第 2 章 Chapter 2

大数据安全、隐私保护和审计技术

2.1 大数据安全

大数据（Big Data）指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策目的的信息。大数据是当下最火热的 IT 行业的词汇，随之数据仓库、数据安全、数据分析、数据挖掘等围绕大数据的商业价值的利用将逐渐成为行业人士争相追捧的利润焦点。本节介绍大数据安全的意义和重要作用、大数据面临的问题与挑战，以及大数据安全防护的主要技术。

2.1.1 大数据安全的意义和重要作用

为什么要研究大数据？在开始了解大数据安全之前，需要先搞清楚这个问题。当今，社会信息化和网络化的发展导致数据的爆炸式增长，据统计，平均每秒有 200 万用户在使用谷歌搜索，各行业也在不断产生大量数据。在科学界，《Nature》和《Science》都推出了大数据专栏对其展开探讨，这意味着大数据将成为云计算之后的信息技术领域的另一个信息产业增长点。

现阶段，国家十分支持大数据的发展，国务院以及各级地方政府从 2012 年开始颁布了大量政策来扶持大数据产业。从现有的政策来看，大数据的发展已经被列为国家发展战略了，大数据的重要性不言而喻。大数据已经得到政府高层、互联网企业以及其他各个行业企业的认可，对大数据的开发和应用的力度也相应加大。近年来，我国高度重视大数据发展，仅 2015 年最高层面就发出了多次重视大数据的声音。2015 年 5 月，李克强总理提出，大数据产业是中国推动“互联网 +”战略的重要支撑。2015 年 6 月，习近平主席考察贵阳，调研贵阳大数据交易所时说发展大数据确实有道理。在政策层面，2015 年 9 月，国务院通

过《关于促进大数据发展的行动纲要》，这是支持大数据发展的第一部正式国家层面文件，对大数据的规范化发展起到了至关重要的作用。

大数据对企业的影响也是巨大的，正是大数据对企业所产生的立竿见影的效果，现在已经得到更多公司的重视。首先，大数据能够彻底改变企业内部运作模式，以往的管理是“领导怎么说?”，现在变成“大数据的分析结果是什么?”。这是对传统领导力的挑战，也推动了对企业管理岗位人才的重新定义。企业管理人才不仅要懂企业的业务流程，还要成为数据专家，跨专业的要求改变了过去领导力主要体现在经验和过往业绩上，如今新的要求是熟练掌握大数据分析工具，善于运用大数据分析结果，并结合企业的销售和运营管理实践。当然大数据对企业的作用中一个不可回避的关键因素是数据的质量，有句话叫“垃圾进，垃圾出”，是说如果采集的是大量垃圾数据，则会导致产生的分析结果也是毫无意义的垃圾。

大数据也在影响着我们每个人的生活，使得一些服务更加贴近大家的生活。打开浏览器上网，广告弹窗推荐的商品可能正好就是你最近想买的东西。翻阅自己的微博，查看定位信息就能够准确回忆起一年前的今天你在哪里，做了什么。在搜索引擎中输入几个关于自己的关键词，也许可以重温你在 10 年前写下的网络日志。进入淘宝网，它就能根据你的历史浏览记录为你贴心地推荐你想要的商品。大数据现在已经进入人们的生活。

正是因为大数据对国家、企业、个人具有重要的作用，并具有很高的研究价值，所以大数据安全现在成为学术与工业界的研究热点，是人们公认的大数据相关问题中关键的问题之一。没有安全，发展就是空谈，数据安全是发展大数据的前提，必须将它摆在更加重要的位置。我们在使用和发展大数据的同时，也容易出现大数据引发的个人隐私安全、企业信息安全乃至国家安全问题。

1）与大数据安全及个人关系最密切的就是个人隐私安全，在大数据时代，想屏蔽外部数据商挖掘个人信息是不可能的。目前，各社交网站均不同程度地开放其用户所产生的实时数据，这些数据被一些数据提供商收集，还出现了一些监测数据的市场分析机构。通过人们在社交网站中写入的信息、智能手机显示的位置信息等多种数据组合，已经可以以非常高的精度锁定个人，挖掘出个人信息体系，因此，用户隐私问题堪忧。据统计，通过分析用户 4 个曾经到过的位置，就可以识别出 95% 的用户。“你没有隐私，忘记这事吧。”有数据统计，中国 78.2% 的网民个人身份信息被泄露过，包括姓名、学历、家庭住址、身份证号及工作单位等。其中，82.3% 的网民亲身感受到了个人信息泄露给日常生活造成的不良影响。

2）企业迈进大数据时代，信息安全面临多重挑战。企业在获得“大数据时代”信息价值增益的同时，也在不断地累积风险，大数据安全方面的挑战日益增大。首先是黑客窃密与病毒木马对企业信息系统的入侵，大数据在云系统中进行上传、下载、交换的同时，极易成为黑客与病毒的攻击对象。而“大数据”一旦被入侵并产生泄密，就会对企业的品牌、信誉、研发、销售等多方面带来严重冲击，并带来难以估量的损失。通常，那些对大数据

分析有较高要求的企业，会面临更多的挑战，例如电子商务、金融、天气预报的分析预测、复杂网络计算和广域网感知等。任何一个会误导目标信息提取和检索的攻击都是有效攻击，因为这些攻击会对安全厂商的大数据安全分析产生误导，导致其分析偏离正确的检测方向。这些攻击需要我们集合大量数据，进行关联分析才能够知道其攻击意图。大数据安全是与大数据业务相对应的，传统时代的安全防护思路此时难以奏效，并且成本过高。无论是从防范黑客对数据的恶意攻击，还是从对内部数据的安全管控角度，为了保障企业信息安全，迫切需要一种更为有效的方法对企业大数据安全进行有效管理。

3）大数据时代，国家安全将受到信息战与网络恐怖主义的威胁，大数据安全的重要性在国家层面也需要得到重视。如今的信息时代，安全环境发生了质的变化。不管是战争时期还是和平年代，一国的各种信息设施和重要机构等都可能成为打击目标，而且保护它们免受攻击已超出了军事职权和能力的范围。决策的不可靠性、信息自身的不安全性、网络的脆弱性、攻击者数量的激增、军事战略作用的下降和地理作用的消失等，都使国家安全受到了严峻的挑战。此外，大数据也使网络恐怖主义者有了可乘之机。庞大海量的大数据涉及面广泛，将有可能使网络恐怖主义的势力侵入人们生活的方方面面。大数据对国家安全的影响涉及了国家安全内容的诸多方面，我们平时关注比较多的有科技安全、信息安全，其实大数据安全对国民安全、政治安全、意识形态安全、社会公共安全等的影响也很大。

大数据的发展给我们带来了机遇，但是也带来了挑战。大数据已经影响到个人、企业、国家，对整个社会都有很重要的影响，在享受大数据的便利的同时我们必须重视大数据安全。

2.1.2　大数据安全面临的问题与挑战

“世界的本质是数据，大数据开启了一次重大的时代转型，也是一场生活、工作与思维的大变革。”随着世界各国在陆、海、空、天、电、网多维度战略的部署，信息技术爆炸式发展。基于大数据发展对国家、社会的组织结构和治理模式，对商业、企业的决策方式和业务策略，对个人的生活、思维方式等各方面产生的深刻影响，各界逐渐开始关注“信息”本身而不只是“技术”了。在大数据时代，人类信息管理准则也将面临重新定位，而在信息安全问题日益突出的当下，大数据在给信息安全带来新挑战的同时，也为信息安全领域的发展带来新机遇。

1. 大数据成为网络攻击的显著目标

在网络空间中，大数据成为更容易被“发现”的大目标，承载着越来越多的关注度。大数据自身规模大且集中的特点使得其在网络空间中无疑是一个更易被“发现”“命中”的大目标，低成本、高收益的攻击效果对黑客而言是充满诱惑力的。一方面，大数据不仅意味着海量的数据，也意味着更复杂、更敏感的数据，这些数据成为更具吸引力的目标，会吸引更多的潜在攻击者。另一方面，数据的大量聚集，使得黑客通过一次成功的攻击就能够获得更多的数据，无形中降低了黑客的攻击成本，增加了“收益率”。

2. 大数据加大隐私泄露风险

网络空间中的数据来源涵盖非常广阔的范围，例如传感器、社交网络、记录存档、电子邮件等，大量数据的聚集不可避免地加大了用户隐私泄露的风险。一方面，大量数据聚集，包括大量的企业运营数据、客户信息、个人的隐私和各种行为的细节记录，这些数据的集中存储增加了数据泄露风险，而这些数据不被滥用，成为人身安全的一部分。另一方面，一些敏感数据的所有权和使用权并没有明确的界定，很多基于大数据的分析都未考虑其中涉及的个体的隐私问题。

从个人隐私的角度而言，用户在互联网中产生的数据具有累积性和关联性，单点信息可能不会暴露隐私，但如果采用大数据关联性抽取和集成有关某用户的多点信息并进行汇聚分析，其隐私泄露的风险将大大增加，关联性利用类似于现实生活中通过“人肉搜索”将某人或事物暴露。

从企业、政府等大的角度而言，大数据安全标准体系尚不完善，隐私保护技术和相关法律法规尚不健全，加之大数据所有权和使用权出现分离，使得数据公开和隐私保护很难做到友好协调。在数据的合法使用者利用大数据技术收集、分析和挖掘有价值信息的同时，攻击者也同样可以利用大数据技术最大限度地获取他们想要的信息，这无疑增加了企业和政府敏感信息泄露的风险。

