《大数据导论（通识课版）》

教材官网： http://dblab.xmu.edu.cn/post/bigdataintroduction/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第3章 大数据技术

（PPT版本号：2019年秋季学期）

扫一扫访问教材官网

林子雨

厦门大学计算机科学系

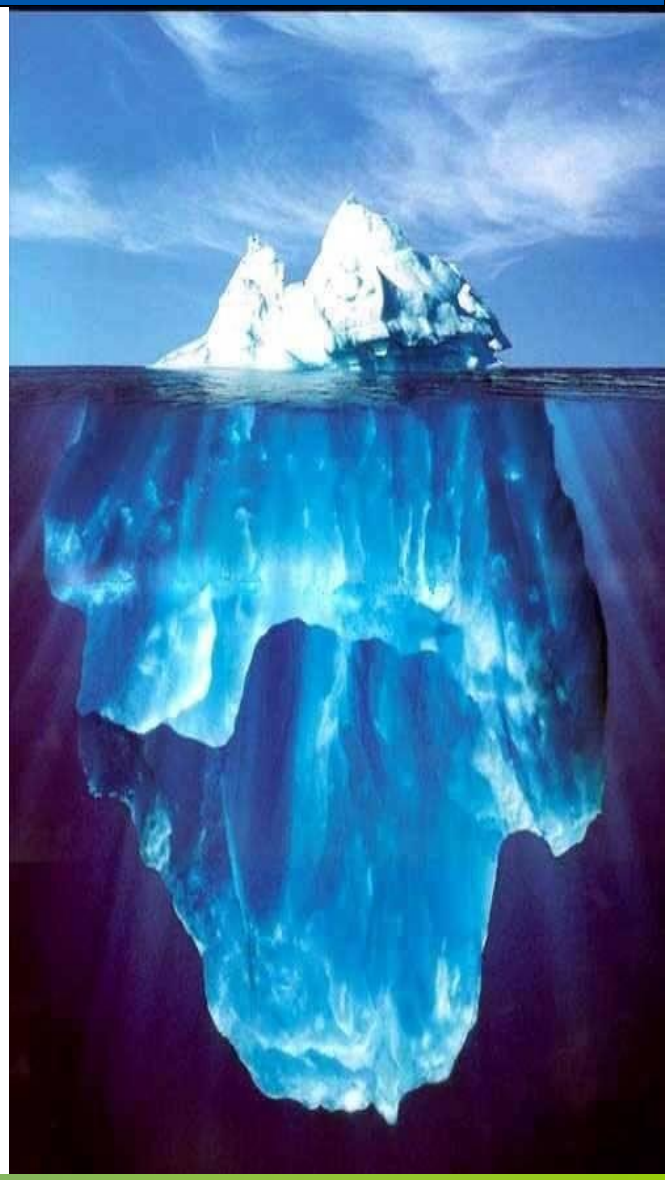
E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

课程教材

- 林子雨 编著《大数据导论——数据思维、数据能力和数据伦理（通识课版）》
- 高等教育出版社，2019年11月

提纲



高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

3.1概述

表　　大数据技术的不同层面及其功能

技术层面	功能
数据采集与预处理	利用ETL工具将分布的、异构数据源中的数据，如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；利用日志采集工具（如Flume、Kafka等）把实时采集的数据作为流计算系统的输入，进行实时处理分析；利用网页爬虫程序到互联网网站中爬取数据
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析
数据可视化	对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据安全和隐私保护	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

