项目 1　搭建单节点 Hadoop 整合平台

近年来，“ 大数据 ”一词已经越来越被人们所熟知，它引领着新一轮的信息技术革命，人类也步入了一个崭新的大数据时代。随着大数据的概念、技术、应用不断地深入到社会中的各个方面，它在迅速而深刻地改变着我们工作方式和生活方式。

Hadoop 作为大数据行业中的一位元老级成员，可以说是专门为大数据而生的。它提供了一种可以高效处理海量规模数据的方式，为大数据的发展提供了巨大帮助。这里以单节点 Hadoop 整合平台的搭建为例，让大家对大数据以及 Hadoop 平台有一个初步的了解。Hadoop 整合平台指的是 Hadoop 软件自身，其中包括了 HDFS 分布式文件系统和 MapReduce 分布式计算框架，同时搭配上 HBase 分布式数据库，以及 HBase 所需要使用到的 Zookeeper 分布式协调服务。

任务 1.1　认识大数据

随着计算机硬件性能的发展、互联网传输速度的提升、智能便携移动终端设备的出现和快速普及，人类所产生的数字信息量也越来越大，近几年来甚至是以爆发式的速度增长。以 2017 年的天猫 “ 双十一 ”购物狂欢节活动为例，仅一天的成交额就达到了 1 682 亿元，产生的交易记录达到 14.8 亿条，产生的物流订单超过 8 亿单。而在最近几年，这样会产生海量规模数据的应用数不胜数。2017 年，新浪微博的每日活跃用户数量达到 1.65 亿，而微博内容的总条数已累积超过千亿。YouTube 在 2017 年中平均每天视频播放的总时长超过了 10 亿小时，这个时长的视频如果由一个人来看的话需要约 10 万年的时间，这足够使用光速旅行从银河系的一端飞到另一端，并且 YouTube 上的新视频还在以每分钟 300 小时的数量增长。而 2017 年春节期间，使用微信收发的红包总数量达到了 460 亿个，同时通过微信的音视频通话功能进行新年问候的总时长也是达到了 21 亿分钟。全球最大的搜索引擎谷歌（Google）每分钟的搜索量可达 278 万次，全球最大的社交网站脸书（Facebook）每分钟的点赞数量超过 400 万次。预计

到了 2020 年，全球的数据总量将会达到 44 ZB，这换算成我们更为熟悉的单位钛字节 TB 和吉字节 GB，那就是 440 亿 TB 和 44 万亿 GB。

2013 年被定义为世界大数据元年，它是世界步入大数据时代的一个标志。想要深入了解大数据，可以从理论（Theory）、技术（Technology）、实践（Utilization）三个方面来进行，如图 1-1-1 所示。

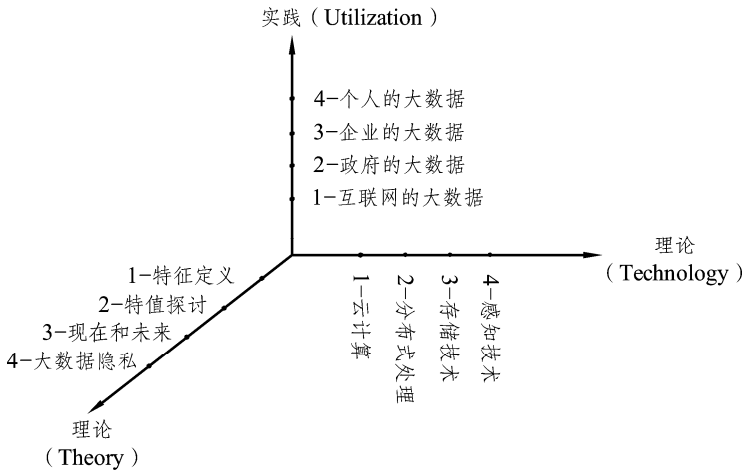

图 1-1-1　深入理解大数据

1.1.1　从理论上了解大数据

最早提出大数据时代的麦肯锡公司在其有关大数据研究的报告中将大数据定义为大小超出常规的数据库工具获取、存储、管理、分析能力的数据集。而 IBM 公司将大数据的特征归纳为 5 个“ V ”，分别为 Volume、Velocity、Variety、Value、Veracity。

