大数据安全标准现状和思考

叶晓俊,金涛,刘璘

大数据系统软件国家工程实验室;清华大学大数据研究中心,软件学院,北京 100084

摘要 随着中国《网络安全法》《网络产品和服务安全审查办法(试行)》《数据安全管理办法(征求意见稿)》等法律法规的陆续实施,对大数据运营商提出了诸多合规要求。如何应对大数据时代日益显著的数据安全风险,确保其符合网络安全法律法规政策,需要对网络运营者数据业务及安全管控措施进行规范化。在明确了大数据安全内涵、指出了大数据产业面临的安全挑战后,对照工业界大数据平台和大数据应用安全解决方案阐述了大数据安全目标及其大数据平台与应用关键技术与机制。

关键词 大数据;数据安全;隐私保护;安全标准;安全保护技术

2015年9月国务院印发的《促进大数据发展行动纲要》(简称《纲要》)指出,“数据已成为国家基础性战略资源,大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响”。因此《纲要》提出要“推进大数据产业标准体系建设,建立健全大数据安全保障体系,建立大数据安全评估体系”。

为此,全国信息安全标准化技术委员会(简称信安标委)在2016年3月组织大数据安全技术与标准研讨会,就落实《纲要》中的大数据安全保障体系、强化大数据安全支撑技术与机制等进行讨论,对大数据安全标准化工作进行规划,并成立了大数据安全标准特别工作组(SWG-BPS),连续两年发布了大数据安全标准化白皮书,基本规划了中国大数据安全标准体系[1]。

为了维护国家安全、社会公共利益,保护公民、法人和其他组织在网络空间的合法权益,保障个人信息和重要数据安全,国家互联网信息办公室在2019年5月发布了《数据安全管理办法(征求意见稿)》(简称《办法》)。《办法》第六条要求网络运营者应当按照有关法律、行政法规的规定,参照国家网络安全标准,履行数据安全保护义务,建立数据安全管理责任和评价考核制度,制定数据安全计划,实施数据安全技术防护。

为促进大数据安全标准在中国各种大数据工程项目中的落地,不仅需要从信息安全技术和数据

收稿日期:2019-11-08;修回日期2019-12-31

基金项目:国家重点研发计划项目(2016YFB0501504);国家自然科学基金项目(U1509213)

作者简介:叶晓俊,教授,研究方向为数据库安全,电子信箱:yexj@tsinghua.edu.cn

引用格式:叶晓俊, 金涛, 刘璘. 大数据安全标准现状和思考[J]. 科技导报, 2020, 38(3): 94-102; doi: 10.3981/j.issn.1000-7857.2020.03.006

安全管理上理解这些法律法规和标准规范要求,促进大数据平台和大数据应用安全保护技术研发,还应该从大数据产业发展角度坚持保障数据安全与产业发展并重,护航中国大数据产业健康发展。

围绕大数据安全标准化工作及标准应用需求,本文明确了大数据安全内涵、指出了大数据产业面临的安全挑战风险,对照目前工业界大数据平台和大数据应用安全解决方案阐述了大数据安全目标和大数据平台与应用相关的关键技术与机制,为未来中国大数据产业化发展、大数据安全标准建设和大数据安全评估工作提供参考。

1　大数据安全内涵

在《信息技术 大数据 术语》(GB/T 35295—2017)中,大数据被定义成"具有数量巨大、种类多样、流动速度快、特征多变等特性,并且难以用传统数据体系结构和数据处理技术进行有效组织、存储、计算、分析和管理的数据集"。该定义只是从信息技术角度指出了大数据处理活动中的4V(大体量、多样性、速度快和多变性)特征,未能体现人们期望的"数据-信息-知识-价值"的数据价值特征,以及在数据"收集、准备、分析和行动"等数据增值过程中数据生命周期相关活动面临的安全挑战及相关法规遵从性需求。因此有必要从数据语义、数据生命周期和数据处理的信息技术(IT)空间等维度去分析和理解大数据产业发展中需要采用的信息安全技术[2]。

在传统数据管理体系中,数据是指反映客观事物属性的记录。因此数据本身没有意义,只有在具体应用中数据对实体行为产生影响时才成为信息。相应地,数据安全只是信息安全的一个组成部分,主要聚焦于数据物理保护、数据存储加密、数据残留等安全技术与机制。换句话说,信息安全是防止未经授权的访问、使用、破坏、修改或销毁信息,数据安全是防止未经授权的访问、使用、破坏、修改或销毁存储和通信中的数据。

