第 1 章

大数据概述

学习目标

了解大数据的产生过程，理解大数据的定义，掌握大数据的特征，理解大数据思维的原理，掌握大数据的处理流程，了解数据采集的方法，掌握数据预处理的几个环节，了解大数据的发展现状及应用领域，了解大数据的发展趋势。

1.1 大数据的概述

1.1.1 大数据的产生

大数据是信息通信技术发展积累至今，按照自身技术发展逻辑，从提高生产效率向更高级智能阶段的自然生长。无处不在的信息感知和采集终端为人们采集了海量的数据，而以云计算为代表的计算技术的不断进步，则为我们提供了强大的计算能力。

从采用数据库作为数据管理的主要方式开始，数据产生方式大致经历了被动产生、主动产生和自动生成三个阶段，这种数据产生方式的巨大变化最终导致大数据的产生。

1. 运营式系统阶段

数据库的出现使得数据管理的复杂度大大降低。在实际使用中，数据库大多为运营系统所采用，作为运营系统的数据管理子系统，如超市的销售记录系统、银行的交易记录系统、医院病人的医疗记录系统等。

人类社会数据量的第一次大的飞跃正是在运营式系统开始广泛使用数据库时开始的。这个阶段的最主要特点是，数据的产生往往伴随着一定的运营活动，而且数据是记录在数据库中的，例如，商店每售出一件产品就会在数据库中产生一条相应的销售记录。这种数据的产生方式是被动的。

2. 用户原创内容阶段

互联网的诞生促使人类社会数据量出现第二次大的飞跃，但是真正的数据爆发产生于

Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容。这类数据近几年一直呈现爆炸式的增长。一是以博客、微博和微信为代表的新型社交网络的出现和快速发展，增强了用户产生数据的意愿；二是以智能手机、平板电脑为代表的新型移动设备的出现，使人们在网上发表自己意见的途径更为便捷。这个阶段的数据产生方式是主动的。

3. 感知式系统阶段

感知式系统的广泛应用促使人类社会数据量出现第三次大的飞跃，最终导致了大数据的产生，今天我们正处于这个阶段。

随着技术的发展，人们已经有能力制造极其微小的带有处理功能的传感器，并将其广泛地布置于社会的各个角落，通过它们来对整个社会的运转进行监控。这些设备会源源不断地产生新数据，这种数据的产生方式是自动的。

简单来说，数据产生经历了被动、主动和自动三个阶段。这些被动、主动和自动的数据共同构成了大数据的数据来源，但其中自动式的数据才是大数据产生的最根本原因。

1.1.2 大数据的定义

通俗地讲，大数据是指在一定时间内无法用常规软件工具进行抓取、管理、处理和分析的数据集合。最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡，它在研究报告 *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数据库工具获取、存储、管理和分析能力的数据集，即大数据是现有数据库管理工具和传统数据处理手段很难处理的大型、复杂的数据集，涉及采集、存储、搜索、共享、传输和可视化等方面。

1.1.3 大数据的特征

大数据有 5 个特征，归纳为 5V：Volumn（数据容量）、Variety（数据种类）、Velocity（数据处理速度）、Value（数据价值）和 Veracity（数据真实性），如图 1.1 所示。

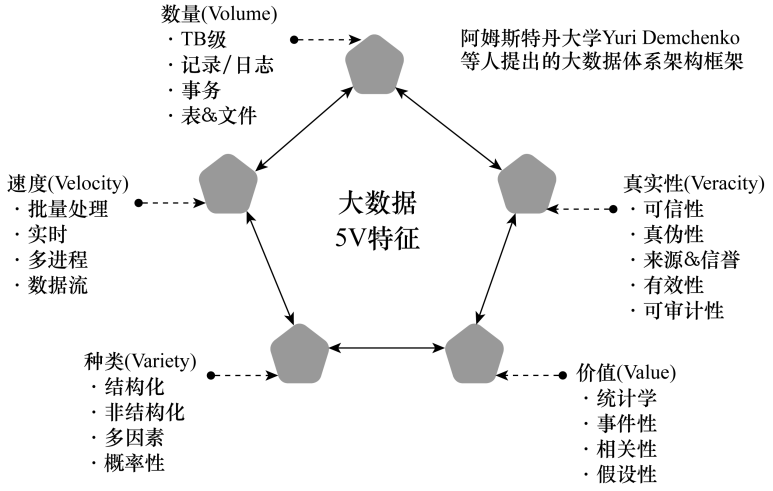

图 1.1 大数据的 5V 特征

1. Volumn(数据容量)

大数据的特征首先表现为数据量巨大，包括采集、存储和计算量都非常大。存储单位至少是 TB、PB、EB 或 ZB。随着网络及信息技术的高速发展，数据开始爆发性增长。人和物的轨迹通过社交网络、移动网络、各种智能终端等被记录下来，产生大量的数据。微博、照片、录像等都成为数据的来源，企业也面临着数据量的大规模增长自动化传感、生产监测、环境监测等大量自动或人工产生的数据。全球云指数报告中预测称，到 2021 年全球云数据中心流量将达到每年 19.5ZB，云数据中心流量将占总数据中心流量的 95%。

2. Variety(数据种类)

大数据的数据类型和来源多样化，包括多种不同格式和不同类型的数据。数据来源包括人与系统交互时与机器自动生成，来源的多样性导致数据类型的多样性。根据数据是否具有一定的模式、结构和关系，数据可分为三种基本类型：结构化数据、非结构化数据、半结构化数据。

(1) 结构化数据，指遵循一个标准的模式和结构(conform to a data model or schema)，以二维表格的形式存储在关系型数据库里的行数据，如信息管理系统数据、医疗系统数据等数据间因果性较强的结构化数据。结构化数据是先有结构，后产生数据。由于关系型数据库发展较为成熟，因此结构化数据的存储、分析方法也发展得较为全面，有大量的工具支持结构化数据分析，分析方法大都以统计分析和数据挖掘为主。其中，关系型数据库(relational database)是创建在关系模型基础上的数据库，关系模型即二维表格模型，因此一个关系型数据库包括一些二维表且这些表之间具有一定的关联。关系型数据库可运用 SQL 语言通过固有键值提取相应信息。

(2) 非结构化数据，是指不遵循统一的数据结构或模型的数据(如文本、图像、视频、音频等)，不方便用二维逻辑表来表现。这部分数据在企业数据中占比大，且数据间没有因果关系，增长速率更快。非结构化数据更难被计算机理解，不能直接被处理或用 SQL 语句进行查询。非结构化数据常以二进制大型对象(BLOB，将二进制数据存储为一个单一个体的集合)形式，整体存储在关系型数据库中；或存储在非关系型数据库中(NoSQL 数据库)，其处理分析过程要更为复杂。