从大数据基础技术的角度而言，无论是被公认为大数据标准开源软件的 Hadoop，还是大数据依托的数据库基础 NoSQL，其本身均存在数据安全隐患。Hadoop 作为一个分布式系统架构对数据的汇聚增加数据泄露风险的同时，作为一个云平台也存在着云计算面临的访问控制问题，其派生的新数据也面临加密问题。NoSQL 技术将不同系统、不同应用和不同活动的数据进行关联，加大了隐私泄露风险。又由于数据的多元非结构化，使得企业很难对其中的敏感信息进行定位和保护。

3. 大数据对现有的存储和安防措施提出挑战

大数据存储带来新的安全问题。大数据集中的后果是复杂多样的数据存储在一起，例如开发数据、客户资料和经营数据存储在一起，可能会出现违规地将某些生产数据放在经营数据存储位置的情况，造成企业安全管理不合规。大数据的大小影响到安全控制措施能否正确运行。对于海量数据，常规的安全扫描手段需要耗费过多的时间，已经无法满足安全需求。安全防护手段的更新升级速度无法跟上数据量非线性增长的步伐，大数据安全防护存在漏洞。

4. 大数据技术被应用到攻击手段中

在企业用数据挖掘和数据分析等大数据技术获取商业价值的同时，黑客也在利用这些大数据技术向企业发起攻击。黑客最大限度地收集更多有用信息，比如社交网络、邮件、微博、电子商务、电话和家庭住址等信息，为发起攻击做准备，大数据分析让黑客的攻击更精准。此外，大数据为黑客发起攻击提供了更多机会。黑客利用大数据发起“僵尸网络

攻击”，可能会同时控制上百万台“傀儡机”并发起攻击，这个数量级是传统单点攻击不具备的。

5. 大数据成为高级可持续攻击的载体

黑客利用大数据将攻击很好地隐藏起来，用传统的防护策略难以检测出来。传统的检测是在单个时间点进行的基于威胁特征的实时匹配检测，而高级可持续攻击（APT）是一个实施过程，并不具有能够被实时检测出来的明显特征，无法被实时检测。同时，APT 攻击代码隐藏在大量数据中，很难被发现。此外，大数据的价值低密度性，让安全分析工具很难聚焦在价值点上，黑客可以将攻击隐藏在大数据中，给安全服务提供商的分析造成很大困难。黑客设置的任何一个会误导安全厂商目标信息提取和检索的攻击，都会导致安全监测偏离应有的方向。

6. 大数据技术为信息安全提供新支撑

大数据在带来新安全风险的同时也为信息安全的发展提供了新机遇。大数据正在为安全分析提供新的可能性，对于海量数据的分析有助于信息安全服务提供商更好地刻画网络异常行为，从而找出数据中的风险点。对实时安全和商务数据结合在一起的数据进行预防性的分析，以便识别钓鱼攻击，防止诈骗和阻止黑客入侵。网络攻击行为总会留下蛛丝马迹，这些痕迹都以数据的形式隐藏在大数据中，利用大数据技术整合计算和处理资源有助于更有针对性地应对信息安全威胁，使得网络攻击行为无所遁形，有助于找到发起攻击的源头。

7. 大数据对信息安全的合规性要求

大数据时代出现数据拥有权和使用权的分离，数据经常脱离数据拥有者的控制范围而活跃着，这就对数据需求合规性和用户授权合规性提出新的要求，包括数据形态和转移方式的合规性。数据需求方为精准开展一个业务，要求数据拥有者提供原始敏感数据或未脱敏的统计类数据，显然这违背了信息安全的本意。就算数据需求遵循最小级原则，对数据的提供未超出合理范围，用户授权仍是数据服务的前提，包括转移数据使用的目的、范围、方式以及授权信息的保存等各个环节。

在对信息安全提出合规性要求的同时，引入第三方的标准符合性审查服务也很必要。如通过针对数据提供者和接受者双方的审查，包括文档资料安全规范的审查，技术辅助现场审查，在供方和需方之间做扫描和数据检测，提供第三方公平的数据安全审查服务。

2.1.3　大数据安全防护技术

1. 数据发布匿名保护技术

数据发布匿名保护技术是对大数据中的结构化数据实现隐私保护的关键技术手段。匿名化的处理过程可以用图 2-1 简单表示。具体来说，数据库对所有人都是公开的，任何人都可以自由访问，但是却不能将数据库中的任一记录对应到具体某一个体上。为了对数据

表中的数据进行隐私保护，自由访问型隐私保护通常采取的办法是对原始数据实施“数据匿名化”操作。所谓“数据匿名化”就是数据发布者在数据发布前需要对真实数据表实施一定的预处理，使攻击者无法从经过匿名变换后的数据表中唯一推导出某个具体个体对应的敏感信息，从而实现对个体隐私信息的隐藏。

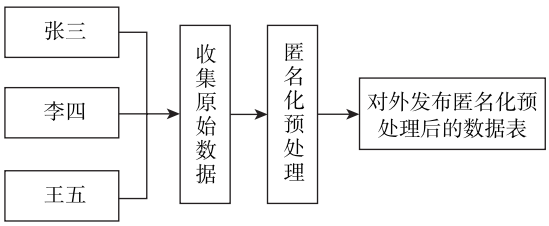

图 2-1 匿名化处理过程

2. 社交网络匿名保护技术

因为用户的个性化信息与用户隐私密切相关，所以互联网服务提供商一般会对用户数据进行“数据匿名化”之后再提供共享或对外发布。表面上看，活跃于社交网络上的信息并不泄露个人隐私。但事实上，几乎任何类型的数据都如同用户的指纹一样，能通过辨识找到其拥有者。在当今社会，一旦用户的通话记录、电子邮件、银行账户、信用卡信息、医疗信息等大规模数据被无节制地搜集、分析与交易利用，那么用户都将“被透明”，不仅个人隐私荡然无存，还将引发一系列社会问题。因此深入理解社交网络的匿名化和去匿名化这一对相互依存的博弈过程，才能更好地在社交网络活动中保护好用户的隐私，这个问题已成为当前大众关注的焦点。社交网络中典型的匿名保护技术如下：

1）用户标识匿名与属性匿名保护，在数据发布时隐藏了用户的标识与属性信息。属性数据在社交网络上变化最频繁，内容最丰富，它生动地描述了用户的个性化特征，能够帮助系统建立完整的用户轮廓，提高推荐系统的准确性。然而，用户往往不希望将所有属性信息对外公开。例如：用户观看私密视频的记录被曝光，会对用户的网络形象造成最直接的破坏，甚至影响用户的正常生活。属性隐私保护要求对社交网络的属性信息进行匿名化处理，阻止攻击者对用户的属性隐私进行窥探。

2）用户间关系匿名保护，在数据发布时隐藏了用户间的关系。社交关系数据本身蕴含着巨大的价值。互联网服务提供商可基于用户现有的社交结构分析用户的交友倾向、向用户推荐朋友等，有助于保持社交群体的活跃度和黏性。但是与此同时，分析者也可以挖掘出用户不愿公开的社交关系、交友群体特征，从而导致用户的社交关系隐私暴露。为此，社交关系隐私保护要求节点对应的社交关系保持匿名，使攻击者无法确认特定用户拥有哪些社交关系。

3. 数据水印技术

数据水印技术是指将标识信息以难以察觉的方式嵌入数据载体内部且不影响其使用的方法，多见于多媒体数据的版权保护，也有针对数据库和文本文件的水印方案。当然，实现数据水印技术的前提是，数据中存在冗余信息或可容忍一定的精度误差。数据水印技术按照不同的划分方法有不同的分类，在大数据领域，比较常用的是按照特性划分为鲁棒数字水印和易损数字水印两类，鲁棒数字水印可用于大数据起源证明，易损数字水印可用于

证明数据的真实性。

鲁棒数字水印主要用在数字作品中标识著作权信息，利用这种水印技术可以在多媒体内容的数据中嵌入创建者、所有者的标识信息，或者嵌入购买者的标识（即序列号）。在发生版权纠纷时，创建者或所有者的信息用于标识数据的版权所有者，而序列号用于追踪违反协议而为盗版提供多媒体数据的用户。用于版权保护的数字水印要求有很强的鲁棒性和安全性，除了要求在一般图像处理（如滤波、加噪声、替换、压缩等）中生存外，还需能抵抗一些恶意攻击。

易损水印与鲁棒水印的要求相反，它主要用于完整性保护。这种水印同样是在内容数据中嵌入不可见的信息，当内容发生改变时，这些水印信息会发生相应的改变，从而可以鉴定原始数据是否被篡改。易损水印应对一般图像处理（如滤波、加噪声、替换、压缩等）有较强的免疫能力（鲁棒性），同时又要有较强的敏感性，既允许一定程度的失真，又要能将失真情况探测出来。它必须对信号的改动很敏感，人们才能根据易损水印的状态判断出数据是否被篡改过。

4. 数据溯源技术

数据溯源技术的目标是帮助人们确定数据仓库中各项数据的来源，也可用于文件的溯源与恢复。数据溯源技术的意义在于根据追踪路径重现数据的历史、状态和演变过程，实现数据历史档案的追溯。目前数据溯源的基本方法包括标注法和反向查询法[⊖]。

标注法是一种简单且有效的数据溯源方法，使用非常广泛。它通过记录相关的信息来追溯数据的历史状态，即用标注的方式来记录原始数据的一些重要信息，如背景、作者、时间、出处等，并将标注和数据一起传播，通过查看目标数据的标注来获得数据的溯源。

反向查询法，有的文献也称为逆置函数法。由于标注法并不适合细粒度数据，特别是大数据集中的数据溯源，于是，有人提出了反向查询法，此方法是通过逆向查询或构造逆向函数对查询求逆，或者说根据转换过程反向推导，由结果追溯到原数据。这种方法是在需要时才计算，所以又叫 lazzy 方法。

标注法和反向查询法各有优缺点，在实际使用时需要根据具体的情况进行选择。标注法的优点是实现简单，容易管理；缺点是只适合小型系统，对于大型系统而言很难为细粒度的数据提供详细的数据溯源信息，因为很细可能导致元数据比原始数据还多，需要额外的存储空间，会对存储造成很大的压力，而且效率低。反向查询法的优点是追踪比较简单，只需存储少量的元数据就可实现对数据的溯源，不需要存储中间处理信息、全过程的注释信息；缺点是用户需要提供逆置函数（并不是所有的函数都具有可逆性）和相对应的验证函数，构造逆置函数具有一定局限性，实现起来相对比较复杂。

5. 访问控制技术

访问控制（Access Control）指系统限制用户身份及其所属的预先定义的策略组使用某

⊖ 明华，张勇，符小辉 . 数据溯源技术综述 [J]. 小型微型计算机系统，2012, 33(9):47-53.