3.2数据采集与预处理

3.2.1数据采集的概念
3.2.2数据采集的三大要点
3.2.3数据采集的数据源
3.2.4数据清洗

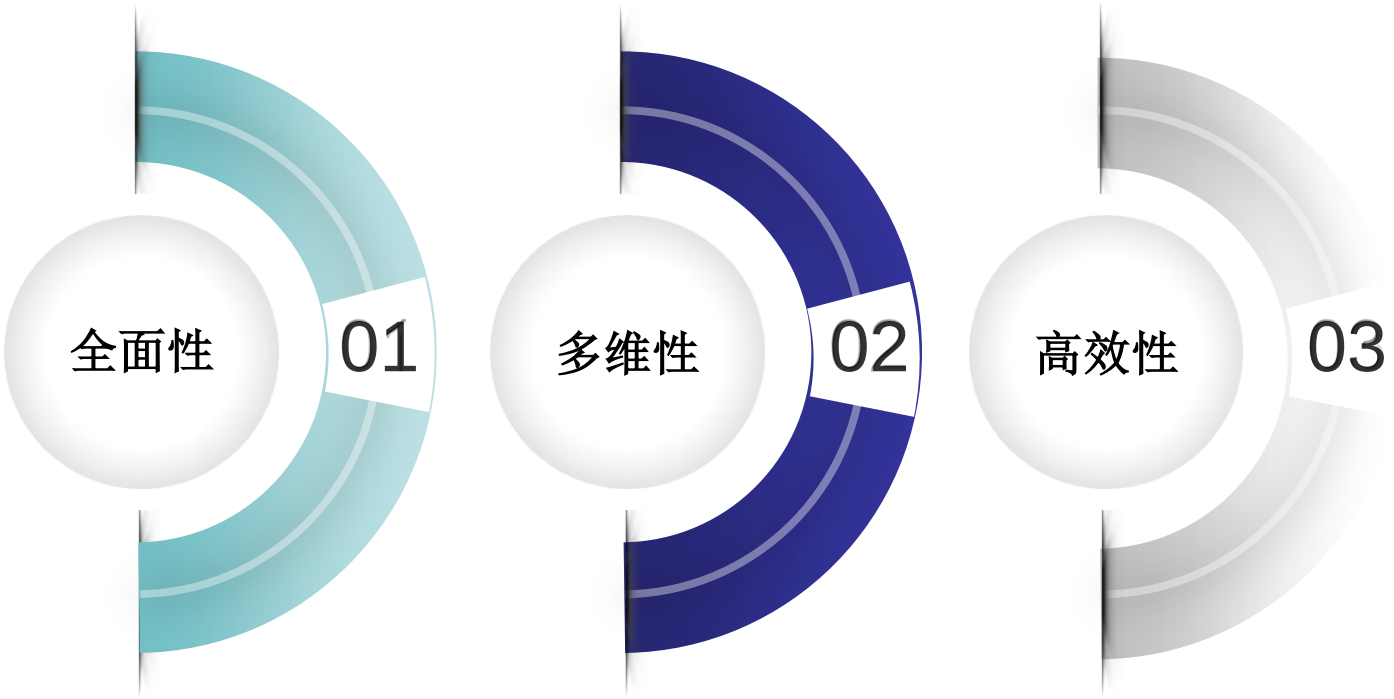
3.2.1数据采集的概念

数据采集，又称“数据获取”，是数据分析的入口，也是数据分析过程中相当重要的一个环节，它通过各种技术手段把外部各种数据源产生的数据实时或非实时地采集并加以利用。

表 传统的数据采集与大数据采集区别

	传统的数据采集	大数据采集
数据源	来源单一，数据量相对较少	来源广泛，数据量巨大
数据类型	结构单一	数据类型丰富，包括结构化、半结构化和非结构化
数据存储	关系数据库和并行数据仓库	分布式数据库，分布式文件系统

3.2.2数据采集的三大要点



3.2.3数据采集的数据源

传感器

互联网数据

企业业务
系统数据

日志文件

3.2.3数据采集的数据源

企业可以借助于ETL（Extract-Transform-Load）工具，把分散在企业不同位置的业务系统的数据，抽取、转换、加载到企业数据仓库中，以供后续的商务智能分析使用

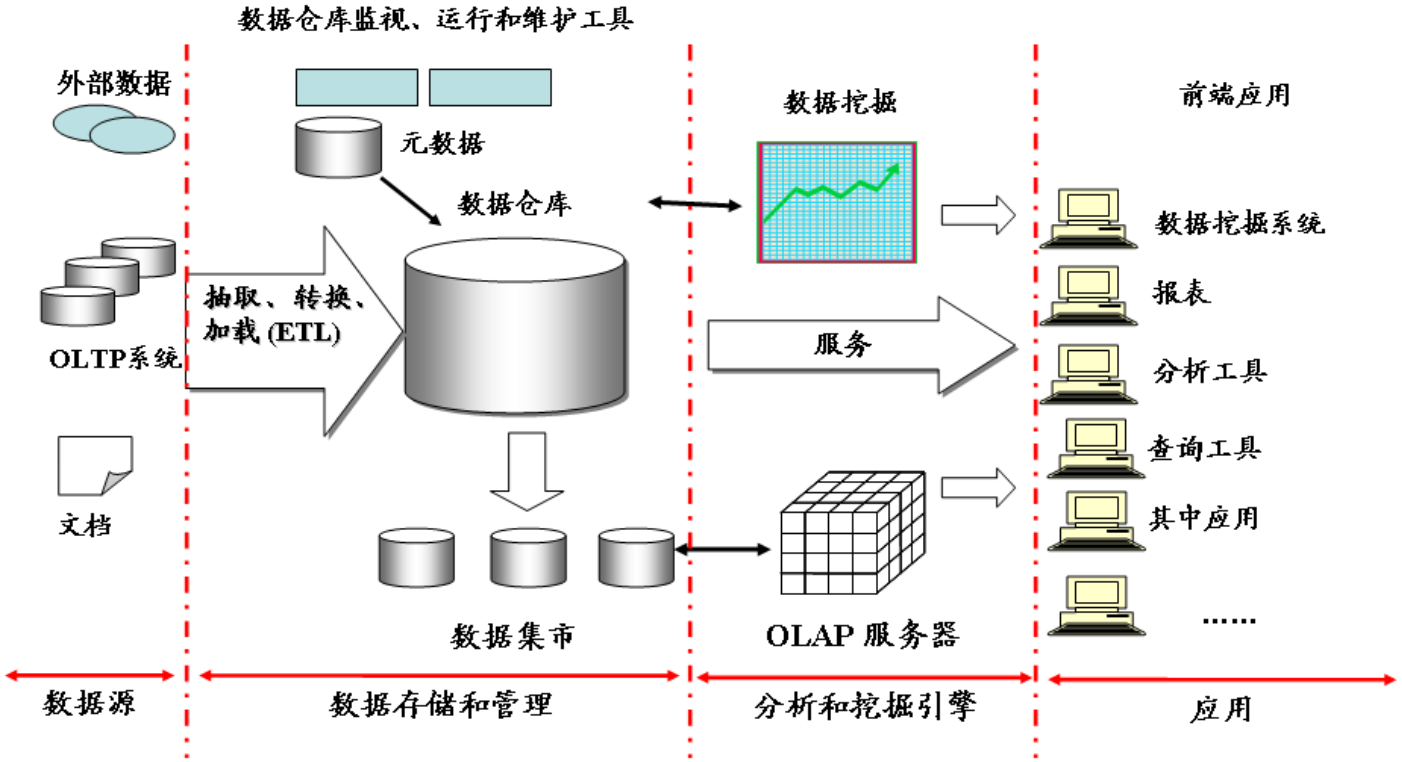


图 数据仓库体系架构

3.2.4数据清洗

1.需要清洗的数据的主要类型

- 残缺数据
- 错误数据
- 重复数据

2.数据清洗的内容

- 一致性检查
- 无效值和缺失值的处理
- 估算
- 整例删除
- 变量删除
- 成对删除

3.3数据存储和管理

3.3.1 传统的数据存储和管理技术
3.3.2 大数据时代的数据存储和管理技术

3.3.1 传统的数据存储和管理技术

1.文件系统

•文件系统是操作系统用于明确存储设备（常见的是磁盘，也有基于NAND Flash的固态硬盘）或分区上的文件的方法和数据结构，即在存储设备上组织文件的方法。操作系统中负责管理和存储文件信息的软件机构称为文件管理系统，简称“文件系统”

•我们平时在计算机上使用的WORD文件、PPT文件、文本文件、音频文件、视频文件等，都是由操作系统中的文件系统进行统一管理的

3.3.1 传统的数据存储和管理技术

2.关系数据库

- 一个关系数据库可以看成是许多关系表的集合，每个关系表可以看成一张二维表格
- 目前市场上常见的关系数据库产品包括Oracle、SQL Server、MySQL、DB2等

