CAICT 中国信通院

大数据安全白皮书

（2018 年）

中国信息通信研究院

安全研究所

2018 年 7 月

引言

当前，全球大数据产业正值活跃发展期，技术演进和应用创新并行加速推进，非关系型数据库、分布式并行计算以及机器学习、深度挖掘等新型数据存储、计算和分析关键技术应运而生并快速演进，大数据挖掘分析在电信、互联网、金融、交通、医疗等行业创造商业价值和应用价值的同时，开始向传统第一、第二产业传导渗透，大数据逐步成为国家基础战略资源和社会基础生产要素。

与此同时，大数据安全问题逐渐暴露。大数据因其蕴藏的巨大价值和集中化的存储管理模式成为网络攻击的重点目标，针对大数据的勒索攻击和数据泄露问题日趋严重，全球大数据安全事件呈频发态势。相应的，大数据安全需求已经催生相关安全技术、解决方案及产品的研发和生产，但与产业发展相比，存在滞后现象。

习近平主席在中共中央政治局就实施国家大数据战略第二次集体学习时指出，要构建以数据为关键要素的数字经济，推动实体经济和数字经济融合发展，推动互联网、大数据、人工智能同实体经济深度融合。同时，要切实保障国家数据安全。这要求我们必须坚持国家总体安全观，树立正确的网络安全观，坚持“以安全保发展，以发展促安全”，充分发

挥大数据在推动产业转型升级、提升国家治理现代化水平等方面重要作用的同时，深刻认识大数据安全的重要性和紧迫性，认清大数据安全挑战，积极应对复杂严峻的安全风险，坚持安全与发展并重，加速构建大数据安全保障体系，保障国家大数据发展战略顺利实施。

本报告首先从大数据带来的变革出发，探讨了大数据安全区别于传统安全的特殊内涵；然后聚焦技术领域，给出大数据安全技术总体视图，分别从平台安全、数据安全和个人隐私安全三个方面梳理了大数据环境下面临的安全威胁以及相应的安全保障技术的发展情况；最后基于大数据安全技术发展现状，提出大数据安全技术未来发展方向与建议，为大数据产业和安全技术发展提供依据和参考。

目录

一、对大数据安全的认识和思考

大数据在数量规模、处理方式、应用理念等方面都呈现了与传统数据不同的新特征。大数据是具有体量大、结构多样、时效强等特征的数据；处理大数据需采用新型计算架构和智能算法等新技术；大数据的应用强调以新理念应用于辅助决策、发现新知识，更强调在线闭环的业务流程优化。从安全视角看，大数据这些新特性，产生了哪些影响？我们认为：

（一）大数据已经对经济运行机制、社会生活方式和国家治理能力产生深刻影响，需要从“大安全”的视角认识和解决大数据安全问题

大数据发展过程中，资源、技术、应用相依相生，以螺旋式上升的模式发展。无论是商业策略、社会治理、还是国家战略的制定，都越来越重视大数据的决策支撑能力。但也要看到，大数据是一把双刃剑，大数据分析预测的结果对社会安全体系所产生的影响力和破坏力可能是无法预料和提前防范的。例如，美国一款健身应用软件将用户健身数据的分析结果在网络上公布，结果涉嫌泄露美国军事机密，这在以往是不可想象的。未来，基于大数据的智能决策将会在经济运行、社会生活、国家治理方面发挥更重要的作用，大数据可能会对国家“11 种安全”的方方面面产生更加深远的影响。

因此，必须从“大安全”的视角审视大数据安全问题，必须站在国家总体安全观的高度，打破传统的重技术的安全保护思维模式，建立涉及经济、法律、技术等多角度全方位的大数据安全保障体系。

（二）大数据正逐渐演变为新一代基础性支撑技术，大数据平台的自身安全将成为大数据与实体经济融合领域安全的重要影响因素

目前来看，大数据正在成为一种通用的数据处理技术，除推动人工智能、虚拟现实等新兴信息技术应用创新之外，互联网、大数据通过与实体经济的深度融合，正加速推进传统制造业向数字化、网络化、智能化发展。然而，在信息化和工业化融合业务繁荣发展的背后，安全问题如影随形。针对大数据平台的网络攻击手段正在悄然变化，攻击目的已经从单纯地窃取数据、瘫痪系统转向干预、操纵分析结果，攻击效果已经从直观易察觉的系统宕机、信息泄露转向细小难以察觉的分析结果偏差，造成的影响可能从网络安全事件上升到工业生产安全事故。目前，传统基于监测、预警、响应的网络安全技术难以应对上述攻击变化，需要进行理念创新，针对不断变化演进的网络攻击形态，设计建构更加完善的大数据平台安全保护体系，为上层跨行业跨领域的业务应用提供基础性安全保障。

（三）大数据时代，数据在流动过程中实现价值最大化，需要重构以数据为中心、适应数据动态跨界流动的安全防护体系

大数据时代，数据作为一种特殊的资产，能够在流通和使用过程中不断创造新的价值。因此，在大数据应用场景下，数据流动是“常态”，数据静止存储才是“非常态”。同时，可以预见到，未来大数据业务环境将更加开放，业务生态将更加复杂，参与数据处理的角色将更多元，系统、业务、组织边界将进一步模糊，导致数据的产生、流动、处理等过程比以往更加丰富和多样。数据的频繁跨界流动，除可能导致传统的数据泄露风险外，还会引发新的安全风险。特别是在数据共享环节中，传统数据访问控制技术无法解决跨组织的数据授权管理和数据流向追踪问题，仅靠书面合同或协议难以实现对数据接收方的数据处理活动进行实时监控和审计，极易造成数据滥用的风险，最典型的案例即是今年曝光的“剑桥分析”事件。未来，数据共享和流通将成为刚性业务需求，传统的静态隔离安全保护方法将彻底不能满足数据流动安全防护的需求，必须通过动态变化的视角分析和判断数据安全风险，构建以数据为中心的动态、连续的数据安全防护体系。

（四）大数据推动数字经济新业态新模式蓬勃发展，广

大民众却面临享受便捷化泛在化信息服务与保护个人信息权利之间的两难抉择

近年来，我国网络购物、移动支付、共享经济等数字经济新业态新模式发展迅猛，基于互联网、移动互联网、物联网的信息服务已经渗透到社会生活的方方面面，为广大民众提供便捷、高效、全天候的服务。以普惠金融为例，利用大数据对个人数据的挖掘和分析，能够帮助金融科技公司更好的理解用户需求，提供个性化定制服务；利用大数据进行金融风险控制，能够实现流水线操作，减少经营成本，提高服务效率，提升用户体验。例如，某互联网金融服务企业推出的“310”个人信贷服务模式，即“3 分钟填表、1 分钟批贷、0 人工干预”，为用户提供了传统信贷服务无法比拟的业务体验，同时将业务成本从每单 2000 元降至 2.3 元。然而，用户享受便捷服务的代价是出让自己的个人信息权利。每日推荐、个人日报、免押租车等信息服务，都是基于大数据技术对用户个人数据进行挖掘分析，形成用户画像，进而提供的定制化服务。但大数据应用场景下，无所不在的数据收集技术、专业化多样化的数据处理技术，使得用户难以控制其个人信息的收集情境和应用情境，用户对其个人信息的自决权利自然被削弱。特别是，企业间的数据共享日益频繁，利用大数据的超强分析能力对多源数据进行处理，能够将经过匿名化处理的数据再次还原，导致现有数据脱敏技术“失灵”，直接威

