第一部分　数据分析与处理简介

在信息时代人们接触最多的是数据，然而数据应该如何使用，有哪些工具可以帮助人们更好地使用已有的数据，本书将逐步展开介绍。在此之前需要先简单了解数据分析与处理的一些基本内容。

第 1 章　数据分析与处理概述

正如大家所能感知到的，现在已经进入大数据时代，身处这个时代，周围很多事物都被数字化。几乎每个处在这个时代的人，每天都会产生大量的数据。虽然产生了大量的数据，但其中大部分都被抛弃了，只有少部分被加以利用并转化为新的产能。这是因为当前缺乏技术人员，能做数据分析与处理的人才很少，所以很多企业想做也做不到，更多详情可参考下面的讲解。

1.1　了解大数据

在阐述数据分析与处理之前，需要先了解大数据。

“大数据”是伴随数据信息的存储、分析等技术进步，而被人们所收集、利用的超出以往数据体量，且具有更高价值的数据集合、信息资产。

“大数据”仍然是数据信息的一类，之所以称为“大数据”，是因为其具有不同于传统数据信息的特征。

“大数据”这一理念直到最近几年才在国内受到高度关注，也是近几年才得到大部分企业的认可。实际上，早在 20 世纪 80 年代，未来学家、社会思想家阿尔文·托夫勒（Alvin Toffler）已在《第三次浪潮》（*The Third Wave*）中提出了“大数据”，并称“大数据”为“第三次浪潮的华彩乐章”。

《自然》（*Nature*）于 2008 年 9 月推出了名为“大数据”的封面专栏，从科学及社会经济等多个领域描述了“数据信息”在其中所扮演的越来越重要的角色，让人们对“数据信息”的广阔前景有了更多的期待，对身处或即将来临的大数据时代充满了好奇。

真正让“大数据”成为互联网信息时代科技界热词的是麦肯锡全球研究院（MGI）。麦肯锡全球研究院在 2011 年 5 月发布了一份名为《大数据：下一个创新、竞争和生产力的前沿》（*The next frontier for innovation，competition and productivity*）的研究报告，这是第一份从经济和商业等多个维度阐述大数据发展潜力的研究成果的报告，并对大数据的概念进行了描述，列举了大数据相关的核心技术，分析了大数据在各行业的应用，同时为政府和企业的决策者提出了应对大数据发展的策略。该报告的发布，极大地推动了大数据的发展。此后，大数据迅速成为科技热词，并引起了各国政府及商业巨头的广泛关注。

2012 年 1 月，瑞士达沃斯世界经济论坛将大数据作为论坛的主题之一，并发布了名为《大数据，大影响：国际发展新机遇》（*Big Data，Big Impact：New Possibilities for International Development*）的报告。

2012 年 3 月，美国奥巴马政府颁布《大数据的研究和发展计划》，启动了一项耗资超过 2 亿美元、涉及 12 个联邦政府部门，以及共计 82 项与大数据相关的研究和发展计划，希望通过提高大型复杂数据的处理能力，加快美国科技发展的步伐。

2012 年 4 月，成立于 2003 年的 SPLUNK 公司成为大数据处理领域第一家成功上市的公司，在 NASDAQ 上市的首个交易日以 109%的涨幅让人们对大数据充满了想象空间。

2012 年 5 月，英国建立了世界上首个关于政府数据信息开放的研究所。

2013 年，澳大利亚、法国等先后将大数据上升到国家战略层面，这是继美国和英国之后，新一轮关于大数据国家发展战略的动向。

从 2012 年开始，我国以 BAT（阿里巴巴、腾讯、百度）为首的互联网企业以及传统的运营商企业也纷纷启动了关于大数据的研发和应用。

2014 年 3 月，大数据首次进入国家政府工作报告；2015 年年初，李克强总理在政府工作报告中提出“互联网+”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合。

大数据（Big Data）的概念目前并没有得到学术界或实业界一致公认的十分确切的界定。目前对大数据的定义比较有说服力的主要有如下两种。

第一种对大数据的定义如下：大数据，或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模大到无法通过人工，在合理时间内实现截取、管理、处理，并整理成人类所能解读的信息。