表 学生信息表

学号	姓名	性别	年龄	考试成绩
95001	张三	男	21	88
95002	李四	男	22	95
95003	王梅	女	22	73
95004	林莉	女	21	96

3.3.1 传统的数据存储和管理技术

3.数据仓库

数据仓库（Data Warehouse）是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策

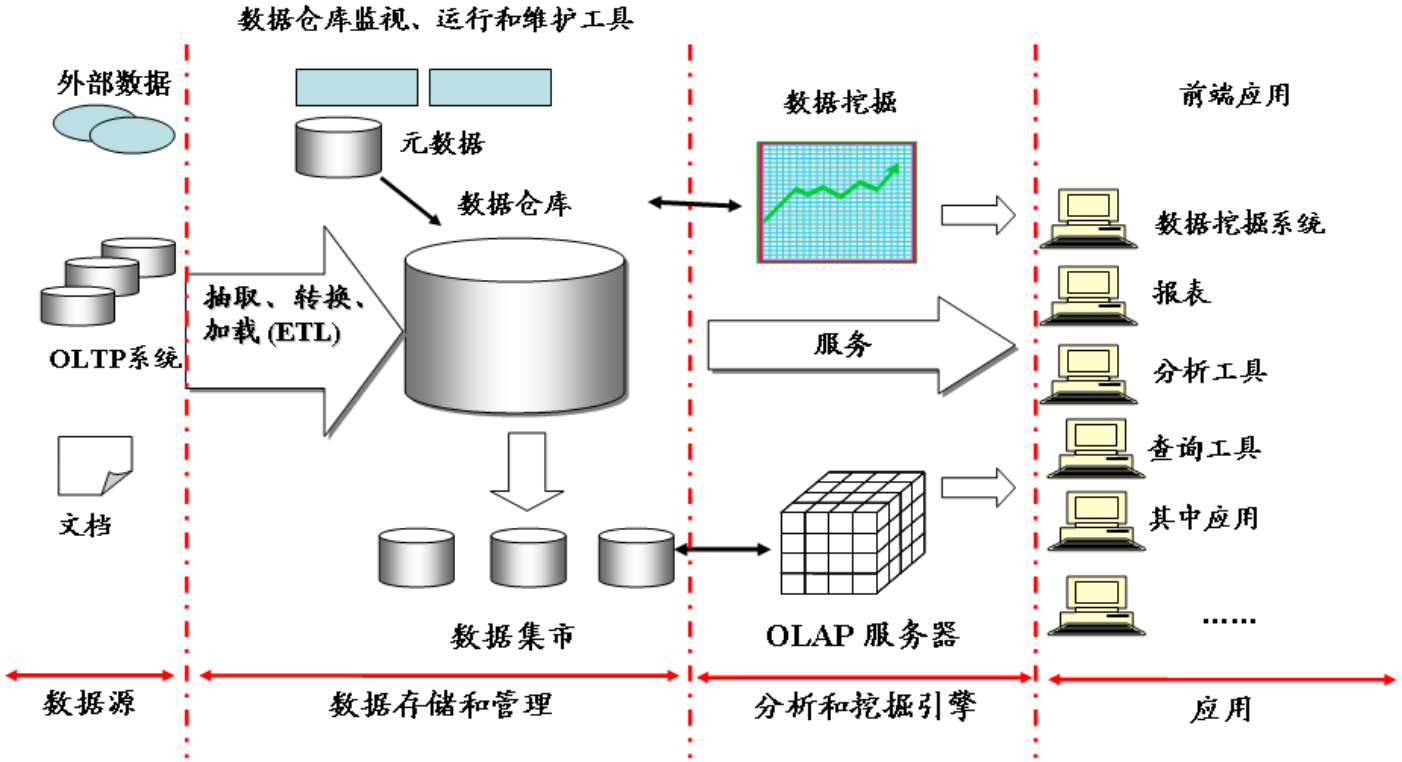


图 数据仓库体系架构

3.3.1 传统的数据存储和管理技术

4.并行数据库

- 并行数据库是指那些在无共享的体系结构中进行数据操作的数据库系统
- 这些系统大部分采用了关系数据模型并且支持SQL语句查询，但为了能够并行执行SQL的查询操作，系统中采用了两个关键技术：关系表的水平划分和SQL查询的分区执行
- 并行数据库系统的目标是高性能和高可用性,通过多个节点并行执行数据库任务，提高整个数据库系统的性能和可用性