些数据资源的手段。这在数据库领域已经是很成熟的技术，在大数据安全领域也有很多访问控制技术得到了广泛的应用，主要包括基于角色的访问控制、基于属性的访问控制和风险自适应的访问控制技术。

基于角色的访问控制技术是应用最广泛的技术，该方法给不同角色赋予不同的访问控制权限。其基本思想是，不是将系统操作的各种权限直接授予具体的用户，而是在用户集合与权限集合之间建立一个角色集合，每一种角色对应一组相应的权限。一旦用户被分配了适当的角色，该用户就拥有此角色的所有操作权限。这样做的好处是，不必在每次创建用户时都进行分配权限的操作，只要分配用户相应的角色即可，而且角色的权限变更比用户的权限变更要少得多，这样将简化用户的权限管理，减少系统的开销。

基于属性的访问控制技术是通过综合考虑各类属性（如用户属性、资源属性、环境属性等）来设定用户的访问权限。基于属性的访问控制技术实现了细粒度的权限控制，所有实体的描述都采用同一种方式——属性来进行描述，但不同实体的属性权限可能不同，这使得访问控制判定功能在判定时能够进行统一处理。在基于属性的访问控制中，访问判定是基于请求者和资源具有的属性，请求者和资源在基于属性的访问控制技术中通过特性来标识，而不是通过 ID 来标识，这使得传统的基于身份的访问控制具有足够的灵活性和可扩展性，同时使得安全的匿名访问成为可能，这在大型分布式环境下是十分重要的。

风险自适应的访问控制是针对大数据场景推荐的一种访问控制方法。风险自适应的访问控制针对的是在大数据场景中，安全管理员可能缺乏足够的专业知识，无法准确地为用户指定其可以访问的数据的情况。在大数据场景中，数据种类和来源复杂，用户角色也十分复杂，往往无法准确地为用户预先指定其可以访问的数据，最好是在某个访问行为发生时针对具体上下文进行判断，自适应的访问控制正是这样一种上下文敏感的动态系统安全访问技术。

2.2 大数据隐私保护

大数据是当前学术界以及产业界的研究热点，它正在潜移默化地影响着人们日常的工作、学习以及娱乐方式。但是目前大数据在收集、存储和处理的过程中面临着诸多挑战，大数据时代隐私保护问题日渐突出。收集和分析用户浏览数据，可以提高广告效应；收集用户位置数据，可为附近商户引流。人们能想到的用户行为和隐私数据，事实上都有可能在利益驱使下被多方知名或不知名团体收集。即使下载一款简单应用程序，也有可能在授权时被具有不同目的的机构获取隐私数据，用于商业或其他目的。大数据所导致的隐私泄露给用户带来了严重困扰，也给各商家带来了信任危机。

2.2.1 大数据隐私保护的意义和重要作用

随着互联网技术的迅速发展，整个社会被迫进入“大数据”时代。不管人们愿意与否，

用户的个人数据总是无意中被动地被某些个人和公司收集并利用。大数据爆炸式的发展席卷了全球 IT、零售、交通等行业，给这些行业带来了巨大的变革，同时也改变了人们的生活。大数据时代的一个特点就是可以将用户保留在互联网中的敏感数据转化为有价值的资源，通过对这些数据的分析，无论是商家、保险公司以及其他以服务为导向的公司，都可以提供更贴心、更多的个人服务，他们甚至比我们都“了解”自己。个人数据共享化以及透明化已成为不可阻挡的趋势。过去，具有公共权力的政府机构合法地掌握着大量公民的个人隐私数据，但现在很多公司和个人拥有了大量来自互联网的个人隐私数据，在某些方面甚至可能比政府机构掌握的数据还要细致和全面。对企业而言，用户数据是宝贵的资源，因为可以通过数据挖掘和机器学习从中获取丰富的有用价值。同时，用户的隐私数据也是危险的“潘多拉的盒子”，一旦数据发生泄露，用户的个人隐私将被侵犯。近年来不断发生的隐私泄露事件提醒着我们，公民个人隐私保护已面临严重的挑战。大量数据的分析和使用，使得人们的生活更方便，同时，越来越多的人担心隐私泄露的问题。

大数据的快速发展面临着与个人隐私保护需求发生冲突的矛盾。根据 Wikibon 的报告，在 2017 年，美国大数据产业的市场规模将达到 500 亿美元，如图 2-2 所示。这充分证明了大数据将为美国的很多行业带来巨大的价值。

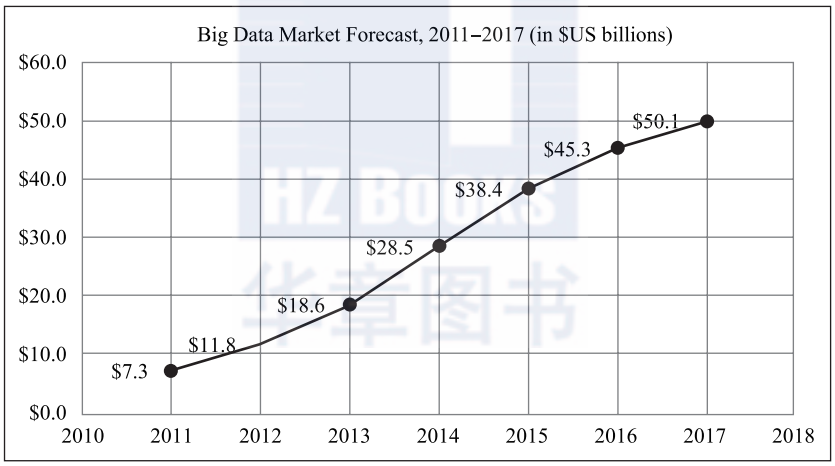

图 2-2　美国大数据市场的预测（来源：Wikibon 2014）

大数据所带来的巨大经济价值吸引着各行各业的人，但是大数据是把双刃剑，频繁出现的隐私泄露事件不仅给相关从业者敲响了警钟，也时刻提醒着用户在享受大数据应用带来的各种便利的同时，需要更加注重保护个人隐私。一些机构通过大数据收集个人信息并使用，已经造成了一系列的问题。

近年来，大数据安全事件呈现高发之势。中国互联网协会发布的《2015 中国网民权益保护调查报告》显示，63.4% 的个人网上活动信息被泄露过，78.2% 的个人身份信息被泄露过，因个人信息泄露、垃圾信息、诈骗信息等现象导致的总体经济损失达到了 805 亿元。

2015 年 8 月，线上票务营销平台大麦网被发现存在安全漏洞，600 余万用户账户和密码遭到泄露，并且这些隐私数据已被黑产行业进行售卖与传播。

2016 年 10 月，我国警方在广东破获一起高科技经济犯罪案件，犯罪嫌疑人攻破了多个商业银行网站，并且窃取了储户的身份证号、银行卡号、支付密码等数据，同时组织一批人在网上大肆盗刷别人的信用卡，涉案金额近 15 亿元，涉及银行 49 家。

以上案例说明：一些企业和个人利用职务和技术之便，蓄意搜集公民个人隐私数据，然后出售给有需要的客户，达到自己赚钱的目的，或由自己直接实施欺诈行为以获利。而购买他人数据隐私信息者，则利用购得的信息发送广告，进行产品推广，或者进行诈骗等犯罪活动。

大数据环境下，类似的大数据隐私泄露安全事件不胜枚举。这些安全事件不仅对用户造成了严重的安全隐患，同时也给相关行业带来了巨大的经济损失。保护大数据隐私在提倡“开放、共享”的大数据时代显得格格不入，但是对于大数据隐私的保护已经迫在眉睫。大数据隐私保护不仅对于用户有重大意义，同时相关企业也可以从用户群中获取更多的信任感，从而获取更多用户的数据，并且发掘出用户的潜在需求，不断改善产品，获取用户好感。大数据隐私带来的保护措施，保证了数据获取与利用的可靠性，企业可以更容易地从用户中获取数据，也能更好地从数据中发掘出对用户有价值的信息，反馈给用户，实现大数据的良性循环利用。

2.2.2 大数据隐私保护面临的问题与挑战

大数据研究需要数据开放与共享。但是在现实情况中，大量数据处于闲置、无人问津的状态。因为数据的开放和共享可能会导致隐私的泄露，很多数据拥有者或者管理者不敢或不愿开放、共享数据，导致工业界有数据、缺技术，而学术界有技术、缺数据。

与此同时，人们面临的威胁并不仅限于个人隐私泄露，还有基于大数据对人们状态和行为的预测。一个典型的例子是，国外某零售商通过对历史记录进行分析，比家长更早知道其女儿已经怀孕的事实，并向其邮寄相关广告信息。此外，社交网络分析研究也表明，可以通过其中的群组特性发现用户的属性。例如通过分析用户的社交平台信息，可以发现用户的生活方式、消费习惯以及业务爱好等。