在大数据时代下,人们对多种数据源进行采集和汇聚存储,并采用分布式处理技术及各种机器学习技术对数据进行组织、存储和分析处理,目的是为了从海量数据中挖掘潜在价值,驱动组织业务价值实现,实现组织的各种使命[3]。因此数据管理具有分布式、无中心服务器、多组织协调等特点,数据安全管理面临新的挑战[1]。理解大数据安全需要从时间、空间和语义3个方面明确大数据安全相关的技术和机制(图1)。

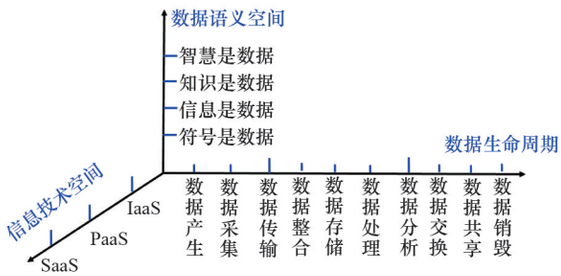

图1　大数据时代下的数据管理维度

1)时间维度。传统的数据安全是假设在封闭环境(有安全边界)下讨论数据分类分级、数据组织、存储和访问权限,以及数据在传输和存储中的加密技术,并未从数据生命周期角度考虑数据从产生,经过数据采集、数据传输、数据存储、数据处理(包括计算、分析、可视化等)、数据交换,直至数据销毁等各阶段演变过程中的数据安全需求。因此大数据安全必须考虑处于生命周期不同阶段数据在不同安全域间交换与共享、数据发布、不可信主体之间的数据安全保护技术与机制需求。

2)空间维度。依照《信息技术 大数据 技术参考模型》(GB/T 35589—2017),大数据生态系统是由系统协调者、数据提供者、大数据应用提供者、大数据平台(框架)提供者和数据消费者等5个逻辑功能构件组成。因此从横向空间来看,大数据安全覆盖数据预处理、数据处理和分析、数据可视化等数据增值活动空间相关的安全技术与机制,从纵向空间看,大数据安全覆盖分布式文件存储、大数据管理平台和不同数据处理类型的分布式计算相关的分布式节点间可信计算、多版本数据一致性、数据互操性规范等安全技术和机制。

3)语义维度。在网络空间环境下,大数据既包括原始采集的未经加工的无语义的微观数据集,

也包括经过加工处理得到的各种汇总统计的宏观数据,包括驱动组织决策相关的规则等知识数据集和由大数据分析驱动产生的各种价值数据集等。因此大数据安全需要考虑由多源数据派生、聚合、关联分析等数据分析过程中的数据资源操作安全策略与规范,也需要考虑数据分析结果输出的安全授权控制机制,并采取必要的技术手段和管控措施保证共享数据分析结果不泄露个人信息、重要数据等敏感信息,包括对数据分析过程数据操作进行记录,以备对数据分析质量及各种数据源真实性进行追踪溯源。

安全既是形容词,表述的是一个复杂信息技术系统的动态过程或状态,目标是使组织与个体、数据业务和IT系统不会受到伤害或遭受损失;安全也是名词,具有保密性、完整性、可用性(简称CIA)等安全属性。在大数据语境下,CIA属性概念需要上升的网络空间维度,既需要从信息技术(IT)角度考虑信息安全管控措施,也需要从操作技术(OT)角度防范数据驱动决策带来的不安全状态,包括用户个人信息相关隐私保护问题[1-2]。

1)信息安全(security)。泛指对个人、组织和财产等数据资产和IT资产的保护,使它们免遭外部威胁和一些犯罪活动的损害。因此在这个含义下的大数据安全是社会政治性的安全问题。如社会安全、国家安全、网络安全等,需要一方面分析和识别大数据潜在的安全威胁及相关的安全攻击,另一方面分析和研究大数据资产及其保护大数据的安全措施是否能抵抗这些攻击和威胁。

