大数据分类模型和算法研究

刘宝锺 ◎ 著

云南大学出版社

大数据分类模型和算法研究

刘宝锺 ◎著

云南大学出版社
YUNNAN UNIVERSITY PRESS

图书在版编目（CIP）数据

大数据分类模型和算法研究 / 刘宝锺著 . -- 昆明 : 云南大学出版社 , 2019

ISBN 978-7-5482-3613-9

Ⅰ . ①大… Ⅱ . ①刘… Ⅲ . ①数据处理 Ⅳ . ① TP274

中国版本图书馆 CIP 数据核字（2019）第 008888 号

策划编辑： 王翌沣
责任编辑： 王翌沣
封面设计： 黄伟娟

大数据分类模型和算法研究

刘宝锺 著

出版发行： 云南大学出版社
印　　装： 昆明理煜印务有限公司
开　　本： 787mm × 1092mm　1/16
印　　张： 29
字　　数： 535 千字
版　　次： 2020 年 1 月第 1 版
印　　次： 2020 年 1 月第 1 次印刷
书　　号： 978-7-5482-3613-9
定　　价： 120.00 元
社　　址： 昆明市一二一大街 182 号
（云南大学东陆校区英华园内）
邮　　编： 650091
电　　话： （0871）65033244　65031071
E-mail： market@ynup.com

若发现本书有印装质量问题，请与印厂联系调换，联系电话：0871-64167045。

/前　言/

新时代，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。在以云计算为代表的技术创新“大幕”的衬托下，原本看起来很难收集和使用的数据开始被利用起来了，通过各行各业的不断创新，逐步为人类创造更多的价值。作为继云计算、物联网之后 IT 行业又一颠覆性的技术，大数据备受人们关注。大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的各行各业，都存在着大数据的印迹。

海量的数据在日积月累中不断地爆发式增长，为了探求如何在大数据中获得更多的价值，对海量数据的处理和分析的需求迫在眉睫。大数据的主要特点有海量（volume）、高速（velocity）、准确（veracity）、多样（variety）等，在大数据技术发展的起步阶段，国内外研究的主要侧重点是处理海量数据和处理多样的数据类型。然而，在当前互联网时代下的大数据大多存在于金融股票、运营商网络流量、网站实时请求、交通数据流等业务中，数据的形式大多是以高速的流式数据形态传递。与存储在传统数据库中的静态数据不同，流式数据作为一种新的数据形态，对数据分析过程的高速性和准确性的要求更加严格。对于流式数据的分析处理需要我们能够快速地记录实时数据流信息，并更加准确地保证信息的时效性。

本书强调了大数据的宝贵价值，论述了常用的数据分析技术与方法，在此基础上设计对应的大数据分类模型（线性分类模型和分类分析模型）；阐述了人工神经网络的相关理论，涉及的具体的大数据算法包括关联规则分析算法、分布式算法、聚类算法等，并对大数据分析算法的并行化进行了相关研究；阐述了各个算法的应用场景及算法复杂度，从应用的角度提供了大量实例，使读者能够快速、高效进阶各类算法，并能够将之熟练应用到将来的工作实践中。

本书部分内容参考和借鉴了国内外学者的一些相关理论研究成果，并引用了互联网中的相关理论，在这里对他们一并表示衷心感谢！作者在撰写过程中，虽极力丰富本书内容，力求著作的完美无瑕、但仍难免存在疏漏和错误之处，还望各位同仁斧正。

作　者

2018 年 6 月

目录
CONTENTS

第1章 绪论

随着科学、技术和工程的迅猛发展，近20年来，许多领域（如光学观测、光学监控、健康医护、传感器、用户数据、互联网和金融公司以及供应链系统）都产生了海量的数据，大数据的概念也随之被再次重视。与传统的数据相比，除了大容量等表象特点，大数据还具有其他独特的特点，例如，大数据通常是无结构的，并且需要得到实时分析，因此大数据的发展需要全新的体系架构，用于处理大规模数据的获取、传输、存储和分析。

1.1 大数据的概念和特征

1.1.1 大数据的概念

“大数据”的概念起源于2008年9月《自然》（*Nature*）杂志刊登的名为“BigData”的专题。2011年《科学》（*Science*）杂志也推出专刊“Dealing With Data”对大数据的计算问题进行了讨论。谷歌、雅虎、亚马逊等著名企业在此基础上，总结了它们利用积累的海量数据为用户提供更加人性化服务的方法，进一步完善了“大数据”的概念。