胁用户的隐私安全。

综上，大数据安全是涉及技术、法律、监管、社会治理等领域的综合性问题，其影响范围涵盖国家安全、产业安全和个人合法权益。同时，大数据在数量规模、处理方式、应用理念等方面的革新，不仅导致大数据平台自身安全需求发生变化，还带动数据安全防护理念随之改变，同时引发对高水平隐私保护技术的需求和期待。

二、大数据安全技术总体视图

如前所述，大数据安全是一个跨领域跨学科的综合性问题，可以从法律、经济、技术等多个角度进行研究。本报告以技术作为切入点，梳理分析当前大数据的安全需求和涉及的技术，提出大数据安全技术总体视图，如图 1 所示。在绘制大数据安全技术总体视图的过程中，我们参考了 NIST 等国内外关于大数据技术参考架构的研究成果。考虑到大数据平台为上层应用系统提供存储和计算资源，是对数据进行采集、存储、计算、分析与展示等处理的工具和场所，因此，我们以大数据平台为基本出发点，形成了大数据安全总体视图。

在总体视图中，大数据安全技术体系分为大数据平台安全、数据安全和个人隐私保护三个层次，自下而上为依次承载的关系。大数据平台不仅要保障自身基础组件安全，还要

为运行其上的数据和应用提供安全机制保障；除平台安全保障外，数据安全防护技术为业务应用中的数据流动过程提供安全防护手段；隐私安全保护是在数据安全基础之上对个人敏感信息的安全防护。

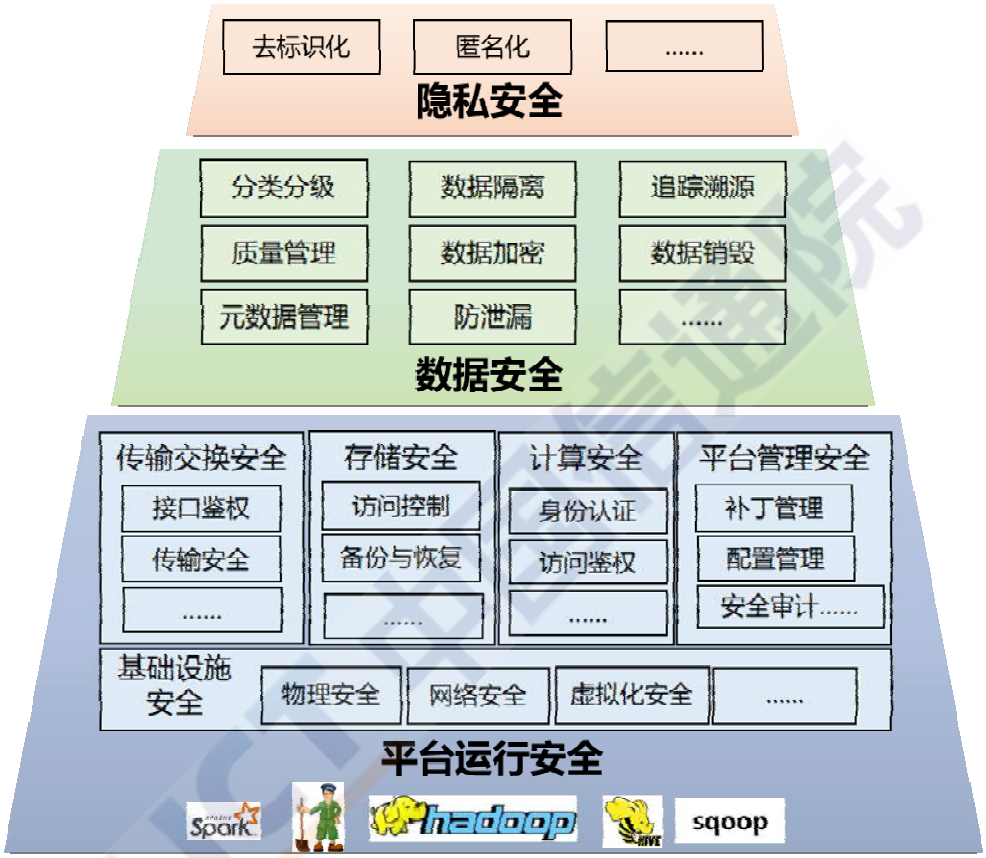

图 1.大数据安全技术总体视图

（一）大数据平台安全

大数据平台安全是对大数据平台传输、存储、运算等资源和功能的安全保障，包括传输交换安全、存储安全、计算安全、平台管理安全以及基础设施安全。

传输交换安全是指保障与外部系统交换数据过程的安全可控，需要采用接口鉴权等机制，对外部系统的合法性进

行验证，采用通道加密等手段保障传输过程的机密性和完整性。存储安全是指对平台中的数据设置备份与恢复机制，并采用数据访问控制机制来防止数据的越权访问。计算组件应提供相应的身份认证和访问控制机制，确保只有合法的用户或应用程序才能发起数据处理请求。平台管理安全包括平台组件的安全配置、资源安全调度、补丁管理、安全审计等内容。此外，平台软硬件基础设施的物理安全、网络安全、虚拟化安全等是大数据平台安全运行的基础。

（二）数据安全

数据安全防护是指平台为支撑数据流动安全所提供的安全功能，包括数据分类分级、元数据管理、质量管理、数据加密、数据隔离、防泄露、追踪溯源、数据销毁等内容。

大数据促使数据生命周期由传统的单链条逐渐演变成为复杂多链条形态，增加了共享、交易等环节，且数据应用场景和参与角色愈加多样化，在复杂的应用环境下，保证国家重要数据、企业机密数据以及用户个人隐私数据等敏感数据不发生外泄，是数据安全的首要需求。海量多源数据在大数据平台汇聚，一个数据资源池同时服务于多个数据提供者和数据使用者，强化数据隔离和访问控制，实现数据“可用不可见”，是大数据环境下数据安全的新需求。利用大数据技术对海量数据进行挖掘分析所得结果可能包含涉及国家安全、经济运行、社会治理等敏感信息，需要对分析结果的共享和

