大数据背景下网络安全问题研究

洪运国 ◎ 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

大数据背景下网络安全问题研究

洪运国 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

图书在版编目（CIP）数据

大数据背景下网络安全问题研究 / 洪运国著. —北京 ：北京理工大学出版社，2021.3

ISBN 978-7-5682-9618-2

Ⅰ. ①大… Ⅱ. ①洪… Ⅲ. ①计算机网络-网络安全-研究
Ⅳ. ①TP393.08

中国版本图书馆 CIP 数据核字（2021）第 043177 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 /（010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址 / http：//www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 唐山富达印务有限公司
开　　本 / 710 毫米 ×1000 毫米　1/16
印　　张 / 8.5　　责任编辑 / 李　薇
字　　数 / 161 千字　　文案编辑 / 李　薇
版　　次 / 2021 年 3 月第 1 版　2021 年 3 月第 1 次印刷　　责任校对 / 周瑞红
定　　价 / 49.80 元　　责任印制 / 施胜娟

前　言

信息科技的发展为大数据时代的到来提供了技术支撑，而数据产生方式的变革是推动大数据时代到来至关重要的因素。大数据关键技术涵盖数据存储、处理、应用等多方面的技术，根据大数据的处理过程，可将其分为大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘等环节。当前，许多国家政府、国际组织和企业等都意识到了大数据的重要性，纷纷投身大数据市场，大数据已经形成产业规模，并上升到国家战略层面，大数据技术和应用呈现纵深发展。

网络安全是指网络系统的硬件、软件及其系统中的数据受到保护，不因偶然或者恶意的原因而遭到破坏、更改、泄露，系统连续可靠正常地运行，网络服务不中断。其具有保密性、完整性、可用性、可控性、可审查性等特性。可以说，凡是涉及网络上信息的保密性、完整性、可用性、真实性和可控性的相关技术和理论都是网络安全所要研究的领域。

网络安全是一个关系国家安全和主权、关系社会稳定、关系民族文化继承和发扬的重要问题。随着全球信息化步伐的加快，网络安全变得越来越重要，随着信息技术的深入发展，网络安全问题日益严峻，如利用网络干涉其他国家内政以及大规模网络监控、窃密等严重危害国家政治安全和用户信息。

进入大数据时代，数据的爆炸式增长让数据安全和隐私保护问题变得更加复杂，也让各行各业陷入了大数据安全防护的迷思之中。而当前，网络安全法律保护还未跟上技术发展的步伐，深刻理解大数据安全的内涵，针对行业特征分析问题、制定安全需求，寻找防护对策才是最为妥当的做法。

本书重点阐述了大数据时代网络信息安全面临的挑战、大数据时代网络信息安全防范对策等方面问题。全书共分六章，各章内容相对独立，有基础理论，也有具体实例和实际应用。

全书组织结构如下：第 1 章为大数据概述，包括大数据对社会生活的影响、大数据关键技术、大数据的现状与发展、大数据的应用等内容；第 2 章为网络安全概述，包括网络安全的层次划分、网络安全的重要性、网络安全现状分析等内容；第 3 章为大数据时代网络信息安全面临的挑战，包括大数据时代的安全形势分析、大数据时代的安全需求等内容；第 4 章为国外大数据安全发展的主要经验及思考，包括国外大数据发展现状、国外大数据安全形势、国外大数据安全发展

的经验等内容；第 5 章为大数据时代网络信息安全防范对策，包括完善大数据时代信息安全管理体系、推进互联网信誉体系建设和网络道德建设、提高大数据安全防范技术等内容；第 6 章为大数据与云安全，包括云安全架构及关键技术、云安全架构下的大数据安全等内容。

