数据来源不可靠的情况。例如，互联网上有一份含有硅胶的洗发水品牌名单，而这正是厂家故意攻击对手而伪造的。但在 A 收到错误数据后，A 好心地传给 B，B 传给 C，C 再传给更多的人。就像曾参杀人的故事一样，到最后真假难辨，连曾参的母亲也不再相信儿子的清白了。

（3）时序错乱。一位父亲焦急地在网上发了一份邮件，要求大家献出他们的爱心，帮助寻找失踪的 8 岁女儿，最后留下了联系电话。大多网友在收到了这封邮件之后好心转发，但

事实上，他的女儿在迷路后不到 2 小时就独自回家了。然而，这封于 2005 年发出的追查信件在互联网上流传了 6 年之久。2011 年，这个 14 岁的女孩通过电视镜头感谢了网友，并要求大家不要再找她。女孩的父亲说，在过去的 6 年里，她的家人平均每天接到 20 多个电话，最远的来自越南和乌克兰。这是因为时序错乱导致了不准确的数据。

3. 分析结果不可靠

这就是为什么天气预报系统尽管变得越来越复杂，但仍然不是 100%准确的原因。

因此，当试图收集大量的数据以期发现一些规律和趋势时，必须考虑这些复杂数据中存在哪些不确定因素，否则这些不可靠的数据将形成不确定的分析结果，从而影响后续决策的价值。

1.4 大数据的分类和采集方法

大数据获取是指从传感器和智能设备、企业在线系统、企业离线系统、社交网络、互联网平台等获取数据的过程。数据包括射频识别（radio frequency identification，RFID）数据、传感器数据、用户行为数据、社交网络交互数据、移动互联网数据及其他类型的结构化、半结构化和非结构化的海量数据。

由于数据源种类繁多，数据类型复杂，数据量大，数据生产速度快，传统的数据采集方法完全不能满足要求。因此，大数据采集技术面临着许多技术挑战。一方面要保证数据采集的可靠性和效率，另一方面要避免数据的重复。

1.4.1 大数据的分类

传统数据分为业务数据和行业数据，新数据包括内容数据、线上行为数据和线下行为数据三大类。因此，在大数据体系中，数据共分为以下 5 种。

（1）业务数据：账户数据、客户关系数据、买卖双方数据、库存数据等。

（2）行业数据：车流量数据、能耗数据、PM2.5 数据等。

（3）内容数据：应用日志、电子文档、机器数据、语音数据、社交媒体数据等。

（4）线上行为数据：页面数据、交互数据、表单数据、会话数据、反馈数据等。

（5）线下行为数据：车辆位置和轨迹、用户位置和轨迹、动物位置和轨迹等。

大数据的主要来源有以下几种。

（1）企业系统：客户关系管理系统、企业资源计划系统、库存系统、销售系统等。

（2）机器系统：智能仪表、工业设备传感器、智能设备、视频监控系统等。

（3）互联网系统：电商系统、服务行业业务系统、政府监管系统等。

（4）社交系统：微信、QQ、微博、博客、新闻网站等。

在大数据体系中，数据源与数据类型的关系如图 1.1 所示。

大数据系统从传统企业系统中获取相关的业务数据。

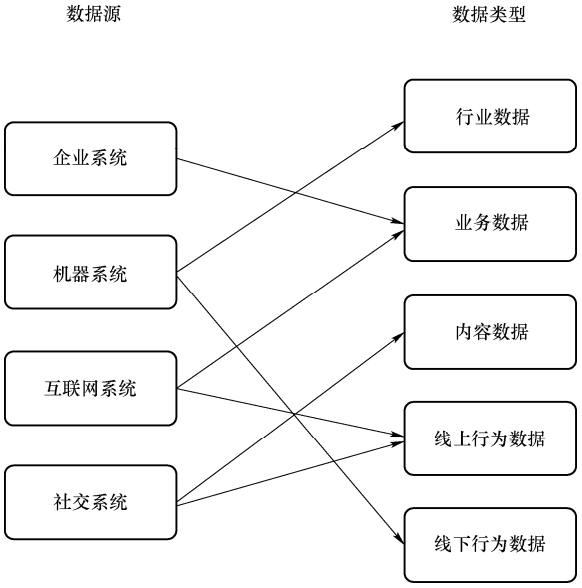

图 1.1　数据源与数据类型的关系

机器系统产生的数据分为两大类。一类是通过智能电表和传感器获取行业数据，如高速公路卡口设备获取车流数据、智能电表获取用电量等。另一类是通过各种监控设备获取线下行为数据，如人、动物、物体的位置和跟踪信息。

