大数据安全保障关键技术分析综述

王 丹, 赵文兵, 丁治明
(北京工业大学计算机学院, 北京 100124)

摘 要: 针对作为各个行业信息化建设重要基础支撑的大数据在收集、存储、传输和使用过程中面临的诸多安全风险,分析了大数据在这些过程中面临的安全威胁. 综述了大数据相关系统平台和大数据应用中安全保障的关键技术及最新进展,包括用户访问控制、数据隔离、数据完整性、隐私保护、安全审计、高级持续性攻击(advanced persistent threat,APT)防范等,以应对云计算、物联网、移动互联等新技术的快速发展对大数据带来的安全挑战和更高的安全要求. 同时也对大数据的安全保障技术的发展趋势进行了展望.

关键词: 大数据; 安全; 关键技术
中图分类号: TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2017)03 - 0335 - 15
doi: 10.11936/bjutxb2016020025

Review of Big Data Security Critical Technologies

WANG Dan, ZHAO Wenbing, DING Zhiming
(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: For big data has been pervasively applied in many industries as crucial information supports and confronted with lots of security risks in the procedure of collection, storage, transmission and application, this paper analyzes the security threats in big data's life cycle, summarizes the security critical technologies and new developments according to big data system platforms and big data applications, including user access control, data isolation, data integration, privacy protection, security audit, precaution advanced persistent threat etc., to meet the new security challenges and requests proposed by the rapid developments of cloud computing, Internet of Things, mobile internet, etc. It also prospects the security tendency for big data.

Key words: big data; security; critical technologies

继云计算、物联网之后,大数据成为信息技术领域的又一热点[1-3],在数据挖掘[4-6]、人工智能[7]、社会计算[8-9]、生物与化学[10-11]等领域的应用日渐深入. 大数据体量巨大、高速产生、类型多样、分布协同的特征,使其在收集、存储、传输和使用的过程中面临着诸多安全风险[12-14]. 首先,由于数据集中、体量庞大、数据价值稀疏,安防工具难以聚焦于价值点上. 其次,分布式处理增加了大数据泄露的风险. 同时,大数据技术同样能够帮助黑客最大限度地收集相关信息,实施更具精准性的攻击. 针对这些问题,现有的信息安全技术不能在大数据环境中完全有效的运用[15-17]. 因此,大数据系统的所有者和管理者极为关注大数据的安全保障问题,对大数据进行有效的安全防护已经成为计算机领域的热点研究之一[18-20].

本文首先分析大数据环境下的安全威胁. 然后论述现有保障大数据安全的关键技术和有关应对策略,最后展望大数据安全的未来.

收稿日期: 2016-02-26
基金项目: 国家自然科学基金资助项目(91546111); 北京市教育委员会资助项目(PXM2015_014204_500221)
作者简介: 王 丹(1969—), 女, 教授, 主要从事计算机软件、Web安全、可信软件方面的研究, E-mail:wangdan@bjut.edu.cn

1 大数据面临的安全威胁

伴随 Web 应用的普及,交易、互动、对话等越来越多的社会生活都是在网络空间内进行的,相伴而生的大数据已成为开放网络中主要的被攻击目标. 大数据多依托于云存储,大数据之间具有很强的关联性,同时,多数大数据的安全保护措施很简单,大数据蕴含信息的价值带来巨大的吸引力,使其更易成为被攻击目标.

大数据和云计算的出现,导致数据所有权和使用权分离,产生了数据所有者、提供者、消费者 3 种角色,而非传统技术时代数据一直处于所有者的可控范围之内,并且数据所有者存在数据安全合规性的要求. 因此,大数据应用出现后,数据的生命周期从传统的产生、存储、使用、传输和销毁 5 阶段变为产生、收集、存储、使用、传输、共享、发布和销毁 8 个阶段,每个环节都面临着新的安全威胁和挑战. 结合大数据的特征,本文从大数据的产生、收集、存储、传输和使用几个阶段分析大数据面临的安全威胁,介绍大数据处理平台和针对大数据的新型攻击与防御.

1.1 收集阶段

在大数据收集阶段,数据源存在潜在被攻击的威胁. 例如,传感器网络中,被俘获的节点可以轻易绕过网络路由信任机制的核验,使目标区域的用户节点接收不到数据或者接收虚假数据,实现阻断、篡改网络数据或者旁路至敌方节点的不良企图. 应对该问题需要在节点之间增加额外的互认机制:当节点收到其他节点发来的消息时,需要确认监测区内发送数据包的节点的有效性,而非虚假节点冒充[21-23]. 网络技术和数据挖掘的广泛使用,使得大数据能够被高效自动收集、初步进行智能动态分析,同时也增加了大数据的安全风险. 此外,用户的关键隐私数据可能会被采集、流转至非信任区域,从而失去对这些关键数据的控制,产生隐私数据泄露等安全风险[24].

1.2 存储阶段

1.2.1 大数据所依托的云存储自身面临的安全威胁

大数据基本以云计算作为存储的架构,而此架构没有清晰定义的安全边界,为安全保护措施增加了难度,数据传输时的完整性和保密性受到很大威胁,数据的完整性、容错性、可恢复性等方面也存在一些安全问题. 同时,需要用更高的数据安全保障标准来要求多客户连接、高交互性以及作为公共数据中心的云存储[25],也出现了混合云存储安全的跨域身份认证问题,不可信云服务商提供存储服务时的数据加密问题等新情况. 服务的多级模式带来统一安全监控问题. 云存储带来数据安全与服务的信任问题[26].

针对存储服务比较常见的攻击方式有篡改、删除和盗用. 通过数据完整性检测可以帮助用户或者用户委托的第三方判别云服务器中数据的原始性保持情况,相关攻击有:1) 替代攻击. 当服务器丢失用户关注的数据时,仍然返回未丢失标识以及相应证明来欺骗用户;2) 重放攻击. 当服务器收到数据检测请求时,旋即将此前检测的计算结果发给用户而非按照协议重新计算,造成数据完整的假象;3) 伪造攻击. 服务器通过伪造用户要求检测的数据以及相关证明来欺骗用户[27].

1.2.2 大数据类型的多样性对存储安全提出了挑战

80% 以上的大数据是非结构化的,通常采用 NoSQL 的形式存储. 此种方式存在存储模式不成熟,代码漏洞较多,自身安全机制尚不完善等问题,并且在数据安全方面未设置严格的访问控制和隐私管理,尚需时间来检验各种措施的安全性. 用作服务器软件的 NoSQL 未有足够的安全内置,以致客户端应用程序需要内建安全措施,因此产生授权过程、身份验证和输入验证等安全问题. 时至今日 NoSQL 尚未形成业界标准,虽然各种产品层出不穷,但由于各自为政、自成体系,使得企业很难采取统一的安全策略来保护机密信息[28-29].

1.3 传输阶段

数据传输需要各种协议相互配合,有些协议缺乏专业的数据安全保护机制,因此,跨平台传输可能给大数据带来安全风险. 数据采集过程中存在的误差造成数据本身的失真和偏差. 不够完善的数据版本管理也可能造成信息的误判. 数据传播过程中的泄漏、破坏或拦截,会带来隐私泄露、谣言传播等安全管理失控的问题. 同时,大数据平台中的节点交互、分布式存储架构和协同计算可能引发云环境下数据存储的安全与完整性,密钥管理与可销毁性,安全验证与安全组合等问题. 大数据系统可能被恶意用户或者恶意节点注入恶意信息,迅速并且全方位地扩散,严重危害广大用户和整个系统[30-31]. 因此,大数据传输中,信道安全、数据防破坏、数据防篡改、设备物理安全等几个方面需要着重考虑.

1.4　使用阶段

1.4.1　非授权访问和数据的非常规使用

数据的非授权访问可能出现在数据分析、数据挖掘的过程中. 系统维护可能丢弃原始数据或加工数据，导致关键数据的损失. 数据交付过程中可能发生数据挖掘结果的不合规发布，缺乏应对数据泄露的预案、无法追溯取证等问题. 大数据的隐私保护包括个人隐私的保护和隐私数据在存放、传输和使用过程中避免泄露[32-33]. 大数据中的相当比例包含位置信息，个人位置隐私保护的重要性也日益凸显[34-35].