3.3.2 大数据时代的数据存储和管理技术

1.分布式文件系统

分布式文件系统（Distributed File System）是一种通过网络实现文件在多台主机上进行分布式存储的文件系统

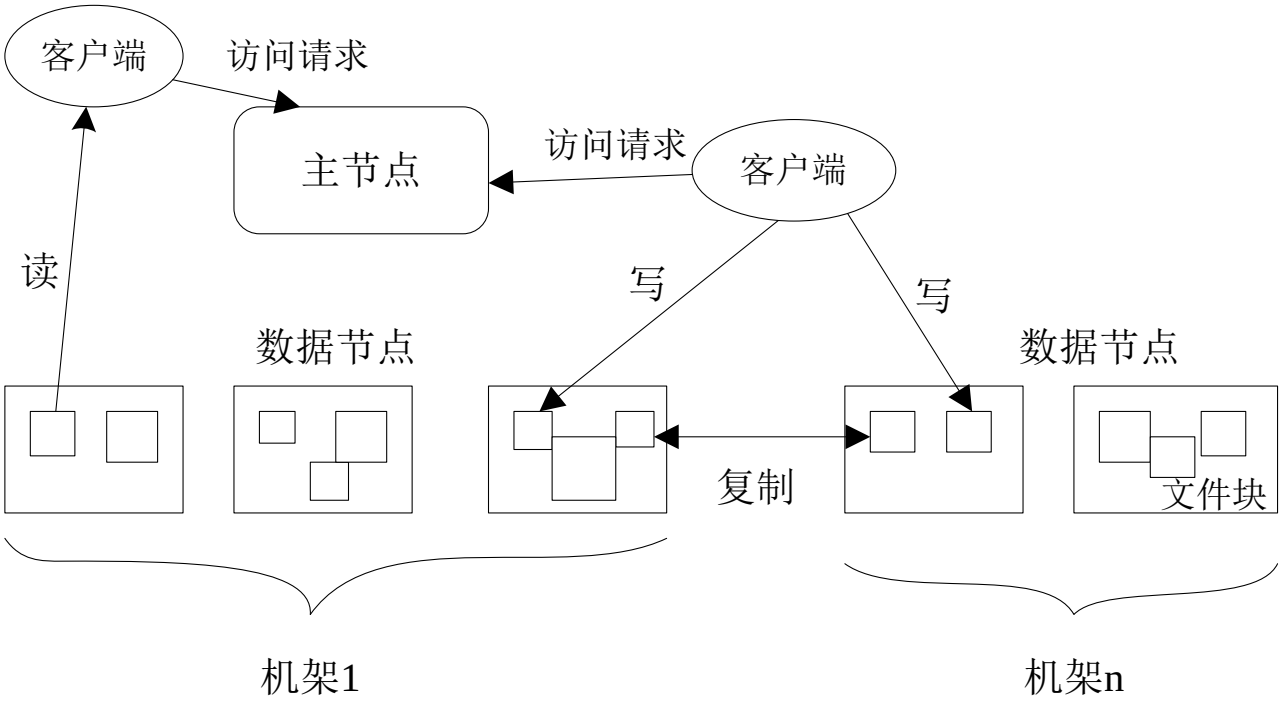


图 分布式文件系统的整体结构

3.3.2 大数据时代的数据存储和管理技术

2.NewSQL和NoSQL数据库

（1）NewSQL数据库

•NewSQL是对各种新的可扩展、高性能数据库的简称，这类数据库不仅具有对海量数据的存储管理能力，还保持了传统数据库支持ACID和SQL等特性

•目前具有代表性的NewSQL数据库主要包括Spanner、Clustrix、GenieDB、ScalArc、Schooner、VoltDB、RethinkDB、ScaleDB、Akiban、CodeFutures、ScaleBase、Translattice、NimbusDB、Drizzle、Tokutek、JustOne DB等

3.3.2 大数据时代的数据存储和管理技术

2.NewSQL和NoSQL数据库

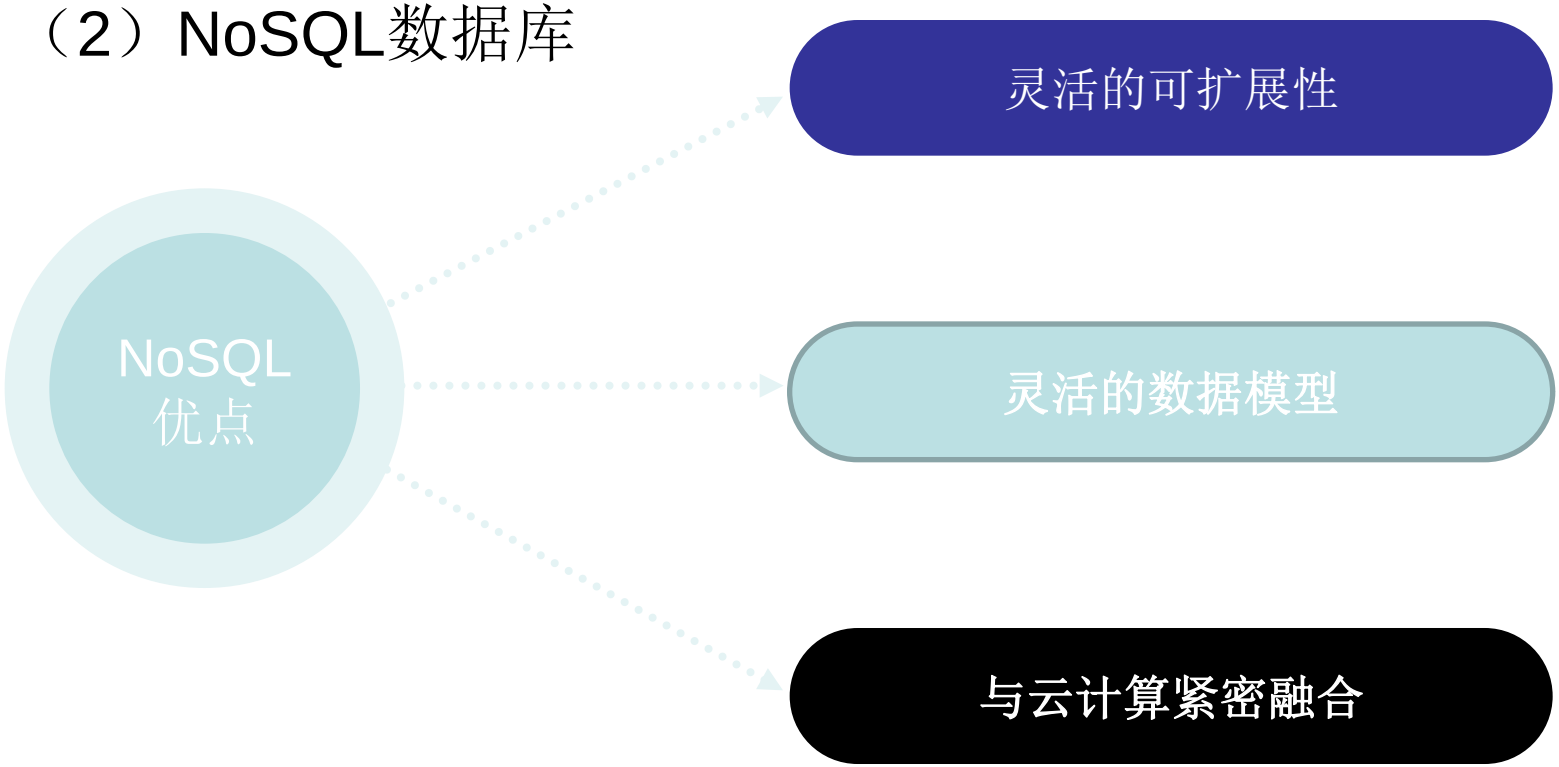
（2）NoSQL数据库

- NoSQL是一种不同于关系数据库的数据库管理系统设计方式，是对非关系型数据库的统称，它所采用的数据模型并非传统关系数据库的关系模型，而是类似键/值、列族、文档等非关系模型
- NoSQL数据库没有固定的表结构，通常也不存在连接操作，也没有严格遵守ACID约束，因此，与关系数据库相比，NoSQL具有灵活的水平可扩展性，可以支持海量数据存储