互联网系统将生成相关的业务数据和线上行为数据，如用户反馈和评价信息、用户购买的产品和品牌信息等。

社交系统产生大量的内容数据及线上行为数据，比如微博、照片。

因此，大数据的收集与传统数据的收集有很大的不同。

从数据来源的角度来看，传统的单一数据源来自传统企业的客户关系管理系统、企业资源规划系统和相关业务系统，而大数据还需要从社交系统、互联网系统、各种类型的机器和设备获得。在数据量方面，互联网系统和机器系统生成的数据量要远远大于企业系统生成的数据量。从数据结构的角度来看，传统的数据采集系统采集的是结构化数据，而大数据采集系统还需要采集大量的视频、音频、照片等非结构化数据，以及网页、博客、日志等半结构化数据。

1.4.2　数据采集的步骤与方法

数据采集是指收集符合数据挖掘研究要求的原始数据。原始数据是研究人员获得所需数据的主要或次要来源。数据采集不仅可以从现有的、可用的海量数据中收集和提取想要的二手数据，还可以通过问卷调查、访谈、沟通等手段获取一手数据。用任何一种方法获取数据的过程都可以称为数据采集。数据采集就是如何获取原始数据。如果把采集数据比喻成准备一顿饭，可以亲自下厨做饭（用一手数据），也可以叫外卖（用二手数据）。

数据采集是数据挖掘的基础，是数据准备的第一步。传统的数据采集来源是单一的，存储、管理和分析的数据量相对较小。其中大部分可以通过关系数据库和并行数据库进行处理。

在依靠并行计算提高数据处理速度方面，传统的并行数据库技术追求高一致性和容错性，难以保证其可用性和可伸缩性。

根据不同类型的数据源，大数据的采集方式也不同。为了满足大数据采集的需要，采用大数据的处理模式，即 MapReduce 分布式并行处理模式或基于内存的流处理模式。

针对 4 种不同类型的数据源，大数据的采集方法有以下几大类。

1. 数据库采集

传统企业使用 MySQL、Oracle 等传统关系数据库来存储数据，在大数据时代，Redis、MongoDB、HBase 等 NoSQL 数据库也经常被用于数据采集。通过在采集端部署大量数据库，并在这些数据库之间进行负载均衡和分片，企业可以完成大数据采集工作。

2. 系统日志采集

系统日志采集主要是收集公司业务平台每天产生的大量日志数据，供离线和在线大数据分析系统使用。高可用性、高可靠性和可扩展性是日志采集系统的基本特征。许多互联网企业都有自己的海量数据收集工具用于系统日志采集，如 Hadoop 的 Chukwa、Cloudera 的 Flume、Facebook 的 Scribe 等。这些工具都采用分布式架构，能够满足日志数据采集和每秒几百兆的传输需求。

3. 网络数据采集

网络数据采集是指通过网络爬虫或网站公共 API 从网站获取数据信息的过程。网络爬虫将开始从一个或多个初始 Web 页面的 URL 获取每个网页的内容，在爬行网页的过程中，它会不断地从当前页面中提取新的 URL，并将它们放入队列中，直到满足停止条件。这样就可以从 Web 页面中提取非结构化数据和半结构化数据，并存储在本地系统中。除了网络中包含的内容外，还可以使用带宽管理技术来处理网络流量的收集，如 DPI（deep packet inspection）或 DFI（deep /dynamic flow inspection）。

4. 感知设备数据采集

感知设备数据采集是指通过传感器、摄像头等智能终端对信号、图片或视频进行数据的自动采集。大数据智能感知系统需要对结构化、半结构化和非结构化的海量数据进行智能识别、定位、跟踪、访问、传输、信号转换、监控、前期处理和管理，其关键技术包括大数据源的智能识别、感知、适应、传输和访问。

1.5 大数据的基本概念

1.5.1 大数据技术的主要内容

大数据技术是指与大数据的收集、传输、处理和应用相关的技术，即利用非传统工具对大量结构化、半结构化和非结构化数据进行处理，获得分析和预测结构的一系列数据处理技术。

近年来，互联网、云计算、移动计算和物联网发展迅速，无处不在的移动设备、射频识别、无线传感器及互联网服务生成大量数据，对数据处理的实时性和有效性提出了更高的要求，传统的常规技术已让位于新的技术，包括分布式缓存、基于 MPP 的分布式数据库、分

布式文件系统、各种 NoSQL 分布式存储方案等。数据库、数据仓库、数据集市等信息管理技术，在很大程度上也是为了解决伴随大规模数据而出现的问题。被称为数据仓库之父的 Bill Inmon 早在 20 世纪 90 年代就提出了大数据技术。大数据技术的内容框架如图 1.2 所示。