根据维基百科的定义，大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中，大数据指的是不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。

“大数据”研究机构Gartner将“大数据”定义为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。

1.1.2 大数据的特征

大数据是相对于一般数据而言的，目前对大数据尚缺乏权威的严格定义，通常大家用“4V”来反映大数据的特征。

1.1.2.1 Volume（规模性）

大数据之“大”，体现在数据的存储和计算均需要耗费海量规模的资源上。规模大是大数据最重要的标志之一，事实上，数据只要有足够的规模就可以称为大数据。数据的规模越大，通常对数据挖掘所得到的事物演变规律越可信，数据的分析结果也越具有代表性。如美国宇航局收集和处理的气候观察、模拟数据达到 32PB；而 FICO 的信用卡欺诈检测系统要监测全世界超过 18 亿个活跃信用卡账户。不过，现在也有学者认为，社会对大数据的关注，应更多地引导到对数据资源获得与利用的重视上来，因为对于某些中小数据的挖掘也有价值，目前报道的一些大数据挖掘的应用例子，不少只是 TB 级的规模。

1.1.2.2 Velocity（高速性）

大数据的另一特点在于数据增长速度快，亟须及时处理。如大型强子对撞机实验设备中包含 15 亿个传感器，平均每秒钟收集超过 4 亿的实验数据；同样在 1 秒钟里，有超过 3 万次用户查询提交到谷歌，3 万微博被用户撰写。而人们对数据处理的速度的要求也日益严格，力图跟上社会的节奏。有报道称，美国中情局就要求利用大数据将分析搜集数据的时间由 63 天缩短为 27 分钟。

1.1.2.3 Variety（多样性）

在大数据背景下，数据在来源和形式上的多样性愈加突出。除以结构化形式存在的关系数据，网络上也存在大量的位置、图片、音频、视频等非结构化信息。其中，视频等非结构化数据占很大比例，有数据表明，到 2016 年，全部互联网流量中，视频数据将达到 55%，那么，有理由相信，大数据中 90% 都将是非结构化数据。并且，大数据不仅仅在形式上表现出多元化，其信息来源也表现出多样性，大致可将其分为网络数据、企事业单位数据、政府数据和媒体数据等几种。

1.1.2.4 Value（高价值性）

大数据价值总量大，但价值稀疏，即知识密度低。大数据以其高价值吸引了全世界的关注，据全球著名咨询公司麦肯锡报告：“如果能够有效地利用大数据来提高效率和质量，预计美国医疗行业每年通过数据获得的潜在价值可超过 3000 亿美元，能够使得美国医疗卫生支出降低 8%。”然而，大数据的知识密度非常低，IBM 副总裁表示：“可以利用 Twitter 数据获得用户对某个产品的评价，但是往往上百万条记录中只有很小的一部分真正讨论这款产品。”并且，虽然数据规模与数据挖掘得到的价值之间有相关性，但是两者难以用线性关系表达。这取决于数据的价值密度，同一事件的不同数据集即便有相同的规模（如对同一观察对象收集的长时间稀疏数据和短时间密集数据），其价值也可以相差很多，因为数据集“含金量”不同，大数

据中多数数据是重复的，忽略其中一些数据并不影响对其挖掘的结果。

1.2 大数据的发展趋势

1.2.1 大数据的背景

一般来说，大数据泛指巨量的数据集。当今社会，互联网尤其是移动互联网的发展，显著地加快了信息化向社会经济以及大众生活等各方面的渗透，促使了大数据时代的到来。近年来，人们能明显地感受到大数据来势迅猛。有关资料显示，1998年，全球网民平均每月使用流量是1MB，2003年是100MB，而2014年是10GB；全网流量累计达到1EB（即10亿GB）的时间在2001年是一年，在2004年是一个月，而在2013年仅需要一天，即一天产生的信息量可刻满1.88亿张DVD光盘。事实上，我国网民数居世界首位，产生的数据量也位于世界前列，其中包括淘宝网站每天超过数千万次的交易所产生的超50TB的数据，包括百度搜索每天生成的几十PB的数据，也包括城市里大大小小的摄像头每月产生的几十PB的数据，甚至还包括医院里CT影像抑或门诊所记录的信息。总之，大到学校、医院、银行、企业的系统行业信息，小到个人的一次百度搜索、一次地铁刷卡，大数据存在于各行各业，存在于民众生活的边边角角。