1.4.2　访问控制机制需要加强

新的应用资源源源不断地接入大数据系统平台，因此，访问控制是实现数据受控共享的有效手段. 进行访问权限细粒度划分，构造用户权限和数据权限(只读、只写、读写)的复合组合控制方式，是对访问权限的有效控制途径，在一定程度上能够提高敏感数据的安全性[15]. 然而，大数据应用中用户多样性带来的权限多样性要求，超过了自主访问控制能够实现的安全级别，强制访问控制无法满足权限的动态性需求，角色访问控制不能将角色和权限有效地对应起来[32,36-37]. 因此，在大数据架构下的访问控制机制还需要对这些新问题进行分析和探索.

1.5　大数据处理平台面临的安全威胁

大数据应用需要在强大的基础计算能力支持下实施. 基于云计算高性价比、分布式处理的优势，很多大数据应用都部署在云平台上. 在云平台未有周全的保护或者监管措施的情境下，云服务可能给大数据带来账户或服务劫持、数据泄漏、数据滥用等风险，甚而造成云计算风险等同于大数据风险的局面. 例如，Hadoop 作为一个大数据的基础处理平台，构建了高度分布式的、冗余和弹性数据存储池，以完成大规模并行计算. 但是，大数据堆栈各层的集成、数据节点与客户端/资源管理机构之间的通信等潜在风险处，都需要分析、设计有效的安全机制，以便在无数据泄露之虞的前提下处理海量数据的聚合[38-40].

1.6　新型安全攻击

针对大数据的新型安全攻击中最具代表性的是高级持续性攻击(advanced persistent threat，APT). APT 手段多样、目标明确、持续时间长，是一种有组织、有确定目标、隐蔽性强、破坏力大的攻击. 由于 APT 的潜伏性，使其持续性成为一个不确定的实时过程，令管理员难以察觉，实时检出率比较低. 基于代码的传统安全方对检测 APT 无效，APT 往往能够绕过传统防御体系，潜伏在系统中[41]. 大数据应用为入侵者实施可持续的数据分析和攻击提供了极好的隐藏环境. 传统的基于内置攻击事件库的实时匹配分析检测，是对单时间点进行威胁特征攻击的检测技术，对于 APT 很难奏效[42-44]. 从近年来互联网上的用户账号的信息失窃大事件及其后续连锁反应可以看出，大数据更容易吸引黑客，而一旦攻击得手，失窃的信息量甚至是难以估量的.

2　大数据系统平台安全保障的关键技术

大数据处理平台通常由数据存储层、数据处理层、应用接口层和业务支撑层组成，安全解决方案需要综合考虑这 4 个层面. 安全关键技术包括：面向不同数据类型的存储和处理技术、大数据系统的统一策略管理、配置基线检查和监控技术、大数据并行去隐私化技术、策略化抽取和集成技术、多维度大数据审计技术、访问监控和报警技术、访问行为追踪技术等. 很多应用于传统环境的安全技术也可以在大数据环境中使用. 但是，由于大数据应用环境的特殊性，还需要其他的一些新的安全技术来确保大数据系统平台的安全性. 云端大数据的安全防护措施常见的有以下几种类型.

2.1　虚拟化安全

大数据应用所需的分布式处理能力依赖于虚拟化技术，虚拟化大幅提升了基础计算资源的利用率[45-47]，同时也带来了一些安全漏洞[48]. 比如，开源的虚拟化软件内核虚拟机(kernel virtual machine，KVM)绕过安全权限控制的问题. 虚拟化安全是云端大数据防护的基础，目前，用以设计虚拟化安全方案的思路有 2 种：1) 通过虚拟化层本身的安全改造，从底层实现虚拟资料的安全防护，此种方案实现难度较大；2) 在虚拟机上加载安全模块，对虚拟机进行数据加密、完整性保护等防护措施，此种方案要求较高的大规模部署能力.

通常情况下，云计算中不同用户的多个虚拟机建立在同一物理资源上，必须采用有效的隔离措施，才能防止数据泄漏. 虚拟机扫描技术是行之有效的安全解决方案之一，即直接扫描虚拟机或通过虚拟机中安装软件监控用户的虚拟机，以确保当前用户的虚拟机正常运行、未进行非法计算或访问[49].

同时，云服务通过虚拟化安全集中管控平台来统一管理所有的虚拟机和安全组件[50]，具体技术

有：1）信息同步. 虚拟化集中管控平台定时或按需自动同步虚拟化防火墙和云平台信息，为其他组件提供云平台、虚拟机的信息及状态变化，及时了解云平台健康状况和虚拟化防火墙的状态. 2）云平台完整性监控. 如果云平台组件被他人恶意篡改，则有可能造成云平台的不稳定甚至数据泄露，所以保障云平台的完整性对于云服务来说是至关重要的. 通过云平台信息及状态对其完整性进行监控，及时发现系统平台、组件的变化并通知管理员，从而保证云平台的正常运行. 3）虚拟机补丁管理. 为了修补虚拟机操作系统的漏洞，虚拟化集中管控平台对补丁进行管理，将运维人员测试过的系统补丁根据预先设置的策略在适当时间下发，在尽量不影响业务运营的同时完成补丁更新，实现安全管控. 4）防火墙的集中管控和策略下发. 随着云规模的增大，虚拟化集中管控平台通过信息同步数据获取云平台中所有虚拟化防火墙信息，以便运维人员使用 Web 界面监控其状态，及时发现虚拟化防火墙的状态异常，并且支持策略配置及策略的集中下发，同时对迁移的虚拟机进行虚拟化防火墙迁移或安全策略迁移，削减运行维护云服务成本.

2.2 虚拟网络安全

在软件定义型网络（software definition network，SDN）、网络功能虚拟化（network function virtual，NFV）等技术的冲击下，虚拟设备逐渐取代了传统的硬件网元设备，传统的网络运维面临崭新的安全挑战. 例如，开源虚拟交换机 Open Vswitch 数次被发现目录权限存在漏洞，开源云平台 Openstack 的网络模块 Neutron 存在信息泄露问题. 有鉴于此，开源网络虚拟化技术应用之前需要对虚拟网络进行安全加固，可以采用安全代理、安全加密等方式[51-52].

虚拟交换机网络是虚拟网络的核心，它所带来的问题是虚拟网络的安全边界不清晰，传统的交换机、防火墙等设备无法监测和控制虚拟机间的数据流. 虚拟网络安全防护的一种重要技术是划分安全域[53-55]. 通过分析虚拟资源的使用状况，划分为彼此隔离的安全域. 不同安全域之间不能交换信息，用以保证域内数据只能由同域的虚拟资源分享，不致外泄其他域. 安全域与物理主机之间的映射是多对多的联系，既扩大了安全域可能的疆界，使多个物理主机上的虚拟资源能够划分到一个安全域内，跨物理主机实现安全域，又方便虚拟资源划分到不同的安全域内，实现物理资源的共享. 在安全域的基础上，虚拟网络安全技术包括：1）安全域的划分. 使用 VLAN 的网络隔离技术划分虚拟网络，确保不同域内虚拟机通信数据的保密性和通道的独立性；2）虚拟机准入控制. 在不能保证虚拟机安全的情况下，将不符合安全策略要求和可疑的虚拟机放入隔离环境中，限制或者禁止其访问网络，将虚拟机的安全状态信息与网络准入控制捆绑在一起，增强虚拟网络环境的安全性；3）域间访问控制. 通过对安全域间通信实施不同粒度的访问控制和加密策略，防止有恶意的虚拟机监控虚拟网络数据，保护用户通信的安全性[56].

虚拟机本身具有良好的隔离性，但是位于虚拟机管理器（hypervisor）之上的虚拟机之间可能通过某种隐蔽通道进行相互攻击，含有安全风险的虚拟机也可能威胁到宿主机的安全. 这些问题可以通过对虚拟机之间的通信进行更为精细的访问控制加以解决，通过建立在 Bell-La Padula 模型上针对虚拟机的多级安全强制访问控制框架，避免虚拟机之间的隐蔽通道攻击和虚拟机对宿主机的安全威胁[50,57].

2.3 大数据平台防护措施的加固

Hadoop 已经成为大数据时代存储和处理海量数据的热门技术，因此，改造大数据处理平台成为了大数据安全防护的重心之一，特别是 Hadoop 的安全加固上. 较为常见的 Hadoop 加固方案如图 1 所示[58].

其基本工作原理是：Hadoop 平台上加载了海量数据之后，首先需要对数据进行解析与清洗，通过数据分析与挖掘发现异常，其后才能进入大数据处理阶段. 大数据处理流程的安全审计与回溯需要在系统旁路部署监控与管理模块加以实施[59].