披露加强安全管理。

（三）隐私保护

本报告所提的隐私保护是指利用去标识化、匿名化、密文计算等技术保障个人数据在平台上处理、流转过程中不泄露个人隐私或个人不愿被外界知道的信息。隐私保护是建立在数据安全防护基础之上的保障个人隐私权的更深层次安全要求。然而，我们也意识到大数据时代的隐私保护不再是狭隘地保护个人隐私权，而是在个人信息收集、使用过程中保障数据主体的个人信息自决权利。实际上，个人信息保护已经成为一个涵盖产品设计、业务运营、安全防护等在内的体系化工程，不是一个单纯的技术问题。但由于本报告重点聚焦大数据安全技术，因此在谈及数据主体的个人权益保护时，我们选择去繁从简，从研究方向更为清晰的隐私保护技术入手开展研究。

三、大数据安全面临的技术问题和挑战

大数据安全威胁渗透在数据生产、采集、处理和共享等大数据产业链的各个环节，风险成因复杂交织；既有外部攻击，也有内部泄露；既有技术漏洞，也有管理缺陷；既有新技术新模式触发的新风险，也有传统安全问题的持续触发。本报告将聚焦于大数据本身面临的安全威胁，从大数据平台安全、数据安全和个人信息安全三个方面展开分析，确定大

数据安全需求。

（一）平台安全问题与挑战

1、大数据平台在 Hadoop 开源模式下缺乏整体安全规划，自身安全机制存在局限性

目前，Hadoop 已经成为应用最广泛的大数据计算软件平台，其技术发展与开源模式结合。Hadoop 的最初设计是为了管理大量的公共 web 数据，假设集群总是处于可信的环境中，由可信用户使用的相互协作的可信计算机组成。因此最初的 Hadoop 没有设计安全机制，也没有安全模型和整体的安全规划。随着 Hadoop 的广泛应用，越权提交作业、修改 JobTracker 状态、篡改数据等恶意行为不断出现，Hadoop 开源社区开始考虑安全需求，并相继加入了 Kerberos 认证、文件 ACL 访问控制、网络层加密等安全机制，这些安全功能可以解决部分安全问题，但仍然存在局限性。在身份管理和访问控制方面，依赖于 Linux 的身份和权限管理机制，身份管理仅支持用户和用户组，不支持角色；仅有可读、可写、可执行三个权限，不能满足基于角色的身份管理和细粒度访问控制等新的安全需求。安全审计方面，Hadoop 生态系统中只有分布在各组件中的日志记录，无原生安全审计功能，需要使用外部附加工具进行日志分析。另外，开源发展模式也为 Hadoop 系统带来了潜在的安全隐患。企业在进行工具研发的过程中，多注重功能的实现和性能的提高，对代码的

质量和数据安全关注较少。因此，开源组件缺乏严格的测试管理和安全认证，对组件漏洞和恶意后门的防范能力不足。据 Common Vulnerabilities and Exposures（以下简称“CVE”）漏洞列表显示，从 2013 年到 2017 年，Hadoop 暴露出来的漏洞数量共计 18 个，其中有 5 个是关于信息泄露的漏洞，并且漏洞数量逐年增长，这五年的具体漏洞数量如图 2 所示。

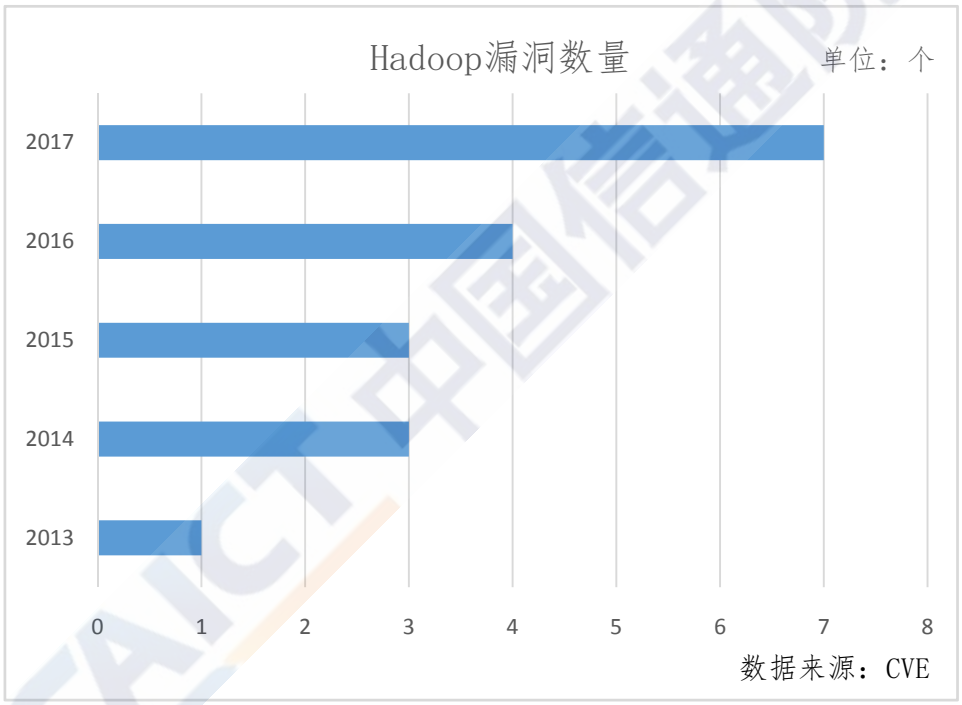

图 2.2013-2017 年 Hadoop 漏洞统计图

2、大数据平台服务用户众多、场景多样，传统安全机制的性能难以满足需求

大数据场景下，数据从多个渠道大量汇聚，数据类型、用户角色和应用需求更加多样化，访问控制面临诸多新的问题。首先，多源数据的大量汇聚增加了访问控制策略制定及

授权管理的难度，过度授权和授权不足现象严重。其次，数据多样性、用户角色和需求的细化增加了客体的描述困难，传统访问控制方案中往往采用数据属性（如身份证号）来描述访问控制策略中的客体，非结构化和半结构化数据无法采取同样的方式进行精细化描述，导致无法准确为用户指定其可以访问的数据范围，难以满足最小授权原则。大数据复杂的数据存储和流动场景使得数据加密的实现变得异常困难，海量数据的密钥管理也是亟待解决的难题。

3、大数据平台的大规模分布式存储和计算模式导致安全配置难度成倍增长

开源 Hadoop 生态系统的认证、权限管理、加密、审计等功能均通过对相关组件的配置来完成，无配置检查和效果评价机制。同时，大规模的分布式存储和计算架构也增加了安全配置工作的难度，对安全运维人员的技术要求较高，一旦出错，会影响整个系统的正常运行。据 Shodan 互联网设备搜索引擎的分析显示，大数据平台服务器配置不当，已经导致全球 5120TB 数据泄露或存在数据泄露风险，泄露案例最多的国家分别是美国和中国[1]。本年初针对 Hadoop 平台的勒索攻击事件，在整个攻击过程中并没有涉及常规漏洞，而是利用平台的不安全配置，轻而易举地对数据进行操作。