此外，大数据因自身可挖掘的高价值而受到重视。在国家宽带化战略的实施、云计算服务的起步、物联网的广泛应用和移动互联网崛起的同时，数据处理能力也迅速发展，数据积累到一定程度，其资料属性将更加明晰，显示出开发的价值。同时，社会的节奏越来越快，要求快速反应和精细管理，亟须借助对数据的分析和科学的决策，这样，我们便需要对上面所说的形形色色的海量数据进行开发。也就是说，大数据的时代来了。

有学者称，大数据将引发生活、工作和思维的革命；《华尔街日报》将大数据称为引领未来繁荣的三大技术变革之一；麦肯锡公司的报告指出，数据是一种生产资料，大数据将是下一轮创新、竞争、生产力提高的前沿；世界经济论坛的报告认为大数据是新财富，价值堪比石油；等等。因此，大数据的开发利用将成为各个国家抢占的新的制高点。

1.2.2 大数据现存的问题

1.2.2.1 速度方面的问题

传统的关系型数据库管理系统（RDBMS）一般都是集中式的存储和处理，没有采用分布式架构，在很多大型企业中的配置往往都基于IOE（IBM服务器，Oracle数据库，EMC存储）。在这种典型配置中，单台服务器的配置通常都很高，可以多达

几十个 CPU，内存也能达到上百 GB，数据库的存储放在高速大容量的磁盘阵列上，存储空间可达 TB 级。这种配置对于传统的管理信息系统（MIS）需求来说是可以满足的，然而面对不断增长的数据量和动态数据使用场景，这种集中式的处理方式就日益成为瓶颈，尤其是在速度响应方面捉襟见肘。在面对大数据量的导入导出、统计分析、检索查询方面，由于依赖于集中式的数据存储和索引，性能随着数据量的增长而急速下降，对于需要实时响应的统计及查询场景更是无能为力。比如，在物联网中，传感器的数据可以多达几十亿条，对这些数据需要进行实时入库、查询及分析，传统的关系数据库管理系统 RDBMS 就不再适合应用需求了。

1.2.2.2 种类及架构问题

RDMBS 对于结构化的、固定模式的数据，已经形成了相当成熟的存储、查询、统计处理方式。随着物联网、互联网以及移动通信网络的飞速发展，数据的格式及种类在不断变化和发展。在智能交通领域，所涉及的数据可能包含文本、日志、图片、视频、矢量地图等来自不同数据采集监控源的、不同种类的数据。这些数据的格式通常都不是固定的，如果采用结构化的存储模式将很难应对不断变化的需求。因此，对于这些种类各异的多源异构数据，需要采用不同的数据存储处理模式，结合结构化和非结构化数据存储。在整体的数据管理模式和架构上，也需要采用新型的分布式文件系统及分布式 NO-SQL 数据库架构，才能适应大数据量及变化的结构。

1.2.2.3 体量及灵活性问题

如前所述，大数据由于总体的体量巨大，采用集中式的存储，在速度、响应方面都存在问题。当数据量越来越大，并发读、写量也越来越大时，集中式的文件系统或单数据库操作将成为致命的性能瓶颈，毕竟单台计算机的承受压力是有限的。我们可以采用线性扩展的架构和方式，把数据的压力分散到很多台计算机上，直到可以承受，这样就可以根据数据量和并发量来动态增加和减少文件或数据库服务器，实现线性扩展。

在数据的存储方面，需要采用分布式可扩展的架构，比如，大家所熟知的 Hadoop 文件系统和 HBase 数据库。同时在数据的处理方面，也需要采用分布式的架构，把数据处理任务分配到很多计算节点上，同时还须考虑数据存放节点和计算节点之间的位置相关性。在计算领域中，资源分配、任务分配实际上是一个任务调度问题。其主要任务是根据当前集群中各个节点上面的资源（包括 CPU、内存、存储空间和网络资源等）的占用情况，和各个用户作业的服务质量要求，在资源和作业或者任务之间做出最优的匹配。由于用户对作业服务质量的要求是多样化的，同时资源的状态也在不断变化，因此，为分布式数据处理找到合适的资源是一个动态调度问题。