(3) 半结构化数据，是指有一定的结构性，但本质上不具有关系性，介于完全结构化数据和完全非结构化数据之间的数据。它可以说是结构化数据的一种，但是结构变化很大。因此，为了了解数据的细节，不能将数据简单按照非结构化数据或结构化数据进行处理，需要特殊的存储(化解为结构化数据/用 XML 格式来组织并保存)和处理技术。半结构化数据包含相关标记，用来分隔语义元素以及对记录和字段进行分层。因此，它也被称为自描述的结构(以树或者图的数据结构存储的数据)，先有数据，再有结构。常见的半结构化数据：XML 文件、HTML 文档和 JSON 文件等，主要来源包括电子转换数据(EDI)文件、扩展表、RSS 源、传感器数据。

(4) 元数据，一种用于描述其他数据的数据。元数据可说明已知数据的一些属性信息(如数据长度、字段、数据列、文件目录等)，提供数据系谱信息(包含数据的演化过程)

和数据处理的起源。元数据可分为三种不同的类型，分别为记叙性元数据、结构性元数据和管理性元数据，主要由机器生成并添加到数据集中。例如数码照片中提供文件大小和分辨率的属性文件。元数据的作用也类似于数据仓库中的数据字典。

数据源不同导致非结构化数据越来越多，需要进行数据清洗、整理筛选等操作，将其变为结构化数据。这意味着要在海量、种类繁多的数据间发现其内在关联。互联网时代，各种设备通过网络连成了一个整体。进入以互动为特征的时代，用户不仅可以通过网络获取信息，还成了信息的制造者和传播者。这个阶段，不仅是数据量开始了爆炸式增长，数据种类也变得繁多。这必然促使人们对海量数据进行分析、处理和集成，找出原本看似毫无关系的那些数据之间的关联性，把似乎没有用的数据变成有用的信息，以帮助人们做出判断。

结构化数据、非结构化数据、半结构化数据的比较如表 1.1 所示。

表 1.1　结构化数据、非结构化数据、半结构化数据的比较

对比内容	定义	结构与数据的关系	示例
结构化数据	具有数据结构描述信息的数据	先有结构，后有数据	各类表格
非结构化数据	不遵循统一的数据结构或模型的数据	只有数据，没有结构	图形、图像、音频、视频等
半结构化数据	处理结构化和非结构化数据之间的数据	先有数据，再有结构	HTML 文档等，一般是自描述的，数据的内容与结构混在一起

3. Velocity（数据处理速度）

数据增长速度快，对数据处理速度和时效性也越来越高。数据来源广，导致数据不断被添加、处理和分析，才能迎接新信息的大量涌入，比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。在网络时代，通过高速的计算机和服务器，创建实时数据流已成为流行趋势。企业不仅需要了解如何快速创建数据，还必须知道如何快速处理、分析并返回给用户，以满足他们的实时需求。

4. Value（数据价值）

相比于传统的小数据，大数据最大的价值在于通过从大量不相关的各种类型的数据中，挖掘出对未来趋势与模式预测分析有价值的数据，并通过机器学习方法、人工智能方法或数据挖掘方法进行深度分析，发新规律和新知识，并运用于农业、金融、医疗等各个领域，从而最终达到改善社会治理、提高生产效率、推进科学研究的效果。但是大数据的价值密度低，大概 80%的数据都是无效数据，比如，在视频监控过程中，可能获取到的有用数据只有一两秒。因此，如何使用数据分析与挖掘算法快速地提取出数据中的有效价值信息，是当今面临的研究难题。

5. Veracity（数据真实性）

真实性就是数据的准确性和可信赖程度，说明数据是准确的而不是冒充的。对于不真实的数据需要通过数据清洗、集成和归约等数据处理流程，将其转换为高质量的数据，再进行后续的数据分析与挖掘工作。

1.1.4　大数据思维

随着大数据技术的快速发展，大数据所创造的价值深刻改变了人们的生活、工作和思维方式。大数据时代，人们的思维也从传统思维方式向大数据思维转变。大数据思维关键的转变在于从自然思维转向智能思维，使得大数据像具有生命力一样，获得类似于人脑的智能、甚至智慧。大数据思维是客观存在，是新的思维观。用大数据思维方式思考问题，解决问题是当下企业潮流。大数据思维开启了一次重大的时代转型。

1. 从“流程”核心转变为“数据”核心

大数据时代，由大数据运营到运营大数据转变，以数据为核心，用数据核心思维方式思考问题，解决问题。计算模式也发生了转变，从“流程”核心转变为“数据”核心。

Hadoop 体系的分布式计算框架已经是以“数据”为核心的范式。非结构化数据及分析需求，将改变 IT 系统的升级方式：从简单增量到架构变化。例如，IBM 将使用以数据为中心的设计，目的是降低在超级计算机之间进行大量数据交换的必要性。

科学进步越来越多地由数据来推动，海量数据给数据分析既带来了机遇，也构成了新的挑战。大数据往往是利用众多技术和方法，综合源自多个渠道、不同时间的信息而获得的。大数据在存储和计算上都体现了数据为核心的理念，数据比流程更重要。云计算为大数据提供了有力的工具和途径，大数据为云计算提供了很有价值的用武之地。云计算机可以从数据库、记录数据库中搜索出你是谁，你需要什么，从而推荐给你需要的信息。

2. 从功能价值转变为数据价值

互联网的发展，使大数据一直在线。数据一定有它的价值，因此，可以用数据价值思维方式思考问题、解决问题。数据为“王”的时代出现了，大数据的价值也在扩大，原有的功能为价值也转变为数据为价值。数据被解释是信息，信息常识化是知识，所以说数据解释、数据分析能产生价值。

大数据的真正价值在于创造，在于填补无数个还未实现过的空白。例如，有人把数据比喻为蕴藏能量的煤矿，煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”，价值含量、挖掘成本比数量更为重要。不管大数据的核心价值是不是预测，但是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。

信息总量的变化导致了信息形态的变化，量变引发了质变。例如，每一分钟拍摄一张人在骑马的照片，随着处理速度越来越快，从 1 分钟一张到 1 秒钟 1 张，再到 1 秒钟 10 张后，就产生了电影。当数量的增长实现质变时，就从一张照片变成了一部电影。

3. 从抽样思维转变为总体思维

在人类无法获得总体数据信息的条件下，采样一直是主要的数据获取手段。在大数据时代，随着数据收集、存储、分析技术的突破性发展，人们可以更加方便、快捷、动态地获得研究对象的所有数据，而不再因诸多限制不得不采用样本研究方法。相应地，对数据的思维方式也应该从样本思维转向总体思维，从而能够更加全面、立体、系统地把握全局。