3.3.2 大数据时代的数据存储和管理技术

2.NewSQL和NoSQL数据库

（2）NoSQL数据库



3.3.2 大数据时代的数据存储和管理技术

2.NewSQL和NoSQL数据库

（3）大数据引发数据库架构变革

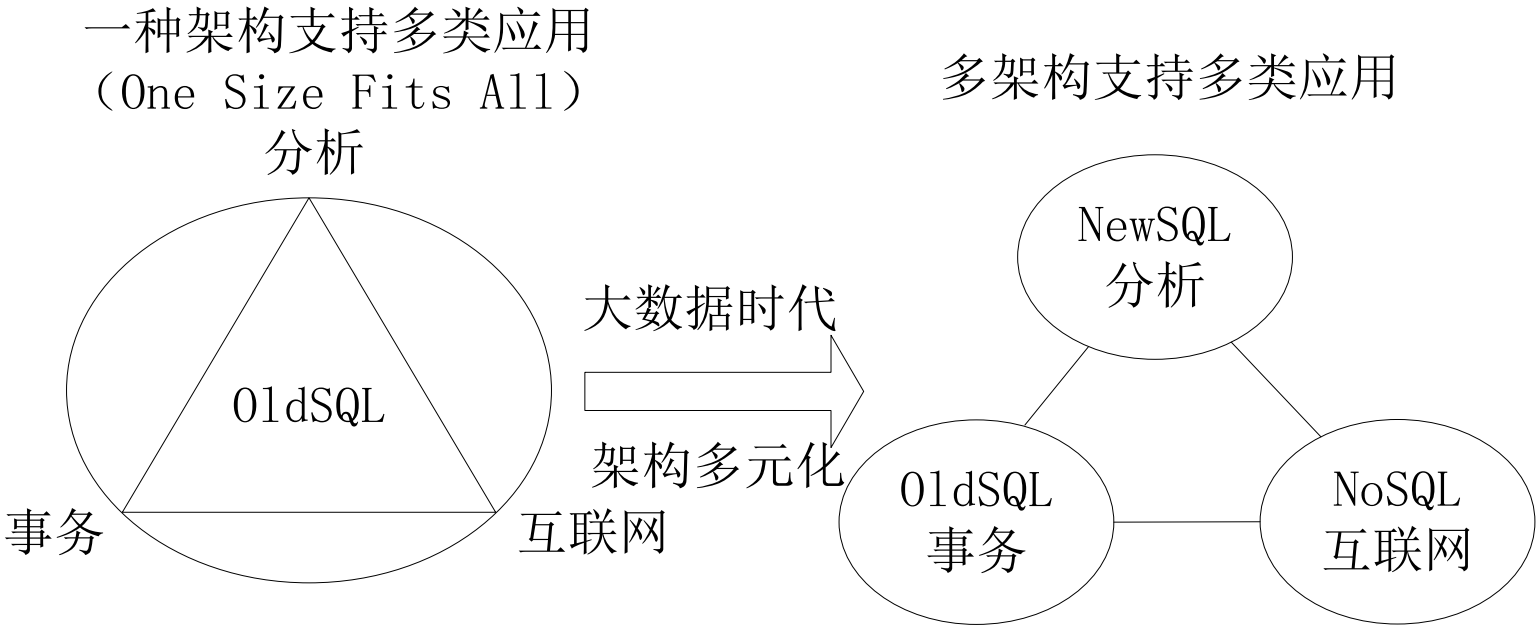


图 大数据引发数据处理架构变革

3.4数据处理与分析

3.4.1 数据挖掘和机器学习算法
3.4.2 大数据处理与分析技术

3.4.1 数据挖掘和机器学习算法

•数据挖掘和机器学习是计算机学科中最活跃的研究分支之一。机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科，专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。

•数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘可以视为机器学习与数据库的交叉，它主要利用机器学习界提供的算法来分析海量数据，利用数据库界提供的存储技术来管理海量数据。

3.4.1 数据挖掘和机器学习算法

典型的机器学习和数据挖掘算法

- 分类
- 聚类
- 回归分析
- 关联规则

3.4.2 大数据处理与分析技术

大数据处理分析技术类型及其代表产品

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

3.5数据可视化

3.5.1 什么是数据可视化
3.5.2数据可视化的重要作用
3.5.3数据可视化案例

3.5.1 什么是数据可视化

- 数据可视化是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程
- 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素表示，大量的数据集构成数据图像，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