4、针对大数据平台网络攻击手段呈现新特点，传统安

[1]https://www.easyaq.com/news/334943503.shtml

全监测技术暴露不足

大数据存储、计算、分析等技术的发展，催生出很多新型高级的网络攻击手段，使得传统的检测、防御技术暴露出严重不足，无法有效抵御外界的入侵攻击。传统的检测是基于单个时间点进行的基于威胁特征的实时匹配检测，而针对大数据的高级可持续攻击（APT）采用长期隐蔽的攻击实施方式，并不具有能够被实时检测的明显特征，发现难度较大。此外，大数据的价值低密度性，使得安全分析工具难以聚焦在价值点上，黑客可以将攻击隐藏在大数据中，传统安全策略检测存在较大困难。因此，针对大数据平台的高级持续性威胁（APT）攻击时有发生，大数据平台遭受的大规模分布式拒绝服务（DDoS）攻击屡见不鲜。Verizon 公司《2018 年数据泄露调查报告》显示，48%的数据泄露与黑客攻击有关，其中，DDoS、钓鱼攻击以及特权滥用是主要的黑客攻击方式，具体数据如图 3 所示。

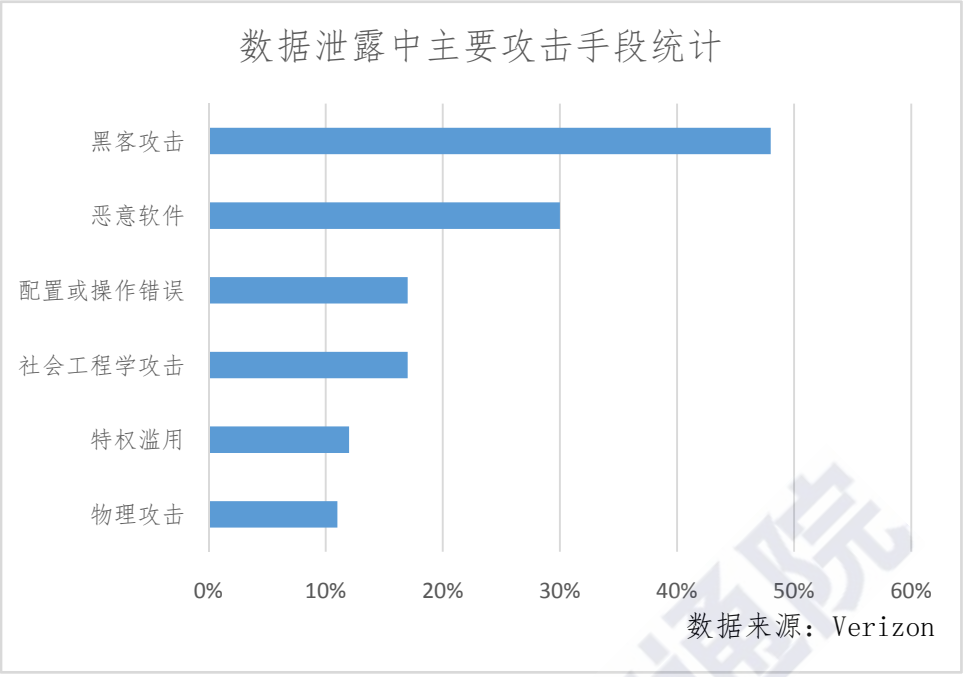

图 3.数据泄露中主要攻击手段统计图

（二）数据安全问题和挑战

除数据泄露威胁持续加剧外，大数据的体量大、种类多等特点，使得大数据环境下的数据安全出现了有别于传统数据安全的新威胁。

1、数据泄露事件数量持续增长，造成的危害日趋严重

大数据因其蕴藏的巨大价值和集中化的存储管理模式成为网络攻击的重点目标，针对大数据的勒索攻击和数据泄露问题日趋严重，重大数据安全事件频发。Gemalto《2017 数据泄露水平指数报告》显示，2017 年上半年全球范围内数据泄露总量为 19 亿条，超过 2016 年全年总量(14 亿)，比 2016 年下半年增长了 160%多，从 2013 年到 2017 年全球数据泄露的具体数目如图 4 所示，从图中可以看出数据泄露

的数目呈现逐年上涨的趋势。仅 2017 年，全球发生了多起影响重大的数据泄露事件，美国共和党下属数据分析公司[2]、征信机构[3]先后发生大规模用户数据泄露事件，影响人数均达到亿级规模。我国数据泄露事件也时有发生。2017 年 3 月，京东试用期员工与网络黑客勾结，盗取涉及交通、物流、医疗等个人信息 50 亿条，在网络黑市贩卖。此外，数据泄露的潜在隐患同样不容乐观，据 Shodan 统计，截至 2017 年 2 月 3 日，中国有 15046 个 MangoDB 数据库暴露在公网，存在严重安全问题。

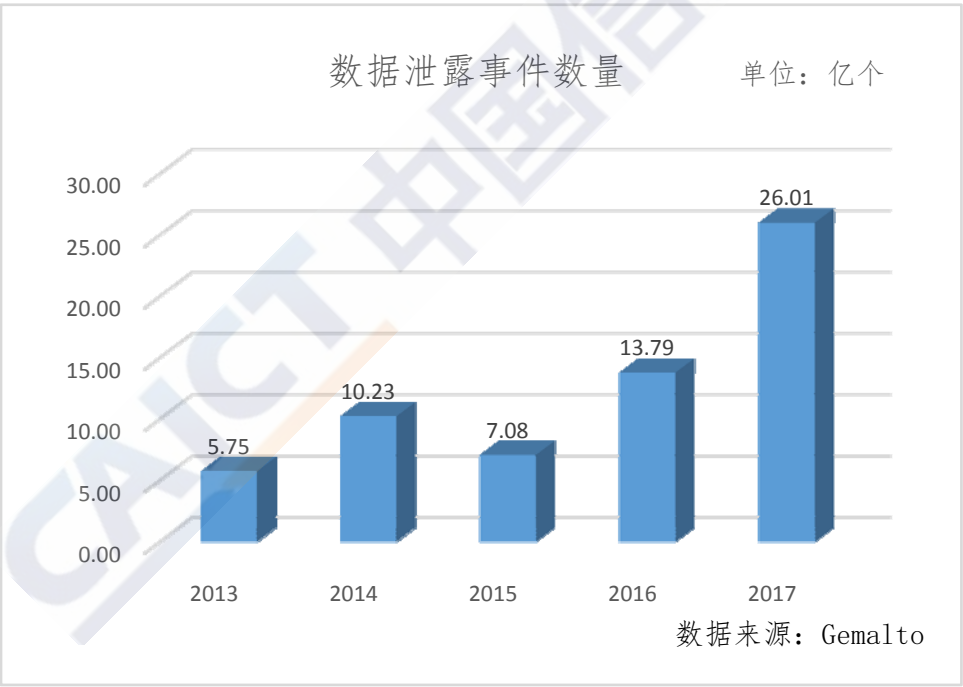

图 4.2013-2017 年数据泄露数量统计图

2、数据采集环节成为影响决策分析的新风险点

在数据采集环节，大数据体量大、种类多、来源复杂的

[2] 2017 年 6 月，美国共和党国家委员会下属的数据分析公司 Deep Root Analytics 被曝泄露 1.98 亿美国公民的个人信息。

[3] 2017 年 5 月至 7 月，美国三大征信机构之一 Equifax 的数据库遭受攻击，1.43 亿用户个人信息遭窃取。

特点为数据的真实性和完整性校验带来困难，目前，尚无严格的数据真实性、可信度鉴别和监测手段，无法识别并剔除虚假甚至恶意的数据。若黑客利用网络攻击向数据采集端注入脏数据，会破坏数据真实性，故意将数据分析的结果引向预设的方向，进而实现操纵分析结果的攻击目的。