用总体思维方式思考问题、解决问题需要全部数据样本而不是抽样，全部样本才能找出规律。在大数据时代，无论是商家还是信息的搜集者，会比我们自己更知道你可能会想干什么。如果数据被真正挖掘的话，通过信用卡消费的记录，可以成功预测一个人未来 5 年内的情况。

4. 从关注精确度转变为关注效率

关注效率而不是精确度，大数据标志着人类在寻求量化和认识世界的道路上前进了一大步，过去不可计量、存储、分析和共享的很多东西都被数据化了，拥有大量的数据和更多不那么精确的数据为我们理解世界打开了一扇新的大门。大数据能提高生产效率和销售效率，原因是大数据能够让我们知道市场的需要。大数据让企业的决策更科学，由关注精确度转变为关注效率的提高，大数据分析能提高企业的效率。例如，在互联网大数据时代，企业产品迭代的速度在加快，三星、小米手机制造商半年就推出一代新智能手机。利用互联网、大数据提高企业效率的趋势下，快速就是效率、预测就是效率、预见就是效率、变革就是效率、创新就是效率、应用就是效率。

竞争是企业的动力，而效率是企业的生命，效率低与效率高是衡量企业成败的关键。一般来讲，投入与产出比是效率，追求高效率也就是追求高价值。手工、机器、自动机器、智能机器之间效率是不同的，智能机器效率更高，已能代替人的思维劳动。智能机器核心是大数据制动，而大数据制动的速度更快。在快速变化的市场，快速预测、快速决策、快速创新、快速定制、快速生产、快速上市成为企业行动的准则，也就是说，速度就是价值，效率就是价值，而这一切离不开大数据思维。

5. 从关注因果转变为关注相关性

在大数据世界未出现时，人们往往执着于现象背后的因果关系，试图通过有限样本数据来剖析其中的内在关联。数据量小的另一个缺陷就是有限的样本数据无法反映出事物之间的普遍性的关联。在大数据时代，思维方式要从因果思维转向相关思维，关注相关性而不是因果关系，因此人们需要放弃对因果关系的渴求，而仅需关注相关关系，即只需要知道是什么，而不需要知道为什么。这就颠覆了千百年来人类形成的传统思维模式和固有偏见。

传统的因果思维是从原因推出结果，而大数据没有必要找到原因，只需要通过大数据挖掘技术挖掘与分析出事物之间隐蔽的关系，获得更多的认知与洞见，运用这些认知与洞见就可以帮助我们捕捉现在和预测未来。例如，通过关注线性的关联及复杂的非线性关联，可以帮助人们看到很多以前不曾注意的数据之间存在的某些联系，还可以掌握以前无法理解的复杂技术和社会动态，关联性甚至可以超越因果关系，成为我们了解这个世界的更好视角。

6. 从不能预测转变为可以预测

用大数据预测思维方式来思考问题、解决问题。大数据的核心就是预测，大数据能够预测体现在很多方面。大数据不是要教机器像人一样思考，相反，它是把数学算法运用到海量的数据上来预测事情发生的可能性。正因为在大数据规律面前，每个人的行为都跟别人一样，没有本质变化，所以商家会比消费者更了解消费者的行为。

互联网、移动互联网和云计算机保证了大数据实时预测的可能性，也为企业和用户提供了实时预测的信息。相关性预测的信息，能让企业和用户抢占先机。由于大数据的全样本性，云计算机软件能预测的效率和准确性也大大提高。

例如，大数据助微软准确预测世界杯。微软大数据团队巴西世界足球赛前设计了世界杯模型，该预测模型正确预测了赛事最后几轮每场比赛的结果，包括预测德国队将最终获胜。预测成功归功于微软在世界杯进行过程中获取的大量数据。到淘汰赛阶段，数据如滚雪球般增多，微软掌握了有关球员和球队的足够信息，适当校准模型并调整对接下来比赛的预测，从而准确预测了比赛结果。

预测模型的诀窍就是在预测中去除主观性，让数据说话。预测性数学模型几乎不算新事物，但它们正变得越来越准确。在这个时代，数据分析能力终于开始赶上数据收集能力，分析师不仅有比以往更多的信息可用于构建模型，也拥有在很短时间内通过计算机将信息转化为相关数据的技术。

7. 从人找信息转变为信息找人

互联网和大数据的发展，是一个从人找信息到信息找人的过程。它要求我们用信息找人的思维方式思考问题，解决问题，我们听收音机，我们看电视，它是信息推给我们的，但是有一个缺陷，不知道我们是谁，后来互联网反其道而行，提供搜索引擎技术，让我们知道如何找到我们所需要的信息，所以搜索引擎是一个很关键的技术。

例如，从搜索引擎向推荐引擎的转变。今天，后搜索引擎时代已经正式来到。什么叫作后搜索引擎时代呢？使用搜索引擎的频率会大大降低，使用的时长也会大大缩短，为什么使用搜索引擎的频率在下降？时长在下降？原因是推荐引擎的诞生。就是说从人找信息到信息找人越来越成为一个趋势，推荐引擎就是说它很懂我，知道我想知道的，所以是最好的技术。乔布斯说，让人感受不到技术的技术是最好的技术。

从人找信息到信息找人，是交互时代的一个转变。信息找人预示着大数据时代可以让信息找人，原因是企业懂用户，机器懂用户，你需要什么信息，企业和机器提前知道，而且主动提供你需要的信息。

8. 从自然思维转变为智能思维

计算机的出现极大地推动了自动控制、人工智能和机器学习等新技术的发展，智能机器人的技术研发也取得了突飞猛进的成果并开始一定应用。应该说，自进入信息社会以来，人类社会的自动化、智能化水平已得到明显提升，但始终面临瓶颈而无法取得突破性进展，机器的思维方式仍属于线性、简单、物理的自然思维，智能化水平仍不尽如人意。但是，大数据时代的到来，可以为提升机器智能带来契机，其中一个核心目标是要从体量

巨大、结构繁多的数据中挖掘出隐藏在背后的规律，从而使数据发挥最大的价值。由计算机代替人去挖掘信息，获取知识，推动思维方式由自然思维转向智能化思维。提升计算机从各种各样的数据(包括结构化、半结构化和非结构化数据)中快速获取有价值信息的能力，这才是大数据思维转变的关键所在和核心内容。

众所周知，人脑之所以具有智能、智慧，就在于它能够对周遭的数据信息进行全面收集、逻辑判断和归纳总结，获得有关事物或现象的认识与见解。同样，在大数据时代，随着物联网、云计算、可视技术等的突破发展，大数据系统也能够自动地搜索所有相关的数据信息，并进而类似人脑一样主动、立体、逻辑地分析数据、做出判断、提供洞见，那么，无疑也就具有了类似人类的智能思维能力和预测未来的能力。