Volume 代表的是大容量，即数据体量巨大。大数据的计量单位符号一般都是 PB（1 024 TB）或 EB（1 024 PB）级别，甚至现在在朝着 ZB（1 025 EB）级别发展。

Velocity 代表的是高速度，即处理速度快。由于数据量巨大，只有达到了一定的处理速度，才能实时地从数据中发掘出有用的信息。

Variety 代表的是多样性，即数据类型的多样性。这里的数据类型包括了视频、音频、图片、地理信息、日志等多种多样的类型，不再只是传统关系型数据库中的结构化数据。

Value 代表的是价值，即以低成本创造高价值。现今随着互联网和计算机硬件的飞速发展，数据的产生和收集变得越来越简单，数据的获取和存放的成本也越来越低。而当很多不全面、不连续，甚至看似无关联的数据，在达到一定规模时，便可以从中产生出不可估量的价值。

Veracity 代表的是真实，指数据的质量真实可靠。只有从用户和业务系统产生的真实数据才是有价值的，才能从中挖掘出有用的信息。

当下，大数据应用价值已在各行各业凸显。大数据帮助政府实现市场经济调控、公共卫生安全防范、灾难预警、社会舆论监督、犯罪预防；大数据帮助城市实现智慧交通、智慧楼宇、智慧公共设施；大数据帮助医疗机构建立患者的疾病风险跟踪机制，帮助医药企业提升药品的临床使用效果，帮助艾滋病研究机构为患者提供定制的药物；大数据帮助航空公司节省运营成本，帮助电信企业实现售后服务质量提升，帮助保险企业识别欺诈骗保行为，帮助快递公司检测分析运输车辆的故障险情以便提前预警维修，帮助电力公司有效识别预警即将发生故障的设备。不管大数据的核心价值是不是预测，基于大数据形成决策的模式已经为不少的企业带来了盈利和声誉。

而大数据时代也有其面临的问题，其中最主要的就是侵犯隐私问题。当在不同网站上注册了个人信息之后，可能这些信息就已经被自动扩散出去了。当人们莫名其妙地接收到各种邮件、电话、短信的滋扰时，不会想到自己的电话号码、邮箱、生日、购买记录、收入水平、家庭住址、亲朋好友等私人信息早就被各种商业机构非法存储或出售给其他任何有需要的企业和个人。即使用户在某个地方删除了相关信息，也许这些信息已经被其他人转载或保存，更有可能已经被存为网页快照，早已提供给任意用户搜索。因此在大数据背景下，很多人都在积极抵制无底线的数字化，这种大数据和个体之间的博弈还会一直持续下去。

当很多互联网企业意识到隐私对于用户的重要性时，为了继续得到用户的信任，采取了很多办法。如谷歌承诺保留用户的搜索记录时间为 9 个月，而有的浏览器厂商提供了无痕冲浪模式，还有社交网站拒绝公共搜索引擎的爬虫，并将提供出去的数据全部采取匿名方式处理等。被誉为大数据商业应用第一人的维克托 · 迈尔 · 舍恩伯格在《大数据时代》一书中提出了一些如何有效保护大数据背景下隐私权的建议，如减少个人信息的数字化、建立完善隐私权保护法、增强数字隐私权基础设施建设等。

1.1.2　从技术上了解大数据

大数据常和云计算联系在一起，因为实时的大型数据集的分析需要分布式处理框架来向数十、数百甚至上万的计算机分配工作，它的特色在于对海量数据的挖掘。如今在谷歌、亚马逊、Facebook 等一批互联网企业引领下，创建了一种行之有效的模式，即云计算提供基础架构平台，大数据应用可以运行在这个平台之上，如图 1-1-2 所示。业内这样认为两者的关系：没有大数据的信息积淀，云计算的处理能力再强大，也难以找到用武之地；而没有云计算的处理能力，大数据的信息积淀再丰富，也同样没有用武之地。大数据和云计算两者之间有效结合之后，便可以提供更多基于海量业务数据的创新型服务，并通过云计算技术的不断发展降低大数据业务的创新成本。