2)功能安全(safety)。指的是大数据生态系统安全功能无论在正常情况或者有故障存在的情况下都应该保证正确实施。因此在这个含义下的大数据安全是指大数据平台和大数据应用的一种状态,大数据系统协调员可以掌控有可能造成大数据系统伤害的威胁,或确保大数据系统的脆弱性不一定会造成伤害。例如基础设施运行安全、数据平台操作安全、数据驱动决策安全等,需要一方面分析和研究大数据系统导致人员伤亡、财产损失、环境破坏等负面影响的各种致因和风险,另一方面还需要对大数据系统的可靠性、可信性、可控性、完整性等进行渗透性测试或安全性分析。

3)隐私保护(privacy)。指数据使用者不愿为他人所知或不愿公开或知悉的、与公共利益无关的纯个人秘密/私事,例如家庭关系、个人收入、社交关系等个人信息或日常生活、社会交往、网络行为等个人私事。在大数据生态系统中包含的个人信息越来越多,因此个人信息安全已经成为大数据安全保护的重要内容[3-4]。

大数据生态系统是所有IT产品和系统与多数据源的集合,是人类未来生存的网络空间环境,人在其中与数据和系统相互作用、相互影响,并由此产生人与人、人与社会的更深层次的交互,因此需要从数据安全、软件安全、组件安全、连接安全、系统安全、人员安全、组织安全和社会安全等不同角度考虑网络空间数据安全性、安全保护和个人隐私相关的大数据安全技术和机制[5],需要从上述大数据内涵和大数据安全内涵出发,采用新思维、新机制和新方法去安全的保护大数据和安全的使用大数据。

2 大数据安全挑战

安全是发展的前提,发展是安全的保障,大数据安全一定要与大数据产业发展同步推进。因此在《大数据安全标准化白皮书(2018版)》中指出了大数据产业化发展面临的安全挑战。

1)法律法规与相关标准挑战。国家安全深受大数据影响,社会治理面临大数据挑战,数据安全相关的法律法规和标准尚需完善。

2)数据安全和个人信息保护挑战。大数据生态系统中的数据安全保护难度加大,个人信息泄露风险加剧,数据真实性保障更困难,数据所有者(例如数据权属)权益难以保障等,国家、行业及其企业与个人都需将个人隐私保护纳入大数据生态系统的规划和实施。

3)大数据技术和平台安全挑战。目前的大数据平台安全机制严重不足,传统的安全控制措施难以适应开放环境下数据安全需求,面向数据生命周期的大数据应用访问控制愈加困难,基础密码技术

亟待突破等。

针对前两个挑战,依据《网络安全法》,中国主管部门颁布了《银行业金融机构数据治理指引》、《证券期货业数据分类分级指引》等数据安全指引,全国信息安全标准化技术委员会委制定了一些大数据安全国家标准,国家互联网信息办公室在2019年颁发了《数据安全管理办法》《个人信息出境安全评估办法》《网络安全审查办法》等数据安全法规征求意见稿,全国人大也正在制定《数据安全法》《个人信息安全保护法》。

针对第3个挑战,大数据系统协调者需要解决以下4个安全问题[1]。

1)大数据平台安全机制不足:目前很多组织采用的开源大数据平台在整体安全规划方面考虑不足,大都采用安全加固的解决方案,缺乏有效的漏洞管理和恶意后门防范能力;而像华为、阿里、亚马逊等特定大数据服务组织从业务方面整合的大数据平台和应用安全技术与机制存在不透明,大数据使用者难以对其安全策略规范、安全实现技术和风险管控措施进行验证;最重要的是目前国内外还没有对大数据平台安全架构形成共识,因此缺乏对大数据平台安全功能(内生安全)技术和机制建设和大数据安全评估的指导性框架。

2)传统安全措施难以适应大数据平台和大数据应用。一方面大数据生态系统边界变得模糊,传统基于边界的访问控制措施、密钥保护机制、安全审计等将变得不再有效,需要引入零信任、属性基等基于密码技术的控制措施;另一方面大数据生态系统涉及的软件和硬件较多,保护的数据分布在不同安全域多个节点中,任何一点遭受故障或攻击,都可能导致整体安全出现问题。因此需要按照功能安全概念解决大数据生态系统生命周期(包括需求规划、设计、实施、集成、验证、确认和配置)和数据业务生命周期内各种脆弱性,给出IT空间不同层面和数据业务不同阶段数据活动安全攻击面上各种潜在威胁的安全防护方案。