3.5.2数据可视化的重要作用

在大数据时代，可视化技术可以支持实现多种不同的目标：

（1）观测、跟踪数据

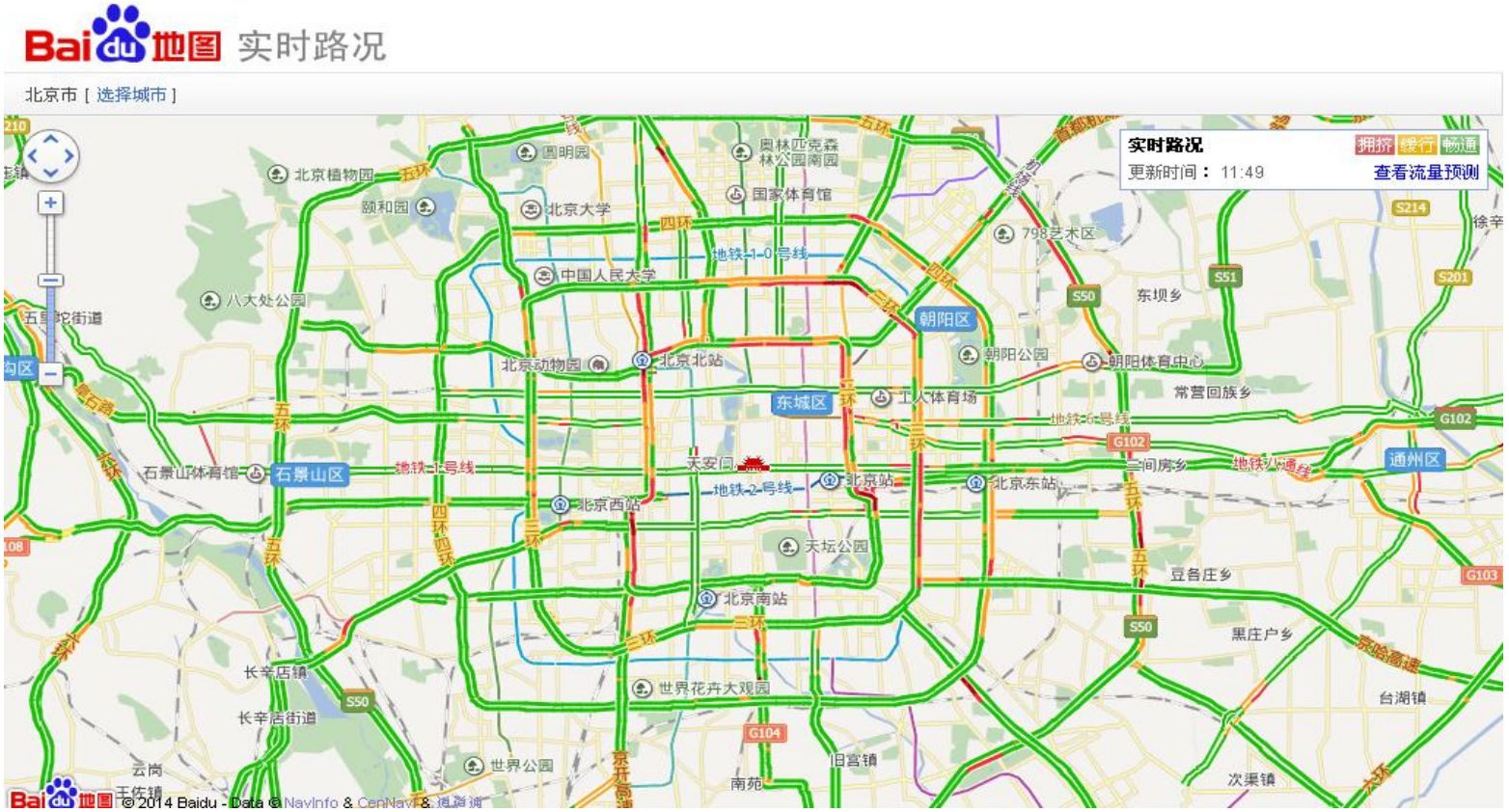


图 百度地图显示的北京市实时交通路况信息

3.5.2数据可视化的重要作用

（2）分析数据

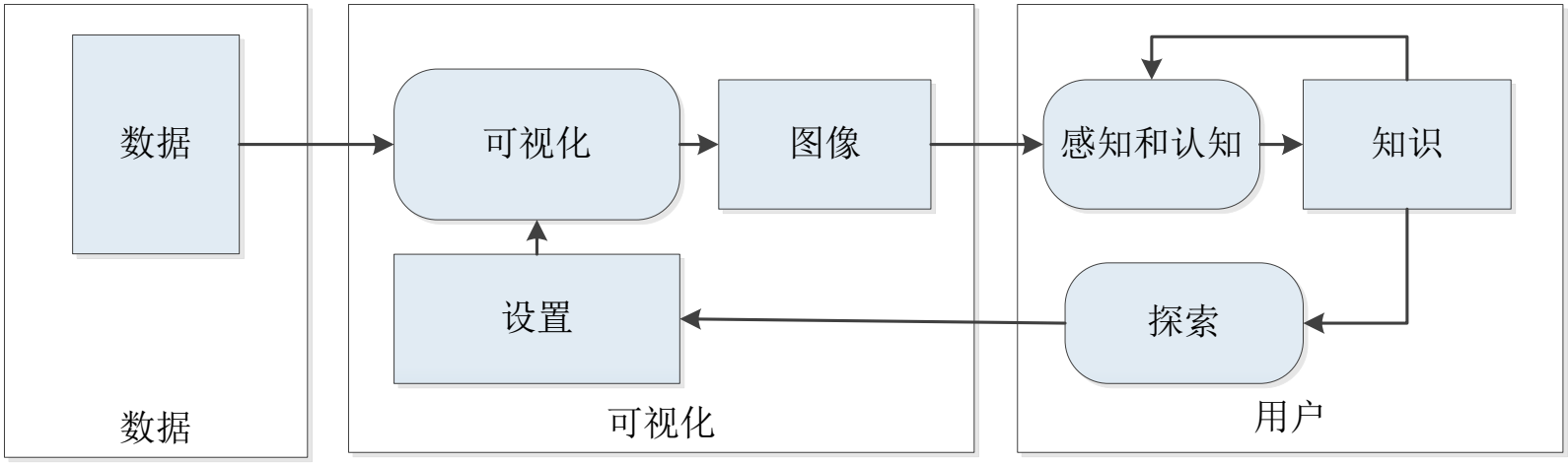


图 用户参与的可视化分析过程

3.5.2数据可视化的重要作用

（3）辅助理解数据

图 微软“人立方”展示的人物关系图

3.5.2数据可视化的重要作用

（4）增强数据吸引力

图 一个可视化的图表新闻实例

3.5.3数据可视化案例

1.全球黑客活动
2.互联网地图
3.编程语言之间的影响力关系图
4.世界国家健康与财富之间的关系

3.5.3数据可视化案例

1.全球黑客活动

安全供应商Norse打造了一张能够反映全球范围内黑客攻击频率的地图（http://map.ipviking.com），它利用Norse 的“蜜罐”攻击陷阱显示出所有实时渗透攻击活动。如图10-11所示，地图中的每一条线代表的都是一次攻击活动，借此可以了解每一天、每一分钟甚至每一秒世界上发生了多少次恶意渗透。