本书由大连职业技术学院洪运国撰写，由于水平有限，加之时间仓促，书中难免有不妥之处，恳请读者批评指正。

洪运国

目　录

第 1 章

大数据概述

1.1 大数据时代

第三次信息化浪潮涌动，大数据时代全面开启。信息科技的发展为大数据时代的到来提供了技术支撑，而数据产生方式的变革是促进大数据时代到来至关重要的因素。

1. 第三次信息化浪潮

根据 IBM（国际商业机器公司）前首席执行官郭士纳的观点，IT（互联网技术）领域每隔 15 年就会迎来一次重大变革。1980 年前后，PC（个人计算机）开始普及，计算机走入企业和千家万户，社会生产力大大提高，人类迎来了第一次信息化浪潮，Intel（英特尔）、IBM、苹果、微软、联想等企业是这个时期的标志。1995 年前后，人类开始全面进入互联网时代。互联网的普及把世界变成“地球村”，每个人都可以自由徜徉于信息的海洋，由此，人类迎来了第二次信息化浪潮，这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔 15 年，在 2010 年前后，云计算、大数据、物联网的快速发展，拉开了第三次信息化浪潮的大幕。大数据时代已经到来，也必将涌现出一批新的市场标杆企业。

2. 信息科技为大数据时代提供技术支撑

信息科技的发展需要解决信息存储、信息传输和信息处理三个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

首先，存储设备的容量不断增加。数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中。随着科学技术的不断进步，存储设备的制造工艺不断升级，容量大幅增加，运行速度不断提升，价格却在不断下降。早期的存储设备容量小、价格高、体积大。例如，IBM 在 1956 年生产的一个早期的商业硬盘，容量只有 5MB，不仅价格昂贵，而且体积有一个冰箱那么大。今天容量为 1TB 的硬盘，大小只有约 8.89cm，读写速度可以达到 200MB/s，且价格仅为 400 元左右。廉价、高性能的硬盘存储设备，不仅提供了海量的存储空间，同时大大降低了数据存储成本。与此同时，以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存作为永久性存储设备，具有体积小、质量轻、能耗低、抗振性

好等优良特性。

总体而言，数据量和存储设备容量二者之间是相辅相成、互相促进的。一方面，随着数据的不断产生，需要存储的数据量不断增加，对存储设备的容量提出了更高要求；另一方面，更大容量的存储设备进一步加快了数据量增长的速度。随着单位存储空间价格的不断降低，人们开始倾向于把更多的数据保存起来，以便在未来的某个时刻可以用更先进的数据分析工具从中挖掘价值。

其次，CPU（中央处理器）的处理能力大幅度提升。CPU 处理速度的不断提升也是促进数据量不断增加的重要因素。性能不断提升的 CPU，大大提高了处理数据的能力，使我们可以更快地处理不断累积的海量数据。

最后，网络带宽不断提升。1977 年，世界上第一条光纤通信系统在美国芝加哥市投入使用，数据传输速率为 45Mbit/s，从此，人类社会的信息传输速度不断被刷新。进入 21 世纪，世界各国更是纷纷加大对宽带网络的建设力度，不断扩大宽带网络的覆盖范围并提高其传输速率。以我国为例，截至 2012 年 6 月，92.6% 的固定宽带用户接入速率达到或超过 2Mbit/s，国际互联网出口带宽达到 1.48Tbit/s，是 2005 年的 11.4 倍。与此同时，移动通信宽带网络迅速发展、3G 网络基本普及、4G 网络覆盖范围不断扩大、5G 网络进入人们的生活，各种终端设备可以随时随地传输数据。大数据时代，信息传输不会再遭遇网络发展初期的瓶颈和制约。

3. 数据产生方式的变革推动大数据时代的来临

数据是我们通过观察、实验或计算得出的结果。数据和信息是两个不同的概念。信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等；而数据则是构成信息的基本单位，离散的数据没有任何实用价值。

