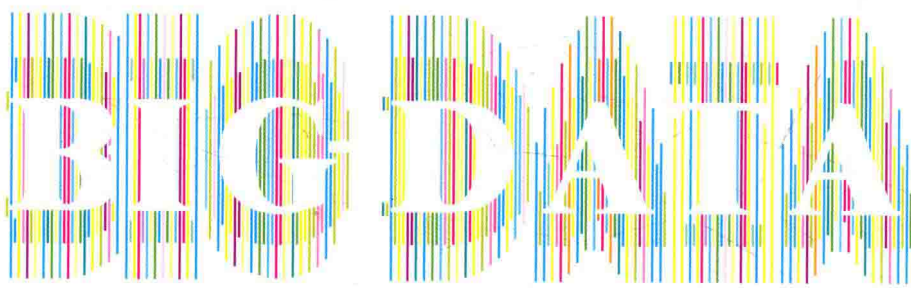
和你一起揭开大数据神秘的面纱

去伪存真、抽丝剥茧、由表及里、循序渐进，让你告别入门客，成为精通者！

大数据技术概论

从虚幻走向真实的数据世界

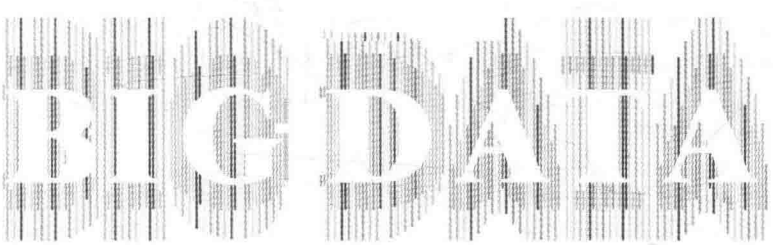
娄岩 ◎ 编著



清華大學出版社

大数据技术概论

娄岩 ◎编著　　徐东雨 ◎参编



清華大學出版社
北京

内 容 简 介

本书从初学者易于理解的角度，以通俗易懂的语言、丰富的实例、简洁的图表、传统和现代数据特征的对比，将大数据这一计算机前沿科学如数家珍地娓娓道来。既介绍了大数据和相关的基础知识，又与具体应用有机结合起来，并借助可视化图表的画面感立体地为读者剖析了大数据的技术和原理，非常便于自学。

本书内容包括大数据概论、大数据采集及预处理、大数据分析、大数据可视化、Hadoop 概论、HDFS 和 Common 概论、MapReduce 概论、NoSQL 技术介绍、Spark 概论、云计算与大数据、大数据相关案例等内容。

本书既可以作为想了解大数据技术和应用的初学者的教材，也适合作为培训中心、IT 人员、企业策划和管理人员的参考书。

图书在版编目(CIP)数据

大数据技术概论/娄岩编著. —北京：清华大学出版社，2017.1
ISBN 978-7-302-45051-1

Ⅰ. ①大…　Ⅱ. ①娄…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 218544 号

责任编辑：付弘宇　薛　阳
封面设计：刘　键
责任校对：焦丽丽
责任印制：何　芊

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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：13　　**字　　数**：315 千字
版　　次：2017 年 1 月第 1 版　　**印　　次**：2017 年 1 月第1 次印刷
印　　数：1～2000
定　　价：35.00 元

产品编号：069372-01

前言

FOREWORD

IT产业在其发展历程中，经历过几次技术浪潮。如今，大数据浪潮正在迅速朝我们涌来，并将触及各个行业和生活的许多方面。大数据浪潮将比之前发生过的浪潮更大、触及面更广，给人们的工作和生活带来的变化和影响也更大。

毋庸置疑，大数据的应用激发了一场思想风暴，也悄然改变了我们的生活方式和思维习惯。大数据正以前所未有的速度颠覆人们探索世界的方法，引起工业、商业、医学、军事等领域的深刻变革。因此，在当前大数据浪潮的猛烈冲击下，人们迫切需要充实和完善自己原有的IT知识结构，掌握两种全新的技能：一是掌握大数据基本技术与应用，使大数据为我们所用的技能；二是掌握数据之间隐藏的规律与关系，以及可视化方法，使大数据更好地服务于社会发展的技能。