同时，搭建大数据环境，无论是硬件还是软件或者其他细节，都需要投入大量的资金，因此大部分公司不会去花费精力搭建自己的大数据基础设施，所以微软的 Azure 和阿里的阿里云等成了一部分公司搭建云计算平台的首选。在这种情况下，云计算的安全性也需要考虑。[⊖]

在这样的背景下，隐私保护被提出，也是大数据应用成功的关键因素之一。可以通过隐私保护技术，寻求一些使用数据的方法，既不妨碍第三方从带有敏感信息的数据集中获

⊖ http://www.business.com/technology/privacy-and-security-issues-in-the-age-of-big-data/

取信息，同时又可避免隐私资料的泄露。

大数据时代的隐私是由于数据融合、数据分析、数据过度收集等造成的，这与传统的隐私泄露问题有本质的不同。大数据隐私管理要服务于数据治理的需要，其本质是要保证数据的正确使用和交易。目前大数据隐私保护面临以下几个问题：

1）大数据依托的 NoSQL（非关系型数据库）缺乏数据安全机制。从基础存储技术角度来看，大数据依托的基础技术是 NoSQL。当前广泛应用的 SQL（关系型数据库）技术，经过长期改进和完善，在维护数据安全方面已经形成严格的权限访问控制和隐私管理工具。而在 NoSQL 技术中，并没有这样严格的要求。大数据的数据来源和承载方式多种多样，数据处于分散的状态，使企业很难定位和保护所有这些私密数据。NoSQL 允许不断对数据记录添加属性，其前瞻安全性变得非常重要，同时也对数据库管理员提出了新的要求。

2）社会工程学攻击带来的安全问题。社会工程学的特点是：不需要专业技术，成本低，效率高。该攻击与其他攻击的最大不同是其攻击手段不是利用高超的攻击技术，而是利用受害者的心理弱点进行攻击。因为不管大数据多么庞大总也少不了人的管理，如果人的信息安全意识淡薄，那么即使在技术上防护手段已做到无懈可击，也没办法有效保障数据安全。由于大数据的海量性、复杂性，以及攻击目标不明确，因此攻击者为了提高效率，经常采用社会工程学攻击。此类攻击的案例很多，如黑客先攻击某论坛的网站，使用户无法正常登录。然后再假冒管理员，以维护网站名义向用户发送提醒信息，索要用户的账号和密码，一般安全意识不强的用户此时会将密码和账号发送给黑客。除此以外，还有采用冒充中奖、假冒社交好友、信用卡挂失等欺诈手段获得合法用户信息。

3）软件后门。在软件定义世界的时代，软件是 IT 系统的核心，也就是大数据的核心，所有的后门可能都是开放在软件上面的。据了解，IBM、EMC 等各大巨头生产制造的存储、服务器、运算设备等硬件产品，几乎都是全球代工的，在信息安全的监听（硬件）方面是很难做手脚的。换句话说，软件才是信息安全的软肋所在。软件供应方在软件上设计了特殊的路径处理，测试人员只按照协议上的功能进行测试，根本就无法察觉软件预留的监听后门。换言之，如果没有自主可控的信息安全检测方案，各种安全机制和加密措施就都形同虚设。此类安全事件繁多，比如，2015 年，由于部分 iOS 开发者使用了非官方渠道的带后门的开发软件 Xcode，导致微信、网易云音乐等 iOS 版本的应用软件存在后门程序，该程序可以上传用户的系统信息到黑客指定的服务器，对大量应用软件造成了巨大的安全隐患。所以，近期代码审计在安全领域是非常重要的。对于现代信息安全而言，最危险的行为是将自主控制的权力交给“他人”。这就好比将自家的钥匙全部交到了外人手里，安全问题又从何谈起呢？

4）大数据存储问题。大数据会使数据量呈非线性增长，而复杂多样的数据集中存储在一起，多种应用的并发运行以及频繁无序的使用状况，有可能会导致数据类别存放错位的情况，造成数据存储管理混乱或信息安全管理不合规范。现有的存储和安全控制措施无法满足大数据安全需求，安全防护手段如果不能与大数据存储和应用安全需求同步升级，就

会出现大数据存储安全防护的漏洞。

5）文件的安全面临极大挑战。文件是整个数据和系统运行的核心。大多数的用户文件都是在第三方的运行平台中存储和处理的，这些文件往往包含了很多部门和个人的敏感信息，其安全性和隐私性自然成为重要的问题。尽管文件的保护提供了对文件的访问控制和授权，例如 Linux 自带的文件访问控制机制，通过文件访问控制列表来限制程序对文件的操作。然而大部分文件保护机制都存在一定程度上的安全问题，它们通常使用操作系统的功能来实现完整性验证机制，因此只依赖于操作系统本身的安全性。但是作为网络攻击，操作系统才是最大的一个攻击点。

6）大数据安全传输的问题。大数据安全传输的问题涉及通信网络的安全、用户兴趣模型的使用安全和私有数据的访问控制安全，既包括传统搜索过程中可能出现的网络安全威胁，比如相关信息在网络传输时被窃听，以及恶意木马、钓鱼网站等，也包括服务器端利用通信网络获取用户隐私的危险。

7）大数据支撑平台云计算安全。云计算的核心安全问题是用户不再对数据和环境拥有完全且直接的控制权，云计算的出现彻底打破了地域的概念，数据不再存放于某个确定的物理节点，而是由服务商动态提供存储空间。这些空间有可能是现实的，也可能是虚拟的，甚至可能分布在不同国家及地区。用户对存放在云中的数据不能像从前那样具有完全的管理权，相比传统的数据存储和处理方式，云计算时代的数据存储和处理，对于用户而言变得非常不可控。云计算环境中用户数据安全与隐私保护难以实现。

8）大数据分析预测带来的用户隐私挑战。从核心价值角度来看，大数据关键在于数据分析和利用，但数据分析技术的发展，对用户隐私带来极大的威胁。在大数据时代，想屏蔽外部数据商挖掘个人信息几乎是不可能的。目前，各社交网站均不同程度地开放其用户所产生的实时数据，这些数据被一些数据提供商收集，甚至是一些监测数据的市场分析机构。例如国外的 Yelp 网站提供用户脱敏后的数据，供学者分析[⊖]。通过人们在社交网站中写入的信息、智能手机显示的位置信息等多种数据组合，已经可以以非常高的精度锁定个人，挖掘出个人信息体系，使得用户隐私安全问题失去保障。

9）大数据共享所带来的安全性问题。我们不知道该如何分享私人数据，才能既保证数据隐私不被泄露，又保证数据的正常使用。真实数据大部分不是静态的，而是越变越大，并且随着时间的变化而变化。许多在线服务要求人们共享私人信息，但是，在记录级的访问控制之外，人们根本不知道共享数据意味着什么，不知道共享后的数据会怎样被连接起来，更不知道如何让用户对共享后的数据仍能进行细粒度控制。

10）大数据访问控制的安全性问题。访问控制是实现数据受控共享的有效手段，由于大数据可能被用于多种不同场景，其访问控制需求十分突出，难以预设角色，实现角色划分。由于大数据应用范围广泛，它通常要被来自不同组织或部门、不同身份与目的的用户

⊖ https://www.yelp.com/dataset_challenge

所访问，实施访问控制是基本需求。然而，在大数据的场景下，有大量的用户需要实施权限管理，且用户具体的权限要求未知。面对未知的大量数据和用户，预先设置角色十分困难。同时，难以预知每个角色的实际权限。面对大数据，安全管理员可能无法准确地为用户指定其可以访问的数据范围，而且这样做效率不高。不同类型的大数据存在多样化的访问控制需求。例如，在 Web 2.0 个人用户数据中，存在基于历史记录的访问控制；在地理地图数据中，存在基于尺度以及数据精度的访问控制需求；在流数据处理中，存在数据时间区间的访问控制需求等。如何统一描述与表达访问控制需求是一个挑战。

2.2.3　大数据隐私保护技术

目前用户数据的收集、存储、管理与使用等均缺乏规范，更缺乏监管，主要依靠企业的自律。用户无法确定自己的隐私在何时何地被何人使用。在实际的商业化场景中，用户应有权决定自己的信息被如何利用。实现用户可控的隐私保护包括：数据在采集时的隐私保护，如数据精度控制处理；数据在共享、发布时的隐私保护，如数据的匿名处理、人工干扰等；数据在分析时的隐私保护；数据生命周期的隐私保护；隐私数据可信销毁。面对频发的隐私泄露事件，隐私保护问题需要得到有效的解决。解决的途径包括研发技术方法和制定法律法规。

在技术方面，隐私保护的研究方向包括威胁发现技术、大数据认证技术、数据真实性分析技术、数据失真处理技术、数据加密技术和限制发布技术。

（1）威胁发现技术

利用该技术，企业可以超越以往的保护（Protection）—检测（Detection）—响应（Reaction）—恢复（Recovery）（PDRR）模式，更主动地发现潜在的安全威胁。相比于传统技术，基于大数据的威胁发现技术具有分析的内容范围更大的优点。

企业信息资产包括数据资产、软件资产、实物资产、人员资产、服务资产和其他为业务提供支持的无形资产。由于传统威胁检测技术并不能覆盖这 6 类信息资产，因此所能发现的威胁有限。而通过在威胁检测方面引入大数据分析技术，能全面发现针对这些信息资产的攻击，而且分析内容的时间跨度更长。现有威胁分析技术具有内存关联性，即实时收集数据，采用分析技术发现攻击。分析窗口通常受限于内存大小，无法应对持续性和潜伏性攻击。而引入大数据分析技术后，威胁分析窗口可以横跨若干年的数据，因此威胁发现能力更强，可以有效应对 APT 类攻击。