智能、智慧是大数据时代的显著特征，大数据时代的思维方式也要求从自然思维转向智能思维，不断提升机器或系统的社会计算能力和智能化水平，从而获得具有洞察力和新价值的东西，甚至类似于人类的智慧。

9. 从企业生产产品转变为由客户定制产品

大数据改变了企业的竞争力，从企业生产产品转变为由客户定制产品。用定制产品思维方式思考问题，解决问题。大数据时代让企业找到了定制产品、订单生产、用户销售的新路子。为大量客户定制产品和服务，成本低、又兼具个性化。比如消费者希望他买的车有红色、绿色，厂商有能力满足要求，但价格又不至于像手工制作那般让人无法承担。因此，在厂家可以负担得起大规模定制带去的高成本的前提下，要真正做到个性化产品和服务，就必须对客户的需求有很好的了解。

企业产品直接销售给用户，省去了中间商流通环节，使产品的价格可以以出厂价销售，让消费者获得了好处，网上产品便宜成为用户的信念，网购市场便形成了。在大数据规律面前，每个人的行为都跟别人一样，没有本质变化。所以商家会比消费者更了解消费者的行为。也许你正在想，工作了一年很辛苦，要不要去哪里度假？打开 E-mail，就有航空公司、旅行社的邮件。这就要依靠大数据技术。

大数据开启了一个重大的时代转型。大数据技术正在改变我们传统的生活以及理解世界的方式，成为新发明和新服务的源泉，而更多的改变正蓄势待发。大数据时代将带来深刻的思维转变，大数据不仅将改变每个人的日常生活和工作方式，改变商业组织和社会组织的运行方式，而且将从根本上奠定国家和社会治理的基础数据，彻底改变长期以来国家与社会某些领域存在的“不可治理”状况，使得国家和社会治理更加透明、高效和智慧。

1.2 大数据的处理流程

大数据不仅数据处理规模巨大，而且数据处理需求多样化，超过了当前计算机存储和处理的能力，因此，数据处理能力成为核心竞争力。大数据处理流程主要包括数据采集、数据预处理、数据存储、数据分析与挖掘、数据展示与应用等环节，如图 1.2 所示。数据质量贯穿于整个大数据流程，每一个数据处理环节都会对大数据质量产生影响。通常，大数据处理要有大量的数据规模、快速的数据处理、精确的数据分析与预测、优秀的可视化

图表以及简练易懂的结果解释。

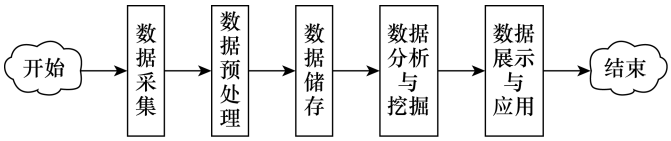

图 1.2　大数据处理流程

1.2.1　数据采集

数据采集，又称数据获取，指从传感器和其他待测设备等模拟和数字被测单元采集数据信息的过程，它是数据处理与分析的基础。数据采集一般有两种解释：一是数据从无到有的过程(Web 服务器打印的日志、自定义采集的日志等)；二是通过使用 Flume 等工具把数据采集到指定位置的过程。

在数据源多样、数据量巨大、采集工具众多的背景下，数据采集并不是将所有能够采集到的数据都采集来，而是将需要采集的数据和能够采集到的数据的“交集”获取到，这才是数据采集工作的意义。

大数据采集过程的挑战主要是并发率高，可能有成千上万的用户同时对数据进行访问和操作，例如火车票售票网站和淘宝网，它们并发的访问量在峰值时达到上百万，所以需要在采集端部署大量数据库才能支撑用户的访问，并且还需在数据库之间进行负载均衡和分片设计。常用的数据采集方法有以下几种。

1. 系统日志的采集方法

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Cloudera 的 Flume、Hadoop 的 Chukwa、Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百兆字节的日志数据采集和传输需求。

Flume 是 Apache 旗下的一款开源、高可靠、高扩展、容易管理、支持客户扩展的数据采集系统。Flume 使用 JRuby 来构建，所以依赖 Java 运行环境。Flume 最初是由 Cloudera 的工程师设计，用于合并日志数据的系统，后来逐渐发展用于处理流数据事件。Flume 数据采集结构如图 1.3 所示。

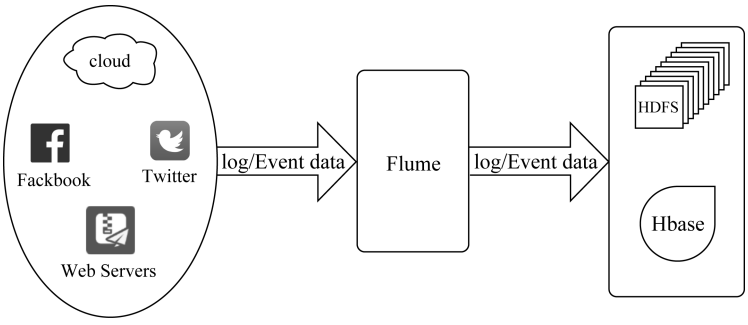

图 1.3　Flume 数据采集结构图

Chukwa 是基于 Hadoop 的 HDFS 和 Map Reduce 来构建(显而易见，它是用 Java 来实现的)，提供扩展性和可靠性。它同时提供对数据的展示、分析和监视。Chukwa 的部署架构如图 1.4 所示。