本书注重实用性，围绕大数据及其相关技术这一主题，采用深入浅出、图文并茂的叙述方式，简明扼要地阐述了大数据及其相关技术的基本理论和发展趋势，使广大读者通过阅读本书，深入了解和掌握大数据的理论和应用，从而更好地把握时代发展的脉搏和历史赋予的机遇。

本书的目标是给广大读者提供一个既通俗易懂，又具有严谨、完整、结构化特征的书籍。其独到之处是既阐明了大数据技术的系统性和理论性，又对传统数据和大数据在来源、结构、特征、存储方式、使用方法等方面，通过大量的表格和图形方式进行了有针对性的对比和阐述，使读者对两者的区别一目了然，对理解和掌握大数据理技术具有事半功倍的效果。另外，考虑到大数据技术涉及许多新名词和专业性极强的词汇，故在全书的每一章中均附有相关术语的注释，方便读者查阅和自学。

本书还力求将大数据技术晦涩难懂的理论知识以通俗易懂的语言和方式，由浅入深地展现在读者面前，便于读者理解和掌握。本书内容重点突出，语言精练易懂，非常便于自学，可作为想了解、使用大数据技术的相关人员，如工程技术人员、IT工作者、企业策划和管理人员的参考书，也可作为相关学习班的培训教材。

全书共分成11章：第1章大数据概论，第2章大数据采集及预处理，第3章大数据分析概论，第4章大数据可视化，第5章Hadoop概论，第6章HDFS和Common概论，第7章

MapReduce 概论，第 8 章 NoSQL 技术介绍，第 9 章 Spark 概论，第 10 章云计算与大数据，第 11 章大数据解决方案相关案例。

本书在写作过程中参阅了大量的中外书籍和相关资料，在此对各位作者表示真诚的谢意。另外本书得到了中国医科大学沙宪政教授和东北大学杨广明教授的大力支持，清华大学出版社对这本书的出版做了精心策划及充分论证，特此感谢！由于作者水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评斧正！

娄　岩

2016 年 6 月

目录

CONTENTS

CHAPTER 1

第 1 章

大数据概论

导 学

内容与要求

大数据是继物联网之后 IT 产业又一次颠覆性的技术变革。本章主要对大数据技术进行概述、对大数据技术的架构、大数据的整体技术和关键技术、大数据分析的典型工具以及大数据未来发展趋势进行介绍，使读者更好地了解什么是大数据技术。

大数据技术的概述包含了大数据的基本概念、大数据的来源、产生阶段、特点、大数据处理的基本流程、特征和应用领域。了解大数据的来源和应用领域，掌握大数据的特点和大数据处理的基本流程。

大数据技术的架构中了解 4 层堆栈式技术架构，包括基础层、管理层、分析层和应用层。

大数据的整体技术和关键技术中了解大数据的整体技术一般包括数据采集、数据存取、基础架构、数据处理、统计分析、数据挖掘、模型预测和结果呈现等。关键技术一般包括大数据采集、大数据预处理、大数据存储及管理、开发大数据安全，大数据技术、大数据分析及挖掘、大数据展现和应用。

大数据分析的 5 种典型工具简介中简单介绍了 5 种工具，包括 Hadoop、Spark、HPCC、Storm 和 Apache Drill。

大数据未来发展趋势中了解数据资源化，随着大数据应用的发展，大数据资源成为重要的战略资源，数据成为新的战略制高点。

重点、难点

本章重点是了解大数据的特点、特征和大数据未来发展趋势。本章的难点是了解大数据技术架构、整体技术和关键技术。

由于各种网络技术的发展、科学数据处理、商业智能数据分析等具有海量需求的应用变得越来越普遍，面对如此巨大的数据量，无论从形式上还是内容上，已无法用传统的方式进行采集、存储、操作、管理、分析和可视化了。而找出数据源，确定数据量，选择正确的数据处理方法，并将结果可视化的过程就变得非常现实和迫切。而无论是分析专家还是数据科学家最终都会探索新的、无法想象的庞大数据集，以期发现一些有价值的趋势、形态和解决问题的方法。我们完全有理由说，大数据是继物联网之后 IT 产业又一次颠覆性的技术变革。