相比传统技术，基于大数据的威胁发现技术还有以下优点：

1）攻击威胁的预测性。传统安全防护技术大多是在攻击发生后对攻击行为进行分析和归类，并做出响应。而基于大数据的威胁分析，可进行超前的预判，对未发生的攻击行为进行预防。

2）对未知威胁的检测。传统的威胁分析常由经验丰富的专业人员根据企业需求和实际情况展开，威胁分析结果在很大程度上依赖于个人经验，分析所发现的威胁是已知的。而

大数据分析的特点是侧重于普通的关联分析，而不侧重因果分析，因此通过采用恰当的分析模型，可发现未知威胁。

（2）大数据认证技术

大数据认证技术指的是收集用户行为和设备行为数据，并对这些数据进行分析，获得用户行为和设备行为的特征，进而通过鉴别操作者行为及其设备行为来确定其身份。这与传统认证技术利用用户所知秘密、所持有凭证或具有的生物特征来确认其身份有很大不同。该技术具有如下优点：

1）攻击者很难模拟用户行为特征来通过认证，因此更加安全。利用大数据技术所能收集的用户行为和设备行为数据是多样的，可以包括用户使用系统的时间、经常采用的设备、设备所处物理位置，甚至是用户的操作习惯等数据。通过这些数据的分析，能够为用户勾画一个行为特征的轮廓。而攻击者很难在方方面面都模仿用户行为，因此其与真正用户的行为特征轮廓必然存在一个较大偏差，无法通过认证。

2）减小了用户负担。用户行为和设备行为特征数据的采集、存储和分析都由认证系统完成。相比于传统认证技术，这极大地减轻了用户负担。例如，用户无须记忆复杂的口令，或随身携带硬件 USBKey。可以更好地支持各系统认证机制的统一。基于大数据的认证技术可以让用户在整个网络空间采用相同的行为特征进行身份认证，可避免传统的不同系统采用不同认证方式，且用户所知秘密或所持凭证各不相同而带来的种种不便。

（3）数据真实性分析技术

目前，基于大数据的数据真实性分析被广泛认为是最为有效的方法。许多企业已经开始了这方面的研究工作，如 Yahoo 和 Thinkmail 等利用大数据分析技术来过滤垃圾邮件；Yelp 等社交点评网络用大数据分析来识别虚假评论；新浪微博等社交媒体利用大数据分析来鉴别各类垃圾信息等。

基于大数据的数据真实性分析技术能够提高垃圾信息的鉴别能力：一方面，引入大数据分析可以获得更高的识别准确率。例如，对于点评网站的虚假评论，可以通过收集评论者的大量位置信息、评论内容、评论时间等进行分析，鉴别其评论的可靠性。如果某评论者对某品牌多个同类产品都发表了恶意评论，则其评论的真实性就值得怀疑。另一方面，在进行大数据分析时，通过机器学习技术，可以发现更多具有新特征的垃圾信息。然而该技术仍然面临一些困难，主要是虚假信息的定义、分析模型的构建等。

（4）数据失真处理技术

通过添加噪声等方法，使敏感数据失真但同时保持某些数据或数据属性不变，即仍然保持某些统计方面的性质。

（5）数据加密技术

采用加密技术在数据挖掘过程中隐藏敏感数据，即使得两个或多个站点通过某种协议完成计算后，每一方都只知道自己的输入数据和所有数据计算后的最终结果。此外，还包括分布式匿名化，即保证站点数据隐私、收集足够的信息实现利用率尽量大的数据匿名。

（6）限制发布技术

有选择地发布原始数据，不发布或者发布精度较低的敏感数据，实现隐私保护。当前这类技术的研究集中于“数据匿名化”，保证对敏感数据及隐私的披露风险在可容忍范围内，包括 *k*-anonymity、L-diversity、T-closeness。

最早被广泛认同的隐私保护模型是 *k*-anonymity（*k*- 匿名），它由 Samarati 和 Sweeney 在 2002 年提出，作者正是美国马萨诸塞州医疗数据隐私泄露事件的攻击者。为应对去匿名化攻击，*k*-anonymity 要求发布的数据中每一条记录都要与其他至少 *k*–1 条记录不可区分（称为一个等价类）。当攻击者获得 *k*- 匿名处理后的数据时，将至少得到 *k* 个不同人的记录，进而无法做出准确的判断。参数 *k* 表示隐私保护的强度，*k* 值越大，隐私保护的强度越强，但丢失的信息也会越多，数据的可用性越低。

然而，美国康奈尔大学的 Machanavajjhala 等人在 2006 年发现了 *k*- 匿名的缺陷，即没有对敏感属性做任何约束，攻击者可以利用背景知识攻击、再识别攻击和一致性攻击等方法来确认敏感数据与个人的关系，从而导致隐私泄露。例如，攻击者获得 *k*-anonymity 的数据，如果被攻击者所在的等价类中都是艾滋病病人，那么攻击者很容易做出被攻击者肯定患有艾滋病的判断（上述就是一致性攻击的原理）。为了防止一致性攻击，新的隐私保护模型 L-diversity 改进了 *k*-anonymity，保证任意一个等价类中的敏感属性都至少有一个不同的值。T-closeness 在 L-diversity 的基础上，要求所有等价类中敏感属性的分布尽量接近该属性的全局分布。(*a*, *k*)- 匿名原则则在 *k*- 匿名的基础上，进一步保证每一个等价类中与任意一个敏感属性值相关记录的百分比不高于 *a*。

然而，上述隐私保护模型依然有缺陷，需要被不断改进，但同时又有新的攻击方法出现，使得基于 *k*- 匿名的传统隐私保护模型陷入这样一个无休止的循环中。从根本上来说，传统隐私保护模型的缺陷在于对攻击者的背景知识和攻击模型都给出了过多的假设。但这些假设在现实中往往并不完全成立，因此攻击者总是能够找到各种各样的攻击方法来进行攻击。直到差分隐私的出现，这一问题才得到较好的解决。

在法律法规方面，欧美早在 20 世纪 70 年代就有专门的隐私保护法。香港地区在回归之前就颁布实施了个人数据条例。该条例于 1995 年颁布，1996 年 12 月 20 日生效。条例的执行由个人数据隐私专员监督。该条例管理个人、企业、公共机构和政府部门对于在世人士的相关数据的使用（如果这些数据可以有效识别该在世人士）。香港的个人数据条例主要强调了数据保护的六大原则：个人数据收集的目的和方式、个人数据的准确性和数据保留的时间、个人数据的使用、个人数据的安全性、信息基本有效可用、个人数据的访问。我国内地虽然没有专门的隐私保护法，但在多个法律法规的条文中涉及了隐私保护，对保护个人隐私做了间接的、原则性的规定。例如，《中华人民共和国宪法》第三十八条、第三十九条、第四十条明确了对公民的人格尊严、住宅、通信自由和通信秘密的保护，这是我国法律对隐私权进行保护的最根本依据。第三十八条规定：“中华人民共和国公民的人格尊严不受侵犯。禁止用任何方法对公民进行侮辱、诽谤和诬告陷害。”这些法律规定对于保

护公民的隐私权具有重要意义。

2.3 大数据治理审计

2.3.1 大数据治理审计概述

1. 审计的含义

审计是由国家授权或接受委托的专职机构和人员，依照国家法规、审计准则和会计理论，运用专门的方法，对被审计单位的财政、财务收支、经营管理活动及其相关资料的真实性、正确性、合规性、合法性、效益性进行审查和监督，评价经济责任，鉴证经济业务，用以维护财经法纪、改善经营管理、提高经济效益的一项独立性的经济监督活动。

审计是对资料做出证据搜集及分析，以评估企业财务状况，然后就资料及一般公认准则之间的相关程度做出结论及报告。进行审计的人员必须有独立性及具相关专业知识。常见的财务审计有以下 3 种。

1）运作审计（作业审计）：检讨组织的运作程序及方法，以评估其效率及效益。

2）履行审计（遵行审计）：评估组织是否遵守由更高权力机构所制订的程序、守则或规条。

3）财务报表审计：评估企业或团体的财务报表是否根据公认会计准则编制，一般由独立会计师进行。

在香港地区，财务报表审计亦称为核数，而会计师事务所则俗称为会计师楼。在台湾地区，财务报表审计亦称为查核、查账。

审计工作一直得到各国政府和社会的重视。传统手工审计是通过对纸质账簿的检查来实现这一职责的。20 世纪 80 年代，以查账为主要手段的审计职业遇到了信息技术的挑战。传统审计面临着“打不开账，进不了门，审不了数”的困境。随着被审计单位信息化趋向普及，审计对象的信息化使得审计信息化成为必然。审计信息化对审计人员和审计工作的开展提出了更高的要求。

2. 大数据时代下的审计发展趋势

随着信息技术的发展，大数据时代为审计提供了机遇和挑战。

当今，大数据伴随着云计算、移动互联网的发展，正在对全球经济社会产生巨大的影响。大数据给现代审计提供了新的技术和方法，要求人们把握大数据的特点，变革现代审计的思维与技术和方法，推动大数据时代审计的发展。

大数据的精髓在于促使人们在采集、处理和使用数据时思维的转变，这些转变将改变人们理解和研究社会经济现象的技术和方法。

1）不再依赖抽样分析方法。在大数据时代，可以收集和处理事物所有的数据。自古以来，当面临大的样本量时，人们都依赖于抽样分析。但是，抽样分析是在信息缺乏和取得

信息受限制的条件下采用的一种方法，这其实是一种人为的限制。如今，科学技术条件已经有了很大的改善，计算机能够处理的数据量已经大大增加，所以现在人们可以收集和处理所有的数据。