3、数据处理过程中的机密性保障问题逐渐显现

数字经济时代来临，越来越多的企业或组织需要参与产业链协同，以数据流动与合作为基础进行生产活动。企业或组织在开展数据合作和共享的应用场景中，数据将突破组织和系统的边界进行流转，产生跨系统的访问或多方数据汇聚进行联合运算。保证个人信息、商业机密或独有数据资源在合作过程中的机密性，是企业或组织参与数据共享合作的前提，也是数据有序流动必须要解决的问题。

4、数据流动路径的复杂化导致追踪溯源变得异常困难

大数据应用体系庞杂，频繁的数据共享和交换促使数据流动路径变得交错复杂，数据从产生到销毁不再是单向、单路径的简单流动模式，也不再仅限于组织内部流转，而会从一个数据控制者流向另一个控制者。在此过程中，实现异构网络环境下跨越数据控制者或安全域的全路径数据追踪溯源变得更加困难，特别是数据溯源中数据标记的可信性、数据标记与数据内容之间捆绑的安全性等问题更加突出。2018 年 3 月的“剑桥分析”事件中 ,Facebook 即是因为对第三方使

用数据缺乏有效的管理和追责机制，最终导致 8700 万名用户资料被滥用，还带来了股价暴跌、信誉度下降等严重后果。

（三）个人隐私安全挑战

大数据应用对个人隐私造成的危害不仅是数据泄露，大数据采集、处理、分析数据的方式和能力对传统个人隐私保护框架和技术能力亦带来了严峻挑战。

1、传统隐私保护技术因大数据超强的分析能力面临失效的可能

在大数据环境下，企业对多来源多类型数据集进行关联分析和深度挖掘，可以复原匿名化数据，进而能够识别特定个人或获取其有价值的个人信息。在传统的隐私保护中，数据控制者针对单个数据集孤立地选择隐私保护技术和参数来保护个人数据，特别是利用去标识、掩码等技术的做法，无法应对上述大数据场景下多源数据分析挖掘引发的隐私泄露问题。

2、传统隐私保护技术难以适应大数据的非关系型数据库

在大数据技术环境下，数据呈现动态变化、半结构化和非结构化数据居多的特性，对于占数据总量 80%以上的非结构化数据，通常采用非关系型数据库（NoSQL）存储技术完成对大数据的抓取、管理和处理。而非关系型数据库目前尚无严格的访问控制机制及相对完善的隐私保护工具，现有的

隐私保护技术，如去标识化、匿名化技术等，多适用于关系型数据库。

四、大数据安全技术发展情况

面对上述大数据安全挑战与威胁，产业各界在安全防护技术方面进行了针对性的实践与探索。本报告从大数据平台安全、数据安全、隐私保护三个方面阐述大数据安全技术的发展现状。

（一）大数据平台安全技术

随着市场对大数据安全需求的增加，Hadoop 开源社区增加了身份认证、访问控制、数据加密等安全机制。商业化 Hadoop 平台也逐步开发了集中化安全管理、细粒度访问控制等安全组件，对平台进行了安全升级。部分安全服务提供商也致力于通用的大数据平台安全加固技术和产品的研发，已有多款大数据平台安全产品上市。这些安全机制的应用为大数据平台安全提供了基础机制保障。

1、Hadoop 开源社区增加了基本安全机制，但安全能力不能满足现实需求

Hadoop 开源系统中提供了身份认证、访问控制、安全审计、数据加密等基本安全功能。身份认证方面，Hadoop 支持两种身份验证机制：简单机制和 Kerberos 机制。简单

机制是默认设置，根据客户进程的有效 UID 确定用户名，只能避免内部人员的误操作。Kerberos 机制支持集群中服务器间的认证和 Client 到服务器的认证。因为 Kerberos 可以实现较强的安全性，同时保证较高的运行性能，目前还没有哪种认证方式可以取代 Kerberos 认证。基于 Kerberos 的认证方式对于系统外部可以实现强安全认证，但 Kerberos 的认证颗粒度基于操作系统用户，无法支持系统内组件之间的身份认证。访问控制方面，目前大数据安全开源技术在访问控制方面主要有基于权限的访问控制、访问控制列表、基于角色的访问控制、基于标签的访问控制和基于操作系统的访问控制等几种方式。POSIX 权限和访问控制列表方式可用于 HDFS、MapReduce、HBase 中，Hive 支持基于角色的访问控制，HBase 和 Accumulo 提供了基于标签的访问控制。在以上几种访问控制方式中，企业主流使用的是基于权限的访问控制和基于角色的访问控制。大数据场景下用户角色众多，用户需求更加多样化，难以精细化和细粒度地控制每个角色的实际权限，导致无法准确为用户指定其可以访问的数据范围，实现细粒度访问控制较为困难。大数据环境访问控制的复杂性不仅在于访问控制的形式多样，另一方面在于大数据系统允许在不同系统层面广泛共享数据，需要实现一种集中统一的访问控制从而简化控制策略和部署。安全审计方面，Hadoop 开源系统各组件均提供日志和审计文件，可以

记录数据访问过程，为追踪数据流向和发现违规数据操作提供原始依据。但 Hadoop 各组件分别进行基本的日志和审计记录，并存储在其内部，实现全系统的安全审计较为困难，需要使用外部的日志聚合系统从集群中所有节点拉取审计日志，放入集中化的位置进行存储和分析。数据加密方面，大数据环境下需要实现数据在静态存储及传输过程的加密保护，其难点在于密钥管理。从 Hadoop2.6 开始，HDFS 支持原生静态加密——应用层加密，是一种基于加密区的透明加密方法，需要加密的目录被分解为若干加密区，当数据写入加密区时被透明地加密，客户端读取数据时被透明地解密。对于动态传输数据，对应 RPC、TCP/IP 和 HTTP，Hadoop 提供了不同的动态加密方法，保证客户端与服务器传输的安全性。目前 Hadoop 开源技术能够支持通过基于硬件的加密方案，大幅提高数据加解密的性能，实现最低性能损耗的端到端和存储层加密。加密的有效使用需要安全灵活的密钥管理和分发机制，目前在开源环境下没有很好的解决方式，需要借助商业化的密钥管理产品。