大数据(Big Data)是指当传统的数据挖掘和处理技术对某些数据无可奈何时使用的处理过程。如数据是非结构化，实时性强或信息量巨大，以至于无法通过关系数据库引擎进行处理的数据，而需要新的技术手段和具有分布式处理数据功能的并行硬件设备来实现。

1.1 大数据技术概述

毋庸置疑，大数据已经走进了我们的生活，且成为整个人类社会关注的热点。什么是大数据，其相关技术、应用领域以及未来的发展趋势将是本章重点介绍的内容。

1.1.1 大数据的基本概念

早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。从技术层面上看，大数据是无法用单台计算机进行处理的，必须采用分布式计算架构。其特色在于对海量数据的挖掘，但它又必须依托一些现有的数据处理方法，如流式处理、分布式数据库、云存储与虚拟化技术，如图 1-1 所示。

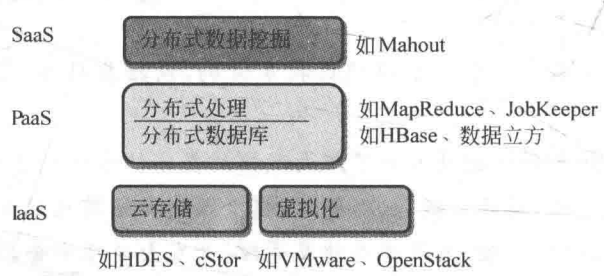


图 1-1 大数据与云技术

网络是大数据的主要载体之一，可以说没有网络就没有今天的大数据技术。美国网络数据中心指出，单就互联网上的数据每年将增长 50%，每两年就将翻一番，而目前世界上 90%以上的数据是最近几年才被人们逐渐认识和产生的。当然数据并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，必然会产生海量的数据

信息。

大数据的意义在于可以通过人类日益普及的网络行为附带生成，并被相关部门、企业所采集，蕴含着数据生产者的真实意图、喜好，其中包括传统结构和非传统结构的数据。

从海量数据中“提纯”出有用的信息，然而这对网络架构和数据处理能力而言无疑是巨大的挑战。在经历了几年的批判、质疑、讨论、炒作之后，人们终于迎来了大数据时代。

大数据的核心在于为客户从数据中挖掘出蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式的研究和探索将是大数据产业健康发展的关键。

1.1.2 IT 产业的发展简史

IT 产业的几个发展阶段如图 1-2 所示，可以说 IT 产业的每一个阶段都是由新兴的 IT 供应商主导的。他们改变了已有的秩序，重新定义了计算机的规范，并为进入 IT 领域的新纪元铺平了道路。

图 1-2 IT 产业的几个发展阶段

20 世纪 60 年代和 70 年代的大型机阶段是以 Burroughs、Univac、NCR、Control Data 和 Honeywell 等公司为首的。在步入 20 世纪 80 年代后，小型机涌现出来，这时为首的公司包括 DEC、IBM、Data General、Wang、Prime 等。

在 20 世纪 90 年代，IT 产业进入了微处理器或个人计算机阶段，领先者为 Microsoft（微软）、Intel、IBM 和 Apple 等公司。从 20 世纪 90 年代中期开始，IT 产业进入了网络化阶段。如今，全球在线的人数已经超过了 10 亿，这一阶段由 Cisco、Google、Oracle、EMC、Salesforce.com 等公司领导。IT 产业的下一个阶段还没有正式命名，人们更愿意称其为云计算/大数据阶段。