3)大数据应用访问控制困难。在开放环境下的大数据应用中,由于用户未知及其数量庞大,数据业务复杂,因此覆盖生命周期的数据权限及用户角色预设困难,基于属性规则的访问控制,包括使用控制权限控制、数据权益保护(包括版权保护)都面临很多挑战。另外多层分布式数据处理、大数据全生命周期内的复杂数据业务活动,都给大数据生态系统中的细粒度数据授权、细粒度安全审计、数据溯源等带来诸多挑战。

4)基础密码技术亟待突破。不可信服务托管环境下的同态加密(功能加密)、完整性校验、密文搜索和密文数据还原等密码算法、密码协议和密码体制实现机制及其效用性都亟待突破才能满足大数据用户的安全使用需求。

3　大数据安全目标

从大数据内涵和大数据安全内涵看出,大数据安全属性不同于传统信息安全领域的保密性、完整性和可用性。这是因为大数据生态系统中的保密性必须同时考虑主体个人隐私和客体数据保密性;完整性必须同时考虑数据传输、分布式存储和处理一致性、主体对数据分析算法真实性及数据生命周期中数据可信性;可用性也需要考虑大数据生态系统的健康运行安全目标。因此从概念上讲,大数据安全目标相关的属性超越了传统信息安全的含义,保密性需要从主客体两个方面考虑个人隐私和数据保密性安全控制措施;完整性概念被真实性概念所替代,需要同时考虑主客体数据可信性、冗余数据间的一致性和生命周期内数据完整性安全控制措施;可用性要求大数据平台提供者和大数据应用提供者监控大数据服务软件和硬件系统资产的健康运行,大数据生态系统无论在正常情况或者有故障存在的情况下都应该保证正确实施,以确保数据生命周期内的数据活动始终满足数据和主体保密性和真实性要求。

从大数据运营者的角度看,大数据生态系统应提供包括大数据应用安全管理、身份鉴别和访问控制、数据业务安全管理、大数据基础设施安全管理和大数据系统应急响应管理等业务安全功能,因此大数据业务目标包括以下5方面[2]。

1)大数据应用安全管理。建立在大数据平台

上的应用服务组件及其支持终端注册、大数据服务组件和大数据服务接口与大数据平台安全对接规范、数据供应链相关的接口等。

2）身份鉴别和访问控制。大数据应用和大数据平台对大数据用户身份进行验证,并提供合适的访问控制授权引擎对用户访问的数据资源进行控制,提供诸如基础设施层用户身份验证、应用程序层身份验证、终端用户层身份管理、服务提供商身份管理、粗粒度、细粒度、属性基等多种访问控制、多租户数据安全管理等。

3）数据业务安全管理。围绕大数据的“数据-信息-知识-价值”数据价值链的数据生命周期活动,提供数据在传输、存储和使用过程中的加密功能和密钥管理功能,提供不同业务和不同应用之间数据隔离与接口封装服务,确保用户业务数据保密性、完整性、可用性和数据业务可管理性。

4）大数据基础设施安全管理。提供网络、计算、存储和环境资源,包括点对点传输、存储转发、大数据交换与通信框架操作和维护相关的大数据安全运行基础设施安全管控措施和隐私保护功能,提供诸如威胁和脆弱性管理、安装与配置管理、系统监测和预警、运行日志和安全审计、网络边界控制和基础设施冗余和恢复等功能。

5）大数据系统应急响应管理。提供大数据服务基础设施和数据管理平台风险和责任相关的问题追责、安全合规、安全取证、安全事件管理、风险控制措施等。

4 大数据安全技术

在传统的数据库系统中,用户数据安全是由数据库管理系统(DBMS)内在的事务特性(ACID)和相关安全机制共同实现。换句话说数据库事务的原子性和持久性保证用户数据在事务处理前后的一致性和在IT不同层次空间不会丢失,隔离性和一致性保证用户数据在事务处理空间(事务、缓存和磁盘)的完整性和一致性,事务日志和多路复用技术保障软件故障和硬件故障下用户数据完整性和可用性。因此数据库安全机制借用了DBMS内在事务特性,并通过内置的或通过接口调用外部的用户身份鉴别、用户数据管理、DBMS安全管理、数据库审计、资源限制、密码操作等安全技术和基准就能有效地对数据库对象进行安全保护。