2）在大数据时代，人们不再热衷于寻找事物的因果关系，而是充分利用事物的相关关系。寻找因果关系是人类长期发展过程中形成的习惯。相关关系也许不能准确地告知某件事情为何会发生，但是它会提醒人们这件事情正在发生，在许多情况下，这种提醒的帮助作用已经足够大了。

3）不再热衷于追求数据的精确度，而是追求利用数据的效率。当测量事物的能力受限制时，人们关注的是获取最精确的结果。但是，在大数据时代，当拥有海量数据时，大数据纷繁多样，优劣掺杂，绝对的精准不再是人们追求的主要目标，更重要的是追求数据的及时性和使用效率。

一个拥有巨大潜力的领域是审计转型，下一代的审计师需要有 IT 相关的知识以及传统的财务审计能力。从传统的审计方法到以无缝的方式充分整合大数据分析，这将是一个巨大的飞跃。但目前大数据治理审计仍处于起步阶段。

面对大数据所带来的新思维、新技术和方法的变革，会计、审计人员需要应时而变，以适应思维模式及数据处理模式的变化。大数据对会计、审计发展的影响主要表现在以下几个方面：

（1）从事后的财务报告向实时财务报告发展

传统会计工作中，会计人员只是在企业生产经营业务发生后才编制财务报告，而且财务报告编制过程漫长，年度财务报告一般用三四个月时间才能完成编制，这严重影响了会计信息的及时性和利用效率。随着信息技术迅速发展，越来越多的人意识到实时财务报告的重要性，而大数据技术使实时财务报告成为可能。实时财务报告是信息技术与大数据技术较好地交叉融合的产物，是信息化条件下会计技术和方法发展的必然产物。尤其对业务数据和风险控制“实时性”要求较高的特定行业，如银行、证券、保险等行业，实施实时财务报告迫在眉睫。

（2）从抽样审计模式向总体审计模式发展

抽样审计模式由于抽取样本的有限性而忽视了大量的业务活动，无法完全发现和揭示被审计单位的重大舞弊行为，隐藏着严重的审计风险。在大数据时代，数据的跨行业、跨企业搜集和分析，可以不用随机抽样方法，而采用搜集和分析被审计单位所有数据的总体审计模式。大数据环境下的总体审计模式是分析与审计对象相关的所有数据，使得审计人员可以建立总体审计的思维模式。

（3）从单一审计报告向综合审计成果应用发展

目前，审计人员的审计成果主要是提供给被审计单位的审计报告，其格式固定，内容单一，包含的信息较少。随着大数据技术在审计中的广泛应用，审计人员的审计成果除了审计报告外，还包括在审计过程中采集、挖掘、分析和处理的大量资料和数据，可以提

供给被审计单位用于改进经营管理，促进审计成果的综合应用，提高综合审计成果的应用效果。

审计人员对大数据技术的应用，促进了审计成果的进一步综合应用。首先，审计人员通过对审计中获取的大量数据进行汇总、归纳，从中找出内在规律、共性问题和发展趋势，为被审计单位投资者和其他利益相关者提供数据证明、关联分析和决策建议。其次，审计人员通过应用大数据技术，从不同的角度、不同的层面整合提炼，以满足不同层次的需求。再次，审计人员将审计成果进行智能化留存，通过大数据技术，将问题规则化并固化到系统中，以便于计算或判断问题发展趋势。最后，审计人员将审计成果与被审计单位进行关联，可以减少实地审计的时间和工作量，提高审计工作的效率。

（4）从精确的数字审计向高效的数据审计发展

直到今天，审计人员的数字审计技术依然建立在精准的基础上。这种思维方式适用于掌握“小数据量”的情况，因为需要分析的数据很少，所以审计人员必须尽可能精准地量化被审计单位的业务。相比依赖于小数据和精确性的时代，大数据因为更强调数据的完整性和混杂性，帮助审计人员进一步接近事情的真相，“局部”和“精确”将不再是审计人员追求的目标，审计人员追求的是事物的“全貌”和“高效”。

在大数据环境下，传统的很多审计技术和方法显得效率低下和无法实施，大数据时代的超大数据体量以及占相当比例的半结构化和非结构化数据的存在，已经超越了传统数据库的管理能力，必须使用新的大数据存储、处理和检索方法。围绕大数据，一批新兴的数据挖掘、数据存储、数据处理与分析技术涌现出来。在实施审计时，审计人员应使用分布式拓扑结构、云数据库、联网审计、数据挖掘等新型的技术手段和工具，以提高审计的效率。

3. 什么是大数据治理

在各行各业中，随处可见因数量、速度、种类和准确性结合带来的大数据问题，为了更好地利用大数据，大数据治理逐渐提上日程。在传统系统中，数据需要先存储到关系型数据库 / 数据仓库后再进行各种查询和分析，这些数据称为静态数据。而在大数据时代，除了静态数据以外，还有很多数据对实时性要求非常高，需要在采集数据时就进行相应的处理，处理结果存入关系型数据库 / 数据仓库、MPP 数据库、Hadoop 平台、各种 NoSQL 数据库等，这些数据称为动态数据。比如高铁机车的关键零部件上装有成百上千个传感器，每时每刻都在生成设备状态信息，企业需要实时收集这些数据并进行分析，当发现设备可能出现问题时及时告警。再比如在电信行业，基于用户通信行为的精准营销、位置营销等，都会实时地采集用户数据，并根据业务模型进行相应的营销活动。

数据治理是指在企业数据整个生命周期（从数据采集到数据使用直至数据存档）制定由业务推动的数据政策、数据所有权、数据监控、数据标准以及指导方针。数据治理的重点在于，要将数据明确地作为企业的一种资产看待。

大数据治理的核心是为业务提供持续的、可度量的价值。大数据治理人员需要定期与

企业高层管理人员进行沟通，保证大数据治理计划可以持续获得支持和帮助。相信随着时间的推移，大数据将成为主流，企业可以从海量的数据中获得更多的价值，而大数据治理的范围和严格程度也将逐步上升。

4. 大数据治理审计的含义与目的

大数据治理审计是指独立于审计对象的审计人员，以第三方的客观立场对大数据治理过程进行综合检查与评价，向审计对象的最高领导层提出问题与建议的一连串活动。

大数据治理审计的目的是了解组织大数据治理活动的总体状况，对组织是否实现大数据治理目标进行审查和评价，充分识别与评估相关治理风险，提出评价意见及改进建议，促进组织实现大数据治理目标。

5. 大数据治理审计的特点

大数据治理审计除了具有传统审计的权威性、客观性、公正性特点之外，还具有一些独有的特点，主要包括：

1）与传统审计的目的不同。传统审计的目的是“对被审计单位会计报表的合法性、公允性及会计处理方法的一贯性发表审计意见”。上面提到过，大数据治理审计的目的是对组织是否实现大数据治理目标进行审查和评价，充分识别与评估相关风险，提出评价意见及改进建议。

2）大数据治理审计是事前、事中和事后审计的结合体。传统审计中的财务报表审计往往是年度审计，属于事后审计，而大数据治理审计是事前、事中和事后审计兼而有之。由审计人员所进行的大数据治理规划审计属于事前审计，大数据治理实施过程中的审计属于事中审计，而对其在一定期间的运作情况所进行的审计属于事后审计。

3）大数据治理审计促使传统审计模式发生改变。抽样审计是传统的审计模式，即在不可能收集和分析被审计单位全部数据的情况下，主要依赖于抽样技术针对抽取的样本进行审计，并由此推断审计对象的整体情况。大数据时代能够收集和分析组织的所有相关数据，审计模式发生了改变，已从抽样审计向总体审计模式发展，即对大数据总体进行多角度的深层次分析，以发现其中隐藏的更具价值的信息及判断总体的特征，克服了抽样审计模式的不足。

4）运用了大数据分析技术。运用大数据分析技术是大数据治理审计特征之一，即依托大数据分析平台，开展组织内部业务数据与财务数据的治理审计工作。

5）更重视大数据信息的安全性。在信息化高度发展的时代，大数据的安全性关系着组织的命运、社会的稳定及国家的安全。组织应采取各种措施（如管理措施、技术措施及物理措施）保护大数据的安全。

2.3.2　大数据治理审计内容

大数据治理审计是从审计的视角对大数据治理进行监督和评价。简单来说，大数据治

理是为了保证数据质量，而大数据治理审计是为了保证大数据治理的质量。

大数据治理审计的内容如下。

1. 大数据治理战略目标审计

不同行业、不同公司的大数据治理战略目标不同。对大数据治理战略目标进行审计，一方面可以确保大数据治理的目标符合行业或者企业需求，另一方面可以向管理层提供大数据治理战略规划，使得大数据治理过程得到控制和监督。例如，大数据治理战略目标的审计内容为：

1）本次大数据治理的目标是什么？

2）本次大数据治理的目标是否合理？

3）本次大数据治理的目标是否符合该企业的状况？

4）是否制定了大数据治理的计划？

5）该计划是否合理？

2. 大数据治理内容审计

大数据治理内容审计非常重要，大数据治理的目标是什么决定着大数据治理的内容。对大数据治理内容进行审计，可以更好地建立健全大数据治理审计系统。大数据治理审计内容如下：

