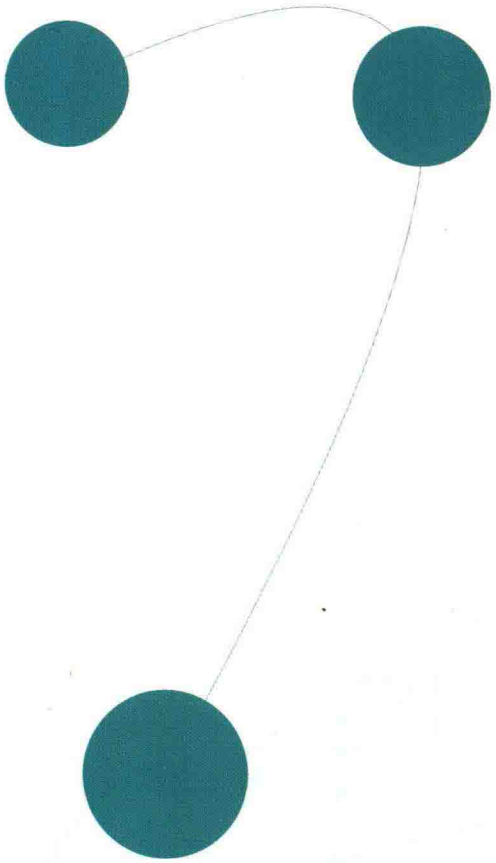
实战 Hadoop

大数据处理

曾刚 编著



清華大學出版社

实战 Hadoop

大数据处理

曾刚 编著

清華大學出版社
北京

内 容 简 介

本书以“大数据”为起点，较详细地介绍了 Hadoop 的相关知识。全书共分为 9 章，介绍了大数据的基本理论、Hadoop 生态系统、Hadoop 的安装、HDFS 分布式文件系统、MapReduce 的原理及开发、HBase 数据库、Hive 数据仓库、Sqoop 数据转换工具，最后结合实际介绍了大数据在智能交通和情报分析中的应用。本书力求用浅显的语言、生动的案例、详细的操作步骤向广大读者介绍 Hadoop；力求深入浅出，把复杂的理论与实际案例相结合，用平实的语言把深奥的原理简单化；力求图文并茂，通过适当的图表把零乱的知识点有序地展现在读者面前；力求紧跟时代步伐，尽量结合较新版本的软件阐述大数据处理的相关知识。

本书适合作为 Hadoop 技术的初学者、工程技术人员、大专院校研究生或高年级本科生的学习用书或参考书。

图书在版编目(CIP)数据

实战 Hadoop 大数据处理/曾刚编著. --北京：清华大学出版社，2015
ISBN 978-7-302-41144-4

Ⅰ. ①实…　Ⅱ. ①曾…　Ⅲ. ①数据处理软件　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 183368 号

责任编辑：田在儒
封面设计：王跃宇
责任校对：刘　静
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：17.25　　**字　　数**：419 千字
版　　次：2015 年 8 月第 1 版　　**印　　次**：2015 年 8 月第1次印刷
印　　数：1～1500
定　　价：39.00 元

产品编号：064835-01

前言
FOREWORD

随着社会的进步和计算机技术的发展，人类社会产生的数据正呈爆炸式增长。数据是人类社会重要的战略资源，大数据是“未来的新石油”，大数据对未来的科技与经济发展将带来重大影响，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家和企业间争夺的焦点。大数据如此重要，但大数据人才却十分短缺，据预测，到2018年美国大数据分析人才缺口是19万人，中国作为全球第二大经济体，拥有的数据占全球总量的13%，增长速度保持在50%左右，明显高于全球的增长速度。如此巨大的市场，大数据处理的技术人才必将成为炙手可热的人才，未来几年内我国将需要十几万的大数据人才。

大数据处理大体上分为：批处理系统、流式处理系统、交互式数据处理系统、图数据处理系统。Hadoop是目前应用最成功也是最广的批处理平台，国内外的企业和机构的数据处理系统纷纷向Hadoop处理平台过渡和转型，Hadoop已经成为大数据处理的工业标准。而其他的处理模式尚未形成完整的生态系统。