数字信息每天在无线电波、电话电路和计算机电缆等媒介中川流不息。我们周围到处都是数字信息，在高清电视机上看数字信息，在互联网上听数字信息，自己也在不断制造新的数字信息。例如，每次用数码照相机拍照后，都产生新的数字信息；通过电子邮件把照片发给朋友和家人，又制造了更多的数字信息。不过，没人知道这些流式数字信息有多少、增加速度有多快、其激增意味着什么。正如中国人在发明文字前就有了阴阳学说，并用其解释包罗万象的宇宙世界一样，西方人用制造、获取和复制的所有 1 和 0，通过计算机处理组成了数字世界。人们通过拍摄照片和共享音乐制造了大量的数字信息，而公司则组织和管理这些数字信息的访问、存储，并为其提供强有力的安全保障。

目前世界上有三种类型模拟数字转换方式：

(1) 为数字信息量的增长提供动力和服务；

(2) 胶片影像拍摄转换为数字影像拍摄，模拟语音转换为数字语音；

(3) 模拟电视转换为数字电视。

从数码照相机、可视电话、医用扫描仪到保安摄像头，全世界有 10 亿多台设备在拍摄影像，这些影像成为数字海洋中最大的组成部分，通过互联网、企业内部网在个人计算机(PC)、服务器及数据中心中复制，通过数字电视广播和数字投影银幕播放。

2007 年是有史以来人类创造的信息量第一次在理论上超过可用存储空间总量的一年。然而，这并不可怕，调查结果强调现在人类应该也必须合理调整数据存储和管理。如三十多年前，通信行业的数据大部分还是结构化数据。如今，多媒体技术的普及导致非结构化数据如音乐和视频等的数量出现爆炸式增长。虽然三十多年前的一个普通企业用户文件也许表现为数据库中的一排数字，但是如今的类似普通文件可能包含许多数字化图片和文件的影像或者数字化录音内容。现在，92%以上的数字信息都是非结构化数据。在各组织和企业中，非结构化数据占到了所有信息数据总量的 80%以上。

另外可视化是引起数字世界急速膨胀的主要原因之一。由于数码照相机、数码监控摄像机和数字电视内容的加速增长及信息的大量复制趋势，使得数字世界的容量和膨胀速度超过此前估计。个人日常生活的“数字足迹”大大刺激了数字世界的快速增长。通过互联网及社交网络、电子邮件、移动电话、数码照相机和在线信用卡交易等多种方式，每个人的日常生活都在被“数字化”。数字世界的规模在 2006—2011 年五年间约膨胀了 10 倍，如图 1-3 所示。

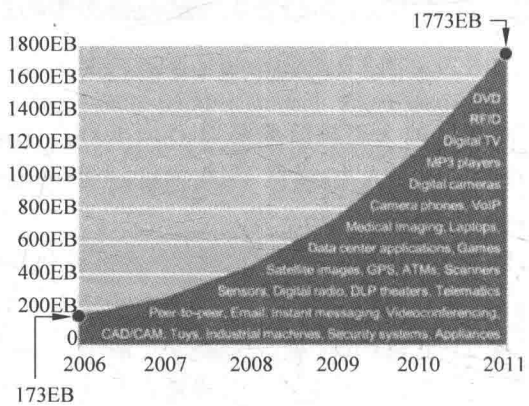


图 1-3　2006—2011 年全球数字信息的增长

大数据快速增长的原因之一是智能设备的普及，如传感器、医疗设备及智能建筑(如楼宇和桥梁)。此外，非结构化信息，如文件、电子邮件和视频，将占到未来 10 年新生数据的 90%。非结构化信息增长的另一个原因是由于高宽带数据的增长，如视频。

用户手中的手机和移动设备是数据量爆炸的一个重要原因。目前，全球手机用户共拥有 50 亿台手机，其中 20 亿台为智能手机，相当于 20 世纪 80 年代 20 亿台 IBM 的大型机在消费者手里。

大数据正在以不可阻拦的磅礴气势，与当代同样具有革命意义的最新科技进步(如虚拟

现实技术、增强现实技术、纳米技术、生物工程、移动平台应用等)一起,揭开人类新世纪的序幕。

对于地球上每一个普通居民而言,大数据有什么应用价值呢?只要看看周围正在变化的一切,你就可以知道,大数据对每个人的重要性不亚于人类初期对火的使用。大数据让人类对一切事物的认识回归本源,其通过影响经济生活、政治博弈、社会管理、文化教育科研、医疗、保健、休闲等行业,与每个人产生密切的联系。