大数据平台的组成中通常包括繁杂的网络拓扑结构、种类繁多的硬件设备、各种操作系统与应用软件，这样，复杂的系统可以通过安全策略进行加固[30,50]，包括：

1）采用统一的安全策略描述语言，制订集中式的安全策略，配合处理语义丰富的网络安全事件. 网络安全性策略通过事件消息来实现，也就是说在策略被违反时做出主动响应. 事件承载对设备的策略请求和与该请求有关的交互，例如网络事件管理模型、策略模型和规范等.

2）采用统一的自驱动安全策略模型，对大数据平台中的安全产品和设备进行统一配置和管理，使得这些安全产品通过自驱动实现安全功能. 大数据平台中安全策略可由多个类型的策略节点构成，根据前序节点的执行结果选择驱动下一节点的执行，

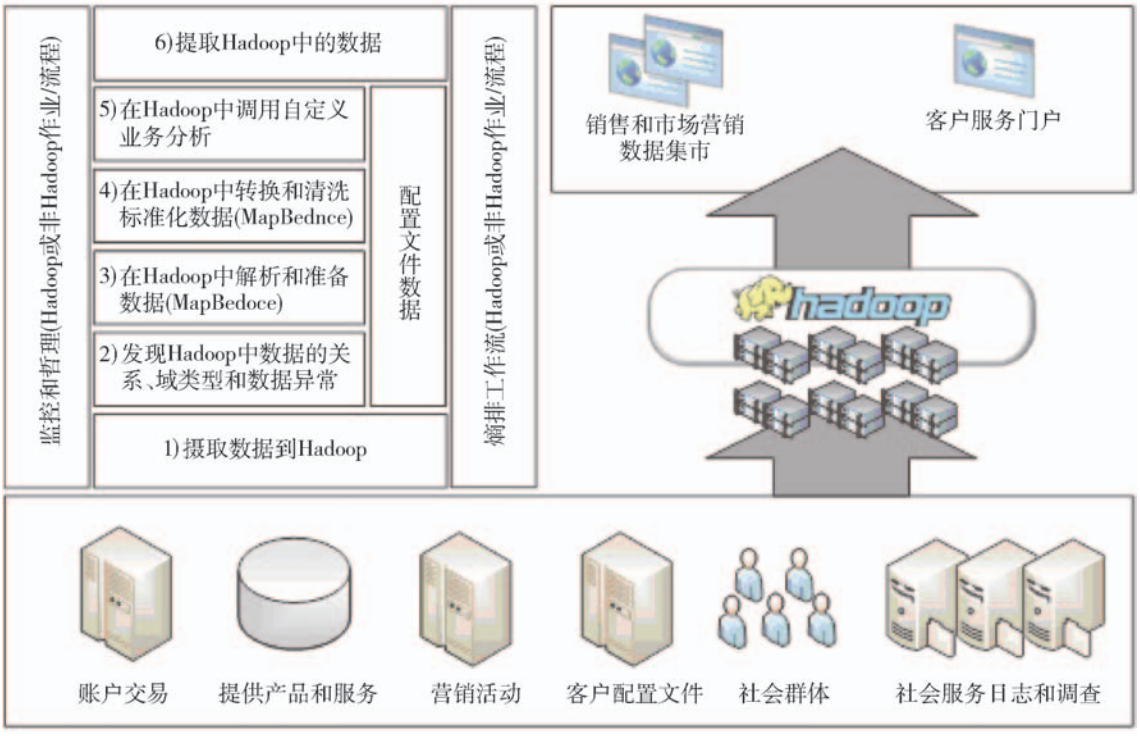

图 1　Hadoop 安全加固方案

Fig. 1　Framework for secure Hadoop

从而产生以工作流方式执行的、跨资源的、自驱动的执行链.

传统的事件关联技术难以对大数据应用中海量安全事件进行高效的分析和处理. 大事件应用中的海量安全事件,要求在平台范围内进行统一的实时网络安全整体架构,需要将安全事件的时域和空域归并关联处理,利用安全事件关联算法,提高归并关联效率.

安全事件的关联可以将大数据平台上的多源头、多级别、多类型的安全信息进行汇聚和降维,消除冗余的安全事件;利用数理统计和模型推理,对相互关联的安全事件进行归并处理,建立安全事件的生命周期模型,帮助安全管理人员从纷繁芜杂的安全事件中快速定位首要威胁;汇聚多源安全事件的类型、来源和目的,综合处理、分析事件的起因、发生的时段和对平台造成的威胁,实现安全事件的自动关联. 安全事件的归并关联可以采用基于多特征的事件关联、基于攻击序列的启发式关联、攻击场景重建等技术[60].

3) 虚拟化电子取证追溯. 大数据平台上的存储和计算资源均已虚拟化,资源动态迁移会产生信息加工、存储的节点变化. 这种情况下,传统方式无法完成异常数据的分析取证,需要设计相应的虚拟化取证方法.

其中一种思路是将云平台视作多个虚拟机构成的系统,对其上运行的虚拟机实例进行取证分析,进行云计算建模;之后利用现场迁移技术,在虚拟化软件层中进行虚拟机实例的信息保全,确保迁移镜像文件内容的完整性和一致性.

针对虚拟化环境中复杂的身份转换机制,按照数字证据的特点和法律要求,通过事前内置虚拟身份追踪与取证机制、事中增强丰富的虚拟身份相关行为审计、事后完善虚拟身份相关证据保全方法,从而构建面向云环境的立体化虚拟身份取证机制,为云计算环境提供符合法律要求的虚拟身份相关证据的获取、保全、分析、展示等一体化功能[61].

2.4　大数据存储安全

大数据安全储存的根本目标是保证存储数据的安全,主要采用虚拟化海量储存技术. 大数据存储安全可以通过固件获得[62],也可以通过软件来实现,既包括传统的存储加密和信息安全技术,也覆盖大数据依赖的云存储所带来的特殊安全问题和技术[63].

2.4.1　数据加密以保证私密性

传统的数据安全存储首先要把数据集进行适当的加密保护,之后通过加密措施把数据的使用与保

管过程加以分离,以实现密钥与要保护的数据的隔离[64]. 然而,鉴于大数据的体量,使用传统方法加密则要消耗大量 CPU 的计算时间,严重影响大数据处理系统的性能,为了解决这个问题,同态加密是大数据加密的可选技术之一[65].

同态加密的核心思路是构造等价的加密函数,即对明文进行环上的加法和乘法运算后再加密,与加密后对密文进行加法和乘法运算等价. 基于不同运算的同态性,可以委托未经信任审核的第三方对数据进行处理而不泄露信息,极大地缓解了大数据平台的计算压力. 不仅如此,使用全同态加密技术的数据检索方法可以直接对加密的数据进行检索,既保证了被检索的数据不被统计分析,还能使得被检索数据经过简单运算后,仍然保持对应的明文顺序[66]. 因此,同态加密成为大数据私密性保护中的研究热点[67-68].

2.4.2 数据备份以保证数据可用性

数据的可用性是大数据应用研究中的热点. 现有的大数据系统中,数据的可用性多通过数据块备份的方案来加以保障,即将数据分块,把数据块复制备份,分别存储到多个数据节点,以便在某些数据节点失败的情况下,仍然能够保证数据的可用性[69-70]. 例如,在 Hadoop 中,Hadoop 文件系统通过机架感知技术管理文件系统的元数据(NameNode)确定每个实际存储的数据(DataNode)所属的机架 ID,采用机架感知策略来改进数据的可靠性、可用性和网络带宽的利用率,实现了文件块的多副本备份[71].

云环境下的多副本管理大多依托于现在的云存储技术,例如常用分片存储和可擦写代码技术. 数据分片的目的之一是提高并行数据写入和读取性能,二是可以与副本、可擦写代码一起实现数据保护. 数据分片后存储的方式有轮询式存储、最大化存储方式和均衡性协约方式. 分片数据存储后,会根据读的请求还原成原始文件,但这会消耗较多的资源,因而需要综合考查需求和资源约束来决定是否要分片[72-73].

数据分片处理之后,可以进一步对不同分片进行分布式存储. 其中,可擦写代码(erasure code)模式的特点是没有复制技术,也能提高数据冗余性,控制数据保护带来的开销. 在这个机制中采用 M∶K 的冗余容错机制来保护数据,如 9∶3、10∶6. 数据通过分片和生成校验码进行存储和读取. 在 9∶3 的模式中,数据分片以 $M=9$ 的模式进行,校验码为 $K=3$. 数据分片的每个分片都存在不同硬盘上,数据能从任意一个分片上恢复[74-76]. 这样,即使丢失部分校验码,数据一样能被恢复而且整个解码过程消耗很小[77].