大数据平台提出旨在解决数据4V特征问题,且主要目标是赋能数据开发人员,提高数据分布式处理效率。因此在大数据平台建设初期,开放人员假设平台是运行在可信环境中,所以只考虑了分布式处理节点间的可信计算安全问题,其他安全技术与机制并没有在分布式文件系统或分布式数据处理平台架构设计中得到重点关注,而在后期大数据平台部署和大数据应用提出的通过安全加固方案,必然会降低大数据平台的便利性和安全技术和机制实施的有效性。所以早期的大数据服务提供者的平台本身安全性主要借助组织层面的网络安全及物理或逻辑隔离来得到保证,有关用户数据使用权限主要在大数据应用中解决或借助第三方数据加固安全组件(例如CDH的Sentry、HDP的Ranger、华为的FusionInsight等大数据平台安全加固组件)等数据中台(中间件)的安全能力来实现用户数据安全,大数据平台只提供了数据脱敏、数据加密等简单的数据安全服务组件[6]。

上述这种附加式的安全加固方案的有效性难以满足大数据服务组织的安全要求。因此业界提供大数据服务的互联网企业如亚马逊、阿里云、腾讯云等都是通过在内部各业务系统上构建数据安全保护能力,以解决用户业务安全和数据安全问题。这种不断从组织内部数据业务内生长出的安全技术和机制,能伴随企业大数据业务的发展而持续提升,能持续保证组织的业务安全需求。现在人们开始将这些业务领域的安全技术和实现机制进行抽象,希望在大数据平台内置这些安全能力,因此出现了零信任安全模型、内生安全(功能安全)这样的新模型或新概念,并且已经研发了具备事务特性和数据生命周期管理能力的数据湖(data lake)这样的内置数据安全保护能力的开源大数据平台。

考虑到大数据平台一般是基于分布式处理技术,多采用云计算技术和多租户架构,以及大数据平台的安全持续运行、用户数据安全和隐私保护需

求,业界希望大数据平台具有内生安全功能实现大数据安全目标。例如针对保密性属性,由于用户担心失去对大数据平台中的安全控制,担心大数据服务商是否提供了足够的机制以阻止敏感数据泄漏,大数据服务商是否诚实可信,不会截取客户敏感数据等,因此透明加密、密钥互操作性管理等成为大数据平台必备的安全功能;另外,用户也想知道大数据服务商数据处理/计算是否正确,如何能确保大数据服务商真的没有篡改它存储客户数据等数据真实性问题,包括是否提供足够的弹性(可伸缩性)确保大数据平台能健康运行。因此在大数据平台层,围绕大数据安全目标必须具备的大数据安全技术和机制包括以下8方面[2,7-8]。

1) 保密性安全技术与机制。数据提供者、大数据平台提供者和大数据应用提供者应提供数据和数据主体保密性安全控制措施。例如使用安全套接层/传输层安全(SSL/TLS)等安全协议保证数据传输的机密性;使用基于凭证的数据访问策略、基于属性的细粒度访问控制策略、基于虚拟机(VM)技术的边界控制等保证数据存储访问保密性;使用公钥基础设施(PKI)、基于身份/属性基加密(ABE)等密码学方法保证数据托管存储访问保密性;使用支持密文数据搜索和加密数据同态处理等的功能加密技术提供密文数据透明处理;使用集中存储的密钥管理互操作服务保证数据分布式存储和分布式处理的安全访问;使用数据匿名化处理技术、数据扰动技术、差分隐私技术等保障发布数据主体敏感信息的安全性等。

2) 真实性安全技术与机制。大数据平台提供者和大数据应用提供者应确保大数据服务中数据和主体真实性。例如使用终端输入验证方式来保证采集过程中采集的数据来自可信的数据源;使用领域相关的语义约束条件验证数据语义或用户操作满足典型业务规则,确保数据操作过程中数据完整性;使用数字签名等密码技术从数学角度来验证数据和主体真伪;在传输过程中使用安全传输层协议等保证数据传输完整性;使用数据验证计算技术确保分布式数据关键片段计算确实符合预期的计算结果,启用细粒度或高级安全审计机制以确保大数据服务可追踪能力;使用可信计算保证数据和主体处理值得信赖;使用大数据服务中各种加密机制、安全协议等保证数据完整性和数据主体敏感信息隐私保护等。