数据有很多种，比如数字、文字、图像、声音等。随着信息化进程的加快，我们在日常生产和生活中每天都会产生大量的数据，比如商业网站、政务系统、零售系统、办公系统、自动化生产系统等，每时每刻都在不断地产生数据。当今数据已经渗透到每个行业和业务职能领域，成为重要的生产因素。从创新到决策，数据推动着企业的发展，并使各级组织运营得更为高效。可以这样说，数据将成为每个企业获取核心竞争力的关键要素。数据资源已经和物质资源、人力资源一样成为国家的重要战略资源，影响着国家和社会的安全、稳定与发展，因此，数据也被称为“未来的石油”。

数据产生方式的变革，是推动大数据时代来临的重要因素。总体而言，人类社会的数据产生方式大致经历了三个阶段：运营式系统阶段、用户原创内容阶段和感知式系统阶段。

（1）运营式系统阶段

人类社会最早大规模管理和使用数据，是从数据库的诞生开始的。大型零售

超市销售系统、银行交易系统、股市交易系统、医院医疗系统、企业客户管理系统等大量运营式系统，都建立在数据库的基础之上。数据库中保存了大量结构化的企业关键信息，用来满足企业各种业务需求。在这个阶段，数据的产生方式是被动的，只有当实际的企业业务发生时，才会产生新的数据并存入数据库。

（2）用户原创内容阶段

互联网的出现，使数据传播更加快捷，而不需借助磁盘、磁带等物理存储介质来传播数据。网页的出现进一步促进了大量网络内容的产生，从而使人类社会的数据量开始呈现“井喷式”增长。但是，互联网真正的数据爆发产生于以“用户原创内容”为特征的 Web 2.0 时代。Web 1.0 时代主要以门户网站为代表，强调内容的组织与提供，大量上网用户本身并不参与内容的产生。而 Web 2.0 技术以 Wiki（多人协作的写作系统）、博客、微博、微信等自服务模式力主，强调自服务，大量上网用户本身是内容的生成者，尤其是随着移动互联网和智能手机终端的普及，人们可以随时随地使用手机发微博、传照片，这使数据量开始急剧增加。

（3）感知式系统阶段

物联网的发展最终导致了人类社会数据量的第三次跃升。物联网中包含大量的传感器，如温度传感器、湿度传感器、压力传感器、位移传感器、光电传感器等。此外，视频监控摄像头也是物联网的重要组成部分。物联网中的这些设备，每时每刻都在自动产生大量数据，与 Web 2.0 时代的人工数据产生方式相比，物联网中的自动数据产生方式，将在短时间内生成更密集、更大量的数据，使人类社会迅速步入“大数据时代”。

1.2　大数据的定义及特征

1.2.1　大数据的定义

时至今日，“数据”变身成“大数据”，开启了一次重大的时代转型。“大数据”这一概念的形成，有三个标志性事件：一是 2008 年 9 月，美国《自然》杂志专刊——*The next google*，第一次正式提出“大数据”概念；二是 2011 年 2 月 1 日，《科学》杂志专刊——*Dealing with data*，通过社会调查的方式，第一次综合分析了大数据对人们生活的影响，详细描述了人类面临的“数据困境”；三是 2011 年 5 月，麦肯锡研究院发布报告——*Big data*：*The next frontier for innovation*，*competition*，*and productivity*《大数据：下一个创新、竞争和生产力的前沿》，第一次给大数据做出了相对清晰的定义：大数据是指其大小超出了常规数据库工具获取、储存、管理和分析能力的数据集。

对于大数据，Gartner（全球第一家信息技术研究和分析公司）给出了这样的定义：大数据是需要新的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。大数据的意义不仅仅在于掌握庞大的数据信息，更在于对这些含有重要意义的数据进行专业化处理之后产生的价值。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，并且通过“加工”来实现数据的“增值”。

大数据是个宽泛的概念，上面的定义突出了一个“大”字。因为大数据不仅要用来描述大量的数据，还应涵盖处理数据的速度和能力。前面几个定义都是从大数据本身出发，而人们更关心的是大数据能帮助人们做什么。大数据发展的终极目标是人们从各种类型的海量数据中快速获得高价值的信息，没有价值或者没有发现其价值的大数据从某种意义上来说是一种资源的浪费。

