贵阳市加快经济发展产业知识系列干部读本

GUIYANGSHI JIAKUAI JINGJI FAZHAN CHANYE ZHISHI XILIE GANBU DUBEN

走进大数据时代

ZOUJIN DASHUJU SHIDAI

贵阳市干部培训教材编审指导委员会 / 编

大数据已成为贵阳转型发展的战略性选择。本书涵盖了大数据概念、大数据发展与应用、大数据产业化之路、生态贵阳与大数据等内容，旨在帮助广大干部了解前沿大数据产业发展趋势和发展价值，树立大数据思维，自觉实践，开启贵阳绿色经济发展新征程。

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

贵阳市加快经济发展产业知识系列干部读本

GUIYANGSHI JIAKUAI JINGJI FAZHAN CHANYE ZHISHI XILIE GANBU DUBEN

走进大数据时代

ZOUJIN DASHUJU SHIDAI

贵阳市干部培训教材编审指导委员会 / 编

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

走进大数据时代/贵阳市干部培训教材编审指导委员会编．—北京：电子工业出版社，2014.9
（贵阳市加快经济发展产业知识系列干部读本）
ISBN 978-7-121-24233-5

Ⅰ．①走… Ⅱ．①贵… Ⅲ．①电子商务—产业发展—研究—贵阳市
Ⅳ．①F724.6

中国版本图书馆 CIP 数据核字（2014）第 203786 号

责任编辑：徐蔷薇　　特约编辑：劳嫦娟
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1 000　1/16　印张：9.75　字数：156 千字
版　　次：2014 年 9 月第 1 版
印　　次：2014 年 9 月第 1 次印刷
定　　价：39.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

序　言

刘　俊

中共贵阳市委常委、组织部长，市委党校校长

回顾人类文明的发展时代历程，每一次科技的进步都带来了产业和社会发展的巨大变革。随着云计算、物联网、移动互联网、智慧城市等新技术、新模式的不断发展，数据无时无刻不在产生，如何从大量类型多样的流动数据中挖掘它们的潜在价值，大数据应运而生。信息技术和互联网的新发展带来了大数据的爆发式增长，数据正在成为驱动经济增长和社会进步的重要基础和战略资源。

哈佛大学教授加里金说："这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。"大数据作为下一轮产业革命的主导力量，数据正在成为21世纪的"原油"，不仅能提升社会生产力、创造新的社会价值，更能够提高政府管理效率、服务水平和加快创新能力建设。对于海量数据资源的挖掘和应用催生的大数据产业，蕴含着巨大的商业价值和社会价值，是全球促进创新、角力竞争、提高生产力的前沿领域。发展大数据产业是贵阳市实现后发赶超的创新性产业，是实现产业转型和新型工业化的战略选择，是统筹经济发展与生态文明建设的必由之路。

贵阳发展大数据产业，生态环境好、气候凉爽、地质稳定、电力充沛、交通便利，有着得天独厚的要素保障，随着富士康第四代产业园、中国电信、移动、联通三大运营商数据中心落户贵安新区，中关村贵阳科技园的先进科技等产业资源持续导入，进一步的强化了大数据发展的产业配套支撑能力。近年来，中央和省、市出台了一系列支持大数据产

业的政策意见，贵阳发展大数据迎来了政策环境利好的历史机遇，这些优势机遇为发展大数据产业、推动科技创新创造了良好的条件。贵州省委常委、贵阳市委书记陈刚同志指出："大数据产业已经汇聚成一股势不可挡的发展潮流，唯有顺势而为、追赶潮流，才能创造跨越式发展的美好明天。"贵阳市紧紧依靠改革、开放、创新三大活力，坚持发展和生态两条底线，深化对大数据产业链、产业模式、产业集成等方式的认识和理解，充分挖掘大数据的商业价值和管理价值，强力推进数据中心、云计算服务、服务外包、智慧城市等一系列云建设，努力使贵阳在大数据领域走在前列，实现贵阳的后发赶超、升级发展。

《走进大数据时代》干部读本涵盖了大数据概念、大数据发展与运用、大数据产业化之路、生态贵阳与大数据等内容。希望广大干部通过读本了解前沿大数据产业发展趋势和发展价值，树立大数据思维，把大数据作为贵阳产业发展的重要突破口和增长极，推动治理体系和治理能力现代化建设，提高大数据时代党员干部具备的能力和素养，增强工作水平，更有效率、更为科学地推动大数据的发展，开启贵阳绿色经济发展新征程。