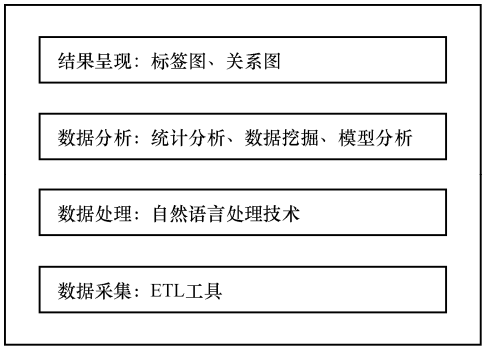

图 1.2　大数据技术的内容框架

（1）数据采集。ETL 是提取—转换—加载的意思，用于描述从源数据提取、转换和加载数据到目标数据的过程。

ETL 工具负责从分布式异构数据源提取数据，如关系型数据和平坦的数据文件，经过临时中间层的清洗、转换和集成，最后加载到数据仓库或数据集市，成为联机分析处理和数据挖掘的基础。

（2）数据存取。利用关系数据库和 NoSQL 存取数据。

（3）基础架构。应用云存储、分布式文件存储等架构。

（4）数据处理。自然语言处理（NLP）是一门研究人机交互语言问题的学科。自然语言处理的关键是使计算机理解自然语言。因此，自然语言处理也被称为自然语言理解或计算语言学。它是语言信息处理的一个分支，也是人工智能的核心课题之一。

（5）统计分析。主要包括假设检验、显著性检验、差异分析、相关分析、t 检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、曲线估计、因子分析、聚类分析、主成分分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）等。

（6）数据挖掘。主要包括分类、估计、预测、相关性分析或关联规则、聚类、描述和可视化、复杂数据类型挖掘等。

（7）模型分析。包括预测模型、机器学习、建模仿真。

（8）结果呈现。大数据处理结果利用云计算、标签云和关系图等方式呈现。大数据处理问题实际上是由信息设施的处理能力与数据处理问题之间的矛盾引起的。大数据的体量、时间、地点客观上加剧了矛盾的演变，加上计算机内存容量有限，造成了传统模式输入/输出（I/O）的压力增加，缓存命中率很低，很难获得最好的平衡性能、最低的功耗与成本，从而使得计算机系统难以处理超过 PB 级的大数据。

1.5.2 大数据的处理过程

一般来说，大数据处理的流程分为 4 步，分别是大数据采集、大数据导入与预处理、大数据统计与分析，最后一步是大数据挖掘。

（1）大数据采集。大数据的采集方法在之前已经简单介绍过了，在此不再赘述。

（2）大数据导入与预处理。采集端本身有一个大型数据库，数据首先会经过一些简单的数据清洗和预处理工作，之后会被导入一个集中的大型分布式数据库或分布式存储集群。部分用户在导入数据时还使用 Twitter 的 Storm 平台对数据进行流计算，以满足部分业务实时计算的要求。大数据导入与预处理的主要特点是导入的数据量大，常常达到 100 MBps，甚至是千兆的数量级。

（3）大数据统计与分析。统计和分析分布式数据库的主要用途是对储存在分布式计算集群中的大数据进行总结和分类。对于特定的需求可以使用不同的处理工具，如使用工具 EMC GreenPlum 以满足共同分析的需求，使用 Oracle Exadata 满足一些实时的应用需求，使用基于 MySQL 储存 InfoBright 列类型数据，使用基于 Hadoop 存储半结构化数据。统计与分析部分的主要特点是分析中涉及的数据量巨大，对系统资源，特别是 I/O 资源占用极大。

（4）大数据挖掘。与上述统计分析过程不同的是，数据挖掘一般没有任何预设的主题，主要是对现有数据进行基于各种算法的计算，以达到预测效果，从而实现一些高层数据分析的需求。典型算法有 K-means 聚类算法、SVM 统计学习算法和 NaiveBayes 分类算法，主要使用的工具有 Hadoop Mahout 等。这个过程的主要特点是用于数据挖掘的算法非常复杂，计算所涉及的数据量和计算量非常大，常用的数据挖掘算法主要是单线程的。

大数据处理过程至少要满足以上 4 个基本步骤，才能成为一个相对完整的处理过程。

1.5.3 大数据的关键问题与关键技术

1. 大数据的关键问题

在大数据环境中，数据源非常丰富，数据类型多样，存储和分析挖掘的数据量巨大，对数据表示的要求很高，强调数据处理的效率和可用性。

1）非结构化和半结构化数据处理

如何利用信息技术处理非结构化和半结构化数据是一个重要的研究课题。如果将通过数据挖掘提取粗糙知识的过程称为首次挖掘，那么可将粗糙知识与量化的主观知识（包括具体经验、常识、直觉、情景知识和用户偏好）相结合生成智能知识的过程称为二次挖掘。从一次挖掘到二次挖掘，是数量上质的飞跃。