国内与Hadoop相关的技术书籍明显存在以下不足。①版本较老。在编者研究大数据技术时，书籍较少且相当多书籍版本为0.20版的Hadoop，虽然能够清晰地讲述Hadoop原理与技术，但已经不能适应时代发展的要求。②内容较单一。有不少书籍是对Hadoop官方技术文档的翻译或者是资料的整理，较少涉及较深层次的Hadoop应用，如架构设计、领域应用等。编者在参考了大量文献的基础上，并结合在专业领域的应用编写了本书。本书图文并茂，深入浅出，用浅显的语言讲解Hadoop原理的同时，结合具体的应用代码以加深读者对Hadoop技术的理解。最后，通过案例，让读者理解应用系统的体系架构，以及Hadoop在整个系统中的位置与作用。

本书适用于以下对象。

- 对大数据感兴趣的读者。
- 想了解Hadoop的初学者。
- 大数据处理的从业人员。
- HBase、Hive的爱好者。
- 开设Hadoop相关课程的大专院校。

学习本书的前提条件是：①具有一定的Java编程基础；②具有数据库相关的基础知识；③具有Linux相关的基础知识。

本书在编排上，加入了一些注意事项和提示，提醒读者注意一些容易被忽视的细节。本书还涉及大量的 Linux 或 Hadoop 命令，为了阅读方便，这些需要读者自己输入的命令均采用加粗字体；机器的输出信息采用小字体编排。

本书共分为 9 章，体系结构如下。

第 1 章为大数据概述。主要讲述了大数据的概念与特点、研究背景、应用示例、研究的意义、相关的关键技术、处理模式、代表性系统的发展前景。

第 2 章为 Hadoop 简介。本章重点介绍了 Hadoop 的起源、由来、相关项目的介绍以及版本的衍化。

第 3 章为 Hadoop 的安装。介绍了 Ubuntu Server、JDK 的安装；SSH 公钥认证的原理、安装、配置以及 SecureCRT 公钥登录；Hadoop 的三种安装模式；Hadoop 2.2 的安装。

第 4 章为 HDFS 文件系统。介绍了互联网时代对存储系统的新要求；HDFS 系统的特点；HDFS 文件系统的组成；HDFS 的两种操作方式：Shell 方式和 API 方式；HDFS 的高可用性以及小文件存储问题。

第 5 章为 MapReduce 原理及开发。介绍了 MapReduce 模型下编程的示例；MapReduce 的工作原理；Shuffle 原理；Shuffle 过程的优化；故障的处理方法、作业的调度方式；五类典型的 MapReduce 应用。

第 6 章为 HBase 数据库。介绍 HBase 数据库的特点、架构、原理；HBase 的安装方法；HBase 的 Shell 和 API 操作方法；MapReduce 操作 HBase 的方法；HBase 的优化方法。

第 7 章为 Hive 数据仓库。介绍了 Hive 的架构、安装方法、HQL 的使用，并介绍了复杂类型以及 Hive 函数。

第 8 章为数据整合。介绍了使用 Sqoop 把关系型数据库表整合到 Hadoop 的 HDFS、HBase、Hive 中的方法。

第 9 章为典型应用案例介绍。介绍了 Hadoop 在智能交通中的应用及在情报分析中的应用。

本书在编写的过程中，得到了同事们的鼓励和支持，也得到妻子和女儿的关心和照顾。同时还要感谢清华大学出版社的编辑，他们在书稿的编辑出版过程中做了大量的工作，感谢他们对我的支持和鼓励。

由于 Hadoop 的发展非常迅速，加之本人的水平有限，书中难免会有错误和遗漏之处，恳请谅解和批评指正，欢迎提出宝贵的意见和建议。编者的电子邮箱为 dlzenggang@126.com。

编　者

2015 年 5 月于大连

目 录
CONTENTS

第1章 大数据概述

1.1 大数据简介

1946年，世界上第一台计算机ENIAC诞生，标志着人类社会进入信息时代。随着计算机技术全面融入社会生活，社会的各个领域产生了大量的信息，并且开始爆炸式增长。

随着互联网的普及，为了满足人们搜索网络信息的需求，搜索引擎抓取并存储了巨大的信息；社交网络把分散于各处的人们联系起来；电子商务在满足人们便捷购物的同时，收集了大量的购物意愿与购物习惯的数据；2007年推特(Twitter)开始独立运营，标志着移动互联网时代的到来。2010年是中国的微博元年，2011年微信(WeChat)开始运营，这些移动应用的运行也产生了海量的数据。随着平安工程的开展，各地纷纷开始安装使用视频监控系统，产生了海量的视频数据；银行、股市、保险等金融部门在运营中产生了大量非常重要的交易数据；安装有各种传感器的物联网中有巨大的数据流在运转；电信部门通过通话、短信等多种业务也在快速地产生着大量数据。