2014 年 9 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目录

第一章 大数据概念

一、大数据的定义

大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的产物，它具有数据量大、类型复杂、内容变化快的特征，蕴含广泛的应用价值和巨大的市场机会，将改变新一轮产业格局，推动经济社会的深刻变革。

维基百科[1]中定义的大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。这个定义是各种学术和应用领域最广泛引用的一个定义。

以大数据的四个特征作为补充，能给出一个更为清晰的大数据的定义。国内一些专家就把大数据定义为数据量大、数据类型多、数据流动快和数据潜在价值大的数据集[2]。

1. 数据量大（Volume）

数据量大是大数据区分于传统数据最显著的特征。

2. 数据类型多（Variety）

大数据所处理的数据类型不仅是单一的文本文件或者数据库中的数据，它还包括了互联网上的交易订单、网络日志、博客、微博、音频、视频等各种复杂结构类型的数据。

3. 数据流动快（Velocity）

速度是大数据区分于传统数据的重要特征。在大量数据面前，需要实时分析出有用信息，处理数据的速度对于数据利用非常重要。

4. 数据潜在价值大（Value）

在基础研究和技术开发领域，上述三个特征已经足够表征大数据的

特征。但在商业应用领域，第四个特征就显得非常关键。投入如此巨大的基础研究和技术开发的努力，就是因为人们洞察到了大数据的潜在的巨大价值。

二、浅谈大数据的“大”

数据量的大小是用计算机存储容量的单位来计算的，基本的单位是字节（Byte），每一数据量的更大量级是按照千分位递进，如下所示：

1Byte(1B)	相当于一个英文字母
1KiloByte(1KB) = 1024B	相当于一则短篇故事的内容
1MegaByte(1MB) = 1024KB	相当于一则短篇小说的文字内容
1GigaByte(1GB) = 1024MB	相当于贝多芬第五乐章交响曲的乐谱内容
1TeraByte(1TB) = 1024GB	相当于一家大型医院中所有的X光图片内容
1PetaByte(1PB) = 1024TB	相当于50%的全美学术研究图书馆藏书信息内容
1ExaByte(1EB) = 1024PB	5EB相当于至今全世界人类所讲过的话语
1ZettaByte(1ZB) = 1024EB	如同全世界海滩上的沙子数量的总和
1YottaByte(1YB) = 1024ZB	1024个像地球一样的星球上的沙子数量的总和

目前，一般中型企业数字化的数据量在TB级以上，一些特大型企业数字数据量达到了PB级，谷歌、百度、新浪、腾讯、淘宝这些互联网公司的数字数据量都在PB级以上。美国科学家吉姆·格雷（Jim Gray）认为，每18个月全球新增的数据量是计算机有史以来全部数据量的总和，也就是说数据量每18个月就能翻一番。据国际咨询机构IDC统计[3]，全球数据量在2010年正式进入了ZB时代，其预计到2020年，全球将总共拥有35ZB的数据量（见图1-1）。但是，令人欣喜的是，在过去的50年，数据存储的成本却每两年就能降一半，而单位存储器上的存储密度却增加了5000万倍。

因此，我们的世界正在变成一个数据的世界，我们正处于大数据时代的边缘。像水、空气和石油这些自然资源一样，数据也正成为这个世界中的一种自然资源。

此为试读，需要完整PDF请访问：www.ertongbook.com

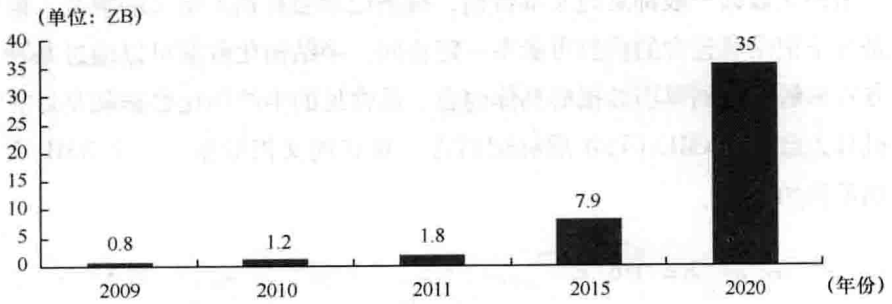

图 1-1 全球数据量的增长

三、复杂多样的数据类型

大数据不仅体现在数量大，也体现在数据类型多。在如此大量的数据中，仅有约 20% 的数据属于传统数据库中的结构化数据，80% 的数据属于更广泛存在于互联网、社交网络、电子商务、物联网等领域的非结构化数据。由信息技术所产生的这些数据已经远远超越了目前人力所能处理的范围。