3 大数据使用中安全保障的关键技术

3.1 用户访问控制

大数据跨平台传输产生的安全风险与用户访问控制密切相关,可以根据数据的密级和用户需求将大数据和用户设定不同的权限等级,严格控制访问权限,加强用户权限管理. 访问控制常用的手段有身份认证、口令加密、文件权限设置、网络设备权限控制等. 由于云计算强大的计算能力,传统数字密码的破解变得更为简单,传统的基于单一凭证的身份认证技术,比如口令核对法,基于智能卡的身份认证、一次性口令认证等认证方法等,已经不能满足大数据用户的安全需求. 生物特征认证更加复杂,难以破解,得到越来越多的关注. 这是因为云计算环境下认证中心的计算能力已经强大到足以支持识别组合的生物特征来进行身份认证,其计算能力的不断提升也使得实现生物特征的预处理、编码、压缩及加密成为可能,终端的智能化使生物特征的采集及鉴别成为可能,促进了生物特征认证的普及[15].

只在授权时核验用户身份的真实性,权限授予后不再监管的缺陷可能造成用户对资源的访问过程中执行非法操作而不被发现的严重问题. 大数据是多用户共享的资源,关卡式的访问控制模型存在对用户后期行为监控盲区的安全隐患. 比如 Hadoop 平台现在的访问控制模型:利用 Kerberos 安全认证协议完成对用户的身份验证,结合平台/内部的访问控制列表(access control list,ALC)访问授权机制,通过与 Delegation Token、BlockAccess Token 等令牌的配合使用,实现对集群中用户的访问控制过程[30]. 但是,该模型以关卡式的模式执行,用户后期即使对集群资源存在非法行为也不会被集群发现. 因此,亟需建立在行为监控基础上的用户信任度评估模型和访问控制模型,使访问权限的授予过程由关卡式的控制模式,变为实时动态的控制过程. 通过分析用户行为记录,实时更新用户信任值,动态控制用户对平台的访问,使得对用户的权限授予控制粒度更加精细化,使访问控制机制方面变得更加安全、灵活[78-80].

3.2 数据隔离技术

虚拟化技术的负面作用之一是削弱了数据间的

物理隔离，致使数据间的边界甚为模糊，每个用户都有成为发起攻击节点的潜在条件，对数据的安全构成了极大的威胁. 基于云计算平台的软件系统大多采用所有用户的数据共享一个软件系统实例的多租户(multi-tenancy)架构，加密虽然有效，但并非万能. 在云计算服务日趋专业化的同时，会因为服务的多层转包，环节过多疏于监控而引发数据安全问题，必须开发面向安全的数据隔离机制来确保用户之间的数据不可见.

目前数据隔离技术中比较成熟的有：

1) 共享表架构(shared schema multi-tenancy)，即通过某些特定字段标示数据的从属关系，所有软件系统共享相同的数据实例和数据库表. 这种架构的优势是最大限度利用了数据实例的存储能力，极大地降低了硬件成本；劣势是同时极大地增加了业务逻辑的复杂程度，相应产生了高昂的容灾备份成本.

2) 分离数据库架构(separated database)，与共享表架构相反，由于每个软件系统拥有单独的数据库实例，这种架构能够高效实现数据隔离和容灾备份，但是硬件成本也相对较高.

3) 分离表架构(shared database separated schema)，这种架构是一种折中方案，即每个客户都拥有自己的一系列数据库表，软件系统只共享相同的数据实例.

实现数据隔离和容灾备份与共享表架构相比要容易一些，与分离数据库架构相比硬件成本要低一些[81-82].

3.3　数据完整性保护

数据完整性检测能够检测出存储数据是否被篡改，当损失的数据量小于一定规模时，可以通过前期的编码恢复出完整的数据.

保证数据完整性的常用方法有信息认证编码和数字签名[67]. 信息认证编码依赖对称密钥产生校验并附加在数据后面，数字签名依赖公共密钥的结构. 由于对称算法较之于非对称算法在速度上占优势，数据的完整性检查大都采用信息认证编码的机制. 例如，普遍应用的分布式文件系统首先将较大的数据卷划分为默认 64 M 或 128 M 的数据块，接着在每个数据块后追加一个数字签名，然后存储起来以备后续完整性测试使用. 由于存在资源不足、任务烦琐、密钥管理等问题，用户通常难以亲自进行数据完整性的验证，即便用户检测出存在完整性问题也无法确定问题根源[83]. 对于这种情况，用户和云服务商通常采用可信第三方来完成检测，即用户将加密数据存储在云端，可信第三方预先计算用于验证的哈希值，以挑战/应答的方式验证云端存储数据的存在性和完整性. 该方法适合于静态数据，对于动态变化的数据则显得开销过大，难以为继. 大数据完整性验证大多采用轻量级完整性验证机制. 其中支持泛在接入和移动计算的大数据完整性验证机制多采用抽样策略，当数据发生位偏转时，用户或可信第三方可能无法及时地发现如此小概率的事件，因而，需要设计更加高效、合理的数据完整性证明机制. 此外，传统的数据完整性校验方法进行校验前需要将数据下载到本地. 由于大数据体量极大，下载数据块将使网络不堪重负，对于海量数据经常束手无策，因此，针对大数据的、能够在云端进行的数据完整性校验方法也是业内研究的重点[84-86].

数据完整性证明机制可以分为数据持有性证明(provable data possession，PDP)机制和数据可恢复证明(proofs of retrievability，POR)机制，其中是否对原数据采用容错预处理技术是二者的分野[25]. 2 种机制各有所长，PDP 机制能够快速判断远程节点上数据是否损坏，证明效率较高. POR 机制不仅能识别数据是否已损坏，而且能恢复已损坏的数据，功能更为全面. 2 种机制分别适用不同的场景，PDP 机制主要用于检测大数据文件的完整性，POR 机制则用于确保重要数据的完整性，如压缩文件的压缩表. 对于这类核心数据，哪怕只是损坏极小一部分，仍然能够造成整个数据文件的失效[87-89].

3.4　隐私保护

目前，用于大数据隐私保护[19,90-91]的技术有：

1) 通过扰动原始数据，造成数据失真，使攻击者不能发现真实的原始数据，比如在位置数据中加入假位置(fake location)和哑元(dummy)用于位置扰乱以保护位置隐私[92-93].

2) 通过安全套接字协议层(secure sockets layer，SSL)对大数据进行加密，实现数据集之间的保护.

3) 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，通过限制发布、数据匿名化实现隐私保护[94-96].

隐私保护最基本的手段有属性控制和匿名 2 种方法. 属性控制由用户自行设置属性的他人可见或不可见. 匿名方法在数据发布时隐去表明用户身份的属性，如姓名、身份证号、地址等. 上述 2 种方法都基于敌手无任何背景知识或其他数据来源的假

设. 当攻击者具有一定的背景知识或者可以利用多个数据源进行交叉连接,匿名方法是不安全的. 为此, k-匿名隐私保护模型应运而生,这种模型对半身份属性进行归纳表示,以保证任意一个归纳组都至少有 k 条记录. 然而,当同一归纳组内隐私的熵过低时,敌手不需要匹配到某一条记录便能获得受害人的隐私. 为解决该问题,又提出了 l-多样性模型,即在保证 k-匿名的基础上进一步归纳,使得同一桶内至少包含 l 个不同的隐私值. 但是当攻击者拥有丰富的背景知识时,结合发布的信息进行关联分析,还是能够推断出某个记录的敏感信息. 大数据环境下,为了减少数据共享或发布时无意间的泄露,在传输前应该对数据匿名化,同时结合其他技术使接受者对收到的数据无法开展关联推断,这样既能利用数据,又能避免涉及具体的个体[96]. k-匿名算法的扩展可以用于位置隐私保护[97-98].

此外,还有差分隐私算法,向包含隐私属性的数据表添加噪音,使得在该表中添加或删除某个元组对特定数据操作和查询产生的影响小于阈值,如此处理之后,敌手在访问不同数据集时,不同数据集中受害人是否存在敏感信息的后验概率不会有明显改变,因此无从推测受害人的隐私信息[99].

在保障云存储安全的推动下,数据隐私保护向着动态数据隐私保护和用户访问行为的隐私保护两方面发展. 用户的隐私数据中需要参与计算的为动态数据,对其隐私保护尚无彻底的解决方案,全同态加密为动态数据隐私保护提供了一种理论支持,但在实用性上还存在相当的差距. 用户对数据的访问行为也在一定程度上涉及用户的隐私. 恶意攻击者可以通过对行为及背景的总结来猜测数据的重要性、相关领域等信息,采用新型的访问控制模型不失为一种较好的途径[100-101].