第二种对大数据的定义如下：大数据，是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

2011 年 5 月，麦肯锡全球研究院在《大数据：下一个创新、竞争和生产力的前沿》中对“大数据”的描述为，“大小超出了典型数据库软件的采集、存储、管理和分析等能力的数据集”，这一界定只是十分基础的定义，仅仅从数据信息的体量上进行了界定。

IT 研究与顾问咨询公司 Gartner 给出的定义如下：大数据是具有更强决策力、洞察发现力和流程优化能力的海量、高增长率、多样化的信息资产。虽然对大数据尚未有公认的界定，但这并不意味着人们对这个概念没有达成较为普遍的共识，从以上定义来看，可以认为大数据是伴随数据信息的存储、分析等技术，而被人们所收集、利用的超出以往数据体量、类型，且具有更高价值的数据集合、信息资产。

目前，大数据的范围一般是从 TB 级到 PB 级。随着信息技术的高速发展，数据体量已从 GB 级（1GB=1024MB）升级到 TB 级（1TB=1024GB）、PB 级（1PB=1024TB），甚至 EB 级（1EB=1024PB）、ZB 级（1ZB=1024EB）。据国际数据公司（IDC）预测，2020 年全球数据量将达到 35.2ZB。

数名计算机科学家和业内高管称，2008 年大数据开始在技术圈内出现。起初，许多科学家和工程师都认为大数据只不过是一个营销术语。2008 年年末，大数据得到部分美国知名计算机科学研究人员的认可，业界组织“计算社区联盟”（Computing Community Consortium）发表了一份有影响力的白皮书——《大数据计算：在商务、科学和社会领域创建革命性突破》，该书由 3 位计算机科学家共同完成，分别是卡内基·梅隆大学的兰道尔·布赖恩特（Randal E. Bryant）、加利福尼亚大学伯克利分校的兰迪·卡兹（Randy H. Katz）、华盛顿大学的爱德华·拉佐斯加（Edward D. Lazowska）。他们的认可对大数据提供了智力支持。而对大数据的发展史来说，2012 年十分重要，大数据由技术圈走入了真正的主流市场。

当前，大数据仍然是数据信息的一个类别，之所以称为大数据，是因为其具有不同于传统数据信息的特征。

Gartner 的分析师道格拉斯 •兰尼（Douglas Laney）于 2001 年首次提出了大数据的“3V”特征，即容量大（Volume）、多样化（Variety）和速度快（Velocity）。

随着技术的进步，以及对大数据研究的深入，人们对大数据特征的认识也发生了一些变化。目前，业界普遍认可的一种理解大致如下。

（1）巨量，即数据体量十分庞大。

（2）多样，即信息类型多样，包括结构化信息（如消费者提交的信息、交易信息等）和大量非结构化信息（如微博、日志、GPS 定位信息等）。

（3）价值，价值密度低，商业价值高，受限于数据体量以非结构性数据的大量存在，相对于传统数据库，其数据价值密度较低；但由于信息关联性更强，所以其挖掘价值较大。

（4）高速，数据处理需要通过高速运算迅速得到分析结果，以满足大数据时代对时效性的要求。

基于大数据多个“V”的特征，维克托 • 迈尔 • 舍恩伯格（Victor Maier Schoen Berg）在《大数据时代：生活、工作与思维的大变革》中提出了 3 个基于大数据特征的重大思维转变：首先，要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本；其次，要乐于接受数据的纷繁复杂，而不再追求精确性；最后，人们的思想发生了转变，不再探求难以捉摸的因果关系，转而关注事物的相关关系。

随着信息的发展，开始有人意识到对海量数据进行研究与分析可以从中提取出极具价值的信息，对商业的发展和社会的进步都具有时代化的意义，这也是促进国家和社会推动大数据分析的核心动力。

1.2 数据分析与处理的需求

1.1 节对大数据的一些概念进行了简单介绍，可以了解到，当今大数据的应用对一家公司的发展颇为重要，而要使用大数据，就要涉及数据的分析与处理，对数据的分析与处理目前主要有如下几方面需求。

（1）大量数据等待专业人员进行处理。每天都有大批量的数据产生，能使用这些数据的公司并不多，而能将这些数据使用好的公司更是少之又少，其中很重要的一个原因是缺乏这方面的人才。目前，综观国际和国内，真正从事数据处理的人才并不多，在从事数据处理工作的人员中，真正受过专业训练的人就更少，所以真正的数据处理方面的专业人才非常稀少，而能将数据处理和业务相联系的跨领域人才更是难以寻找。这就导致了有大量数据的公司并不少，但能用好数据的公司很少。