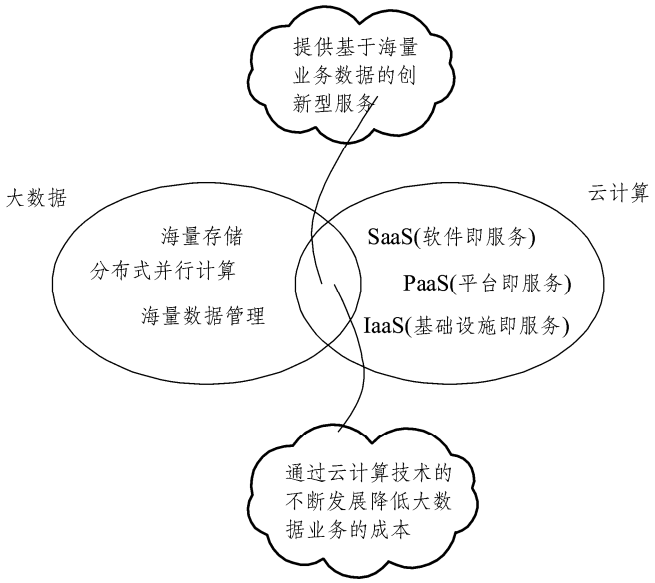

图 1-1-2　大数据与云计算的关系

大数据与云计算的最显著区别表现在两个方面。一方面是概念上的不同，云计算改变了整个 IT（Information Technology）行业，而大数据则是改变了构建在 IT 行业基础上的各种业务，而其中大数据又必须以云计算作为基础架构，才能得以顺畅运行。另一方面是受众群体不同，云计算属于 CIO（首席信息官）等技术相关职位人员所关心的技术层面，是一个进阶的 IT 行业解决方案；而大数据则是属于 CEO（首席执行官）等管理相关人员所关心的业务层面的产品。

大数据可以抽象地分为大数据存储和大数据分析两个部分，其中大数据存储的目的是为了支撑大数据分析业务。目前来看，大数据存储和大数据分析已经发展成为两个截然不同的计算机技术领域。大数据存储致力于研发可以扩展至 PB 甚至 EB 级别的数据存储平台，而大数据分析关注在最短时间内处理大量不同类型的数据集的计算平台。

大数据的技术最核心、最重要的部分就是分布式处理技术。分布式处理系统可以将处于不同地点、具备不同功能、拥有不同数据的多台计算机用通信网络连接起来，然后在控制系统的统一管理下，协调完成信息处理任务。本书所要介绍的 Hadoop 就是大数据分布式处理系统的典型代表。它所包含的 MapReduce 软件框架，能以一种可靠、高效、可伸缩的方式对大数据进行分布式处理。同时 MapRedcue 还利用 HDFS 分布式文件系统来存储数据，采用将数据移动到计算的方式进一步提高大数据分布式处理的效率。

另外，大数据的采集与感知技术的发展也是紧密联系在一起的。以传感器技术、指纹识别技术、RFID 技术、坐标定位技术等为基础的计算机感知能力的提升同样是物联网发展的基石。全世界的工厂自动化生产线、汽车、电表等设备上有着无数的数据传感器，随时测量和传递着有关位置、运动、震动、温度乃至空气中化学物质变化等信息，都会产生海量的数据。而随着智能移动终端设备的发展和普及，感知技术可谓迎来了发展的高峰期。除了地理位置信息被广泛地应用之外，一些新的感知手段也开始出现，如手机中内嵌的指纹传感器，即将面世的可以检测空气污染及危险化学药品的带有嗅觉传感器的智能手机，感知用户当前情绪的手机智能技术等。这些事物被逐渐感知和捕获的过程，其实就是世界被逐渐数据化、信息化的过程。

1.1.3 从实践上了解大数据

大数据的主要来源是互联网。互联网上的数据每年增长 50%，几乎每两年便会翻一番，而目前世界上 90%以上的数据都是最近十年左右才产生的。据预测，到 2020 年，全球将总共拥有 35 ZB 的数据量。互联网是大数据发展的前哨阵地，随着 Web 2.0 时代的发展，人们已经越来越习惯于将自己的生活通过网络进行数据化，方便分享与记录与回忆。