1.2.2 大数据的特征

当前，较为统一的认识来自互联网数据中心对大数据的定义，其包含四个基本特征：Volume（数据量大）、Variety（数据类型繁多）、Velocity（处理速度快）、Value（商业价值高），即所谓的“4V”特征。

第一，数据量大。从2013年至2020年，人类的数据规模扩大了50倍，人类产生的数据量增长到44万亿GB，相当于美国国家图书馆数据量的数百万倍，且每18个月就会翻一番。

第二，数据类型繁多。大数据与传统数据相比有一些不同：数据来源很多，分为社交网络、搜索引擎、传感器数据、通话记录、位置信息等；数据类型多，分为文本、音频、视频、光谱、图片等；数据格式多，分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，这些多样性的数据对数据的处理能力提出了更高的要求。

第三，处理速度快。随着现代感测技术、互联网、计算机技术的发展，数据生成、储存、分析、处理的速度远远超出人们的想象，这是大数据区别于传统数据的显著特征。一方面数据不断更新，增长速度快；另一方面数据访问、处理、交付等速度加快。在每一天的每一分钟里，从网络购物、打电话、浏览网页到访问社交网站等都会产生海量的新数据。随着新数据的不断出现，人们对数据处理的速度提出了越来越高的要求。数据处理的时效性高，才能使大量的数据得到有效的利用。此外，随着移动网络的发展，人们对数据的实时应用需求更加普遍，对数据的响应时间也更加敏感，大多数人希望在第一时间抓住重要的信息。

第四，商业价值高。大数据有巨大的潜在价值，但同其指数呈爆发式增长相比，某一对象或模块数据的价值密度较低，这无疑给我们开发海量数据增加了难度和成本。

大数据的“4V”特征使大数据有别于传统的数据概念。大数据的概念与“海量数据”不同，后者只是单纯强调数据的量，而大数据不仅用来描述大量的数据，还进一步指出了数据的复杂形式、数据的快速时间特性以及对数据进行专业化处理并最终获得有价值信息的能力。

1.3　大数据的来源

美国互联网数据中心报告指出，互联网上的数据每年将增长 50%，每两年便翻一番，目前世界上 90% 以上的数据是最近几年才产生的。此外，全世界的工业设备、汽车、电表上有着无数的数码传感器，以便随时测量和传递有关位置、运动、振动、温度、湿度乃至空气中化学物质的变化，它们也产生了海量的数据信息。

根据数据来源的不同，可以将大数据分为以下几种类型：

1. 来源于使用者

人们在互联网活动以及使用移动互联网的过程中会产生数据，包括文字、视频、图片等。这些数据实时、量大，代表着每个网民的真实想法，反映了他们想要了解和感兴趣的事。虽然价值密度低，但是能够反映出人们的真实想法。

2. 来源于计算机

各类信息系统产生的数据，是以文件、多媒体、数据库的形式存在的，也包括审计日志等自动生成的数据。这些数据基本上来自企业，大多属于结构化数据。

3. 来源于数字设备

各类数字设备采集的数据，比如通过天文望远镜拍摄的图像、视频信息、气象学中的卫星云图数据、移动设备的呼叫详单等。

由于来源不同，类型不同的数据反映的是同一事物的不同方面。例如，消费者的消费记录能够反映出他们的消费能力、消费频率、消费喜好等；消费渠道信息能够反映出他们的渠道喜好；支付信息能够反映出他们的支付渠道等情况。

1.4　大数据对社会生活的影响

大数据将会对社会发展产生深远的影响，具体表现在以下几个方面：

1. 广告投放精准化

据报道，美国 Target（塔吉特）连锁超市创建了一套女性购买行为在怀孕期间变化的模型，通过采集女性用户的购买行为数据并对其进行分析，就能判断女性用户是否怀孕，并进一步向其推送所需的婴儿用品。不仅如此，如果用户从他