图 一张能够反映全球范围内黑客攻击频率的地图

3.5.3数据可视化案例

2.互联网地图

为了探究互联网这个庞大的宇宙，俄罗斯工程师 Ruslan Enikeev 根据 2011 年底的数据，将全球 196 个国家的 35 万个网站数据整合起来，并根据 200 多万个网站链接将这些“星球”通过关系链联系起来，每一个“星球”的大小根据其网站流量来决定，而“星球”之间的距离远近则根据链接出现的频率、强度和用户跳转时创建的链接来确定，由此绘制得到了“互联网地图”（http://internet-map.net）。

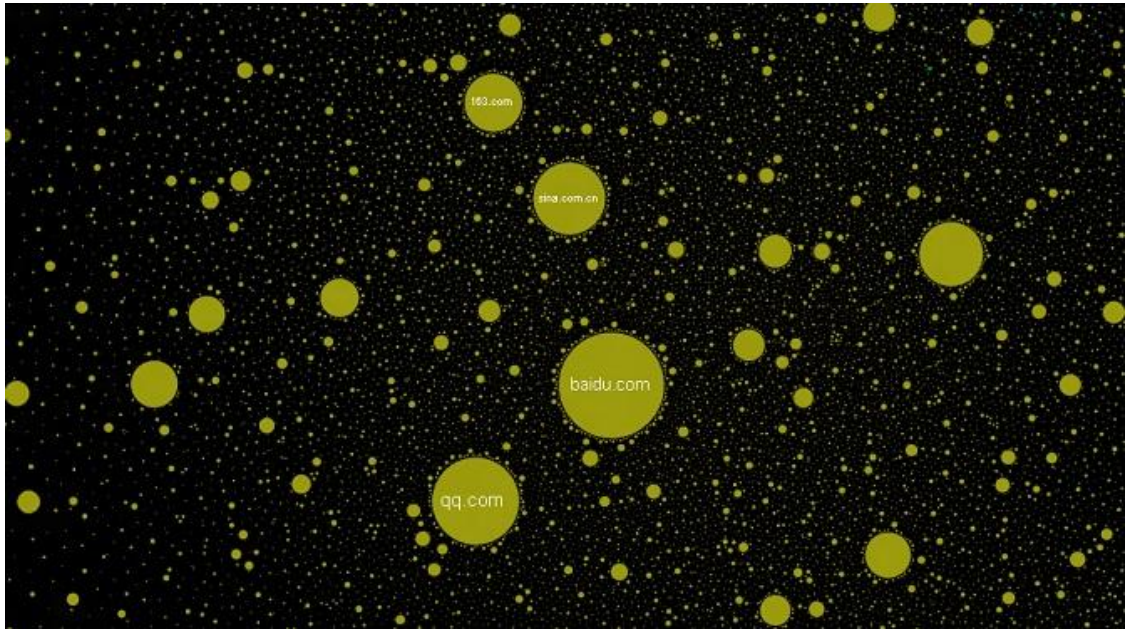


图　俄罗斯工程师绘制的“互联网地图”

3.5.3数据可视化案例

3.编程语言之间的影响力关系图

Ramio Gómez利用来自Freebase上的编程语言维护表里的数据，绘制了编程语言之间的影响力关系图，图中的每个节点代表一种编程语言，之间的连线代表该编程语言对其他语言有影响，有影响力的语言会连线多个语言，相应的节点也会越大。

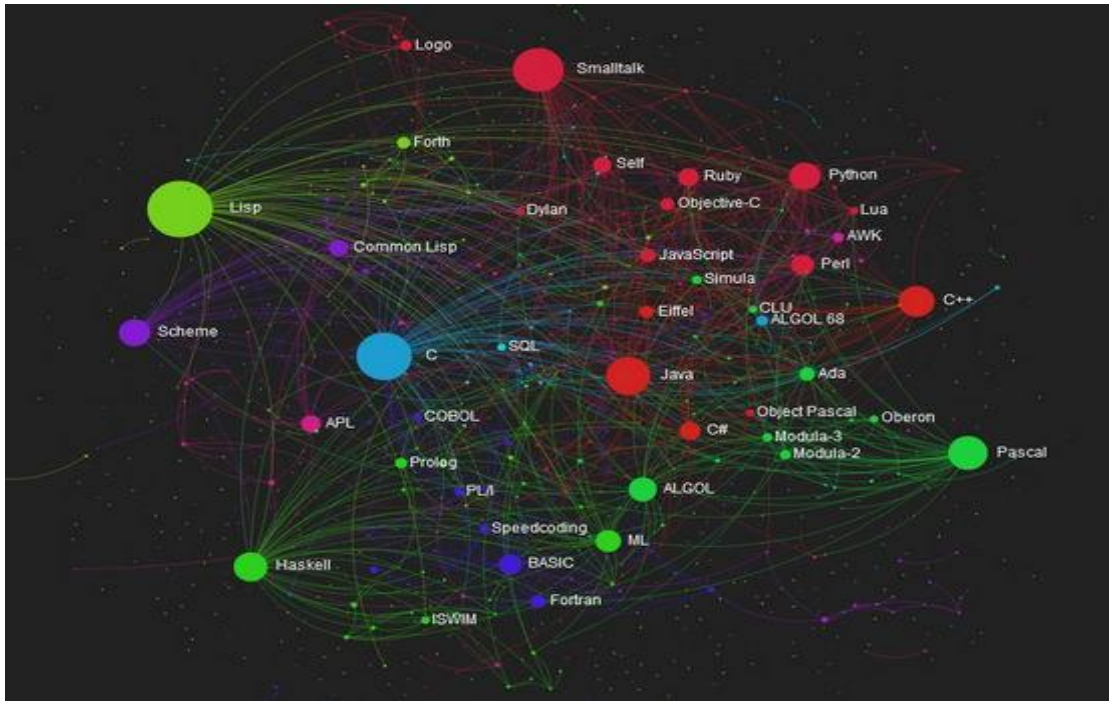


图　编程语言之间的影响力关系图

3.5.3数据可视化案例

4.世界国家健康与财富之间的关系

“世界国家健康与财富之间的关系”利用可视化技术，把世界上200个国家，从1810年到2010年历时200年其各国国民的健康、财富变化数据（收集了1千多万个数据）制作成三维动画进行了直观展示（http://www.moojnn.com/Index/whn）。