互联网上的大数据很难清晰地界定分类界限，大致可以分为以下五类：

（1）用户行为数据。这类数据主要用于精准广告投放、内容推荐、行为习

惯分析、喜好分析、产品优化等。

（2）用户消费数据。这类数据主要用于精准营销、信用记录分析、活动促销、理财推荐等。

（3）用户地理位置数据。这类数据主要用于在线或离线的推广、线上到线下的推广、商家推荐、交友推荐等。

（4）互联网金融数据。这类数据主要用于点对点网络借款、小额贷款、支付、信用分析、供应链金融等。

（5）用户社交数据。这类数据主要用于潮流趋势分析、流行元素分析、受欢迎程度分析、舆论监控分析、社会问题分析等。

近几年来，各国政府也越来越重视大数据的发展。美国政府曾投资 2 亿美元致力于拉动大数据相关产业的发展，将大数据战略上升到国家意志层面。美国政府将大数据定义为“未来的新石油”，并表示一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分。未来，对数据的占有和控制甚至将成为国家核心资产。在国内，政府各个部门都掌握着构成社会基础的原始数据，如气象数据、金融数据、信用数据、电力数据、煤气数据、自来水数据、道路交通数据、客运数据、安全刑事案件数据等。这些数据在每个政府部门里面看起来都是单一的、静态的，但是如果政府可以将这些数据关联起来，并对这些数据进行有效的关联分析和统一管理，这些数据必定将产生无法估量的价值。

而现在的城市也已经逐渐在走向智能化，如智能电网、智慧交通、智慧医疗、智慧环保、智慧城市等。这些都是依托于大数据实现的，可以说大数据是智能化的核心能源。从国内整体的投资规模来看，自 2012 年至今，全国开始进行智慧城市建设的城市超过了 290 个，通信网络和数据平台等基础设施的建设投资规模接近 5 000 亿元。大数据为智慧城市的各个领域提供角色支持。在城市规划方面，通过对城市地理、气象等自然信息，以及经济、社会、文化、人口等人文社会信息的挖掘，可以为城市规划提供决策依据，强化城市管理服务的科学性和前瞻性；在交通管理方面，通过对道路交通信息的实时挖掘，能有效地缓解交通拥堵，并快速响应突发状况，为城市交通的良性运转提供科学的决策依据；在舆情监控方面，通过网络关键词搜索及语义智能分析，能提高舆情，分析的及时性、全面性，快速全面地掌握社情民意，提高公共服务能力，快速高效地应对网络突发的公共事件，打击违法犯罪；在安防和防灾领域，通过大数据的挖掘，可以及时发现人为和自然灾害、恐怖事件等，提高应急处理能力和安全防范能力。另外作为国家的管理者，政府还应该将国家所掌控的数据逐步开放，提供给更多有能力的组织、机构、个人来分析和利用，加快大数据对

社会发展的促进作用。

对于企业来说，其管理者最关注的还是报表曲线的背后所具有的信息，以及通过这些报表曲线应该做出怎样的决策，这些也都需要通过数据来传递和支撑。在理想的状态下，大数据可以改变公司的影响力，它可以节省开支、增加利润、取悦买家、增加用户忠诚度、转化潜在客户、增加吸引力、提高竞争力、开拓市场。对于企业的大数据，随着数据逐渐成为企业的一种资产，数据产业会向传统企业的供应链的模式发展，最终形成数据供应链。这主要表现在两个方面。第一，外部数据的重要性日益超过内部数据，互联网时代单一企业的内部数据与整个互联网数据相比只是沧海一粟。第二，能够提供包括数据供应、数据整合、数据加工、数据应用等多环节服务的公司会有明显的综合竞争优势。