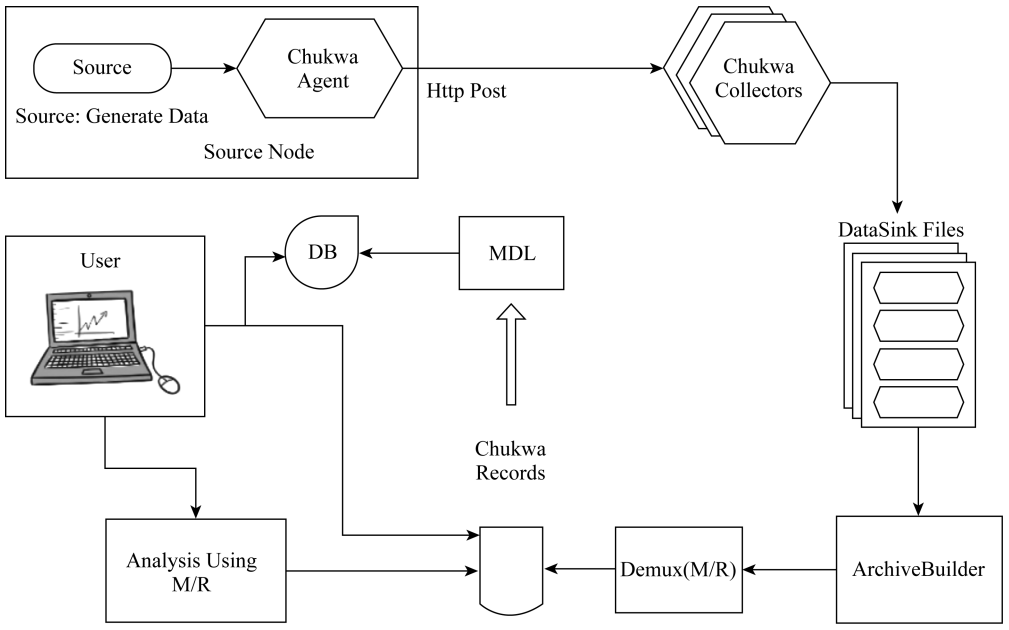

图 1.4　Chukwa 的部署架构

Scribe 是 Facebook 开发的数据(日志)收集系统(如图 1.5 所示)，又被称为网页蜘蛛、网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本，它支持图片、音频、视频等文件或附件的采集。

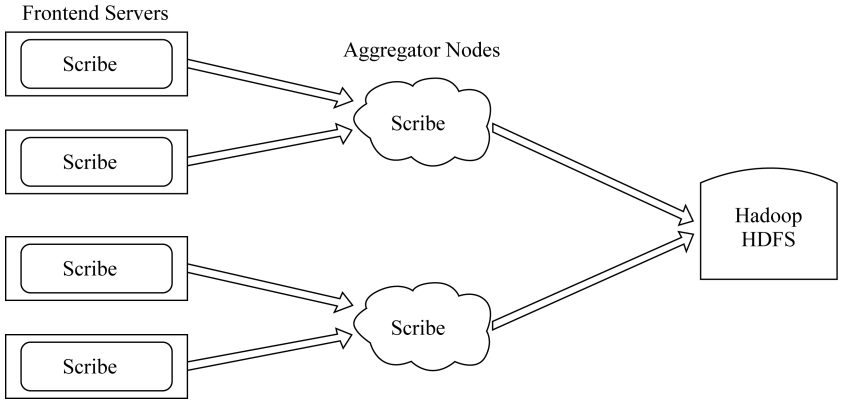

图 1.5　Scribe 数据采集架构图

2. 网络数据的采集方法

网络数据采集是指通过网络爬虫或网站公开 API(应用程序接口)等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文

件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用 DPI 或 DFI 等带宽管理技术进行处理。

3. 其他数据的采集方法

对于企业生产经营数据上的客户数据，财务数据等保密性要求较高的数据，可以通过与数据技术服务商合作，使用特定系统接口等相关方式采集数据。比如八度云计算信息技术有限公司的数企 BDSaaS，无论是数据采集技术、BI 数据分析，还是数据的安全性和保密性，都做得很好。

1.2.2　数据预处理

大数据采集过程中通常有一个或多个数据源，这些数据源包括同构或异构的数据库、文件系统、服务接口等，易受到噪声数据、数据值缺失、数据冲突等影响，因此首先需要对收集到的大数据集合进行预处理，以保证大数据分析与预测结果的准确性与价值性。大数据的预处理环节主要包括数据清理、数据集成、数据归约与数据转换等内容，这可以大大提高大数据的总体质量。

（1）数据清理从字面上可以理解为把“脏”的数据给“洗掉”，泛指发现并纠正数据文件中可识别的错误。它包括对数据的一致性、准确性、真实性和可用性等多方面的数据质量检查。

（2）数据集成则是将多个数据源的数据进行集成，从而形成集中、统一的数据库、数据立方体等。这一过程有利于提高大数据的完整性、一致性、安全性和可用性。

（3）数据归约是在不损害分析结果准确性的前提下降低数据集规模，使之简化，包括维归约、数据归约、数据抽样等技术。这一过程有利于提高大数据的价值密度，即提高大数据存储的价值性。

（4）数据变换是将原始数据转换成适合数据挖掘的形式，包括基于规则或元数据的转换、基于模型与学习的转换等技术，可通过转换实现数据统一，这一过程有利于提高大数据的一致性和可用性。

总之，数据预处理环节有利于提高大数据的一致性、准确性、真实性、可用性、完整性、安全性和价值性等方面质量，而大数据预处理中的相关技术是影响大数据过程质量的关键因素。

1.2.3　数据存储

大数据面对的数据量异常大，不是几个 TB 的硬盘可以容纳的，而且个人电脑上的存储设备一般也无法容纳如此大量的数据。为了能够快速、稳定地存取这些数据，至少得依赖于磁盘阵列，同时还得通过分布式存储的方式将不同区域、类别、级别的数据存放于不同的磁盘阵列中。大数据存储是将这些数据集持久化到计算机中。

以往的关系型数据库受限于设计模式，一般只考虑到了单机的数据存储方式，即不管

数据量大与小，一定会让一台机器存储和管理所有数据。而每台机器上可以承载的存储设备是有限的，一般也不会超过几个 TB。且一旦某个数据库的数据量和文件的尺寸暴增到一定程度后，数据的检索速度就会急剧下降。

为了应对这个问题，很多主流的数据库纷纷提出了一些解决方案。如 MySQL 提供了 MySQL proxy 组件(如图 1.6 所示)，实现了对请求的拦截，结合分布式存储技术，从而可以将一张很大的表中的记录拆分到不同的节点上去进行查询。对于每个节点来说，数据量不会很大，从而提升了查询效率。