2、商业化大数据平台解决方案已经具备相对完善的安全机制

商业化的大数据平台，如 Cloudera 公司的 CDH（Cloudera Distribution Hadoop）、Hortenworks 公司的 HDP(Hortonworks Data Platform) 华为公司的

FusionInsight、星环信息科技的 TDH(Transwarp Data Hub)等，在平台安全机制上，做了如下几个方面的优化。集中安全管理和审计方面 ,通过专门的集中化的组件(如 Manager、Ranger、Guardian）形成了大数据平台总体安全管理视图，实现集中的系统运维、安全策略管理和审计，通过统一的配置管理界面，解决了安全策略配置和管理繁杂的难题。身份认证方面，通过边界防护，保证 Hadoop 集群入口的安全，通过集中身份管理和单点登录等方式，简化了认证机制，通过界面化的配置管理方式，可以方便的管理和启用基于 Kerberos 的认证。访问控制方面，通过集中角色管理和批量授权等机制，降低集群管理的难度，通过基于角色或标签的访问控制策略，实现资源（例如文件、目录、表、数据库、列族等访问权限）的细粒度管理。加密和密钥管理方面，提供灵活的加密策略，保障数据传输过程及静态存储都是以加密形式存在，也可以实现对 Hive、HBase 的表或字段加密，同时提供更好的秘钥存储方案 ,并能提供和企业现有的 HSM（HardwareSecurity Module）集成的解决方案。

商业化大数据安全方案从 2008 年开始起步，经过了大量的测试验证，有众多部署实例，大量的运行在各种生产环境，技术成熟度高。由于这类安全方案的安全机制是只针对特定平台开发，安全保障组件仅适用于该平台，对于其他大数据平台，很难采取此类方案实现平台安全加固。

3、商业化通用安全组件可以为已建大数据平台提供安全加固方案

通用安全组件是指适用于原生或二次开发的 Hadoop 平台的安全防护机制，一般实现方式是通过在 Hadoop 平台内部部署集中管理节点，负责整个平台的安全管理策略设置和下发，实现对大数据平台的用户和系统内组件的统一认证管理和集中授权管理。通过在原功能组件上部署安全插件，对数据操作指令进行解析和拦截，实现安全策略的实施，从而实现身份认证、访问控制、权限管理、边界安全等功能。身份认证方面，在兼容平台原有 Kerberos+LDAP 认证机制的基础上，支持口令、手机、PKI 等多因素组合认证方式，实现外部用户认证和平台内部组件之间的认证，支持用户单点登录。访问控制方面，引入 DAC、MAC、RBAC、DTE 等多种访问控制模式，实现 HDFS 文件、计算资源、组件等细粒度的访问控制，支持安全、审计、操作三权分立。实现平台安全配置基线检查，提高大数据平台自身的安全性。还实现敏感数据的动态模糊化管理等功能。

通用安全组件易于部署和维护、适合对已建大数据系统进行安全加固，可以在不改变现有系统架构的前提下，解决企业的大数据平台安全需求。灵活性强，方便与现有的安全机制集成。这类产品的提供者一般都是专业的安全服务商，专注于安全问题的解决，防护机制的完备性强，精度高，为

开源大数据平台提供了较完备的安全加固方案。

（二）数据安全技术

数据是信息系统的核心资产，是大数据安全的最终保护对象。除大数据平台提供的数据安全保障机制之外，目前所采用的数据安全技术，一般是在整体数据视图的基础上，设置分级分类的动态防护策略，降低已知风险的同时考虑减少对业务数据流动的干扰与伤害。对于结构化的数据安全，主要采用数据库审计、数据库防火墙，以及数据库脱敏等数据库安全防护技术；对于非结构化的数据安全，主要采用数据泄露防护（Data leakage prevention, DLP）技术。同时，细粒度的数据行为审计与追踪溯源技术，能帮助系统在发生数据安全事件时，迅速定位问题，查缺补漏。

1、敏感数据识别技术作为数据安全监控的必要技术条件逐步实现自动化

在敏感数据的监控方案中，基础部分就是从海量的数据中挑选出敏感数据，完成对敏感数据的识别，进而建立系统的总体数据视图，并采取分类分级的安全防护策略保护数据安全。传统的数据识别方法是关键字、字典和正则表达式匹配等方式，通常结合模式匹配算法展开，该方法简单实用，但人工参与的相对较多，自动化程度较低，随着人工智能识别技术的引入，通过机器学习可以实现大量文档的聚类分析，自动生成分类规则库，内容自动化识别程度正逐步提高。

2、数据防泄露技术发展相对成熟并向智能化方向演进

DLP 是指通过一定的技术手段，防止用户的指定数据或信息资产以违反安全策略规定的形式流出企业的一类数据安全防护手段。针对数据泄露的主要途径，DLP 采用的主要技术如下：针对使用泄露和存储泄露，通常采用身份认证管理、进程监控、日志分析和安全审计等技术手段，观察和记录操作员对计算机、文件、软件和数据的操作情况，发现、识别、监控计算机中的敏感数据的使用和流动，对敏感数据的违规使用进行警告、阻断等。针对传输泄露，通常采取敏感数据动态识别、动态加密、访问阻断、和数据库防火墙等技术，监控服务器、终端以及网络中动态传输的敏感数据，发现和阻止敏感数据通过聊天工具、网盘、微博、FTP、论坛等方式泄露出去。目前的 DLP，普遍引入了自然语言处理、机器学习、聚类分类等新技术，将数据管理的颗粒度进行了细化，对敏感数据和安全风险进行智能识别。“智能安全”将会成为 DLP 技术发展的趋势，大数据分析技术、机器学习算法的发展与演进将推动数据泄露防护的智能化发展，DLP 将实现用户行为分析与数据内容的智能识别，实现数据的智能化分层、分级保护，并提供终端、网络、云端协同一体的敏感数据动态集中管控体系。

3、结构化数据库安全防护技术基本成熟，非结构化数据库安全防护亟需加强

结构化的数据安全技术主要是指数据库安全防护技术，可以分为事前评估加固、事中安全管控和事后分析追责三类，其中评估主要是数据库漏洞扫描技术，安全管控主要是数据库防火墙、数据加密、脱敏技术，事后分析追责主要是数据库审计技术。目前数据库安全防护技术发展逐步成熟。而在针对云环境和大数据环境的安全方面，针对非结构化数据库的防护方案已经由一些技术领先的厂商提出，但技术成熟度较低。

4、密文计算技术因多源数据计算机密性需求成为研究热点

随着多源数据计算场景的增多，在保证数据机密性的基础上实现数据的流通和合作应用一直是困扰产业界的难题，同态加密和安全多方计算等密文计算方法为解决这个难题提供了一种有效的解决思路。