们的店铺中购买了婴儿用品，Target 在接下来的几年中会根据婴儿的生长周期情况定期给这些顾客推送相关产品，以此来提高这些客户的忠诚度。

同样，亚马逊和京东商城等购物网站通过数据挖掘技术对用户的行为习惯和喜好进行追踪分析，从大数据背后找到符合用户兴趣和习惯的产品和服务，并向顾客提供个性化的商品推荐。

2. 医疗卫生体系更加精密

通过分析大量用户的搜索记录，比如“咳嗽”“发烧”等特定词条，谷歌公司能准确预测美国冬季流感传播趋势。和官方机构相比，谷歌能提前一两周预测流感爆发，预测结果与官方数据的相关性高达 97%。

对个人而言，大数据可以为个人提供个性化的医疗服务。过去我们看病，医生只能对我们当下的身体情况做出判断，而在大数据的帮助下，将来的诊疗可以对一个患者的累计历史数据进行分析，并结合遗传变异、对特定疾病的易感性和对特殊药物的反应等关系，实现个性化医疗；还可以在患者发生疾病症状前，提供早期的检测和诊断。早期发现和治疗可以显著降低肺癌等疾病给卫生系统造成的负担，因为早期的治疗费用是后期治疗费用的一半。

3. 社会安全管理更为有序

在社会安全管理领域，通过对手机数据的挖掘，可以分析实时动态的流动人口来源、出行、实时交通客流信息及道路拥堵情况。利用短信、微博、微信和搜索引擎，可以收集热点事件，挖掘舆情，还可以追踪造谣信息的源头。美国麻省理工学院通过对十万多人的手机通话、短信和空间位置等信息进行处理，来提取人们行为的时空规律性，进行犯罪预测。

4. 带来新的就业市场

据盖特纳咨询公司预测，大数据将为全球带来 440 万个 IT 新岗位和上千万个非 IT 岗位。麦肯锡公司曾经预测美国到 2018 年需要深度数据分析人才 44 万 ~49 万人，但尚缺口 14 万 ~19 万人；需要既熟悉本单位需求又了解大数据技术与应用的管理者 150 万人，这方面的人才缺口更大。中国人口世界第一，虽然拥有海量的大数据和众多的人才，但能处理与应用大数据的人才却是稀缺资源。

1.5 大数据关键技术

大数据关键技术涵盖数据存储、处理、应用等多个方面，根据大数据的处理过程，可将其分为大数据采集、大数据预处理、大数据存储与管理、大数据分析与挖掘等环节。

1.5.1　大数据采集

大数据采集是大数据生命周期的第一个环节，它通过 RFID 射频数据、传感器数据、社交网络数据、移动互联网数据等方式获得各种类型的结构化、半结构化及非结构化的海量数据。由于可能有成千上万的用户同时进行并发访问和操作，因此，必须采用专门针对大数据的采集方法，其主要包括三种：一是数据库采集，一些企业会使用传统的关系型数据库 MySQL 和 Oracle 等来存储数据，谈到比较多的工具有 Sqoop 和结构化数据库间的 ETL（数据仓库技术），当然当前开源的 Kettle 和 Talend 本身也集成了大数据的集成内容，可以实现和 HDFS（分布式文件系统），HBase（开源数据库）和主流 NoSQL（非关系型数据库）数据库之间的数据同步和集成；二是网络数据采集，网络数据采集主要是借助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的过程，通过这种途径可将网络上非结构化数据、半结构化数据从网页中提取出来，并以结构化的方式将其存储为统一的本地数据文件；三是文件采集，对于文件的采集，谈得比较多的还是使用 Flume（日志收集系统）进行实时的文件采集和处理，虽然 ELK［Elasticsearch（日志存储搜索）、Logstash（日志收集）、Kibana（展示查询）三者的组合］只是处理日志，但是也有基于模板配置的完整增量来实时进行文件采集。如果仅仅是做日志的采集和分析，那么用 ELK 解决方案就完全足够了。