1）大数据治理审计内容是什么？

2）大数据治理审计内容是否有助于实现本组织的大数据治理目标？

3. 大数据治理架构审计

每一个大数据治理系统都有特有的框架结构，通过对大数据治理系统框架的审计，可以在一定程度上使大数据治理系统的架构更加合理。大数据治理架构审计内容如下：

1）该企业建立的大数据治理框架结构是什么？

2）该大数据治理框架结构是否与该企业的大数据治理战略目标一致？

3）该大数据治理框架结构是否满足该企业大数据治理内容需求？

4. 大数据安全审计

数据安全可靠是大数据的根本。如果数据都不可靠，再多的数据量也没有什么意义。大数据安全审计通过对大数据安全相关的评价，确保大数据安全。

5. 大数据生命周期管理审计

数据生命周期管理（Data Life cycle Management，DLM）是一种基于策略的方法，用于管理信息系统的数据在整个生命周期内的流动：从创建和初始存储，到它过时被删除。数据生命周期管理产品涉及过程自动化，通常根据指定的策略将数据组织成各个不同的层，并基于那些关键条件自动地将数据从一个层移动到另一个层。

对大数据生命周期管理进行审计，可以将整个大数据生命周期管理的自动化过程变得

更加透明，更加可靠。对于企业来说，对大数据生命周期管理进行审计，可以使整个大数据生命周期管理机制得到控制。

2.3.3　大数据治理审计方法和技术

1. 大数据治理审计标准规范

与会计审计遵循《审计准则》一样，大数据审计需要有一套共同遵循的审计规范。物联网、云计算快速发展带来大数据审计的需要，各国政府、协会或民间组织也积极关注并推行大数据审计的规范。一般说来，大数据治理审计主要存在于信息审计或云计算的审计规范之中。当前国外主要信息审计的相关标准如下：

信息系统审计与控制基金会在 1996 年制定的 IT 治理模型（CO BIT），是国际公认的、权威的安全与信息技术管理和控制的标准，也是国际上通用的信息系统审计的标准之一。它的宗旨是跨越业务和 IT 控制之间的鸿沟，建立一个面向业务目标的 IT 控制框架。特别是在最新的 CO BIT5.0 版本中，被称为“一个治理和管理企业 IT 的业务框架”，它是 IT 技术人员、用户、企业管理人员和 IT 审计师之间的桥梁[⊖]。

美国国家标准与技术学院（NIST）不仅发布了被广泛引用的《云计算定义》，还发布了《联邦信息系统和机构的信息安全持续监测》（ISC M）报告，提出：通过持续监测，保持其对信息安全、漏洞和威胁的警觉。

美国云安全联盟 CSA 在 2009 年 12 月发布了《云安全指南》。它涵盖了“云计算重点 13 个区域的安全指导”，从云用户角度阐述了可能存在的商业隐患、安全威胁，以及推荐采取的安全措施。

ISACA 是国际信息系统审计协会在 2010 年推出的云计算管理审计、保证程序（Cloud Computing Management Audit/Assurance Program），规定了审计过程中使用的工具、模板以及流程。同时，ISACA 还在程序中规定了审计过程中应该关注的审查点以及遵循的标准，从而保证审计师能够完整、真实地记录有关数据。它主要关注云计算治理的影响、服务供应商以及客户之间的合同履约、云计算控制的具体问题等。如数据审计的审计目标是：为云计算服务提供商的客户提供对服务提供商内部控制的有效性和安全性评估；识别客户组织其他与服务提供商的接口是否存在内部控制缺陷；评估客户的质量和能力情况与服务提供商的内部控制项相关的证明。

其他的信息审计标准还有欧洲网络与信息安全局的《云计算风险评估方法论》、ISO 27001 等。

在我国，由于物联网与云计算等信息化发展相对落后，至今尚未有大数据治理审计的标准，可以参考的主要有 2008 年五部委共同颁布的《企业内部控制规范》和 2009 年银监会颁布的《商业企业信息科技风险管理指引》。

⊖　许金叶，许琳 . 大数据审计：物联网建设的制度保障 [J]. 会计之友，2013(33):118-121.

2. 大数据治理审计方法

审计方法是指为完成审计任务、实现审计目标，所采用的各种技术手段的总称。同时，审计方法也可以说是沟通审计主体和审计客体的桥梁，是审计程序的支柱，是审计过程趋于合理、有效的灵魂。注重审计方法的目的在于：有利于提高审计工作效率和工作质量，抓住问题的实质，以便更好地完成审计任务。

（1）传统审计方法

传统的审计方法又可以细分为：审查书面资料的方法、审查财产物资的方法、审计的分析法、审计抽样方法。

审查书面资料的方法，按照资料形成的顺序，可分为顺查法、逆查法；按照审计的范围，可分为详查法、抽查法；按照资料内容的不同，可分为审阅法、核对法、复算法、查询法、分析法、推理法等。

审计人员在实施审计过程中，经常需要证实被审查事物的性质、形态、数量、价值等是否真实、正确、合理，一般采用一些特殊的方法，如盘点法、调节法、鉴定法、观察法等。人们把这些方法称为证实客观事物的方法。

审计的分析法，也称为分析性复核法，是指审计人员在审计过程中，对审计事项的相关指标进行对比、分析和评价，以便发现其中有无问题或异常情况，为进一步审计提供线索的一种审计方法。常用的分析法有：比较分析法、比率分析法、平衡分析法、趋势分析法、账户分析法、账龄分析法等。

统计抽样审计的基本程序一般分为 3 个阶段，即样本设计阶段、样本选取阶段和抽样结果的评价阶段。

（2）IT 内部审计方法

审计信息化（IT 审计）也可称为信息系统审计或计算机辅助审计。IT 内部审计作为 IT 治理的分支，本质是为了促进 IT 治理目标的实现，实现 IT 资源的价值增值。下面主要是选用商业银行 IT 治理的案例来丰富大数据时代 IT 内部审计概念。

现阶段商业银行 IT 治理雏形可从以下 5 个核心要素进行考察：

1）在 IT 战略部署上，以商业目标、IT 风险及 IT 投资成本为重心；

2）IT 价值交付是指在考虑时间价值的基础上，确保 IT 投资与价值回报效率；

3）在风险管理方面，信息系统风险的控制在 COSO 协议与 Basel 协议双重保障下相对其他行业已发展成熟，对操作风险的控制更为专业；

4）信息系统绩效评价可借助 IT 内部审计与 IT 平衡积分卡来完成；

5）IT 资源包括相关的人力资源及软硬件资源。

（3）大数据审计方法

大数据下的审计主要是指审计人员利用大数据资源，通过大数据方法，找到大数据与被审计单位的联系，验证其经济活动的合法合规性。其具有以下特点：

1）所有数据都将成为被分析的对象。

即使面对大量的样本，也不再使用抽样的方法，降低了审计风险，提高了审计结果的准确性。例如，大数据环境下，中石油所有的收购信息都将以数据的形式保存在数据库中，通过带有限定条件的数据库语言可以将所有收购加油站中的手续不全的商家查找出来，以免其中有些成为漏网之鱼。

2）充分利用外部数据。

大数据为被审计单位获取和利用外部数据创造了条件，可以解决传统审计难以获取、利用外部数据的固有弊端。利用外部数据可以从更多的视角发现可能出现的问题，提高审计效率和准确性。大数据下，通过数据挖掘算法可找出作为企业外部数据的政府土地数据，发现其中的违规行为并及时制止。

3）不需要函证。

大数据下环境，审计人员通过权限可以获取被审计单位的往来单位和往来银行的相关数据。直接通过原始数据便可完成审计工作，不需要通过函证来证实被审计单位的相关经济活动是否真实完整。这就减少了被审计单位与第三方单位舞弊的可能性，也节约了时间和人力、物力。

4）不受时间地点限制。

大数据下的审计工作主要在互联网上进行，审计人员通过权限获取相应的数据，并对其进行分析，得出结论。由于大部分证据都存储在网络数据库中，审计人员不需要在被审计单位工作，也不用固定工作时间，只需一台计算机和网络环境便可进行工作，增加了审计的灵活性，提高了效率。

3. 大数据治理审计技术

大数据环境下，企业能够提供更多、更全面的数据，企业可以充分利用采集来的各方面数据建立集中统一的被审计单位数据中心。在此基础上，借助不同于传统 SQL 关系数据库的新的大数据分析技术，构建审计大数据分析平台和使用更智能的大数据分析技术，通过分析“从数据入口到数据库平台”的更大范围的数据来源，对被审计单位的电子数据进行系统、全面以及跨部门的综合分析，从而解决目前数据分析局限于查找单个问题的缺陷，获得更充分的审计证据，更大地发挥审计的威力。

传统的数据分析技术，如关联规则挖掘、分类、数据聚类、遗传算法、机器学习、自然语言处理、神经网络、预测模型等，也可用于目前的大数据治理审计。但大数据环境下，开展大数据治理审计需要更多的智能技术。目前，为了满足大数据环境下数据分析的需要，一些专门用于处理大数据的关键技术也被研究出来，如 BigTable、云计算、分布式系统、Hadoop、HBase、Map/Reduce、可视化技术等。因此，可借助以上技术进行审计大数据的分析与结果展示。

当用户将数据存储在云服务器中时，就丧失了对数据的控制权。如果云服务提供商不可信，其可能对数据进行篡改、丢弃，却对用户声称数据是完好的。为了防止这种危害，

云存储中的审计技术被提出。云存储审计指的是数据拥有者或者第三方机构对云中的数据完整性进行审计。通过对数据进行审计，确保数据不会被云服务提供商篡改、丢弃，并且在审计的过程中用户的隐私不会被泄露。

当前已有云存储中的审计模型有以下几种：

（1）数据持有（Provable Data Possession，PDP）模型

数据持有模型可以对服务器上的数据进行完整性验证。该模型先从服务器上随机采样相应的数据块，并生成持有数据的概率证据。客户端维持一定数量的元数据，并利用元数据来对证据进行验证。在该模型中，挑战应答协议传输的数据量非常少，因此所耗费的网络带宽较小。