（2）大量数据需要专业人员进行分析。除了大量数据需要专业人员进行处理，大量数据的分析也需要大批的专业人员。大部分数据并不是处理好了就能转化为生产力，还需要对这些处理过的数据进行分析，从中发现相应的商业价值。

（3）很多公司不缺少数据，但缺少能够灵活应用这些已经存在的数据的人才，也就是缺乏有经验的数据处理和分析人才。很多公司已经发展了数十年，在此期间积累了一笔雄厚的

数据资源，但一直都在“沉睡”中，正在等待“从睡梦中被唤醒”，从而展现其价值。但是否真有这样的人才出现，现在不好下定论，毕竟目前很多人还是喜欢新东西、新技术。历史的东西，人们一般都选择敬而远之。

（4）大数据技术已经日渐成熟，很多公司的服务器中正不断涌入大量数据，这些公司应使数据资源流动起来，做数据的处理和分析对它们来说是刻不容缓的。当然，对这些公司来说，首选的问题就是人才，找到能帮它们正确做数据处理和分析的人才是非常关键的，特别是在当下，谁能让数据产生更多的价值，谁就更有生存或发展壮大的可能。

（5）大量传统公司需要做数字化转型，它们的首要任务是如何获取对自己有用的数据，获取数据后如何处理，处理后的数据又如何分析。传统行业，特别是传统制造行业，它们普遍存在并发展了很多年，但一直没有进行转型，所以并没有积累多少数据。对这些传统行业而言，数字化是当务之急，借助一些成熟的工具，数字化转型或许并不那么难，难在如何抛弃传统的思维，尽快转型到通过数据分析与处理带来新的业务增长。另外，对于传统行业，一般的计算机技术人员并不太愿意切入，并且也不容易切入，这方面的数据分析与处理人才自然更加难以找寻。

对数据分析与处理的需求，本书分析的只是市场需求中的冰山一角。数字信息化领域有一块大“蛋糕”等待各个优秀的数据分析与处理人员参与分割，而在当前，这块“蛋糕”尚未成型，有多大亦未可知。在数据时代的浪潮之下，对数据分析与处理人员的需求只会更多，市场的空缺也会逐步变大。

1.3 数据分析与处理的发展前景

1.2 节大致介绍了数据分析与处理的需求。数据分析与处理除了在目前及未来会有庞大的需求，在今后的发展也是非常可观的。

从 2017 年开始，“人工智能”这几个字经常出现在人们的视野中，是近年来比较火热的一个词，人工智能技术也成为谈论比较频繁的话题之一。

还没有接触过人工智能的人们，对人工智能中很多技术点的操作可能并不会有更多的了解，特别是对特征、模型训练更是一无所知。其实，对当今人工智能技术中很多模型的训练，都依赖于大量的数据，而这些数据不能是一些简单粗糙的数据，通常是处理分析后质量比较高的一些数据。高质量的数据可以帮助训练出更高质量的模型，一个高质量的模型能预测出更高精度的结果。

所以，对人工智能的应用，在很大程度上需要依赖专业团队所做的数据分析与处理，如图片、文本、音频等数据，通过提供高质量的数据，可以打造出更贴近人们需要的人工智能应用。

而除了人工智能的应用，在当今信息化不断加剧的时代，各个领域都在不断融合，不同领域的融合过程最根本的是数据交互，有数据交互就需要做数据分析与处理。

每一种新兴技术的兴起，都意味着更加开阔的数据互通。

在云计算、大数据、物联网、人工智能等技术的发展过程中，需要各种不同数据的交互，需要更加便捷、更加高效、更加高质量的数据的应用处理。而伴随这些技术不断融入人们的

生活，对数据分析与处理的需求和要求会不断提出新的挑战。

人们都期望更加便捷和美好的生活方式，但这些都需要大量专业人才的参与才能很好地实现。

为了让更多人更快地走上通往专业人才的道路，下面将详细介绍一些数据分析与处理方面的工具，讲解完工具之后，会通过一些示例讲解这些工具的使用。