大数据时代已悄然来到我们身边,并渗透到我们每个人的日常生活之中,谁都无法回避。它提供了光怪陆离的全媒体,难以琢磨的云计算,无法抵御的虚拟仿真环境和随处可在的网络服务。随着互联网技术的蓬勃发展,我们一定会迎来大数据的智能时代,即大数据技术和生活紧密相连,它再也不仅仅是人们津津乐道的一种时尚,而是成为生活上的向导和助手。中国大数据市场的应用展望如图 1-4 所示。

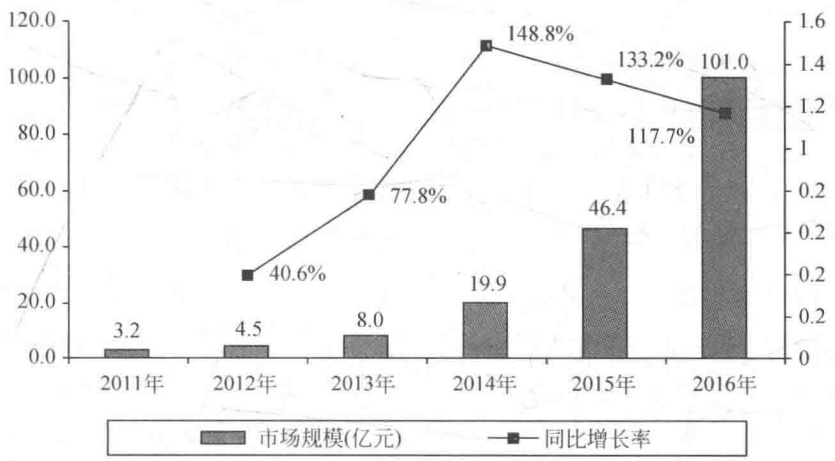


图 1-4 中国大数据市场的应用展望

1.1.3 大数据的来源

大数据的来源非常多,如信息管理系统、网络信息系统、物联网系统、科学实验系统等,其数据类型包括结构化数据、半结构化数据和非结构化数据。

(1) 信息管理系统:企业内部使用的信息系统,包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据,其产生的大数据大多数为结构化数据,通常存储在数据库中。

(2) 网络信息系统:基于网络运行的信息系统即网络信息系统是大数据产生的重要方式,如电子商务系统、社交网络、社会媒体、搜索引擎等都是常见的网络信息系统。网络信息系统产生的大数据多为半结构化或非结构化的数据,在本质上,网络信息系统是信息管理系统的延伸,是专属于某个领域的应用,具备某个特定的目的。因此,网络信息系统有着更独特的应用。

(3) 物联网系统:物联网是新一代信息技术,其核心和基础仍然是互联网,是在互联网基础上延伸和扩展的网络,其用户端延伸和扩展到了任何物品与物品之间,进行信息交换和通信,而其具体实现是通过传感技术获取外界的物理、化学、生物等数据信息。

(4) 科学实验系统：主要用于科学技术研究，可以由真实的实验产生数据，也可以通过模拟方式获取仿真数据。

1.1.4 大数据产生的三个发展阶段

从数据库技术诞生以来，产生大数据的方式主要经过了三个发展阶段。

1. 被动式生成数据

数据库技术使得数据的保存和管理变得简单，业务系统在运行时产生的数据可以直接保存到数据库中，由于数据是随业务系统运行而产生的，因此该阶段所产生的数据是被动的。

2. 主动式生成数据

物联网的诞生使得移动互联网的发展大大加速了数据的产生几率，例如人们可以通过手机等移动终端随时随地产生数据。用户数据不但大量增加，同时用户还主动提交了自己的行为，使之进入了社交、移动时代。大量移动终端设备的出现，使用户不仅主动提交自己的行为，还和自己的社交圈进行了实时互动，因此数据大量产生出来，且具有极其强烈的传播性。显然如此生成的数据是主动的。