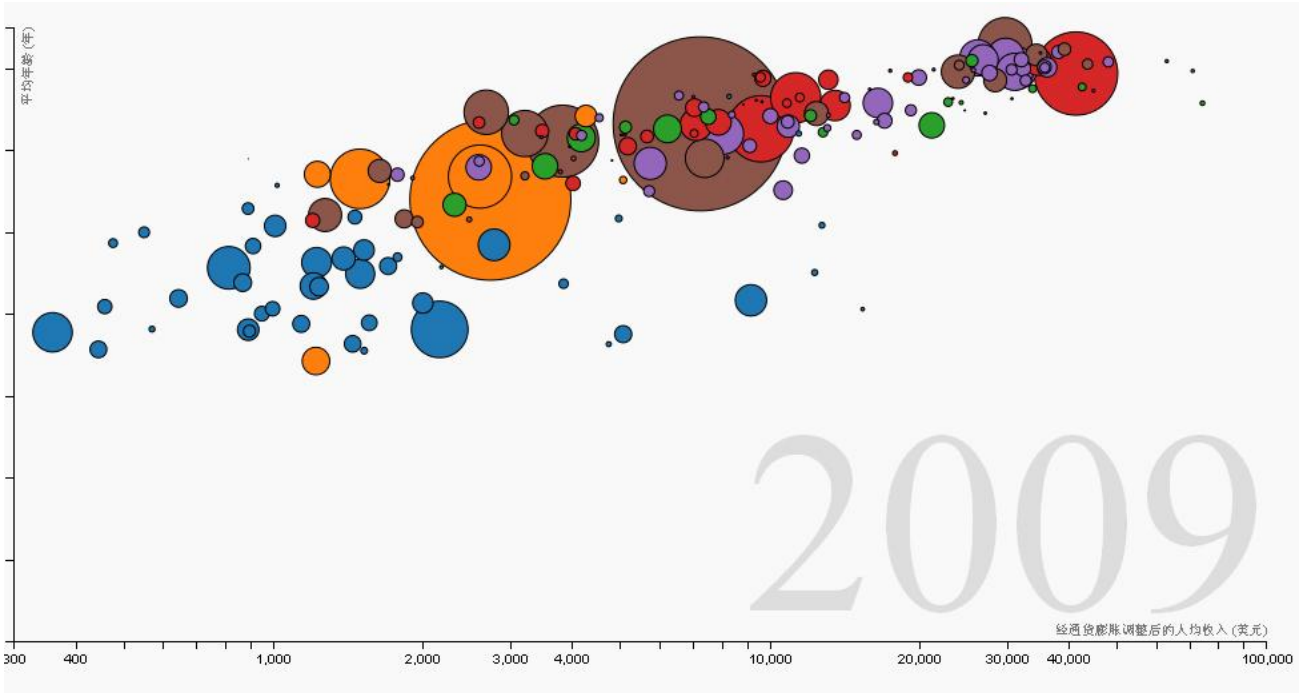


图　世界国家健康与财富之间的关系图

3.6数据安全和隐私保护

3.6.1 数据安全技术
3.6.2 隐私保护技术

3.6.1 数据安全技术

3.6.2 隐私保护技术

- 主要可以借助数据水印的合理性应用，明确用户数据使用的实际需要，并且能够将用户的身份信息加以识别，在不影响用户正常使用数据的前提之下，对数据载体使用检测的方法实现融入，数据水印技术的合理应用能够充分保护原创

- 用户隐私保护的渠道更加众多，同时能够贯穿于数据产生的全过程，主要是针对生产、收购以及加工存储的各项环节，同时能够在数据运输当中实现隐私安全保护体系的构建，在数据的整个生命周期当中，实现对用户信息的保护

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度和2017年度厦门大学教学类奖教金获得者，荣获2017年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过500万字高价值的研究和教学资料，累计网络访问量超过500万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过100万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：《大数据技术原理与应用》教材

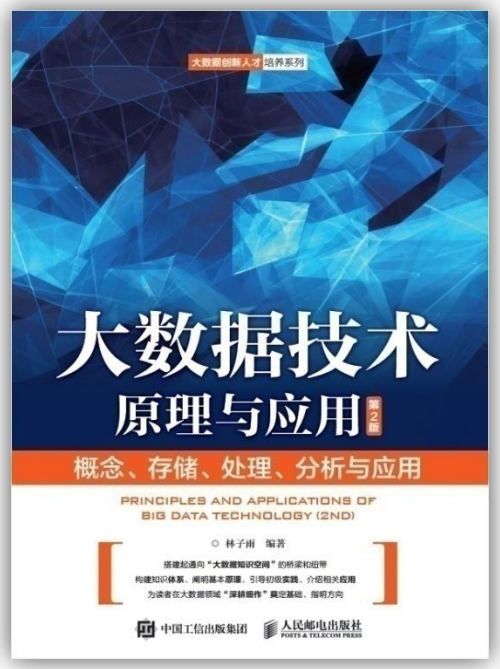
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-44330-4　定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录D：《大数据基础编程、实验和案例教程》

本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

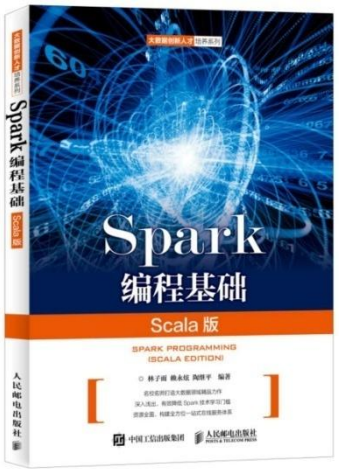
厦门大学林子雨编著

配套实验指导书

清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

附录E：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录F：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

附录G：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《基于协同过滤算法的电影推荐》
《电信用户行为分析》
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》

系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/

扫一扫访问大数据实训课程系列案例教材主页

Department of Computer Science, Xiamen University, 2019