（2）可恢复证明（Proof Of Retrievability，POR）模型

可恢复证明模型主要利用纠错码技术和消息认证机制来保证远程数据文件的完整性和可恢复性。在该模型中，原始文件首先被纠错码编码并产生对应标签，编码后的文件及标签被存储在服务器上。当用户选择服务器上的某个文件块时，可以采用纠错码解码算法来恢复原始文件。POR 模型面临的挑战在于需要构建一个高效和安全的系统来应对用户的请求。有人改进了 POR 模型，他们的模型构建基于 BLS 短签名（BLS short signature），即基于双线性对构造的数字签名方案，该模型拥有很短的查询和响应时间。

上述方案都只能适用于静态数据的审计，无法支持对动态数据的审计。有人改进了 PDP 模型，该模型基于对称密钥加密算法，并且支持数据的动态删除和修改。之后，Erway 等改进了 PDP 模型，提出了 DPDP 模型。该模型扩展了传统的 PDP 模型以支持存储数据的更新操作，该操作的时间复杂度为 O(1)～O(log(n))。后来，又有人改进了前人的 POR 模型，通过引入散列树来对文件块标签进行认证。同时，他们的方法也支持对数据的动态操作，但是此方案无法对用户的隐私进行有效的保护。

第三方审计（Third Party Auditor，TPA）应该满足如下要求：一是第三方审计能够高效地完成对数据的审计，并且不给用户带来多余的负担；二是第三方审计不能为用户隐私带来脆弱性。他们提出的方法基于公钥加密和同态认证，能够在保护用户隐私的情况下完成公开审计。人们提出了一种用于对云中共享数据进行审计的隐私保护策略。他们在对数据的审计过程中利用环形签名来对数据完整性进行验证。此策略能够很好地对用户的隐私进行保护。其不足之处在于通信开销比较大。后来，人们还提出了一种名为 Knox 的云中数据的隐私保护策略。该策略利用群组签名来构造同态认证，使得第三方审计机构不需要从云中获取整个数据即能完成对数据完整性的审计。

随着大数据时代的发展，可以预见到，未来存储在云中的数据会越来越多，这也为大数据审计技术带来了巨大的挑战。在未来的研究中，以下几个方向也许值得研究者们关注：一个是云中数据量越来越大、数据种类越来越丰富，如何提供更加高效、安全的审计服务值得关注；另一个是随着人们在线上的交互越来越频繁，云中数据动态操作可能更加频繁，如何应对如此频繁的数据动态操作也值得研究者们关注。

2.3.4　大数据治理审计流程

大数据治理审计流程和一般的审计流程差不多，大数据治理审计流程一般包括制定大数据治理审计目标、确定大数据治理审计风险领域、制定大数据治理审计计划、搭建大数据治理审计环境、出具审计结果和管理建议。

1. 制定大数据治理审计目标

大数据治理审计目标是指人们在特定的社会生产环境中，期望通过审计实践活动达到的大数据治理最终结果，或者说是指大数据治理审计活动的目的与要求。一般来说，各类大数据治理审计目标都必须满足其服务领域的特殊需要，无论是在公司还是学校，它们都具有各自相对独立的审计目标。大数据治理审计目标的确定，除受审计对象的制约以外，还取决于审计社会属性、审计基本职能和审计授权者或委托者对审计工作的要求。同时，审计目标规定了审计的基本任务，决定了审计的基本过程和应办理的审计手续。

2. 确定大数据治理审计风险领域

大数据治理审计风险是指大数据治理审计过程中可能会产生的风险。大数据治理审计风险包括以下几个方面：

（1）数据采集风险

采集数据是审计分析的第一步，也是关系审计质量的关键一步。特别是大数据环境下审计人员需要采集被审计单位的海量业务数据，在数据采集过程中审计人员主要面临两个风险：一是保证所采集数据的真实性、完整性，满足审计分析的需要；二是保证数据采集过程中被审计单位的系统安全性。

（2）大数据存储和管理的风险

海量的大数据从被审计单位采集回来，在存储和管理方面审计机关和人员面临两方面的风险：一是数据存储风险，海量的大数据如何进行存储，保证数据的完整性，同时可以供审计人员进行审计分析操作；二是数据管理的风险，被审计单位提供的数据包含大量的个人基本信息、敏感信息，审计人员将面对如何对这些数据进行管理，从技术上和制度上保证这些数据没有泄露到社会上的风险。

3. 制定大数据治理审计计划

所谓大数据治理审计计划，是指为了完成各项大数据治理审计业务，达到预期的大数据治理审计目标，在具体执行大数据治理审计程序之前编制的工作计划。大数据治理审计计划通常可分为总体审计计划和具体审计计划两部分。

（1）总体审计计划

总体审计计划是对审计的预期范围和实施方式所做的规划，是负责大数据治理审计人员从接受审计委托到出具审计整个过程基本工作内容的综合计划。

（2）具体审计计划

具体审计计划是依据总体审计计划制定的，对实施总体审计计划所需要的审计程序的

性质、时间和范围所做的详细规划与说明。一般通过编制审计程序表的方式来体现。

一般来说，制定大数据治理审计计划有以下几个作用：

1）为大数据治理审计人员和审计工作明确方向。

现代社会的迅速发展，使审计面临和从事的工作越来越复杂。要切实解决审计面临的问题和所从事的工作，就必须协调各个方面，调动各种资源，使所有审计人员齐心协力完成工作。一份良好的审计计划为审计人员制定了统一目标，使所有审计人员凝聚所有资源朝着一个方向，共同努力来完成同一个任务，从而减少内耗，缩短时间，降低审计成本，促进审计任务顺利实现。

2）减少未来不确定因素的负面影响。

社会在不断发展，审计也在不停地发展。无论是审计组织的外部环境因素还是审计组织内部因素，在未来的发展中都具有一定的不确定性和变化性。审计计划是面向未来的，能够通过周密细致的研究，系统运用各种科学方法手段来预测审计未来的发展变化，尽可能将审计未来的变化和不确定因素转化为确定因素。通过审计计划，将各种不利因素转化为有利因素，减少未来不确定因素的负面影响，促进审计工作的顺利进行，确保审计目标的实现。

3）为大数据治理审计考核工作提供前提条件。

任何一项工作之后都要进行考核，为激励、组织和领导等工作提供前提条件。科学系统的考核工作需要一个科学合理的基础。审计计划能够为审计考核工作提供一个合理前提，也只有审计计划才能作为审计考核的基础，才能促使审计激励工作取得最大的效果。

4）为大数据治理审计控制工作提供标准。

任何一项工作在进行过程中都有可能因种种客观或主观原因而出现偏差，影响工作任务的完成。因此，要随时对审计过程进行检查，加强审计项目过程的控制，促使审计目标的顺利实现。要进行审计控制就需要一个控制标准，否则管理人员就无法实施控制。审计计划是审计控制的基础，它为审计项目控制提供了控制标准。

5）提高审计效率和社会效益。

大数据治理审计计划能够通过各种科学技术方法来制定和选择科学详细的项目方案，能够用科学决策代替经验判断，能够统筹安排审计资源，能够有针对性地根据经济社会发展来科学安排审计项目等。这些都能够有力促进审计效率的提高，充分发挥“经济卫士”和“经济谋士”的功能，从而促进社会效益的提高，促进经济社会的和谐发展。

4. 搭建大数据治理审计环境

搭建大数据治理审计环境是指根据大数据治理审计的目标和计划，设计出可以帮助进行大数据治理审计的环境平台。一般来说，进行大数据治理审计都要依靠审计程序，企业产生的数据不可能都由人工去处理，所以进行大数据治理审计的开源项目应运而生。本书中将要介绍的 Apache Ranger 就是一种可以用于大数据治理审计的软件。

通过搭建大数据治理审计环境，大数据治理审计部分工作可以交由审计程序自动进

行。如图 2-3 所示，审计程序可以跟踪每个相关的用户和系统事件并创建审计日志。例如 Apache Ranger 是用于 Hadoop 的集中式安全管理解决方案，使管理员能够为 HDFS 和其他 Hadoop 平台组件创建和实施安全策略，并且为 Hadoop 的各个零部件提供细粒度的安全权限机制。它可以对 Hadoop 生态系统上的组件如 Hive、HBase 等进行细粒度的数据访问控制，并解决授权和审计。在 Hadoop 生态系统上的组件的操作都会记录在日志当中，这些日志中的数据可以用于定期安全审计。

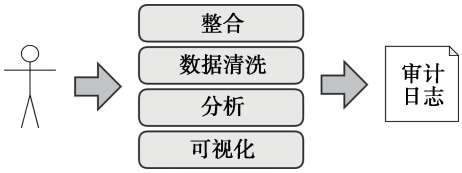

图 2-3　跟踪每个相关的用户和系统事件并创建审计日志

具体来说，Apache Ranger 可以提供这些审计能力。

1）用户行为日志：Apache Ranger 维护与所有相关的用户和系统的事件和信息日志文件。

2）安全审计日志：Apache Ranger 维护一个专门的安全审计日志，捕捉相关的安全调查和审计行动，包括所有身份验证尝试、权限变更等。

5. 出具审计结果和管理建议

接下来需要根据大数据治理审计计划的内容，执行大数据治理审计的计划内容。大数据治理审计计划执行完了之后，最后一个流程是根据大数据治理审计的结果，出具审计的结果和管理建议。审计结果一般以审计报告的形式呈现。审计报告是审计工作情况的全面总结汇报，说明审计工作的结果。审计目标的实现结果是通过审计报告来反映的，审计报告反映委托方的最终要求，也反映审计方完成任务的工作质量，同时也是对被审事项的评价和结论的集中体现。在审计报告的最后一项，一般还会给出一些大数据治理的管理建议。