1.2.2.4 成本问题

集中式的数据存储和处理，在硬件、软件选型时，基本采用的方式都是配置相当高的大型机或小型机服务器，以及访问速度快、保障性高的磁盘阵列，来保障数据处理性能。这些硬件设备都非常昂贵，动辄高达数百万元。同时，软件也经常是国外大厂商如 Oracle、IBM、SAP、微软等的产品，对于服务器及数据库的维护需要专业技术人员，投入及运维成本很高。在面对海量数据处理的挑战时，这些厂商也推出了形似庞然大物的“一体机”解决方案，如 Oracle 的 Exadata、SAP 的 HANA 等，通过把多服务器、大规模内存、闪存、高速网络等硬件进行堆叠，来缓解数据压力，然而这造成在硬件成本上更是大幅跳高，一般的企业很难承受。新型的分布式存储架构、分布式数据库如 HDFS、HBase、Cassandra、MongoDB 等由于大多采用去中心化的、海量并行处理 MPP 架构，在数据处理上不存在集中处理和汇总的瓶颈，同时具备线性扩展能力，能有效地应对大数据的存储和处理问题。在软件架构上也都实现了一些自管理、自恢复的机制，以面对大规模节点中容易出现的偶发故障，保障系统整体的健壮性。因此，对每个节点的硬件配置，要求并不高，甚至可以使用普通的 PC 作为服务器，在服务器成本上可以大大节省，在软件方面开源软件占据非常大的价格优势。

当然，在谈及成本问题时，我们不能简单地进行硬件、软件的成本对比。要把原有的系统及应用迁移到新的分布式架构上，从底层平台到上层应用都需要做很大的调整。尤其是在数据库模式以及应用编程接口方面，新型的 NO-SQL 数据库与原来的 RDBMS 存在较大的差别，企业需要评估迁移及开发成本、周期及风险。除此之外，还须考虑服务、培训、运维方面的成本。但在总体趋势上，随着这些新型数据架构及产品的逐渐成熟与完善，以及一些商业运营公司基于开源基础为企业提供的专业数据库开发及咨询服务，新型的分布式、可扩展数据库模式必将在大数据浪潮中胜出，从成本到性能方面完胜传统的集中式大机模式。

1.2.2.5 价值挖掘问题

大数据由于体量巨大，同时又在不断增长，因此单位数据的价值密度在不断降低。但同时大数据的整体价值在不断提高，大数据被类比为石油和黄金，因此从中可以发掘出巨大的商业价值。要从海量数据中找到潜藏的模式，需要进行深度的数据挖掘和分析。大数据挖掘与传统的数据挖掘模式也存在较大的区别：传统的数据挖掘一般数据量较小，算法相对复杂，收敛速度慢。然而大数据的数据量巨大，在数据的存储、清洗、ETL（抽取、转换、加载）方面都需要能够应对大数据量的需求和挑战，在很大程度上需要采用分布式并行处理的方式。比如 Google、微软的搜索引擎，在

对用户的搜索日志进行归档存储时，就需要多达几百台甚至上千台服务器同步工作，才能应付全球上亿用户的搜索行为。同时，在对数据进行挖掘时，也需要改造传统数据挖掘算法以及底层处理架构，同样采用并行处理的方式才能对海量数据进行快速计算分析。Apache 的 Mahout 项目就提供了一系列数据挖掘算法的并行实现。在很多应用场景中，甚至需要挖掘的结果能够实时反馈回来，这对系统提出了很大的挑战，因为数据挖掘算法通常需要较长的时间，尤其是在大数据量的情况下，需要结合大批量的离线处理和实时计算才可能满足需求。

数据挖掘的实际增效也是我们在进行大数据价值挖掘之前需要仔细评估的问题。并不见得所有的数据挖掘计划都能得到理想的结果。首先，需要保障数据本身的真实性和全面性。如果所采集的信息本身噪声较大，或者一些关键性的数据没有被包含进来，那么所挖掘出来的价值规律也就大打折扣。其次，要考虑价值挖掘的成本和收益。如果对挖掘项目投入的人力物力、硬件及软件平台耗资巨大，项目周期也较长，而挖掘出来的信息对于企业生产决策、成本效益等方面的贡献不大，那么片面地相信和依赖数据挖掘的威力，也是不切实际和得不偿失的。