1.5.2　大数据预处理

数据的世界是庞大而复杂的，同时也会有残缺的、虚假的和过时的数据。想要获得高质量的分析挖掘结果，就必须在数据准备阶段提高数据的质量。大数据预处理可以对采集到的原始数据进行清洗、填补、平滑、合并、规格化以及检查一致性等，将那些杂乱无章的数据转化为相对单一且便于处理的构型，为后期的数据分析奠定基础。大数据预处理主要包括：数据清理、数据集成、数据转换以及数据归约四大部分。

1. 数据清理

数据清理主要包含遗漏值处理（缺少感兴趣的属性）、噪声数据处理（数据中存在着错误或偏离期望值的数据）和不一致数据处理。

2. 数据集成

数据集成是指将多个数据源中的数据合并存放到一个一致的数据存储库中。在这一过程中我们着重要解决三个问题：模式匹配、数据冗余、数据值冲突检测与处理。

3. 数据转换

数据转换是指处理抽取上来的数据中存在的不一致的过程。数据转换一般包括两类：第一类是数据名称及格式的统一，即数据粒度转换、商务规则计算以及统一的命名、数据格式、计量单位等；第二类是数据仓库中存在源数据库中可能

不存在的数据，因此需要进行字段的组合、分割或计算。数据转换实际上还包含了数据清洗的工作，需要根据业务规则对异常数据进行清洗，保证后续分析结果的准确性。

4. 数据归约

数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量。主要包括：数据方聚集、维归约、数据压缩、数值归约和概念分层等。数据归约技术可以用数据集的归约表示，使数据集变小，但同时仍然近于保持原数据的完整性。也就是说，在归约后的数据集上进行挖掘，依然能够得到与使用原数据集近乎相同的分析结果。

1.5.3 大数据存储与管理

大数据存储与管理要用存储器把采集到的数据存储起来，建立相应的数据库，以便管理和调用。大数据存储技术路线最典型的有三种：

1. MPP（大规模并行处理）架构的新型数据库集群

MPP 架构的新型数据库集群，重点面向行业大数据，采用 Shared Nothing（无共享）架构，通过列存储、粗粒度索引等多项大数据处理技术，再结合 MPP 架构高效的分布式计算模式，完成对分析类应用的支撑，其运行环境多为低成本的 PC Server（服务器），具有高性能和高扩展性的特点，在企业分析类应用领域获得了极其广泛的应用。这类 MPP 产品可以有效支撑 PB 级别的结构化数据分析，这是传统数据库技术无法胜任的。

2. 基于 Hadoop（分布式系统基础架构）的技术扩展和封装

基于 Hadoop 的技术扩展和封装，其围绕 Hadoop 衍生出了相关的大数据技术，以应对传统关系型数据库较难处理的数据和场景。例如，针对非结构化数据的存储和计算等，可充分利用 Hadoop 开源的优势，伴随相关技术的不断进步，其应用场景也将逐步扩大，目前最为典型的应用场景就是通过扩展和封装 Hadoop 来实现对互联网大数据存储、分析的支撑。

3. 大数据一体机

这是一种专为大数据的分析处理而设计的软、硬件结合的产品，由一组集成的服务器、存储设备、操作系统、数据库管理系统以及为数据查询、处理、分析用途而预先安装及优化的软件组成，高性能大数据一体机具有良好的稳定性和纵向扩展性。

1.5.4 大数据分析与挖掘

大数据分析与挖掘的主要目的是把隐藏在一大批看来杂乱无章的数据中的信息集中起来，进行萃取、提炼，以找出潜在有用的信息和所研究对象的内在规律的过程。主要由可视化分析、数据挖掘算法、预测性分析、语义引擎以及数据质