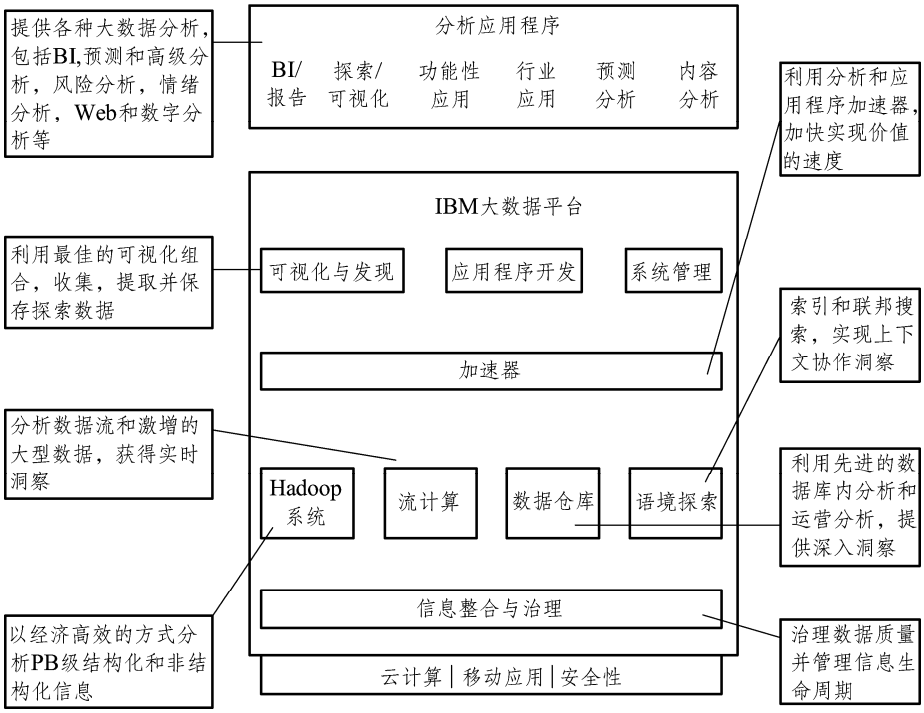

图 1-1-3　IBM 大数据平台架构

从 IT 产业的发展来看，第一代 IT 巨头大多是提供 ToB（面向企业）服务的，如 IBM、微软、Oracle、SAP、HP 这类传统 IT 企业。而第二代 IT 巨头大多是提供 ToC（面向用户）服务的，如雅虎、Google、亚马逊、Facebook 这类互联网企业。大数据到来之前，这两类公司彼此之间基本上是井水不犯河水，相

互之间的竞争关系较少。但大数据时代来临之后，这两类公司开始了直接的竞争。出现这种现象的主要原因是；在互联网巨头的带动下，传统 IT 巨头的客户普遍开始从事电子商务业务，正是由于客户进入了互联网，所以传统 IT 巨头们也被拖入了互联网领域。以 IBM 为例，在上一个十年中成功抛弃了计算机硬件开发与生产，成功转向了软件和服务，并提出了大数据平台架构，如图 1-1-3 所示。

大数据对于个人而言现在还只是需要数据被采集以及享受最终产生的结果，并且这种采集和结果都还是分布在不同的平台和应用之上的。未来，每个用户将可以在互联网上注册个人的数据中心，以存储个人的大数据信息；同时用户可以确定哪些个人数据可被采集，并通过可穿戴设备或植入芯片等感知技术来捕获个人的大数据信息，如牙齿数据、心律数据、体温数据、视力数据、记忆能力、地理位置信息、社会关系数据、运动数据、饮食数据、购物数据等。用户可以将其中的牙齿数据授权给某牙科诊所使用，由他们监控和使用这些数据，进而为用户制订有效的牙齿防治和维护计划；也可以将个人的运动数据授权提供给某个运动健身机构，由他们检测自己的身体运动机能，并有针对性的制订和调整个人的运动计划；还可以将个人的消费数据授权给金融理财机构，由他们帮客户制订合理的理财计划并对收益进行预测。

以个人为中心的大数据将具备以下的特征。首先数据仅留存在个人中心，其他的第三方机构只被授权使用，且必须接受用后即销毁的监管。其次，采集个人数据应该明确分类，除了国家立法明确要求接受监控的数据外，其他类型数据都由用户自己决定是否被采集。最后数据的使用只能由用户授权，数据中心可帮助用户监控个人数据的整个生命周期。个人数据中心的愿望要实现还需要一定的时间，其面临着隐私保护和数据监管这两个最大的难题，这会是异常激烈的博弈。

1.1.4 大数据的处理流程

大数据处理的方法有很多，经过长时间的实践，可以总结出一个基本的大数据处理流程，其中包括了四步，分别为采集、导入和预处理、统计和分析、挖掘。

大数据的采集是指利用多个数据库来接收来自客户端（Web、App、传感器）的数据，并且用户可以通过这些数据库来进行简单的查询和处理工作。大数据采集过程中的主要问题就是高并发数，因为同时有可能会有成千上万的用户来