鉴于大数据的半结构化和非结构化特征，基于大数据的数据挖掘产生的结构化粗糙知识（潜在模式）也伴随着一些新的特征。这些结构化的粗糙知识可以被主观知识加工转化为半结构化和非结构化的智能知识。对智能知识的搜索体现了大数据研究的核心价值。

2）大数据复杂性与系统建模

大数据复杂性和不确定性的表征方法和大数据系统建模的突破是实现大数据知识发现的前提。从长远来看，大数据的个体复杂性和随机性所带来的问题将促进大数据数学结构的形成，从而完善大数据统一理论。短期内，学术界鼓励制定结构化数据与半结构化、非结构

化数据转换的一般原则，支持大数据的交叉应用。

大数据的复杂形式引发了大量关于测量和评估粗糙知识的研究。管理科学中的已知优化、数据包络分析、期望理论和效用理论可以用于研究如何将主观知识融入数据挖掘生成粗糙知识的二次挖掘中，人机交互将发挥关键作用。

3）大数据异构性与决策异构性影响知识发现

数据异构和决策异构之间的关系对大数据知识发现和管理决策具有重要影响。由于大数据本身的复杂性，这无疑是一个重要的科学研究课题，对传统的数据挖掘理论和技术提出了新的问题。在大数据背景下，管理决策面临两个异质性问题，即数据异质性和决策异质性。传统的管理决策模型依赖于商业知识的学习和实践经验的积累，而管理决策则基于数据分析。

大数据的出现改变了传统的管理决策结构模式，大数据对管理决策结构的影响将成为一个开放性的科学研究问题。此外，决策结构的变化要求人们探索如何进行二次挖掘，以支持更高层次的决策。不管大数据带来的数据异构性如何，大数据中的粗糙知识仍然可以看作是一种挖掘。需要搜索二次挖掘产生的智能知识，作为数据异构和决策异构之间的桥梁。探究大数据背景下的决策结构是如何改变的，等同于研究如何让决策者的主观知识参与决策的过程。

寻找对大数据进行科学处理的一般性方法是一种对美的探究，尽管这种探究是困难的，但是如果我们找到结构化数据，则已有的数据挖掘方法将成为很好的大数据挖掘工具。

上述大数据的重要技术问题，仅仅是研究大数据的一个起点。此外，还存在一些数据科学问题，包括基于数据库的知识发现规则和基于开放数据源的知识发现规则，以及大数据挖掘的全局和/或局部解决方案的存在等。在不久的将来，将会取得突破性的科学研究和应用成果。

2. 大数据的关键技术

针对上述大数据的关键问题，大数据的关键技术主要包括流处理、并行化、摘要索引和可视化。

1）流处理

随着业务需求的增长和业务流程变得更加复杂，我们的研究重点是数据流而不是数据集。

决策者感兴趣的是捕获实时结果，需要能够在数据流发生时处理数据流的架构，但当前的数据库技术并不适合数据流处理。

例如，计算一组数据的平均值可以采用传统的方法。但是计算移动数据的平均值，需要更高效的算法。要创建一个数据流统计，需要逐步添加或删除数据块，进行移动平均计算，而目前的数据库技术还不成熟。

此外，在应用数据流技术时，围绕数据流的生态系统还不发达。如果您正在与供应商谈判一个大数据项目，您必须知道数据流处理对项目是否重要，以及供应商是否能够提供它。

2）并行化

小数据的情况类似于桌面环境下，磁盘存储容量为 1～10 GB，中数据的数据量为 100 GB～1 TB，大数据分布存储在多台机器上，包含从 1 TB 到多 P B 的数据。如果在分布式数据环境中短时间内处理数据，那么就需要分布式处理。Hadoop 是一个分布式并行处理平台，由一个支持分布式并行查询的大型分布式文件系统组成。

3）摘要索引

摘要索引是创建预先计算的数据摘要以加速查询执行的过程。摘要索引的关键是必须为要执行的查询进行计划。随着数据的快速增长，对摘要索引的需求将不会停止。无论基于长期还是短期的考虑，摘要索引需求的发展必须有一个明确的战略。

4）可视化

数据可视化包括科学可视化和信息可视化。可视化工具是实现可视化的重要基础，主要分为两大类。

一是探索可视化工具，它可以帮助决策者和分析师探索不同数据之间的联系，这是一种可视化的洞察力。类似的工具包括 Tableau、TIBCO 和 QlikView。

二是叙事可视化工具，它可以以独特的方式探索数据。例如，如果需要以可视化的方式在时间序列中按地区查看企业的销售业绩，可视化格式将会被提前创建。数据将按地区逐月显示，并按照预先定义的公式进行排序。