3.5 安全审计与预测

大数据处理平台也采用安全审计技术来对安全事件进行跟踪,以及时发现安全违规事件,便于进行安全事件追责[60,76]. 安全审计的资料收集阶段首先通过系统 IP 扫描、系统端口扫描、系统漏洞扫描等方法搜集原始的系统状态信息,然后将原始状态信息和已有的安全记录(包括已经发生的安全问题及其他类似系统发生的安全问题)进行汇总整理,以此为基础通过数理统计导出相应的结论. 在结论分析的基础上,制定安全等级,采取相应的安全应对措施,预防可能会发生的安全问题[88,102-103].

同时,预防黑客入侵和病毒传播也是大数据安全防护的重要技术和保障手段[104-106]之一. 根据发生系统异常问题所涉及的数据对象,结合异常问题所发生的监控点、参考相似或类似问题的分析结果,通过对一系列历史数据和当前系统实时数据的场景关联分析,预测将来可能会发生或将要发生的安全问题. 通过分析确定问题事件的性质,预测可能存在的安全威胁,并对此安全威胁进行跟踪分析,做好应对此安全威胁的安全防护措施,提高应对安全威胁的安全防护级别.

对大数据的安全问题也可进行可行性预测分析,识别潜在的安全威胁[13]. 通过系统应用日志对已发生的系统操作或应用操作的合法性进行审核. 通过备份信息审核系统与应用配制信息对比审核,判断配制信息是否被篡改. 通过预测分析的研究,结合机器学习算法,利用异常检测等方法,提升大数据安全识别度. 对安全系统内部系统间或服务间的隐密的存储通道的稽核,即对发送和接收信息进行审核,则可以降低系统安全风险[105,107].

3.6 防范 APT 攻击

3.6.1 APT 攻击的防范策略[43-44]

1) 发现策略:由于入侵阶段对攻击者而言很重要,因此通过对本地和外界的流量交互进行检测,对从同一源(无论本地外界)频繁发出的行为异常进行定位与分析,有助于在入侵阶段发现 APT 攻击. 为了确定何种流量数据是非法的,需要大量的流量数据进行行为建模,精确地定位非法操作,增加发现 APT 攻击的可能性.

2) 对抗策略:对敏感数据存储位置,有可能是木马程序控制信息的流量数据以及不正常的周期性信号进行重点跟踪,尽快发现被劫持的主机和网络域,并第一时间进行隔离. 同时,根据日志记录,判断有可能被劫持的本地用户,并进行深入检测和隔离.

3) 预防策略:关键在于对 APT 时刻保持高度重视和高度警惕,增强自身安全意识,合理化数据存储.

3.6.2 常见 APT 攻击的防范手段

1) 沙箱方法:由于 APT 攻击通常使用零日漏洞等高技术手段进行攻击,对其进行匹配检测有着相当大的难度. 沙箱检测最主要的特点是将容易成为 APT 攻击对象的本地文件系统、注册表等内容同一个虚拟的文件系统相互分离. 当虚拟的本地系统被更改时,真实的本地系统并不会被一并更改,而是会重新定向到另外的路径,将虚拟环境的变化加以

保存. 这样就可以做到一面放任攻击者在沙箱中肆意攻击，一面通过对文件系统变化数据的分析总结其攻击方法，并找到应对方式，从而在真正的本地系统中加以实施，应对并预防相应的 APT 攻击. 遗憾的是，沙箱方法大量消耗资源. 另外由于沙箱系统所形成的虚拟环境根据真实环境中的系统，浏览器等与网络交互的资源不同会有很大的不同，对于在一些环境下用沙箱系统可能能够检测出来的 APT 攻击，在另一些环境中则有可能无法发现[108].

2）异常检测方法：通过对确定未受到攻击时的本地流量日志的数据进行总结分析，形成模型. 之后每当接收新到来的流量数据，就将其与安全流量建模进行匹配对比. 对于 APT 攻击的可能目标而言，网络流量是一笔非常大的数据. 而异常检测法则强调利用少量的数据对整体网络流量的异常进行检测. 发现网络异常之后根据对正常流量数据模型的匹配来确定网络异常的形式，甚至是定位 APT 攻击的目的. 这通常是通过检测一些可疑的加密文件和频繁而不正常的心跳信号传输来完成的. 该方法的关键是安全数据建模的可信度以及整体网络流量的检测，若能够在这 2 点有所突破，异常检测方法不失为相对小规模的有效 APT 防御手段[109-110].

3）基于记忆的检测系统：由于 APT 攻击长期潜伏的特点，为了能够发现其攻击模式和攻击手段，需要记录大量日志并对其进行全面分析，这种方法称为全流量检测. 但对于大量的流量数据，在尚未确定 APT 攻击威胁存在的情况下，全方位分析对资源的消耗是无法估量的. 可以将传统的网络安全检测方式与全流量检测进行结合，应用在基于记忆的检测系统中. 通过对日志的分析以及传统检测系统反馈的信息与常态进行对比，评估 APT 攻击可能发生的概率. 当用户认为 APT 攻击已经存在的时候，则调出可疑数据对应的时间段的流量数据日志进行全流量分析，从而进一步判断系统是否遭到 APT 攻击，以及判断 APT 攻击的手段和目的. 对于基于记忆的检测系统而言，其主要问题是传统检测方法究竟能够以怎样的效率去初步断定攻击的存在. 另一方面则是能否用一个足够有效的、可信的方法来对流量数据进行分析以获取 APT 攻击的具体信息[109-110].

4　结论

现阶段的信息安全防护手段已经不能很好地满足大数据时代的信息安全需求，不适应大数据技术不断发展的背景，因此，应加快研发大数据安全关键技术，尽可能找到新的突破口，以保证大数据技术更好的发展. 具体而言：

1）研究大数据支撑下的网络攻击追踪溯源技术[111-112]，在此基础上建立、完善数据安全管理体系，以期有效保护数据安全，确保国家网络空间安全. 数据安全管理体系，要以规范数据平台的建设作为前提，融合大数据时代先进的数据管理概念，并且对数据信息进行合理的动态监控与管理，从而使大数据安全防护变得规范、科学.

2）研发大数据中心安全防护技术，建立、完善云安全技术框架、云服务安全标准及其测评体系，确保基于云服务的数据中心安全. 云安全框架必须在资源层、虚拟层、服务层、应用层各个层次全面考虑各种安全和隐私保护问题. 着重考虑安全递交、安全存储、安全共享与访问、安全更新和安全销毁等阶段，以保证数据从产生到消亡的全生命周期安全. 同时建立安全指导标准及其测评技术体系也是实现云服务安全的重要措施之一.

3）针对大数据分散存储、共享应用、分头管理等特点，加大力度研发大数据管理系统技术、海量数据融合与集成技术、海量数据可视化分析、海量数据挖掘与预测分析技术、大数据支持的智能驱动安全技术、云安全服务技术、浏览器虚拟化技术、多级安全虚拟化桌面技术等关键技术，加快构建自主可控信息系统，在高速组网、集群计算机编程、扩展云计算能力、广泛应用部署等技术中考虑数据安全和隐私保护等问题，并取得突破，推动新型信息安全技术发展，形成自主核心技术优势，提高中国大数据安全技术水平[18,19].

参考文献：

[1] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.
CHENG X Q，JIN X L，WANG Y Z，et al. Survey on big data system and analytic technology [J]. Journal of Software，2014，25(9)：1889-1908. (in Chinese)

[2] 李学龙，龚海刚. 大数据系统综述[J]. 中国科学：信息科学，2015，45(1)：1-44.
LI X L，GONG H G. A survey on big data systems[J]. Science China：Info Sci，2015，45(1)：1-44. (in Chinese)

[3] LOGANATHAN A，SINHA A，MUTHURAMAKRISHNAN V，et al. A systematic approach to big data [J]. International Journal of Information & Computation

Technology, 2014, 4(9): 869 -878.

[4] HUANG Y Q, ZHU F Z, YUAN M X, et al. Telco churn prediction with big data[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne: ACM, 2015: 607-618.

[5] JORDAN M I. Computational thinking, inferential thinking and "big data"[C]// Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Melbourne: ACM, 2015: 1.

[6] ABITEBOUL S, DONG L, ETZIONI O, et al. The elephant in the room: getting value from big data[C]// Proceedings of the 18th International Workshop on Web and Databases. Melbourne: ACM, 2015: 1-5.