（一）按照数据结构分类

按照数据结构，数据可分为结构化数据、半结构化数据和非结构化数据。

1. 结构化数据

结构化数据是存储在数据库里，并且可以用二维表来表达的数据，例如，一张客户信息表。结构化数据示例如表 1-1 所示。

表 1-1 结构化数据示例

客 户 号	客户姓名	交 易 额	所购产品
200048901	张伟	1000.0	冰箱
200057903	李东	456.0	烤炉

2. 半结构化数据

半结构化数据是介于结构化数据和非结构化数据之间的数据类型。

半结构化数据一般都是纯文本数据，每条记录会有预先定义的规范，但是每条记录所包含的信息可能不一定相同。半结构化数据可以通过某种方式来解析得到每项数据的具体内容。最常见的半结构化数据就是计算机日志数据、XML（可扩展标记语言）格式的文档数据。一个 XML 文档示例如下：

```
<? xml version = "1.0" ? >
< Order >                                    //产品订单
< Product xmlns = "http://market" >          //产品
< Title > The Joshua Tree </Title >          //产品名
< Artist > U2 </Artist >                     //产品创作的艺术家
</Product >
</Order >
```

各种计算机日志文件是在计算机系统运行中由计算机、网络设备或者传感器等生成的机器数据，用于记录这些计算机系统内执行的自动化操作的详细信息。最常见的日志文件是互联网日志，它根据预定义的字段顺序对网页访问行为进行记录，一个互联网日志文件的示例如下：

```
2005-01-0316:44:57218.17.90.60GET/Default.aspx-80-218.17.90.60Mozilla/4.0 +
(compatible; + MSIE + 6.0; + Windows + NT + 5.2; + .NET + CLR + 1.1.4322)20000
```

互联网用户对网站的每一次点击都会被网络服务器记录在日志中，由此产生了所谓“点击流”数据，也是日志的一种。

3. 非结构化数据

非结构化数据指的是那些非纯文本类数据，没有标准格式，无法直接解析出相应的数据值。常见的非结构化数据有网页、多媒体（图像、声音、视频等）。如图 1-2 所示为现实生活中的各种非结构化数据，主要包括互联网上的 Web 网页、电子邮件、多媒体文档、多媒体流文件（包括声音流、视频流、文本流、图像流、动画流等）等；各个行业在信息化应用中产生的大量实时多媒体数据和物联网数据，包括各种视

频、图像和音频文件，如 CAD/CAM 数据、视频会议、视频监控、数字电影、卫星图像、遥感图像、大型实景游戏、扫描仪数据、医学影像、传感器数据、分享视频、分享照片、数字电视等；社交网络中的实时消息数据，如微博、微信等。

图 1-2 现实世界中的非结构化数据

（二）按照产生主体分类

数字数据尽管都是由计算机产生的，但从源头看，生成数据的主体并不相同。第一类是由计算机软件系统运行产生的数据，如企业资源规划（ERP）、办公自动化（OA）等软件系统的数据库中的数据。二是人与人的交互所产生的数据，如微博（各种短文字、图片和视频）、微信（各种短文字、音频、视频）、博客、评论、图片和视频分享、呼叫中心的各种留言或者电话投诉等。三是由各种硬件设备自动产生的数据，如计算机服务器日志、传感器数据（天气、环境、温度、压力等）、图像和视频数据（车间监控的视频数据、交通监控数据、社区安全监控数据）、电子标签（RFID）、二维码或者条形码扫描的数据等（见图 1-3）。

在大数据应用中，需要采集这些来自不同数据源、采用不同格式、跨不同业务的数据。例如，在一个制造企业中，产品创新的创意可能需要采集来自电子商务网站的交易数据和社交网站上关于产品的微博评论