1.2.2.6 存储及安全问题

在大数据的存储及安全保障方面，大数据由于存在格式多变、体量巨大的特点，也带来了很多挑战。针对结构化数据，关系型数据库管理系统 RDBMS 经过几十年的发展，已经形成了一套完善的存储、访问、安全与备份控制体系。由于大数据的巨大体量，也对传统 RDBMS 造成了冲击，如前所述，集中式的数据存储和处理也在转向分布式并行处理。大数据更多的时候是非结构化数据，因此衍生了许多分布式文件存储系统、分布式 NO-SQL 数据库等来应对这类数据。然而这些新兴系统，在用户管理、数据访问权限、备份机制、安全控制等各方面还须进一步完善。至于安全问题，一是要保障数据不丢失，对海量的结构、非结构化数据，需要有合理的备份冗余机制，在任何情况下数据不能丢失。二是要保障数据不被非法访问和窃取，只有对数据有访问权限的用户，才能看到数据，拿到数据。由于大量的非结构化数据可能需要不同的存储和访问机制，因此要形成对多源、多类型数据的统一安全访问控制机制，是亟待解决的问题。大数据由于将更多、更敏感的数据汇集在一起，对潜在攻击者的吸引力更大，若攻击者成功实施一次攻击，将能得到更多的信息，“性价比”更高，这些都使得大数据更易成为被攻击的目标。2012 年 Linkedin650 万用户账户密码泄露；雅虎遭到网络攻击，致使 45 万用户 ID 泄露。2011 年 12 月，CSDN 的安全系统遭到黑客攻击，600 万用户的登录名、密码及邮箱遭到泄露。

1.2.2.7 互联互通与数据共享问题

大数据要发挥威力，需要融合多行业的数据分析决策，这在智慧城市建设中尤其重要。为实现跨行业的数据整合，需要制定统一的数据标准、交换接口以及共享协议，这样不同行业、不同部门、不同格式的数据才能基于一个统一的基础进行访问、交换和共享。对于数据访问，还须规定细致的访问权限，规定什么样的用户在什么样的场景下，可以访问什么类型的数据。在大数据及云计算时代，不同行业、企业的数据可能存放在统一的平台和数据中心之上，需要对一些敏感信息进行保护。比如涉及企业商业机密及交易信息方面的数据，虽然是依托平台来进行处理，但是除了企业自身的授权人员之外，要保证平台管理员以及其他企业都不能访问此类数据。

1.2.3 大数据模式

1.2.3.1 结构化数据

结构化数据遵循一个标准的模型或者模式，并且常常以表格的形式存储。该类型数据通常用来捕捉不同对象实体之间的关系，并且存储在关系型数据库中。诸如ERP和CRM等企业应用和信息系统之中会频繁地产生结构化数据。由于数据库本身以及大量现有的工具对结构化数据的支持，结构化数据很少需要在处理或存储的过程中做特殊的考虑。这类数据的例子包括银行交易信息、发票信息和消费者记录等。

1.2.3.2 非结构化数据

非结构化数据是指不遵循统一的数据模式或者模型的数据。据估计，企业获得的数据有80%左右是非结构化数据，并且其增长速率要高于结构化数据。这种类型的数据可以是文本的，也可以是二进制的，常常通过自包含的、非关系型文件传输。一个文本文档可能包含许多博文和推文。而二进制文件多是包含着图像、音频、视频的媒体文件。从技术上讲，文本文件和二进制文件都有根据文件格式本身定义的结构，但是这个层面的结构不在讨论之中，并且非结构化的概念与包含在文件中的数据相关，而与文件本身无关。

存储和处理非结构化的数据通常需要用到专用逻辑。如要放映一部视频，正确的编码、解码是至关重要的。非结构化数据不能被直接处理或者用SQL语句查询。如果需要存储在关系型数据库中，它们会以二进制大型对象（BLOB）形式存储在表中。当然，NO-SQL数据库作为一个非关系型数据库，能够用来同时存储结构化和非结构化数据。半结构化数据有一定的结构与一致性约束，但本质上不具有关系性。半结构化数据是层次性的或基于图形的。这类数据常常存储在文本文件中。由于文本化的本质以及某些层面上的结构化，半结构化数据比非结构化数据更好处理。半结构化数据的一些常见来源包括电子转换数据（EDI）文件、扩展表、RSS源以及传感器