[7] 李武军, 周志华. 大数据哈希学习: 现状与趋势[J]. 科学通报, 2015, 60(5/6): 485-490.
LI W J, ZHOU Z H. Learning to hash for big data: current status and future trends[J]. Chin Sci Bull, 2015, 60(5/6): 485-490. (in Chinese)

[8] 唐杰, 陈文光. 面向大社交数据的深度分析与挖掘[J]. 科学通报, 2015, 60(5/6): 509-519.
TANG J, CHEN W G. Deep analytics and mining for big social data[J]. Chin Sci Bull, 2015, 60(5/6): 509-519. (in Chinese)

[9] 申学易, 买晓琴, 刘超. 基于互联网平台的大数据收集在社会认知研究中的应用[J]. 科学通报, 2015, 60(11): 986-993.
SHEN X Y, MAI X Q, LIU C. Application of Internet-based big data in social cognitive science[J]. Chin Sci Bull, 2015, 60(11): 986-993. (in Chinese)

[10] 宁康, 陈挺. 生物医学大数据的现状与展望[J]. 科学通报, 2015, 60(5/6): 534-546.
NING K, CHEN T. Big data for biomedical research: current status and prospective[J]. Chin Sci Bull, 2015, 60(5/6): 534-546. (in Chinese)

[11] 刘言, 蔡文生, 邵学广. 大数据与化学数据挖掘[J]. 科学通报, 2015, 60(8): 694-703.
LIU Y, CAI W S, SHAO X G. Big data and chemical data mining[J]. Chin Sci Bull, 2015, 60(8): 694-703. (in Chinese)

[12] FENG M L, GHASSEMI M, BRENNAN T, et al. Management and analytic of biomedical big data with cloud-based in-memory database and dynamic querying: a hands-on experience with real-world data[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1970.

[13] VOGEL K M. Big data and the invisible, social dimensions of science[C]// Proceedings of the 2014 Workshop on Human Centered Big Data Research. Raleigh: ACM, 2014: 1.

[14] MIAO X. Big data and smart grid[C]// Proceedings of the 2014 International Conference on Big Data Science and Computing. Beijing: ACM, 2014: 1-2.

[15] JUELS A. A bodyguard of lies: the use of honey objects in information security[C]// Proceedings of the 19th ACM Symposium on Access Control Models and Technologies. Shanghai: ACM, 2014: 1-4.

[16] MARTY R. Cyber security: how visual analytics unlock insight[C]// Proceedings of the 19th ACM SIGKDD International Conference on knowledge Discovery and Data Mining. Chicago: ACM, 2013: 1139.

[17] LIU Q, NGAI E, HU X P, et al. SH-CRAN: hierarchical framework to support mobile big data computing in a secure manner[C]// Proceedings of the 2015 Workshop on Mobile Big Data. Hangzhou: ACM, 2015: 19-24.

[18] 陈左宁, 王广益, 胡苏太, 等. 大数据安全与自主可控[J]. 科学通报, 2015, 60(5/6): 427-432.
CHEN Z N, WANG G Y, HU S T, et al. Independence and controllability of big data security[J]. Chin Sci Bull, 2015, 60(5/6): 427-432. (in Chinese)

[19] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
FENG D G, ZHANG M, LI H. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258. (in Chinese)

[20] PARYASTO M, ALAMSYAH A, RAHARDJO B. Big-data security management issues[C]// the 2nd International Conference on Information and Communication Technology. Bandung: IEEE, 2014: 59-63.

[21] GUPTA M, HAN J W. Heterogeneous network-based trust analysis: a survey[J]. SIGKDD Explorations Newsletter, 2011, 13(1): 54-71.

[22] SADIK S, GRUENWALD L. Research issues in outlier detection for data streams[J]. SIGKDD Explorations Newsletter, 2013, 15(1): 33-40.

[23] KANUPARTHI A, KARRI R, ADDEPALLI S. Hardware and embedded security in the context of internet of things[C]// Proceedings of the 2013 ACM Workshop on Security, Privacy & Dependability for Cyber Vehicles. Berlin: ACM, 2013: 61-64.

[24] ALQASSEM I. Privacy and security requirements framework for the internet of things (IoT)[C]// Companion Proceedings of the 36th International Conference on Software Engineering. Hyderabad: ACM,

2014：739-741.

[25] 傅颖勋，罗圣美，舒继武. 安全云存储系统与关键技术综述[J]. 计算机研究与发展，2013，50(1)：136-145.
FU Y X，LUO S M，SHU J W. Survey of secure cloud storage system and key technology [J]. Journal of Computer Research and Development，2013，50(1)：136-145.（in Chinese）

[26] 丁滟，王怀民，史佩昌，等. 可信云服务[J]. 计算机学报，2015，38(1)：133-149.
DING Y，WANG H M，SHI P C，et al. Trusted cloud service[J]. Chinese Journal of Computers，2015，38(1)：133-149.（in Chinese）

[27] 胡德敏，余星. 云存储服务中支持动态数据完整性检测方法[J]. 计算机应用研究，2014，31(10)：3056-3060.
HU D M，YU X. Dynamic data integrity detection method in cloud storage service [J]. Application Research of Computers，2014，31(10)：3056-3060.（in Chinese）

[28] ANIELLO L，BONOMI S，BRENO M. Assessing data availability of Cassandra in the presence of non-accurate membership[C] // Proceedings of the 2nd International Workshop on Dependability Issues in Cloud Computing. Braga：ACM，2013：1-6.

[29] PIATETSKY G. Interview：michael brodie，leading database researcher，industry leader，thinker [J]. SIGKDD Explorations Newsletter，2014，16(1)：57-63.

[30] WHITWORTH J，SUTHAHARAN S. Security problems and challenges in a machine learning-based hybrid big data processing network systems [J]. Sigmetrics Performance Evaluation Review，2014，41(4)：82-85.

[31] TORJUSEN A B，ABIE H，PAINTSIL E. Towards run-time verification of adaptive security for IoT in eHealth [C] // Proceedings of the 2014 European Conference on Software Architecture Workshops. Vienna：Springer，2014：1-8.

[32] NASIM R，BUCHEGGER S. XACML-based access control for decentralized online social networks [C] // Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. London：IEEE，2014：671-676.

[33] BALDUZZI M，ZADDACH J，BALZAROTTI D. A security analysis of amazon's elastic compute cloud service [C] // Proceedings of the 27th Annual ACM Symposium on Applied Computing. Riva：ACM，2012：1427-1434.

[34] DAMIANI M L. Location privacy models in mobile applications：conceptual view and research directions [J]. GeoInformatica，2014，18(4)：819-842.

[35] 王璐，孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报，2014，25(4)：693-712.
WANG L，MENG X F. Location privacy preservation in big data era：a survey [J]. Journal of Software，2014，25(4)：693-712.（in Chinese）

[36] TINATI R，WANG X，BROWN I，et al. A streaming real-time web observatory architecture for monitoring the health of social machines[C] // Proceedings of the 24th International Conference on World Wide Web. Florence：W3C，2015：1149-1154.

[37] 黄晶. 面向 Hadoop 大数据处理的访问控制与通信安全性研究[D]. 长沙：湖南大学，2013.
HUANG J. The research of access control and communication security for hadoop big data processing [D]. Changsha：Hunan University，2013.（in Chinese）

[38] GESHER A. Adaptive adversaries：building systems to fight fraud and cyber intruders[C] // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago：ACM，2013：1136.

[39] THERDPHAPIYANAK J，PIROMSOPA K. Applying Hadoop for log analysis toward distributed IDS [C] // Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication. Kota Kinabalu：ACM，2013：3.

[40] HUANG C，ZHU S C，WU D H. Towards trusted services：result verification schemes for MapReduce[C] // Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster，Cloud and Grid Computing. Ottawa：ACM，2012：41-48.

[41] ANDERSON B，STORLIE C，YATES M，et al. Automating reverse engineering with machine learning techniques [C] // Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop. Scottsdale：ACM，2014：103-112.

[42] JAFARIAN J H H，AL-SHAER E，DUAN Q. Spatio-temporal address mutation for proactive cyber agility against sophisticated attackers[C] // Proceedings of the First ACM Workshop on Moving Target Defense. Scottsdale：ACM，2014：69-78.

[43] PARUNAK H V D，NICKELS A，FREDERIKSEN R. An agent-based framework for dynamical understanding of DNS events (DUDE) [C] // Proceedings of the 1st International Workshop on Agents and CyberSecurity. Paris：ACM，2014：1-8.