下面看几个具体的例子。

谷歌(Google)通过网络爬虫搜集的网络数据以及其他应用处理的数据量每个月达400PB以上；百度每天处理的数据量达几十PB；脸谱(Facebook)全球注册用户达10亿多人，每个月上传的照片达10亿张，每天产生约300TB的日志数据；淘宝拥有会员3.7亿，在线商品8.8亿，每天交易量达数千万笔，产生约300TB的数据；劳斯莱斯公司对全世界数以万计的飞机发动机进行实时监控，每年传送的数据量达PB级以上。

为了更精确地度量数据，看一下度量单位的关系。

1Byte=8Bit

1KB=1 024Bytes

1MB=1 024KB=1 048 576Bytes

1GB=1 024MB=1 048 576KB=1 073 741 824Bytes

1TB=1 024GB=1 048 576MB=1 099 511 627 776Bytes

1PB=1 024TB=1 048 576GB=1 125 899 906 842 624Bytes

1EB=1 024PB=1 048 576TB=1 152 921 504 606 846 976Bytes

1ZB=1 024EB=1 180 591 620 717 411 303 424Bytes

1YB=1 024ZB=1 208 925 819 614 629 174 706 176Bytes

可以看出，上面的度量单位比常见的度量单位多了PB、EB、ZB、YB。以上单位都是抽象的，下面来看一个形象的例子。

《红楼梦》含标点 87 万字(不含标点 853 509 字),每个汉字占两个字节:1 汉字=16bit=2Bytes,1GB 约等于 671 部红楼梦,1TB 约等于 631 903 部,1PB 约等于 647 068 911 部。

美国国会图书馆藏书 151 785 778 册(2011 年 4 月:收录数据 235TB)。

中国国家图书馆拥有图书 2 631 万册。

1EB=4 000 倍美国国会图书馆存储的信息量。

通过这个形象的例子,就可以知道"大数据"这个概念中的"大"字有多大了。根据 IDC(Internet Data Center,互联网数据中心)的统计,2012 年全球产生的数据量达 2.7ZB,相比 2011 年的 1.8ZB 增长了 48%,这种增长速度还在加快,预计 2020 年,产生数据的总量将达到 35.2ZB,如图 1-1 所示。

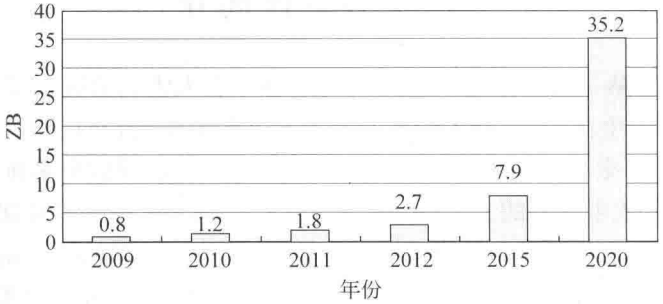


图 1-1 IDC 统计的全球数据量预测

1.1.1 大数据的概念与特点

那么,究竟什么是大数据呢?对这个概念的解释可谓仁者见仁,智者见智,目前还没有一个统一的、大家都认可的定义。不管如何定义,大数据的五大基本特点是大家都认可的,如图 1-2 所示。

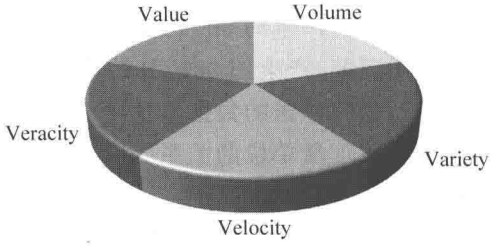


图 1-2 大数据的 5V 特点

1. Volume(体量浩大)

社会生活中产生的数据的体量在不断地扩大,数据集合的规模已经从 MB、GB、TB 到了 PB,在数据中心的数据量甚至以 EB 和 ZB 等单位来度量。IDC 的研究报告称,未来十年,全球大数据将增加 50 倍,管理数据仓库的服务器的数量将增加 10 倍以迎合 50 倍的大数据增长。如此巨大的数据量,带来的是巨大的计算量。