3) 可用性技术与机制。大数据平台提供者和大数据应用提供者应监控大数据服务软件和硬件系统资产的健康运行,以确保数据生命周期各阶段的数据活动一直满足数据和主体机密和真实性安全目标。例如通过系统审计跟踪定位大数据服务过程中的性能瓶颈等,保证数据服务高效性;建立主动抵御拒绝服务攻击协议以保证大数据服务可用性;基于大数据服务运行数据的自动化监控和分析,保证大数据服务自我保护能力(免疫系统),提高大数据服务的自适应能力;提供包含关键安全、性能指标以及趋势指针的仪表盘,以进行不间断的数据安全服务监控;使用机器学习等数据分析算法持续评估数据服务安全基准活动的变化和监测异常事件;提供大数据平台及大数据应用组件配置合规及风险管理插件,提供包括自动监控、合规策略评估以及威胁建模等合规性和违反安全事件检测的系统健康运行管理组件和服务;提供基于日志、网络事件、智能代理的大数据分析等服务接口组件和辅助管理工具。

4) 应用安全支持性。大数据系统应安全地管理接入的大数据应用程序、大数据应用终端和外部潜在的大数据资源,提供诸如大数据应用程序安全注册、大数据应用安全元数据管理、大数据应用开发和部署策略等。大数据应用程序安全注册需登记和管理如物联网网络、移动终端等大数据应用终端设备、数字化产权保护下的各种数据资产及外部服务、应用程序和用户角色。大数据应用安全元数据管理应结构化存储和维护大数据服务安全相关的设备、用户、资产、服务组件等所有数据和主体安全要素,包括数据快速更新、数据结构变化、临时数据存储、数据有效性、大数据服务运行日志、溯源数据等系统运行安全统计数据,以支持应用数据生命周期、合规性控制等复杂应用的安全管理。大数据应用开发和部署实施策略涵盖符合机构信息系统环境建设的大数据应用部署和设施策略、大数据运

行过程中的细粒度审计政策及不同大数据服务角色相关的行为规范等。大数据应用应该提供用户数据导入与导出,用户数据备份、用户行为数据保护等个性化数据安全管理功能。

5) IT空间用户身份鉴别技术与机制。基础设施层用户身份验证应支持分布式计算技术、虚拟计算技术等计算方式的身份验证,例如基于硬件安全模块支持下的可信计算体系,从基础设施层提高大数据服务整体安全性。应用程序层用户身份验证应提供基于公钥基础设施(PKI)等技术的身份认证服务平台,实现对应用层用户的证书、账户、授权、认证和审计的集中管理、整合大数据服务资源、实现应用数据共享和全面集中管控目标。终端用户层身份管理应依据数据提供者和大数据使用者角色自动判断大数据应用中用户的身份信息,保证大数据服务系统中的用户标识和大数据应用用户参考标识与应用层授权信息之间的映射关系。服务提供商身份管理针对大数据系统中数据提供者、大数据服务协调者、大数据平台提供者、大数据应用提供者和大数据使用者,使用安全性断言标记语言来定义数据资源提供者身份(角色),添加安全和隐私保证等要求,扩展传统的用户身份鉴别和授权机制。大数据可聚合多个数据提供者的数据资源,细粒度访问控制使得大数据服务提供者不只是分享数据集和数据服务,同时也分享数据授权策略,因此,大数据系统需要提供基于属性访问控制引擎(例如XACML),提供策略编辑点、策略决策点、策略执行点和策略访问点等面向数据对象的授权管理和访问控制功能。

6) 数据业务安全技术与机制。数据业务安全涉及数据生成、收集、传输、存储、管理、处理、使用、共享及销毁阶段宜采取的信息安全策略和控制措施,目的是降低数据业务活动中的各种数据安全风险,实现大数据应用业务安全目标。因此大数据平台应提供数据存储安全控制措施,包括不同数据副本或数据在不同空间的完整性检测措施,预防数据丢失,保证数据可访问性;部署必要的网络服务网关,确保数据的安全迁移、转换、交换和共享;提供聚合数据管理措施,确保多数据源安全整合;提供服务组件计算可信性验证机制,确保应用服务组件的安全性;具备加密数据的透明计算能力;制定部署、迁移和保留策略、个人信息保护策略、去标识化和匿名化机制,确保数据生命周期中个人隐私与敏感数据的安全管理,包括个人信息重标识风险管理等;提供大数据应用终端验证、数字版权管理、信任管理、数据披露、数据交易伦理、数据治理等数据生命周期相关的数据服务安全。另外大数据平台应支持隐私保持的数据分析和数据挖掘,这是因为大数据服务商存储了大量个人信息,大数据平台为海量数据的隐私挖掘和分析提供了可能。所以大数据平台应提供尽可能多的数据去标识、数据匿名化、数据脱敏等算法和相应的接口,以实现个人信息和行为隐私保护。