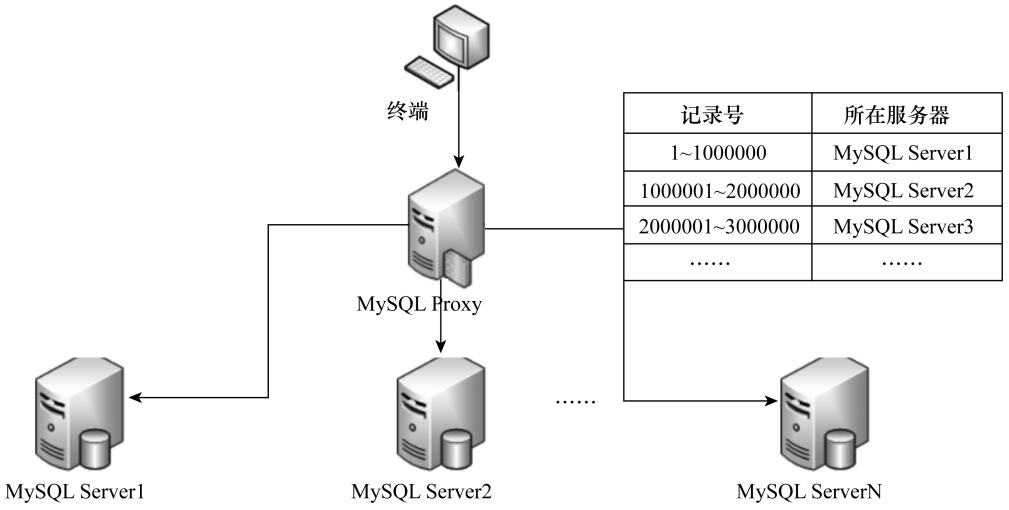

图 1.6　MySQL proxy 组件

Oracle 采取的方案是“大数据机+Hadoop+NoSQL”(如图 1.7 所示)。Oracle 提供了拥有 288 个 CPU、1152G 内存、648T 硬盘的无比豪华的服务器配置，同时结合 Hadoop 和 NoSQL 等技术对其中存储的大数据进行分析。

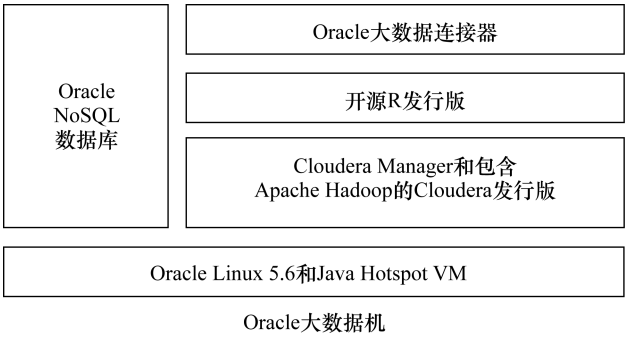

图 1.7　Oracle 大数据机

MongoDB 和 HBase 等非关系型数据库，天生都支持分布式存储，即将一份大的数据分散到不同的机器上进行存储，从而降低了单个节点的存取压力。由于摆脱了表的存储模式，对大数据的响应要比关系型数据库快得多。

无论哪种大数据分布式处理与计算系统，都有利于提高大数据的价值性、可用性、时效性和准确性。大数据的类型和存储形式决定了其所采用的数据处理系统，而数据处理系统的性能与优劣直接影响大数据质量的价值性、可用性、时效性和准确性。因此在进行大数据处理时，要根据大数据类型选择合适的存储形式和数据处理系统，以实现大数据质量的最优化。

1.2.4　数据分析与挖掘

数据分析是大数据处理与应用的关键环节，它决定了大数据集合的价值性和可用性，以及分析预测结果的准确性。通过数据采集、预处理以及存储环节，从异构的数据源中获得了用于大数据处理的原始数据，用户可以根据自己的需求对这些数据进行处理与存储，在数据分析与挖掘环节，根据大数据应用情境与决策需求，选择合适的数据分析技术，将有价值的知识、规律或重要信息从数据库的相关数据集合中提取出来，提高大数据分析结果的可用性、价值性和准确性质量。

数据分析与挖掘技术可以挖掘出隐含在大量数据中的有潜在价值的规则和知识，从而为用户根据数据预测未来可能发生的行为结果提供支持。这些规则包含了数据仓库中数据之间存在的特定联系，并且能够揭示出对预测有用的信息，建立大数据分析规则。我们利用数据挖掘等技术，发现隐含在海量数据中的有价值的信息和潜在规律，构建大数据分析模型，利用可扩展的大数据分布式存储和计算平台，支撑巨量数据的存储和高效分析，进而实现高价值大数据的快速可视化。

与传统信息处理方法相比，数据挖掘技术有其自身特点。

- 信息查询方式是即时随机的，通常由用户提出，对查询也没有精确要求，需要利用数据分析与挖掘技术找出隐含其中容易被忽视的因素。
- 数据挖掘中得到的规则一般是基于大样本的概率统计，不要求其对所有数据都适用，只要在一定条件下有此规律即可使用。

数据分析与挖掘技术主要有统计分析、关联分析、聚类分析、填补缺失值、异常分析等。

1.2.5　数据展示与应用

数据展示，又称数据可视化，指将大数据分析与预测结果以计算机图形或图像的直观方式显示给用户的过程，并可与用户进行交互式处理。数据可视化技术有利于发现大量业务数据中隐含的规律性信息，以支持管理决策。数据可视化环节具有直观性的优点，便于用户理解与使用，是影响大数据可用性和易于理解性的关键因素。

大数据应用是指将经过分析处理后挖掘得到的大数据结果应用于管理决策、战略规划等的过程，它是对大数据分析结果的检验与验证。大数据应用过程直接体现了大数据分析处理结果的价值性和可用性。大数据应用对大数据的分析处理具有引导作用。在大数据收集、处理等一系列操作之前，通过对应用情境的充分调研、对管理决策需求信息的深入分析，可明确大数据处理与分析的目标，从而为大数据收集、存储、处理、分析等过程提供

明确的方向，并保障大数据分析结果的可用性、价值性和用户需求的满足。

1.3 大数据的应用

随着 5G 时代的到来，大数据应用的发展更加迅速，大数据的应用已广泛深入我们生活的方方面面，涵盖医疗、交通、金融、教育、体育、零售等各行各业。

1.3.1 大数据现状分析

全球大数据解决方案不断成熟，各领域大数据应用全面展开，为大数据发展带来强劲动力。大数据逐渐成为全球 IT 支出新的增长点。近年，随着人工智能、AIoT、云计算等技术的推动，全球数据量正在大幅度地扩展和增加。根据国际权威机构 Statista 的统计和预测，全球数据量在 2019 年约达到 41ZB。据 IDC 咨询预计，到 2025 年，全球数据圈将增至 175ZB。如图 1.8 所示。