2. Variety(类型多样)

从数据的组织形式来看,数据包括结构化数据、半结构化数据、非结构化数据。结构化

数据是以二维表的形式来组织数据，例如关系型数据库，数据存储于二维数据表中，数据的类型、格式严格一致，表与表之间通过参照关系建立联系。非结构化数据是指无法通过预先定义的数据模型表述的数据，包括视频、音频、图片、文档、文本等形式。半结构化数据是介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和非结构化数据（如声音、图像文件等）之间的数据，HTML 文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。根据 IDC 的统计，目前 80％的数据为非结构化和半结构化数据，结构化数据仅占总量的 20％。

3. Velocity（生成快速）

随着移动计算、社交媒体和物联网等新技术的不断出现和应用，非结构化数据正在以63％的速度飞速增长着，而结构化数据仅以 32％的速度增长。网络中的数据往往呈现出突发涌现等非线性状态演变现象，因此难以对其变化进行有效评估和预测。另一方面，网络中的数据常常以数据流的形式动态快速地产生，具有很强的时效性，用户只有有效地掌控数据流才能充分利用这些数据。

4. Veracity（真实性高）

随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限被打破，企业越发需要有效的信息之力以确保其真实性及安全性。

5. Value（价值巨大但密度很低）

Value 是大数据的精髓，一方面企业能够利用大数据技术让运算变得更快，另一方面大数据衍生了很多新的商业模式。以保险行业为例，车险公司在车内安装传感器，用以监测司机的驾驶习惯，根据不同的驾驶行为区分司机的安全系数，分别拟定相应的保费标准。信用卡公司也会通过对顾客消费行为、购买模式的分析，制定精准的个性化营销模式。

虽然数据的价值巨大，但是基于传统思维与技术，人们在实际环境中往往面临信息泛滥而知识匮乏的窘态，即大数据的价值利用密度低。因此，要从密度较低的大数据中找到有价值的信息，必须使用某种特定的策略与方法进行数据的挖掘与分析。

从上面的五个特点不难看出，大数据从本质上来讲包含数量、类型、速度三个维度的问题，想要把三个维度从根本上区分开是不可能的，因为大数据概念的提出是源于技术的发展。

对大数据的认识，除了如图 1-2 所示五个基本特点外，社会各界都试图从其他方面对大数据进行解释和定义。亚马逊网络服务（AWS）的大数据科学家 JohnRauser 提出一个简单的定义：大数据就是任何超过了一台计算机处理能力的庞大数据量。AWS 研发小组对大数据的定义：大数据是最大的宣传技术和最时髦的技术，当这种现象出现时，定义就变得很混乱。Kelly 说："大数据是可能不包含所有的信息，但我觉得大部分是正确的。对大数据的一部分认知在于，它是如此之大，分析它需要多个工作负载，当你的技术达到极限时，也就是数据的极限。"

从数据的类别上看，大数据指的是无法使用传统流程或工具进行处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统处理方法的数据集，这是研究机构 Gartner 的观点，他们认为"大数据"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。

互联网周刊也给出了他们的理解：大数据的概念并不只是大量的数据（TB）和处理大

量数据的技术，或者所谓的“五个 V”之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，并且这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让人们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力。

李国杰等人对大数据的理解与定义是：大数据是指无法在一定时间内用常规机器和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合。

尽管人们对大数据的理解与定义各不相同，但有一点是共同的，那就是大数据已经引起社会各界的关注与重视。相对于如何定义大数据这一概念，人们更关注的是如何使用大数据，哪些技术能更好地处理大数据以及大数据的应用情况如何。

1.1.2 大数据研究的背景

大数据所蕴含的社会、经济、科学研究的价值已经引起社会各方面的广泛关注，如果能够有效地利用大数据，必将对社会发展、经济建设、科学研究产生深远的影响。著名的 O'Reilly 公司断言：“未来属于将数据转换成产品的公司和人们。”

谷歌、雅虎、脸谱、亚马逊、IBM 等公司是大数据技术的直接受益者，也是推动者。谷歌由于搜索引擎的需要，首先提出了 GFS(Google File System)文件系统、MapReduce 计算机制以及大型分布式数据库 BigTable。在谷歌的带领下，雅虎、脸谱等公司推出了开源的分布式文件系统 HDFS、MapReduce 机制、分布式数据库 HBase 等工具。