3. 感知式生成数据

物联网的发展使得数据生成方式得以彻底地改变。例如遍布在城市各个角落的摄像头等数据采集设备源源不断地自动采集并生成数据。

1.1.5 大数据的特点

在大数据背景下，数据的采集、分析、处理较之传统方式有了颠覆性的改变，如表 1-1 所示。

表 1-1 传统数据与大数据的特点比较

	传统数据	大数据
数据产生方式	被动采集数据	主动生成数据
数据采集密度	采样密度较低，采样数据有限	利用大数据平台，可对需要分析事件的数据进行密度采样，精确获取事件全局数据
数据源	数据源获取较为孤立，不同数据之间添加的数据整合难度较大	利用大数据技术，通过分布式技术、分布式文件系统、分布式数据库等技术对多个数据源获取的数据进行整合处理
数据处理方式	大多采用离线处理方式，对生成的数据集中分析处理，不对实时产生的数据进行分析	较大的数据源、响应时间要求低的应用可以采取批处理方式集中计算；响应时间要求高的实时数据处理采用流处理的方式进行实时计算，并通过对历史数据的分析进行预测分析

1.1.6 大数据处理流程

大数据的处理流程可以定义为在适合工具的辅助下，对不同结构的数据源进行抽取和集成，结果按照一定的标准统一存储，利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展示给终端用户。大数据处理的基本流程如图 1-5 所示。

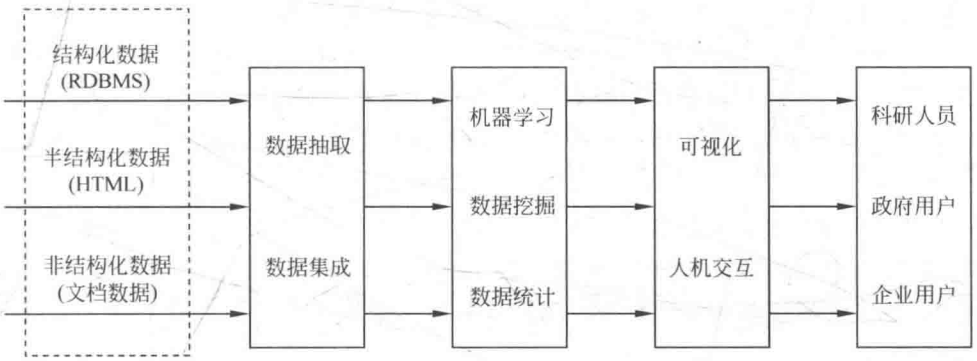


图 1-5 大数据处理的基本流程

1. 数据抽取与集成

由于大数据处理的数据来源类型广泛，而其第一步是对数据进行抽取和集成，从中找出关系和实体，经过关联、聚合等操作，再按照统一的格式对数据进行存储，现有的数据抽取和集成引擎有三种：基于物化或 ETL 方法的引擎、基于中间件的引擎、基于数据流方法的引擎。

2. 大数据分析

大数据分析是指对规模巨大的数据进行分析。大数据分析是大数据处理流程的核心步骤。通过抽取和集成环节，从不同结构的数据源中获得用于大数据处理的原始数据，用户根据需求对数据进行分析处理，如数据挖掘、机器学习、数据统计，数据分析可以用于决策支持、商业智能、推荐系统、预测系统等。

3. 数据可视化

用户最关心的是数据处理的结果及以何种方式在终端上显示结果，因此采用什么方式展示处理结果非常重要。就目前来看，可视化和人机交互是数据解释的主要技术。

数据可视化主要是借助于图形化手段，清晰有效地传达与沟通信息。数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元元素表示，大量的数据集合构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析。而使用可视化技术可以将处理结果通过图形方式直观地呈现给用户，如标签云、历史流、空间信息等；人机交互技术可以引导用户对数据进行逐步分析，参与并理解数据分析结果。