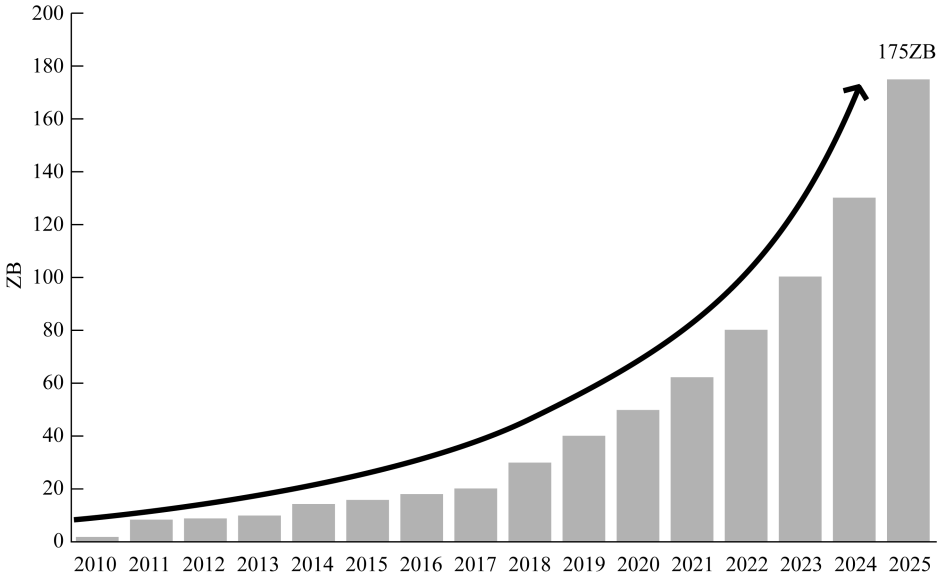

图 1.8 全球大数据圈每年的规模

根据 Wikibon 的研究数据，全球大数据市场规模将从 2018 年的 420 亿美元增长至 2024 年的 840 亿美元，年复合增长率为 12.3%。从细分市场来看，大数据软件市场份额占比将呈逐渐上升趋势。2018 年，大数据软件市场份额占比为 33.3%，到 2024 年，大数据软件市场份额占比将上升至 41.0%；大数据硬件市场比重则呈下降趋势，2018 年大数据硬件市场规模约为 120 亿美元，占比为 28.6%，到 2024 年硬件所占比重预计将下降至 24.1%。如图 1.9 所示。

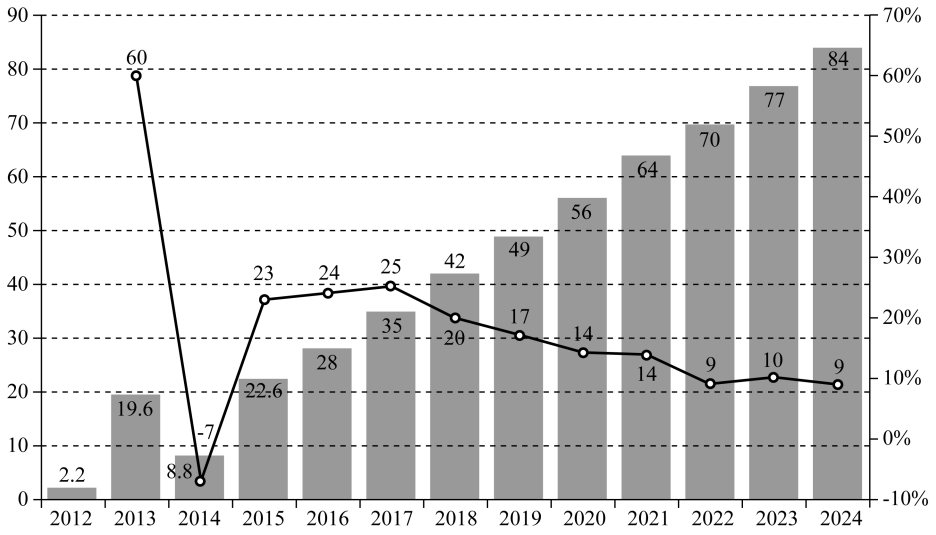

图 1.9　2012—2024 全球大数据市场规模

目前国家政策大力支持发展大数据产业，庞大的数据规模为数据的进一步深入挖掘提供了基础，大数据对各垂直化领域的改造作用凸显，中国的大数据产业在未来的发展潜力更加巨大。根据 IDC《数据时代 2025》报告估算，中国数据圈的规模及全球份额占比将会持续增长，到 2025 年，中国数据圈份额将达到 30%。如图 1.10 所示。

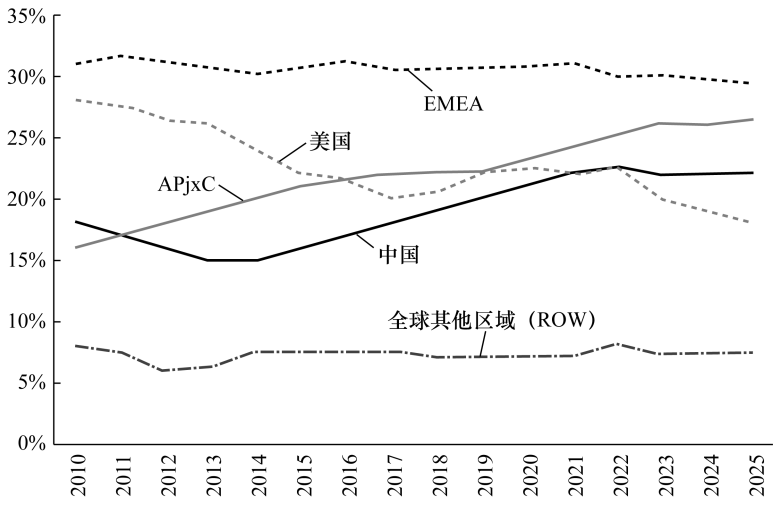

图 1.10　全球数据圈份额

1.3.2　大数据应用领域

大数据技术应用广泛，几乎深入到各个领域，如医疗大数据、电商大数据、金融大数据、安全大数据、能源大数据等等。

1. 医疗领域

大数据分析应用的计算能力可以让我们能够在几分钟内就可以解码整个 DNA，更好地理解和预测疾病，进而制定出最佳的治疗方案，帮助病人进行更好的治疗。

电子病历。大数据最强大的应用就是电子医疗记录的收集。每一个病人都有自己的电子记录，包括个人病史、家族病史、过敏症以及所有医疗检测结果等。这些记录可在不同的医疗机构之间共享。每一个医生都能够在系统中添加或变更记录，而无须再通过耗时的纸质工作来完成。这些记录同时也能帮助病人掌握自己的用药情况，同时也是医学研究的重要数据参考。

早产婴儿预测分析。针对早产婴儿，每秒钟有超过 3000 次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。

传染病疫情防控。传染病疫情一直是人类社会面临的重大生物安全威胁，其有效应对与消除则是考验国家和社会公共卫生应急整体处置和系统疏解能力的核心命题。大数据能够提供大量具有参考价值的“疫情情报”。比如，通过社会网络平台求助信息密度能够识别和研判重点防控和救治社区、通过 POI 等人口出行数据能够研判社区防控执行效果、通过能源和水电消耗及物流信息能研判产能复工水平等等。大数据就是新时期疫情防控需要的高效研判和决策的决策资源。