[44] DUREN M，ALDRIDGE H，ABERCROMBIE R K，et al. Designing and operating through compromise：architectural analysis of CKMS for the advanced metering

infrastructure[C]// Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop. Oak Ridge: ACM, 2013: 48.

[45] BIRKE R, BJORKQVIST M, CHEN L Y, et al. Big data in a virtualized world: volume, velocity, and variety in cloud datacenters[C]// Proceedings of the 12th USENIX Conference on File and Storage Technologies. Santa Clara: USENIX Association, 2014: 177-189.

[46] KO B M, LEE J, JO H S. Toward enhancing block I/O performance for virtualized Hadoop cluster [C] // Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. London: IEEE, 2014: 481-482.

[47] ZHANG W, RAJASEKARAN S, DUAN S H. Minimizing interference and maximizing progress for Hadoop virtual Machines [J]. SIGMETRICS Performance Evaluation Review, 2015, 42(4): 62-71.

[48] XU D Y. Virtualization and security: happily ever after? [C]// Proceedings of the 4th ACM Conference on Data and Application Security and Privacy. San Antonio: ACM, 2014: 73-74.

[49] LIU F, SHU X K, YAO D F, et al. Privacy- preserving scanning of big content for sensitive data exposure with mapreduce[C]// Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. San Antonio: ACM, 2015: 195-206.

[50] YU X Q, NING P, VOUK M A. Securing Hadoop in cloud[C] // Proceedings of the 2014 Symposium and Bootcamp on the Science of Security. Raleigh: ACM, 2014: 1-2.

[51] HIZVER J, CHIUEH T C. Real-time deep virtual machine introspection and its applications [C] // Proceedings of the 10th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments. Salt Lake: ACM, 2014: 3-14.

[52] HE X H, ALVES-FOSS J. A lightweight virtual machine monitor for security analysis on Intel64 architecture[J]. Journal of Computing Sciences in Colleges, 2011, 27 (1): 155-162.

[53] PEARCE M, ZEADALLY S, HUNT R. Virtualization: issues, security threats, and solutions[J]. Computing Surveys (CSUR), 2013, 45(2): 94-111.

[54] BAUMAN E, AYOADE G, LIN Z Q. A survey on hypervisor-based monitoring: approaches, applications, and evolutions[J]. Computing Surveys (CSUR), 2015, 48(1): 1-33.

[55] WIN T Y, TIANFIELD H, MAIR Q, et al. Virtual machine introspection [C] // Proceedings of the 7th International Conference on Security of Information and Networks. Glasgow: ACM, 2014: 405.

[56] ULUSOY H, COLOMBO P, FERRARI E, et al. GuardMR: fine-grained security policy enforcement for mapreduce systems[C] // Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security. Singapore: ACM, 2015: 285-296.

[57] PORTER D E, BOND M D, ROY I. Practical fine-grained information flow control using laminar [J]. Transactions on Programming Languages and Systems, 2015, 37(1): 1-51.

[58] 徐保民, 倪旭光. 云计算发展态势与关键技术进展[J]. 中国科学院院刊, 2015, 30(2): 170-180.
XU B M, NI X G. Development trend and key technical progress of cloud computing [J]. Bulletin of Chinese Academy of Sciences, 2015, 30 (2): 170-180. (in Chinese)

[59] SAHA B, SHAH H, SETH S, et al. Apache Tez: a unifying framework for modeling and building data processing applications [C] // Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne: ACM, 2015: 1357-1369.

[60] BAUMGÄRTNER L, STRACK C, HOBBACH B. Complex event processing for reactive security monitoring in virtualized computer systems[C] // Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems. Oslo: ACM, 2015: 22-33.

[61] 朱劭, 崔宝江. 基于 SAML 的 Hadoop 云计算平台安全认证方法研究[J]. 信息网络安全, 2013, 13(12): 1-11.
ZHU S, CUI B J. Research on authentication method of Hadoop cloud computing platform based on SAML[J]. Netinfo Security, 2013, 13(12): 1-11. (in Chinese)

[62] 田洪亮, 张勇, 许信辉, 等. 可信固态硬盘: 大数据安全的新基础[J]. 计算机学报, 2015, 39(1): 154-168.
TIAN H L, ZHANG Y, XU X H, et al. Trusted SSD: new foundation for big data security[J]. Chinese Journal of Computers, 2015, 39(1): 154-168. (in Chinese)

[63] 马然. 高机密性高可用性的云存储系统研究[D]. 杭州: 浙江大学, 2013.
MA R. The research of high confidentiality and high availability cloud storage systems [D]. Hangzhou: Zhejiang University, 2013. (in Chinese)

[64] 任艳丽, 谷大武, 蔡建兴, 等. 隐私保护的可验证多元多项式外包计算方案[J]. 通信学报, 2015, 36 (8): 23-30.

REN Y L, GU D W, CAI J X, et al. Verifiably private outsourcing scheme for multivariate polynomial evaluation [J]. Journal on Communications, 2015, 36(8): 23-30. (in Chinese)

[65] 曹珍富. 密码学的新发展[J]. 四川大学学报：工程科学版, 2015, 47(1): 1-12.
CAO Z F. New development of cryptography[J]. Journal of Sichuan University: Engineering Science Edition, 2015, 47(1): 1-12. (in Chinese)

[66] SHAFAGH H, HITTHNAWI A, DROESCHER A, et al. Poster: towards encrypted query processing for the internet of things[C]// Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. Paris: ACM, 2015: 251-253.

[67] 谭霜, 贾焰, 韩伟红. 云存储中的数据完整性证明研究及进展[J]. 计算机学报, 2015, 38(1): 164-177.
TAN S, JIA Y, HAN W H. Research and development of provable data integrity in cloud computing[J]. Chinese Journal of Computers, 2015, 38(1): 164-177. (in Chinese)

[68] 谭霜, 何力, 陈志坤, 等. 云存储中一种基于格的数据完整性验证方法[J]. 计算机研究与发展, 2015, 52(8): 1862-1872.
TAN S, HE L, CHEN Z K, et al. A method of provable data integrity based on lattice in cloud computing[J]. Journal of Computer Research and Development, 2015, 52(8): 1862-1872. (in Chinese)

[69] 付印金. 面向云环境的重复数据删除关键技术研究[D]. 长沙：国防科学技术大学, 2013.
FU Y J. Research on key technologies of data deduplication for cloud environment [D]. Changsha: National University of Defense Technology, 2013. (in Chinese)

[70] GUNAWI H S, HAO M Z, LEESATAPORNWONGSA T, et al. What bugs live in the cloud? a study of 3000 + issues in cloud systems[C]// Proceedings of the ACM Symposium on Cloud Computing. Seattle: ACM, 2014: 1-14.

[71] CARDOSA M, WANG C Y, NANGIA A, et al. Exploring MapReduce efficiency with highly-distributed data[C]// Proceedings of the Second International Workshop on MapReduce and Its Applications. San Jose: ACM, 2011: 27-34.

[72] 权一男. 基于节点状态的分布式文件系统存储副本分发策略的研究[D]. 长春：吉林大学, 2013.
QUAN Y N. The research of node's status-based distributed file system storage replication distribution [D]. Changchun: Jilin University, 2013. (in Chinese)

[73] 戚建国. 基于云计算的大数据安全隐私保护的研究[D]. 北京：北京邮电大学, 2015.
QI J G. The study of the security and privacy protection of big data based on cloud computing [D]. Beijing: Beijing University of Posts and Telecommunications, 2015. (in Chinese)

[74] XIA M Y, SAXENA M, BLAUM M, et al. A tale of two erasure codes in HDFS[C]// Proceedings of the 13th USENIX Conference on File and Storage Technologies. Santa Clara: USENIX Association, 2015: 213-226.

[75] CHEN B, AMMULA A K, CURTMOLA R. Towards server-side repair for erasure coding-based distributed storage systems [C]// Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. San Antonio: ACM, 2015: 281-288.

[76] RAJKUMAR M N, KUMAR V V, SIVARAMAKRISHNAN R. Efficient integrity auditing services for cloud computing using raptor codes[C]// Proceedings of the 2013 Research in Adaptive and Convergent Systems. Montreal: ACM, 2013: 75-78.

[77] SARKAR S, SAFAVI-NAINI O, ZHANG L F. RAFR: remote assessment of file redundancy[C]// Proceedings of the 2013 International Workshop on Security in Cloud. Hangzhou: ACM, 2013: 27-32.

[78] KHAN Z, PERVEZ Z, GHAFOOR A. Towards cloud based smart cities data security and privacy management [C]// Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. London: IEEE, 2014: 806-811.