7) 大数据基础设施安全技术与机制。作为关键信息基础设施组成部分的大数据运行支撑环境应提供以下安全技术与机制:威胁和脆弱性管理应识别大数据平台及大数据应用的脆弱性及相关威胁,对分布式拒绝服务攻击、密钥管理、加密协议,以及对脆弱性衍生的问题进行管理;安装与配置管理包括安全参数设置、安全组件部署、安全补丁管理、系统升级等,目的是保护大数据服务基础设施完整性;系统监测和预警需要通过部署大数据运行安全相关的组件和服务实现,它基于大数据基础设施运行数据实现大规模安全情报、复杂事件融合、安全分析、恶意软件监测和修复等安全功能;日志记录和安全审计是通过管理基础设施产生的海量、多样和高速变化的日志大数据,在线分析和统计抽样这些日志信息,为基础设施的安全运营和优化提供安全统计数据;网络边界控制主要为数据源和数据服务不可知的基础设施安全域间建立一条安全连接通道,共享服务网络体系结构,保证在开放环境下基础设施的网络通信安全;基础设施冗余和恢复通过系统复制有计划的维护大数据系统内部软件层次的冗余,以支持故障转移、系统恢复能力或减少大数据基础设施性能延迟,因为从大数据安全失败中恢复系统可能比传统集中式数据管理基础设施需要更加高级的基础设施安装、部署与配置等准备工作。

8）大数据系统合规性和应急响应技术与机制。大数据系统协调者应保证大数据服务的法律法规的遵从性,提供基础设施、数据管理平台和大数据应用风险和责任相关的问题追责、安全合规、安全取证、安全事件管理、风险控制措施等。问题追责主要基于大数据平台和大数据应用之间的信息、流程和角色行为,通过追踪大数据系统的门户和检测点、向前和向后的溯源数据检查等方式实现。大数据系统安全和隐私的合规跨多个领域分类,涉及隐私、行业规范和本国的法律。安全取证可通过大数据安全分析服务组件取证,也可通过在大数据安全失败场景下取证。安全事件管理落实事件处理所需的各类支持资源,为用户处理、报告安全事件提供咨询和帮助。风险控制措施协调应急响应活动与事件处理活动,并与大数据服务运营者相关外部机构(例如IT供应链中的外部服务提供商等)提供事件应急处理机制。

云安全联盟(CSA)在2013年给出了大数据安全和隐私十大挑战,并在2016年出版的《大数据安全与隐私手册100例最佳实践》中给出了相关的技术与机制。全国信息安全标准化技术委员会发布的《大数据安全标准化白皮书》,包括阿里、腾讯等中国大数据服务企业发布的数据安全白皮书都给出了相关的大数据安全技术与机制。在全国信息安全标准化技术委员会计划的国家标准制定中,已经安排了《信息安全技术 健康医疗数据安全指南》《信息安全技术 电信领域大数据安全防护实现指南》等面向大数据应用领域的指南类标准。

5　结论

目前国际标准化组织/国际电工委员会第一联合技术委员会已经有大数据安全、个人信息保护相关的标准制定和研究项目20多项;国际电信联盟电信标准分局的安全工作也有数据安全相关的标准制定和研究项目近20项,但真正形成具有影响力的是美国国家标准化研究院推出的NIST 1500-4《大数据互操作框架》的《第4册 安全和隐私保护》。NIST 1500-4先后已发布3个版本,且相关概念已经转换为ISO/IEC 20547《信息技术 大数据参考架构第4部分:安全与隐私保护》标准制定项目。

中国依照“紧急先行,成熟先上,关注重点”原则,参照《网络安全法》等制定了《信息安全技术 个人信息安全规范》(GB/T 35273—2017)、《信息安全技术 大数据安全服务能力要求》(GB/T 35274—2017)、《信息安全技术 大数据安全管理指南》(GB/T 37973—2019)、《信息安全技术 个人信息去标识化指南》(GB/T 37964—2019)、《信息安全技术 数据安全能力成熟度模型》(GB/T 37988—2019)、《信息安全技术 数据交易服务安全要求》(GB/T 37932—2019)等标准。