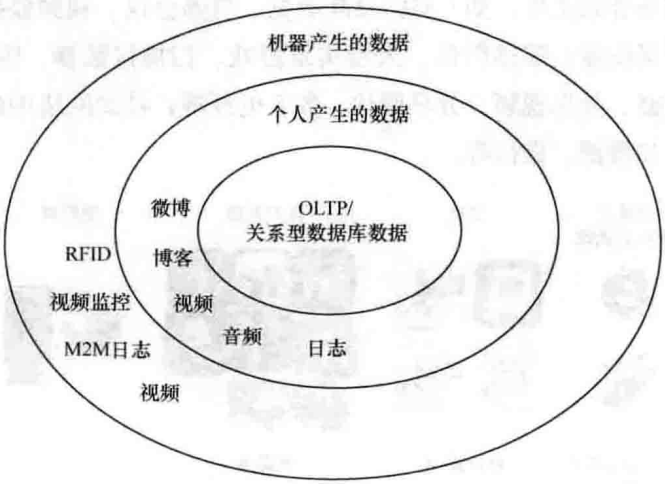

图 1-3 不同的大数据主体

来源：CIOManage（赛智时代），www. ciomanage. com。

和转发信息，产品的设计可能需要采集产品知识库中的二维和三维CAD设计文档以及三维动画原型，产品的市场宣传可能需要采集竞争对手产品的视频短片等。而一家医疗机构，需要分析与患者症状相似的很多病人的电子健康档案和电子病历，需要查阅护士和医生的各种病历记录，需要分析为患者治疗的医疗设备中的日志数据以了解近期就诊趋势，也需要通过远程的家庭医疗设备来分析包含患者健康状态的流媒体数据，这些数据的类型千差万别。

（三）按照数据作用方式分类

按照数据作用的方式，大数据分为交易数据和交互数据。

交易数据是指来自电子商务和企业管理信息系统中的数据，包括来自ERP、企业对企业电子商务（B2B）、企业对个人电子商务（B2C）、个人对个人电子商务（C2C）、团购等信息系统中的数据，这些数据存储在交易系统的数据库中，可以执行联机事务处理（OLTP）和联机分析处理（OLAP）。

交互数据是指社交网络相互作用产生的数据，包括社交媒体交互

（人为生成交互）和机器交互（设备生成交互）的新型数据。

交易和交互数据的有效融合将是大势所趋，大数据应用要能有效集成这两类数据，并实现这些数据的同步处理和分析。

四、瞬息万变的大数据

大数据的速度是指数据被创建、处理和分析的速度。当前，由于计算机处理器和存储等计算技术的不断进步，数据处理的速度越来越快。但是，传统计算技术仍然不能满足大量和复杂类型的大数据的处理速度的要求。例如，在社交网络的计算环境下，大量数据被实时创建，用户需要实时的信息反馈和数据分析，并将这些数据结合到自身高效的业务流程和敏捷的决策过程中，数据处理的速度要求也越来越高。对一次大数据的复杂查询，传统数据库技术可能需要花几个小时，基于大数据技术，处理时间需要逐步降低到分钟级、秒级、毫秒级，甚至完全实时。

五、大数据的潜在价值

大数据是丰富的资源宝藏，它的价值主要体现在三个方面[4]：一是提供更多的信息，挖掘大数据的潜在价值。以往被分析和利用的只是其中20%的结构化数据，今天随着大数据技术的发展已经有能力分析80%的非结构化数据。二是提供更动态的行为信息，挖掘大数据还原真实世界的价值。以往被分析的数据只是流程执行的结果信息、属性描述信息，今天随着大数据应用，已经有能力对流程中的各类行为信息进行获取和分析，包括客户行为信息、员工行为信息、设备行为信息、空间行为信息等。这些信息的获得，是依赖于互联网、物联网、移动互联网等信息基础设施所建立起来的对客观对象行为的跟踪和记录能力。三是提供不同领域数据集相互连接所产生的信息，实现大数据整合创新的价值。以往被分析的数据只是局限于某个范围或某个领域的数据，今天随着技术的发展，不同场景下的数据被连接起来了。连接，让数据产生了网络效应，带来了更大的业务价值。例如，互联网和移动互联网数据的连接，企业数据和社交媒体数据的连接，线上服务和线下服务数据的连

接，网络、社交和空间数据的连接等，不同数据源的连接，使得人类有能力更加全方位地深入还原和洞察真实的、曾经的、复杂的“现实”。这正是大数据的最大价值！

大数据的价值挖掘是与大数据的容量、种类密切相关的。一般来看，数据容量越大，种类越多，信息量越大，获得的知识越多，能够发挥的潜在价值也越大。但这依赖于大数据处理的手段和工具，否则由于信息和知识密度低，可能造成数据垃圾和信息过剩，失去数据的利用价值。

研究表明，数据的价值会随着时间的流逝而降低[5]。简单地看，数据的价值与时间是成反比的。因此，数据处理速度越快，数据价值越能够更好地获得。大数据的价值也与它所传播和共享的范围相关，使用大数据的用户越多，范围越广，信息的价值就越大。这些大数据价值的充分发挥，需要依赖于大数据的分析和挖掘技术。

此为试读，需要完整PDF请访问：www.ertongbook.com