近几年，Nature 和 Science 等国际顶级学术刊物相继出版专刊，专门探讨对大数据的研究。2008 年，Nature 出版专刊 Big Data，从互联网技术、网络经济学、超级计算、环境科学、生物医药等多个方面介绍了海量数据带来的挑战。2011 年，Science 推出关于数据处理的专刊 Dealing with data，讨论了数据洪流(Data Deluge)所带来的挑战。特别指出，倘若能够更有效地组织和使用这些数据，人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用。

2012 年 3 月 22 日，美国政府投资 2 亿美元启动了“大数据研究和发展计划(Big Data Research and Development Initiative)”。这是继美国政府的“信息高速公路”计划之后，又一项重大的科学研究计划。美国政府认为，大数据上升为国家意志后，必将对未来的社会生产和生活带来深远的影响。该计划旨在提高和改进人们从海量和复杂的数据中获取知识的能力，进而加速美国在科学与工程领域发明的步伐，增强国家安全。大数据中蕴含着巨大的价值，美国政府认为，大数据是“未来的新石油”，对未来的科技与经济的发展将产生深远的影响。此外，欧盟也对科学数据基础设施投资 1 亿多欧元，并将数据信息化基础设施作为 Horizon 2020 计划的优先项目之一。

2012 年，联合国发布了题名为《大数据促发展：挑战与机遇》的白皮书，在白皮书中总结了各国政府如何利用大数据响应社会需求，指导本国经济运行，提高本国人民的生活水平，建议各国政府建立 Pulse Labs(脉搏实验室)，研究大数据，挖掘其潜在价值。

我国的研究机构与企业也进行了相应的研究与开发。2012 年，中国计算机学会(CCF)发起并成立了 CCF 大数据专家委员会，CCF 专家委员会还成立了一个“大数据技术发展战略报告”撰写组，撰写并发布了《2013 年中国大数据技术与产业发展白皮书》。2013 年以来，国家自然科学基金、863 计划、973 计划、核高科等研究计划都已经把大数据列为研究的重大

课题。

可以说，一个国家拥有数据的规模和对数据运用的能力将会成为一个国家综合国力的一部分，对数据的占有、控制能力必将成为国家和企业间竞争的制高点。

1.1.3 大数据的应用示例

大数据在许多领域都有应用，例如科学计算、物联网、天文学、天气预报、基因组学、生物学、大社会数据分析、互联网文件处理、制作互联网搜索引擎索引、通信记录明细、军事侦察、社交网络、流行病预测、医疗记录影像的处理、大规模的电子商务等。

1. 在科学计算领域的应用

大型强子对撞机是一座位于瑞士日内瓦近郊欧洲核子研究组织 CERN 的对撞型粒子加速器，它有 1.5 亿个传感器，每秒发送 4 千万次的数据。实验中每秒产生将近 6 亿次的对撞，过滤去除 99.999%的撞击数据后，得到约 100 次的有用撞击数据。将撞击结果数据过滤处理后仅记录了 0.001%的有用数据，全部四个对撞机的数据量复制前为每年产生 25PB，复制后为 200PB。这样年数据增长将达 1.5 亿 PB，也就是相当于每天 500EB，是全世界所有数据源总和的 200 倍。

2. 在政府部门的应用

2012 年，美国奥巴马政府宣布启动大数据研究和发展计划（Big Data Research and Development Initiative），致力于帮助政府部门利用大数据解决重大问题。该计划包括 84 个不同的大数据项目工程和 6 个部门。此外，美国联邦政府还拥有当今世界上顶级的十大超级计算机中的六个。负责气象模拟的 NASA 部门，在其发现者号超级计算机集群中也存储有 32PB 气象观测和模拟数据。这些事例充分说明了美国政府部门对大数据的重视以及为此而展开的应用。

3. 在社会学领域的应用

国际卫生学教授汉斯·罗斯林使用 Trendalyzer 工具软件，呈现了两百多年以来全球的人口统计数据，并将其与其他数据，例如收入、宗教、能源使用量等进行了交叉比对。