2. 电商领域

最早将大数据用于精准营销的就是电商领域，它可以根据消费者的习惯提前生产物料和物流管理，这样有利于美好社会的精细化生产。随着电子商务越来越集中，大数据在行业中的数据量变得越来越大，并且种类非常多。在未来的发展中，大数据在电子商务中将大有作为，其中主要包括预测消费趋势、区域消费特征、顾客消费习惯、消费者行为、消费热点和影响消费的重要因素。

3. 金融领域

大数据在金融行业的使用是非常广泛的，主要使用在交易过程中，如日常的出行、购物、运动、理财等等。现在许多股权交易都是使用大数据算法进行的。这些算法能够越来越多地考虑社交媒体和网站新闻，并且决定接下来的几秒内是选择购买还是出售。

据悉，目前中国的金融行业数据量已经超过 100TB，非结构化数据迅速增长。分析人士认为，中国金融行业正在步入大数据时代的初级阶段。优秀的数据分析能力是当今金融市场创新的关键，资本管理、交易执行、安全和反欺诈等相关的数据洞察力，成为金融企业运作和发展的核心竞争力。金融业面临众多前所未有的跨界竞争对手，市场格局、业务流程将发生巨大改变。未来的金融业将开展新一轮围绕大数据的 IT 建设投资。

4. 安防领域

作为信息时代海量数据的来源之一，视频监控产生了巨大的信息数据。物联网在安防领域应用无处不在，特别是近几年随着平安城市、智能交通等行业的快速发展，大集成、大联网、云技术推动安防行业进入大数据时代。安防行业大数据的存在已经被越来越多的

人熟知，特别是安防行业海量的非结构化视频数据，以及飞速增长的特征数据，带动了大数据应用的一系列问题。

5. 能源领域

能源大数据理念是将电力、石油、燃气等能源领域数据及人口、地理、气象等其他领域数据进行综合采集、处理、分析与应用的相关技术与思想。能源大数据不仅是大数据技术在能源领域的深入应用，也是能源生产、消费及相关技术革命与大数据理念的深度融合，将加速推进能源产业发展及商业模式创新。

1.3.3　大数据发展趋势

大数据成为时代发展一个必然的产物，它在日常生活中，从衣食住行各个层面均有体现。大数据时代，一切可量化，一切可分析。那么，未来大数据发展趋势将是如何呢？

1. 分析应用领域快速发展

数据共享将成为主流。在未来，大数据可能会把不同行业进行细分，更多的数据对行业的分析价值非常巨大。比如医疗，想要获得更大的价值，就要分享和分析，这样才能获取更大的价值，对医疗行业做出贡献。

大数据技术虽已有众多成功应用，但就其效果和深度而言，仍于初级阶段。数据隐藏的价值是非常巨大的，但是也需要 IT 技术不断发现和探索。随着应用层级的发展，企业用户会更加密切关注如何发现数据中的价值，使公司能够得到更快速的发展。IT 基础设施已在不断地发展和完善，大数据分析也会迎来更加快速的发展，深入研究大数据的挖掘技术和方法，根据大数据分析预测未来、指导实践的深层次应用将成为发展重点。

2. 安全与隐私更受关注

数据的价值对企业和行业来说是非常重要的，但是大数据治理体系仍未形成，特别是隐私保护、数据安全与数据共享利用效率之间尚存在明显矛盾，成为制约大数据发展的重要短板。

当前，各界已经普遍认识到了大数据治理的重要意义，大数据治理体系建设已经成为大数据发展重点，但仍处在发展的雏形阶段。关于大数据隐私方面的法律法规并不完善，还需要专门的法规为大数据发展扫除障碍，推进大数据治理体系建设将是未来较长一段时间内需要持续努力的方向。

3. 大数据+人工智能

数据科学与人工智能的结合越来越紧密。人工智能（Artificial Intelligence），英文缩写为 AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。其实就是用大量的数据作导向，让需要机器来做判别的问题最终转化为数据问题。

随着人工智能的发展，在海量数据中挖掘有用信息并形成知识将成为可能。如果说大数据相当于人的大脑存储了海量知识，而人工智能则是吸收了大量的数据，并不断地深度分析创造出更大的价值。人工智能离不开大数据，大数据依托着人工智能。未来大数据技

术将与人工智能技术更紧密地结合，让计算系统具备对数据的理解、推理、发现和决策能力，从而能从数据中获取更准确、更深层次的知识，挖掘数据背后的价值。

4. 大数据+区块链

区块链技术的大数据应用场景逐渐丰富。随着区块链技术的迅速发展，以及不同业务场景区块链的数据融合，数据规模越来越大，数据也会越来越丰富。

区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式，它能够保障数据的私密性，例如，基于区块链技术的英格码系统，在不访问原始数据情况下运算数据，可以对数据的私密性进行保护，杜绝数据共享中的信息安全问题。区块链的可追溯性，使得数据从采集、交易、流通以及计算分析的每一步被记录，留存在区块链上，增强数据的可信性，同时也保证了数据分析结果的正确性和数据挖掘的有效性。

1.4 本章小结

本章重点介绍了大数据的概念和大数据的处理流程。在此基础上介绍了大数据产生的背景、大数据的定义和特征和大数据思维。由此具体介绍了大数据处理流程中数据采集、数据预处理、数据存储、数据分析与挖掘以及数据展示与应用的相关内容。然后，分析了大数据产业现状及应用领域。最后，给出了大数据的发展趋势。

习　　题

一、选择题

1. 数据应包括数据体量巨大，数据类型繁多，(　　)以及处理速度要快等四个基本特点。

   A. 数据结构复杂　　B. 价值密度低而商业价值高

   C. 用户自生成数据比重大　　D. 数据传输量大

2. 数据清洗的方法不包括(　　)。

   A. 缺失值处理　　B. 噪声数据清除

   C. 一致性检查　　D. 重复数据记录处理

3. 智能健康手环的应用开发，体现了(　　)的数据采集技术的应用。

   A. 统计报表　　B. 网络爬虫

   C. API 接口　　D. 传感器

4. 大数据最显著的特征是(　　)。

   A. 数据规模大　　B. 数据类型多样

   C. 数据处理速度快　　D. 数据价值密度高

5. 美国海军军官莫里通过对前人航海日志的分析，绘制了新的航海路线图，标明了大风和洋流可能发生的地点。这体现了大数据思维原理中的(　　)。

A. 在数据基础上倾向于全体数据而不是抽样数据

B. 在分析方法上更注重相关分析而不是因果分析

C. 在分析效果上更追究效率而不是绝对精确

D. 在数据规模上强调相对数据而不是绝对数据

二、简答题

1. 大数据的 5V 特征是什么？
2. 大数据的主要应用领域有哪些？
3. 请简述大数据处理的流程。
4. 请简述大数据思维原理。