同态加密提供了一种对加密数据进行处理的功能，对经过同态加密的数据处理得到一个输出，将这一输出进行解密，其结果与统一方法处理未加密的原始数据得到的输出结果一致。也就是说，其他人可以对加密数据进行处理，但是处理过程不会泄露任何原始内容。同时，拥有密钥的用户对处理过的数据进行解密后，得到的正好是处理后的结果。因为

这样一种良好的特性，同态加密特别适合在大数据环境中应用，既能满足数据应用的需求，又能保护用户隐私不被泄露，是一种理想的解决方案。2009 年，Gentry 提出了第一个全同态加密体制使得该方面的研究取得突破性进展，随后许多密码学家在全同态加密方案的研究上作出了有意义的工作，促进了全同态加密向实用化的发展，但是目前同态加密算法的计算开销过高，尚未应用到实际生产中。

安全多方计算（SecureMulti-PartyComputation, SMPC）是解决一组互不信任的参与方之间保护隐私的协同计算问题，SMPC 要确保输入的独立性，计算的正确性，同时不泄露各输入值给参与计算的其他成员。安全多方计算的这一特点，对于大数据环境下的数据机密性保护有独特的优势。通用的安全多方计算协议虽然可以解决一般性的安全多方计算问题，但是计算效率很低，尽管近年来研究者努力进行实用化技术的研究，并取得一些成果，但是离真正的产业化应用还有一段距离。

5、数字水印和数据血缘追踪技术发展明显滞后于实际需求

以上的数据识别、密文计算、安全监控和防护是“事前”和“事中”的安全保障技术，随着数据泄露事件的频繁发生，“事后”追踪和溯源技术变得越来越重要。安全事件发生后泄露源头的追查和责任的判定是及时发现问题、查缺补漏的关

键，同时，对安全管理制度的执行也会形成一定的威慑作用。目前常用的追踪溯源技术包括数字水印和数据血缘追踪技术。

数字水印技术是为了保持对分发后的数据流向追踪，在数据泄露行为发生后，对造成数据泄露的源头可进行回溯。对于结构化数据，在分发数据中掺杂不影响运算结果的数据，采用增加伪行、增加伪列等方法，拿到泄密数据的样本，可追溯数据泄露源。对于非结构化数据，数字水印可以应用于数字图像、音频、视频、打印、文本、条码等数据信息中，在数据外发的环节加上隐蔽标识水印，可以追踪数据扩散路径。但目前的数字水印方案大多还是针对静态的数据集，满足数据量巨大、更新速度极快的水印方案尚不成熟。

数据血缘(Lineage，Provenance，Pedigree)亦可译为血统、起源、世系、谱系，是指数据产生的链路，数据血缘记载了对数据处理的整个历史，包括数据的起源和处理这些数据的所有后继过程(数据产生、并随着时间推移而演变的整个过程)。通过数据血缘追踪，可以获得数据在数据流中的演化过程。当数据发生异常时，通过数据血缘分析能追踪到异常发生的原因，把风险控制在适当的水平。目前数据血缘分析技术应用尚不广泛，技术成熟度还未达到大规模实际的应用需求。

（三）个人隐私保护技术

大数据环境下，数据安全技术提供了机密性、完整性和可用性的防护基础，隐私保护是在此基础上，保证个人隐私信息不发生泄露或不被外界知悉。目前应用最广泛的是数据脱敏技术，学术界也提出了同态加密、安全多方计算等可用于隐私保护的密码算法，但应用尚不广泛。

1、数据脱敏技术发展成熟，是目前应用最广泛的隐私保护技术

数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现对个人数据的隐私保护，是应用最广泛的隐私保护技术。目前的脱敏技术主要分为如下三种：第一种加密方法，是指标准的加密算法，加密后完全失去业务属性，属于低层次脱敏。算法开销大，适用于机密性要求高、不需要保持业务属性的场景。第二种基于数据失真的技术，最常用的是随机干扰、乱序等，是不可逆算法，通过这种算法可以生成“看起来很真实的假数据”。适用于群体信息统计或（和）需要保持业务属性的场景。第三种可逆的置换算法，兼具可逆和保证业务属性的特征，可以通过位置变换、表映射、算法映射等方式实现。表映射方法应用起来相对简单，也能解决业务属性保留的问题，但是随着数据量的增大，相应的映射表同量增大，应用局限性高。算法映射方法不需要做映射表，通过自行设计的算法来实现数据的变换，这类算法都是基于密码学的基本概念自行设计的，通常的做法是在公开算

法的基础上做一定的变换，适用于需要保持业务属性或（和）需要可逆的场景。数据应用系统在选择脱敏算法时，可用性和隐私保护的平衡是关键，既要考虑系统开销，满足业务系统的需求，又要兼顾最小可用原则，最大限度的保护用户隐私。

2、匿名化算法将成为未来解决隐私保护问题的有效途径

数据匿名化算法可以实现根据具体情况有条件地发布部分数据，或者数据的部分属性内容，包括差分隐私、K 匿名、L 多样性、T 接近等。匿名化算法要解决的问题包括：隐私性和可用性间的平衡问题，执行效率问题，度量和评价标准问题，动态重发布数据的匿名化问题，多维约束匿名问题等。匿名化算法由于能够在数据发布环境下防止用户敏感数据被泄露，同时又能保证发布数据的真实性，这一特性在大数据安全领域受到广泛关注。目前，匿名化算法还有很多挑战性问题亟待解决，算法的成熟度和使用普及程度还不是很高。匿名化相关算法是目前数据安全领域的研究热点之一，目前取得了丰富的研究成果，也得到了一些实际应用，后续匿名化算法会在隐私保护方面得到越来越多的应用。

（四）大数据安全技术发展现状总结

国内外大数据平台安全、数据安全、隐私保护相关的技

术已经取得了一定的进展，能够初步解决本报告第三章提到的安全问题与挑战；但在应对一些新的网络攻击形式、数据应用场景、隐私保护需求方面，大数据安全技术的现有能力和水平还存在一定差距。

平台安全方面，集中的安全配置管理和安全机制部署能够基本满足目前平台的安全需求，大数据平台的漏洞扫描与攻击监测技术相对薄弱。目前的商业化大数据平台和商业化通用安全组件，为 Hadoop 生态系统增加了集中安全管理、准入控制、多因素认证、细粒度访问控制、密钥管理、数据脱敏、集中审计等安全机制，在一定程度上填补了大数据平台安全机制的空缺，基本满足目前平台的安全需求，但 Hadoop 仍处在快速发展的阶段，认证机制依赖 Kerberos，其认证中心可能会成为系统瓶颈。平台防攻击技术方面，目前大数据平台仍然使用传统网络安全的防护手段，对大数据环境下扩大的防护边界和更加隐蔽的攻击方式无法做到全面覆盖，而且行业对大数据平台本身可能的攻击手段关注较少，预防手段不足，一旦有新的漏洞出现，波及范围将十分巨大。