4. 在商业领域的应用

在商业领域，大数据解决方案和应用更是百花齐放、百家争鸣。著名的 Facebook 社交平台，早已开展了基于用户行为分析的数据挖掘和决策分析，能够对其所有用户的 500 亿张照片进行分析处理。沃尔玛每个小时处理的客户交易量超过百万次，这些交易的数据量高达 2.5PB(2560TB)——相当于美国国会图书馆藏书量的 167 倍。

2008 年，淘宝开始投入资源研究基于 Hadoop 的“云梯”数据处理平台，它支撑了淘宝对整个数据的分析工作。目前，集群的节点数达 1 700 个，数据量达 24.3PB，并且以每天 255TB 的速度在不断地增长。

支付宝是国内一个领先的第三方支付平台，为用户和商家提供可信任的第三方担保交易。支付宝目前拥有 7 亿多注册用户，合作商家 45 万家，日交易 3 369 万笔，日交易金额 45 亿元。在支付宝建立的以 Hadoop 为基础的数据处理平台内，“海狗”用于实时搜索，“剑鱼”用于数据查询，“海星”用于数据挖掘，“海豚”用于海量数据计算。

华为公司作为世界范围内著名的电信设备供应商，也积极地参与了 Hadoop 技术的应

用与改进。在 Hadoop 的基础上，扩展了 Hadoop 技术，自主研发了高可用性 Hadoop 平台。在电信领域应用 Hadoop 技术，构建了基于 Hadoop 的信令监测平台。同时对 Hadoop 核心项目与周边项目的改进做出了较大的贡献。

中国移动作为全球最大的移动运营商，其业务涵盖 2G、3G、4G 移动通信及无线宽带接入等多种服务形式，其用户量达 6 亿多。2007 年，开始建立以 Hadoop 技术为基础的“大云”平台，2008 年建立了第一个 256 节点的集群。目前，中国移动已经具有 1 000 多个节点、5 000 个处理器、3PB 数据的大规模数据处理平台，用于进行用户行为分析、客户流失预测、服务关联分析、网络服务质量分析、过滤垃圾短消息等。

1.1.4 大数据研究的意义

(1) 大数据的研究对捍卫国家网络空间的数字主权，维护国家的安全稳定，经济与社会的健康发展具有重大意义。信息时代的数字主权是继海、陆、空、天之后的又一大博弈空间，在大数据领域的落后，意味着失守产业战略制高点，意味着数字主权无险可守，意味着国家安全将出现漏洞。大数据将直接影响国家和社会稳定，是关系国家安全的战略性问题。因此，公安、国保、检察院、法院等关系到国家安全、社会稳定的部门和机关应该加强对大数据技术的研究和学习。

(2) 大数据是国民经济核心产业信息化升级的重要推动力量。以数据为王的大数据时代的到来，使产业界的需求与关注点发生了重大转变：企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为关注大数据处理能力，软件也将从编程为主转变为以数据为中心。大数据处理的兴起也改变了云计算的发展方向，使其进入以分析即服务(AaaS)为主要标志的 Cloud 2.0 时代。

(3) 大数据在科学和技术上的突破将可能诞生出数据服务等战略性新兴产业。数据科学与技术的突破意味着人们能够厘清数据交互连接产生的复杂性，掌握数据冗余与缺失双重特征引起的不确定性，驾驭数据的高速增长与交叉互联引起的涌现性，进而能够根据实际需求从网络数据中挖掘出其所蕴含的信息、知识甚至是智慧，最终达到充分利用网络数据价值的目的。网络数据不再是产业环节上产生的副产品，相反地，网络数据已成为联系各个环节的关键纽带，通过对网络数据纽带的分析与掌握，可以降低行业成本，促进行业效率，提升行业生产力。因此，可以预见，在网络数据的驱动下，行业模式的革新将可能催生出数据服务等一系列战略性的新兴产业。

1.2 大数据处理技术简介

1.2.1 大数据的关键技术

众所周知，大数据所面临的已经不是数据量大的问题了，最重要的问题是分析大数据，只有通过分析才能获取更多智能的、深入的、有价值的信息。当前，越来越多的应用涉及大数据，而这些大数据的属性，包括数量、速度、多样性等都呈现了大数据不断增长的复杂性，所以大数据的分析方法在大数据领域就显得尤为重要，可以说是决定最终信息是否有价值的决定性因素。那么，大数据分析普遍使用的方法与技术有哪些呢？