数据。半结构化数据也常需要特殊的预处理和存储技术，尤其是重点部分不是基于文本的时候。半结构化数据预处理的一个例子就是对 XML 文件的验证，以确保它符合其模式定义。

1.2.4 大数据技术模式

1.2.4.1 批处理计算

批处理计算主要解决针对大规模数据的批量处理，也是我们日常数据分析工作中非常常见的一类数据处理需求。Map Reduce 是最具有代表性和影响力的大数据批处理技术，可以并行执行大规模数据处理任务，用于大规模数据集（大于 1TB）的并行运算。Map Reduce 极大地方便了分布式编程工作，它将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数 Map 和 Reduce 上，编程人员在不会分布式并行编程的情况下，也可以很容易地将自己的程序运行在分布式系统上，完成海量数据集的计算。

Spark 是一个针对超大数据集合的低延迟的集群分布式计算系统，比 Map Reduce 快许多。Spark 启用了内存分布数据集，除了能够提供交互式查询外，还可以优化迭代工作负载。在 Map Reduce 中，数据流经一个稳定的来源进行一系列加工处理后，流出到一个稳定的文件系统（如 HDFS）。而对于 Spark 而言，则使用内存替代 HDFS 或本地磁盘来存储中间结果，因此 Spark 要比 Map Reduce 的速度快许多。

1.2.4.2 流计算

流数据也是大数据分析中的重要数据类型。流数据（或数据流）是指在时间分布和数量上无限的一系列动态数据集合体，数据的价值随着时间的流逝而降低。因此，必须采用实时计算的方式给出秒级响应。流计算可以实时处理来自不同数据源的、连续到达的流数据，经过实时分析处理，给出有价值的分析结果。目前业内已涌现出许多的流计算框架与平台，第一类是商业级的流计算平台，包括 IBM Info Sphere Streams 和 IBM Stream Base 等；第二类是开源流计算框架，包括 Twitter Storm、Yahoo! S4 等；第三类是公司为支持自身业务开发的流计算框架，如 Facebook 使用 Puma 和 HBase 相结合来处理实时数据，百度开发了通用实时流数据计算系统 D-Stream，淘宝开发了通用流数据实时计算系统流数据处理平台。

1.2.4.3 图计算

在大数据时代，许多大数据都是以大规模图或网络的形式呈现，如社交网络、传染病传播途径、交通事故对路网的影响等。此外，许多非图结构的大数据也常常会被转换为图模型后再进行处理分析。Map Reduce 作为单输入、两阶段、粗粒度数据并行的分布式计算框架，在表达多稀疏结构和细粒度数据时，往往显得力不从心，

不适合用来解决大规模图计算问题。因此，针对大型图的计算，需要采用图计算模式，目前已经出现了不少相关图计算产品。Pregel 是一种基于 BSP（Bulk Synchronous Parallel）模型实现的并行图处理系统。为了解决大型图的分布式计算问题，Pregel 搭建了一套可扩展的、有容错机制的平台，该平台提供了一套非常灵活的 API，可以描述各种各样的图计算。Pregel 主要用于图遍历、最短路径、PageRank 计算等，其他代表性的图计算产品还包括 Facebook 针对 Pregel 的开源实现 Giraph、Spark 下的 Graph-X、图数据处理系统 Power Graph 等。

1.2.4.4 查询分析计算

针对超大规模数据的存储管理和查询分析，需要提供实时或准实时地响应，才能很好地满足企业经营管理需求。谷歌公司开发的 Dremel 是一种可扩展的、交互式的实时查询系统，用于只读嵌套数据的分析。通过结合多级树状执行过程和列式数据结构，它能做到几秒内完成对万亿张表的聚合查询。系统可以扩展到成千上万的 CPU 上满足谷歌上万用户操作 PB 级的数据，并且可以在 2~3 秒内完成 PB 级别数据的查询。此外，Cloudera 公司参考 Dremel 系统开发了实时查询引擎 Impala。它提供 SQL 语义，能快速查询存储在 Hadoop 的 HDFS 和 HBase 中的 PB 级大数据。