从这些发布的大数据安全标准看出,在法律法规合规层面已经有了一批大数据安全标准。但相对于大数据平台和大数据服务急需的核心安全技术与机制,包括分布式环境下的数据加密、数据完整性验证、数据标签、区块链、细粒度访问控制、密文透明计算、数据溯源、数据脱敏与安全审计等内在安全相关标准需要加强研究,以便形成一批面向大数据平台和大数据应用的技术标准,特别是支撑大数据平台建设和评估的大数据安全架构需要尽快提出,以推动中国大数据生态系统的产业化应用。所以笔者认为下列大数据平台内生安全相关的技术标准应尽快研制:(1) 覆盖《数据安全管理办法》的术语及安全架构;(2) 指导领域重要数据识别和保护的数据分类分级指南;(3) 开放环境下的边界安全控制技术与机制;(4) 基于属性访问控制、使用控制和属性基的细粒度访问控制技术与机制;(5) 支持分布式加密数据处理的密钥管理框架及协议互操作技术与机制;(6) 不可信环境下数据服务相关的零信任技术与机制;(7) 覆盖数据生命周期和IT层次控制数据处理的数据溯源技术与机制;(8) 支持数据交换与共享的数据脱敏技术与机制;(9) 数据安全测评、数据安全评估指标等。

同时,建议加强大数据技术在大数据生态系统功能安全和网络安全防护方面的研究,包括入侵检测、安全态势感知、网络攻击取证、威胁情报分析等,以利用大数据技术抵御针对大数据生态系统的网络攻击威胁。另外为驱动大数据产业化工作,大

数据服务组织应需分享他们在数据分类分级、数据交换与共享、数据业务连续性等方面的安全最佳实践。目前这些面向组织层面促进大数据产业化发展的安全技术与机制还没有形成统一的共识,需要借助行业或团队标准等对国家标准进行丰富。中国大数据安全技术标准化任重道远。

参考文献(References)

[1] 全国信息安全标准化委员会, 大数据安全标准特别工作组. 大数据安全标准化白皮书(2018版)[R]. 北京: 全国信息安全标准化委员会, 2018.

[2] NIST big data interoperability framework, Volume 4: Security & privacy[R/OL]. [2019-11-15]. https://bigdatawg.nist.gov/V3_output_draft_docs.php.

[3] 杜小勇, 陈跃国, 范举, 等. 数据整理——大数据治理的关键技术[J]. 大数据, 2019, 5(3): 16-25.

[4] 刘贤刚, 孙彦, 胡影, 等. 数据安全国际标准研究[J]. 信息安全与通信保密, 2018(12): 33-49.

[5] Curriculum guidelines for post-secondary degree programs in cybersecurity[EB/OL]. [2019-11-15] https://www.acm.org/.../education/curricula-recommendations/csec2017.pdf.

[6] Smith K T. Big data security: The evolution of Hadoop's security model[R/OL]. [2019-11-15]. https://www.infoq.com/articles/HadoopSecurityModel.

[7] Thangaraj M, Balamurugan S. Survey on big data security framework[C]//International Conference on Knowledge Management in Organizations. Berlin: Springer, 2017: 470-481.

[8] Moreno J, Serrano M A, Fernandez-Medina E, et al. Towards a security reference architecture for big data[C/OL]. [2019-11-15]. https://www.researchgate.net/publication/325218224_Towards_a_Security_Reference_Architecture_for_Big_Data. [2019-11-15]. https://www.researchgate.net/publication/325218224_Towards_a_Security_Reference_Architecture_for_Big_Data.

Big data security standardization and perspectives

YE Xiaojun, JIN Tao, LIU Lin

National Engineering Laboratory of Big Data System Software; Tsinghua Big Data Research Center, School of Software, Tsinghua University, Beijing 100084, China

Abstract Since the launching of the series laws and regulations on network security, the audition of network products and services, data security, there are growing needs for big data service providers for compliance implementation. In order to respond to the data security risks sufficiently and ensure the compliance to network security laws and regulations on big data related business operations, this paper summarizes the key definitions of big data security, the major risks in big data security, and describes the big data security objectives, key security protection technology and mechanisms for big data platforms, as a reference to practitioners in the big data industry.

Keywords big data; data security; privacy; security standard; security protection

(责任编辑 刘志远)