数据安全方面，数据安全监控和防泄露技术相对成熟，数据的共享安全、非结构化数据库的安全防护以及数据泄露溯源技术亟待改进。目前，数据泄露问题在技术上可以得到

较完备的解决，敏感数据自动化识别为防泄露提供了基础技术；人工智能、机器学习等技术的引入，使得数据防泄露向智能化方向演进；数据库防护技术的发展也为数据泄露提供了有力的技术保障。密文计算技术、数据泄露追踪技术的发展仍无法满足实际的应用需求，难以解决数据处理过程的机密性保障问题和数据流动路径追踪溯源问题。具体而言，密文计算技术的研究仍处在理论阶段，运算效率远未达到实际应用的需求；数字水印技术无法满足大数据环境下大量、快速更新的应用需求；数据血缘追踪技术未获得足够的应用验证，其成熟度尚未达到产业化应用水平。

隐私保护方面，技术的发展明显无法满足当前迫切的隐私保护需求，大数据应用场景下的个人信息保护问题需要构建法律、技术、经济等多重手段相结合的保障体系。目前，应用广泛的数据脱敏技术受到多源数据汇聚的严重挑战而可能面临失效，匿名化算法等前沿技术目前鲜有实际应用案例，普遍存在运算效率过低、开销过大等问题，还需要在算法的优化方面进行持续改进，以满足大数据环境下的隐私保护需求。如前所述，大数据应用与个人信息保护之间的突出矛盾不单是技术问题，尤其是在缺乏技术保障的当下，更需要通过加快立法、加强执法规范大数据应用场景下的个人信

息收集、使用行为，尽快构建政府管理、企业履责、社会监督、网民自律等多主体共同参与的个人信息保护制度体系。

五、大数据安全技术未来发展建议

大数据正在成为经济社会发展新的驱动力，日益对经济运行机制、社会生活方式和国家治理能力产生重要影响，大数据安全已上升到国家安全的高度。基于所梳理的大数据安全问题挑战与大数据安全技术发展现状，我们对大数据安全技术的发展提出如下几点建议：

（一）需要站在总体安全观的高度，构建大数据安全综合防御体系

安全是发展的前提，必须全面提高大数据安全技术保障能力，进而构建贯穿大数据应用云管端的综合立体防御体系，以满足国家大数据战略和市场应用的需求。一是建立覆盖数据收集、传输、存储、处理、共享、销毁全生命周期的安全防护体系，综合利用数据源验证、大规模传输加密、非关系型数据库加密存储、隐私保护、数据交易安全、数据防泄露、追踪溯源、数据销毁等技术，与系统现有网络信息安全技术设施相结合，建立纵深的防御体系；二是提升大数据平台本身的安全防御能力，引入用户和组件的身份认证、细粒度的访问控制、数据操作安全审计、数据脱敏等隐私保护机制，从机制上防止数据的未授权访问和泄露，同时增加大数据平

台组件配置和运行过程中隐含的安全问题的关注，加强对平台紧急安全事件的响应能力；三是实现从被动防御到主动检测的转变，借助大数据分析、人工智能等技术，实现自动化威胁识别、风险阻断和攻击溯源，从源头上提升大数据安全防御水平，提升对未知威胁的防御能力和防御效率。

（二）从攻防两方面入手，强化大数据平台安全保护

平台安全是大数据系统安全的基石，基于前面的分析可以看出，针对大数据平台的网络攻击手段正在发生变化，企业面临愈加严峻的安全威胁和挑战，传统的安全监测手段难以应对上述攻击变化，未来大数据平台安全技术的研究不仅要解决运行安全问题，还要进行理念创新，针对不断演进的网络攻击形态，设计大数据平台安全保护体系。在安全防护技术方面，目前无论是开源还是商业化大数据平台，都处在高速发展阶段，在平台安全机制方面的不足之处依然存在，同时，新技术新应用的发展也为平台安全带来未知的安全隐患，需要产业各方在大数据平台安全方面加大投入，从攻防两方面入手，密切关注大数据攻击和防御两方面的技术发展趋势，建立适应大数据平台环境的安全防护和系统安全管理机制，构筑更加安全可靠的大数据平台。

（三）以关键环节和关键技术为突破点，完善数据安全技术体系

大数据环境下，数据在流动中发挥价值，其应用生态环境日益复杂，数据生命周期各环节都面临新的安全保障需求，数据的采集和溯源成为突出的安全风险点，跨组织数据合作的广泛开展触发了多源汇聚计算的机密性保障需求。目前，敏感数据识别、数据防泄露、数据库安全防护等技术发展相对成熟，多源计算中的机密性保护、非结构化数据库安全防护、数据安全预警以及数据发生泄露事件的应急响应和追踪溯源等方面还比较薄弱。应积极推动产学研用结合，加快密文计算等关键技术在运算效率提升方面的研究和应用推广。企业应加强数据采集、运算、溯源等关键环节的保障能力建设，强化数据安全监测、预警、控制和应急处置能力，以数据安全关键环节和关键技术的研究为突破点，完善大数据安全技术体系，促进整个大数据产业的健康发展。

（四）加强隐私保护核心技术产业化投入，兼顾数据利用和隐私保护双重需求

在大数据应用场景下，数据利用和隐私保护是天然矛盾的两端，同态加密、多方安全计算、匿名化等技术可以实现这两者良好的平衡，是解决大数据应用过程中隐私保护问题的理想技术，隐私保护核心技术方面的进展必然会极大推动大数据应用的发展。目前，隐私保护技术的核心问题是效率，存在计算开销大、存储开销大、缺乏评价标准等问题，均处于理论研究阶段，尚未在工程实践中广泛应用，难以应对多

数据源攻击、基于统计的攻击等隐私安全威胁。大数据场景下，个人隐私保护已成为一个备受关注的议题，未来日益膨胀的隐私保护需求将带动专业化隐私保护技术的研发和产业应用。需要鼓励企业、科研机构研究同态加密、多方安全计算等前沿隐私保护算法，同时推动数据脱敏、数据审计等技术手段在大数据环境下的增强应用，提升大数据环境下隐私保护技术水平。

（五）重视大数据安全评测技术的研发，构建第三方安全检测评估体系

当前，国家就大数据安全进行了一系列重大决策部署，习近平主席在十九大报告中指出要推动大数据与实体经济深度融合，并强调切实保障国家数据安全。《"十三五"国家信息化规划》提出实施大数据安全保障工程。可以预见，未来大数据安全政府监管将进一步加强，数据安全相关立法进程将进一步加快，大数据安全监管措施和技术手段将进一步完善，大数据安全监管惩戒力度将进一步加强。同时，构建大数据安全评估体系将成为保障大数据安全的有效举措，通过制定大数据安全技术标准和测评标准，建立大数据平台及大数据服务安全评估体系，推进第三方评估机构和人员资质认证等配套管理制度建设，从平台防护、数据保护、隐私保护等方面切实促进大数据安全保障能力的全面提升。