1.2.5 大数据与大云计算的安全问题

大数据时代给了人们前所未有的数据采集、存储和处理的能力。每一个人都可以把文档、图片、视频等放在云端，享受随时随地同步和查看的便捷性；企业可以将生产、运营、营销等各个环节数字化，还可以收集全行业的信息，通过移动终端就可以轻松地获得企业生产经营的各种报表和趋势预测；政府的服务和社会化管理则可以通过互联网到达每家每户和每个企业。强大的云数据中心和先进的移动互联网技术使得谷歌眼镜、智能手环这样的可穿戴设备及各种多媒体社交工具盛行，发布信息和检索信息都只需在眼睛一眨、指头一动间完成，甚至无须做任何动作就完成了。但是同时，由于大数据的社会化属性，人们在网络空间的任何数据都可能被收集，人们的资料可能被黑客窃取，人们的朋友圈在社交网络上一目了然，人们的言论在微博上历历在目，人们的交易和浏览信息随意地被电商挖掘。大数据和云计算就是一把双刃剑，在方便人们生活的同时，安全和隐私问题也日益凸显。

随着数据中心不断整合以及虚拟化、VDI、云端运算应用程序的兴起，越来越多的运算效能与数据都集中到数据中心和服务器上。不论是个人信息存储在云盘、邮箱，还是企业将数据存储在云端或使用云计算服务，这些都需要安全保护，安全和隐私问题可以说是云计算和大数据时代所面临的最为严峻的挑战。在 IDC 的一项关于“您认为云计算模式的挑战和问题是什么”的调查中，安全以 74.6% 的比例位居榜首，

全球51%的首席信息官认为安全问题是部署云计算时最大的顾虑。趋势科技首席执行官陈怡桦认为："云计算的日益普及已经使越来越多的云计算服务商进入市场。随着在云计算环境中存储数据的公司越来越多，信息安全问题成为大多数IT专业人士最头疼的事情。事实上，数据安全已经是考虑采用云基础设施的机构主要关注的问题之一。"

大数据由于数据集中、目标大，在网络上更容易被盯上；在线数据越来越多，黑客们的犯罪动机也比以往任何时候更强烈；大数据意味着若攻击者成功实施一次攻击，其能得到更多的信息和价值。这些特点都使得大数据更易成为被攻击的目标。

关于网络信息安全，最知名的事件莫过于"棱镜门"了。据美国中情局前职员爱德华·斯诺登披露，"棱镜计划"是一项由美国国家安全局（NSA）于2007年小布什时期开始实施的绝密电子监听计划。该计划能够直接进入美国国际网络公司的中心服务器挖掘数据、收集情报，包括微软、雅虎、谷歌、苹果等在内的9家国际巨头公司参与其中，从音频、视频、图片、文档、邮件和链接信息中分析个人的联系方式和行为。

与此同时，公民的隐私泄露事件也层出不穷，这些泄露大部分是黑客攻击企业数据库造成的。据隐私专业公司PRC（Privacy Rights Clearinghouse）报告称，按保守估计，2011年全球发生了超过500起重大数字安全事故。如2011年4月索尼公司由于系统泄露导致7700万名用户资料遭窃，导致1.7亿美元左右的损失；2011年12月，CSDN的安全系统遭到黑客攻击，600万名用户的登录名、密码和邮箱遭到泄露；Linkedln在2012年被曝650万名用户账户密码泄露；雅虎遭到网络攻击，致使45万名用户ID泄露。

另外一些隐私泄露是因为企业产品功能不完善造成的。比如几年前，腾讯QQ曾经推出朋友圈功能，很多用户的真实名字出现在朋友圈中，引起了用户的强烈抗议，最后腾讯关闭了这一功能。腾讯QQ用户真实姓名在朋友圈中曝光，就是采用了大数据关联分析。由此可见，在大数据的搜集和数据分析过程中，随时可能触及用户的隐私，一旦某一环节存在安全隐患，后果不堪设想。

还有一些则是用户个人不注意造成的隐私泄露。比如，有些用户喜欢在Twitter等社交网站上发布自己的位置和动态信息，结果有几家网站，如"PleaseRobMe.corn""We Know Yrour House"等，能够根据用户所发的信息，推测出用户不在家的时间，找到用户准确的家庭地址，甚至找出房子的照片。这些网站的做法旨在提醒大家，我们随时暴露在公众视野下，如果不培养安全意识和隐私意识，将会给自身带来灾难。

大数据可以光明正大地搜集用户数据，并可以对用户数据进行分析，这无疑让