[79] COHEN S, MONEY W, QUICK M. Improving integration and insight in smart cities with policy and trust [C]// Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics. Craiova: ACM, 2014: 1-9.

[80] SINGAL H, KOHLI S. Escalation of trust analysis in Web[C]// Proceedings of the 12th ACM International Conference on Computing Frontiers. Ischia: ACM, 2015.

[81] 孔兰菊. SaaS 应用交付平台中多租户云数据管理关键技术研究[D]. 济南：山东大学, 2011.
KONG L J. Research on key technology in multi-tenant cloud data management for SaaS application delivery platform[D]. Jinan: Shandong University, 2011. (in Chinese)

[82] 郑永清. 云计算环境面向 SaaS 多租户的可伸缩数据放置研究[D]. 济南：山东大学, 2013.
ZHENG Y Q. Research on elastic data placement for saas multi-tenant in cloud computing [D]. Jinan: Shandong University, 2013. (in Chinese)

[83] 李雪晓, 叶云, 田苗苗, 等. 基于格的大数据动态存储完整性验证方案[J]. 信息网络安全, 2014, 14(4): 46-50.
LI X X, YE Y, TIAN M M, et al. Big data dynamic storage integrity verification scheme based on lattice[J]. Netinfo Security, 2014, 14(4): 46-50. (in Chinese)

[84] COURTIEU P, APONTE M V, CROLARD T, et al. Towards the formalization of SPARK 2014 semantics with explicit run-time checks using coq[C]// Proceedings of the 2013 ACM SIGAda Annual Conference on High Integrity Ianguage Technology. Pittsburgh: ACM, 2013: 21-22.

[85] KUNDU A, BERTINO E. How to authenticate graphs without leaking [C] // Proceedings of the 13th International Conference on Extending Database Technology. Lausanne: ACM, 2010: 609-620.

[86] ZHANG R, SHI J, ZHANG Y C. Secure multidimensional range queries in sensor networks[C]// Proceedings of the Tenth ACM International Symposium on Mobile ad Hoc Networking and computing. New Orleans: ACM, 2009: 197-206.

[87] 肖达, 杨绿茵, 孙斌, 等. 面向真实云存储环境的数据持有性证明系统[J/OL]. 软件学报[2015-10-20]. http: // www. jos. org. cn/1000-9825/4862. htm, 2015.
XIAO D, YANG L Y, SUN B, et al. Provable data possession system for realistic cloud storage environments [J/OL]. Journal of Software [2015-10-20]. http: // www. jos. org. cn/1000-9825/4862. htm, 2015. (in Chinese)

[88] 陈何峰, 林柏钢, 杨旸, 等. 基于 BLS 的多用户多副本数据持有性批量审计[J]. 密码学报, 2014, 1(4): 368-378.
CHEN H F, LIN B G, YANG Y, et al. Public batch auditing for 2M-PDP based on BLS in cloud storage [J]. Journal of Cryptologic Research, 2014, 1(4): 368-378. (in Chinese)

[89] 陈兰香. 一种基于同态 Hash 的数据持有性证明方法[J]. 电子与信息学报, 2011, 33(9): 2199-2204.
CHEN L X. A homomorphic hashing based provable data possession [J]. Journal of Electronics & Information Technology, 2011, 33(9): 2199-2204. (in Chinese)

[90] DARIES J P, REICH J, WALDO J, et al. Privacy, anonymity, and big data in the social sciences [J]. Communications of the ACM, 57(9): 56-63.

[91] 黄刘生, 田苗苗, 黄河. 大数据隐私保护密码技术研究综述[J]. 软件学报, 2015, 26(4): 945-959.
HUANG L S, TIAN M M, HUANG H. Preserving privacy in big data: a survey from the cryptographic perspective[J]. Journal of Software, 2015, 26(4): 945-959. (in Chinese)

[92] DEWRI R. Local differential perturbations: location privacy under approximate knowledge attackers[J]. IEEE Trans on Mobile Computing, 2013, 12(12): 2360-2372.

[93] HUO Z, MENG X F, HU H B, et al. You can walk alone: trajectory privacy-preserving through significant stays protection [C] // Proceedings of the 17th International Conference on Database Systems for Advanced Applications. Busan: ACM, 2012: 351-366.

[94] BABAGUCHI N. Protection and utilization of privacy information [C] // Proceedings of the 1st International Workshop on Information Hiding and Its Criteria for Evaluation. Kyoto: ACM, 2014: 1.

[95] 王璐, 孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报, 2014, 25(4): 693-712.
WANG L, MENG X F. Location privacy preservation in big data era: a survey[J]. Journal of Software, 2014, 25(4): 693-712 (in Chinese)

[96] 周水庚, 李丰, 陶宇飞, 等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5): 847-861.
ZHOU S G, LI F, TAO Y F, et al. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861. (in Chinese)

[97] PAN X, XU J L, MENG X F . Protecting location privacy against location-dependent attacks in mobile services [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(8): 1506-1519.

[98] HASHEM T, KULIK L, ZHANG R. Countering overlapping rectangle privacy attack for moving kNN queries [J]. Information Systems, 2013, 38(3): 430-453.

[99] 丁丽萍, 卢国庆. 面向频繁模式挖掘的差分隐私保护研究综述[J]. 通信学报, 2014, 35(10): 200-209.
DING L P, LU G Q. Survey of differential privacy in frequent pattern mining[J]. Journal on Communications, 2014, 35(10): 200-209. (in Chinese)

[100] NASIM R, BUCHEGGER S. XACML-based access control for decentralized online social networks [C] // Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. London: IEEE, 2014: 671-676.

[101] 李甲帅, 彭长根, 朱义杰, 等. 面向 Hadoop 的风险访问控制模型[J]. 网络与信息安全学报, 2016, 2(1): 0015.
LI J S, PENG C G, ZHU Y J, et al. Risk access

control model for Hadoop [J]. Chinese Journal of Network and Information Security, 2016, 2(1): 0015. (in Chinese)

[102] SOOKHAK M, GANI A, TALEBIAN H, et al. Remote data auditing in cloud computing environments: a survey, taxonomy, and open issues [J]. Computing Surveys, 2015, 47(4): 1-34.

[103] PEDROSA I, COSTA C J. New trends on CAATTs: what are the chartered accountants' new challenges? [C] // Proceedings of the International Conference on Information Systems and Design of Communication. Lisbon: ACM, 2014: 138-142.

[104] ZARRAD A, JALOUD A, ALSMADI I. The evaluation of the public opinion-a case study: MERS-CoV infection virus in KSA[C] // Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. London: IEEE, 2014: 664-670.

[105] SHYAMASUNDAR R K. Security and protection of SCADA: a bigdata algorithmic approach [C] // Proceedings of the 6th International Conference on Security of Information and Networks. Akasary: ACM, 2013: 20-27.

[106] GANDOTRA E, BANSAL D, SOFAT S. Integrated framework for classification of malwares [C] // Proceedings of the 7th International Conference on Security of Information and Networks. Glasgow: ACM, 2014: 417.

[107] 罗恩韬, 胡志刚, 杨杰. 大数据动态安全 SAT 双向防御模型的研究[J]. 计算机应用研究, 2014, 31(5): 1470-1474.
LUO E T, HU Z G, YANG J. Research on bi-directional defense SAT model of big data dynamic safety [J]. Application Research of Computers, 2014, 31 (5): 1470-1474. (in Chinese)

[108] VORAKULPIPAT C, POLPRASERT C, SIWAMOGSATHAM S. Managing mobile device security in critical infrastructure sectors [C] // Proceedings of the 7th International Conference on Security of Information and Networks. Glasgow: ACM, 2014: 65.

[109] GOTO H, TAKADA T. Anomalous network communication detection system by visual pattern on a client computer [C] // Proceedings of the 30th Annual ACM Symposium on Applied Computing. Salamanca: ACM, 2015: 1263-1269.

[110] HOGAN E, JOHNSON J R, HALAPPANAVAR M. Graph coarsening for path finding in cybersecurity graphs [C] // Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop. Oak Ridge: ACM, 2013: 1-4.

[111] SUTHAHARAN S. Big data classification: problems and challenges in network intrusion prediction with machine learning [J]. Acm Sigmetrics Performance Evaluation Review, 2014, 41(4): 70-73.

[112] BROCK J, LUO H, DING C. Locality analysis: a nonillion time window problem [J]. Acm Sigmetrics Performance Evaluation Review, 2014, 41 (4): 102-105.

(责任编辑　杨开英)