



运营商大数据平台的敏感信息保护研究

潘文字¹, 蔡鑫¹, 阮宜龙², 姚晓辉¹

(1. 中国电信股份有限公司上海研究院 上海 200122;

2. 中国电信股份有限公司安徽分公司 合肥 230001)

摘要:探讨了运营商大数据平台的敏感信息范围,给出了敏感信息保护的方法步骤,总结了该研究领域内的关键技术及每类技术的主流技术方法,并针对实际的运营商大数据平台的分层架构,设计了不同敏感信息保护技术的技术实现方案。

关键词:敏感信息保护;运营商大数据平台;匿名化;隐私

doi: 10.3969/j.issn.1000-0801.2014.11.021

Research on Sensitive Information Protection of Carriers' Big-Data Platform

Pan Wenyu¹, Cai Xin¹, Ruan Yilong², Yao Xiaohui¹

(1. Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China;

2. Anhui Branch of China Telecom Co., Ltd., Hefei 230001, China)

Abstract: The scope of sensitive information of carriers' big-data platform was discussed. The methods to protect sensitive information were described. The key technologies and mainstream technical methods in the field of research were summarized. Based on the layered architecture of carriers big-data platform, technical solutions for different sensitive information protection technology were designed.

Key words: sensitive information protection, carriers' big-data platform, anonymization, privacy

1 引言

在运营商的基础网络平台、运营支撑系统以及各类业务平台中蕴含着海量的用户数据,随着大数据概念的兴起,数据价值得到了业界的共识。充分利用好“数据”这一新兴的企业资产,持续不断地从数据“金矿”中挖掘和创造价值,成为国内外运营商一致看好的、实现战略转型和业务创新的突破口。

在全球化开源社区、互联网企业中的先行者等共同推动下,大数据技术已日渐成熟,国内三大运营商都积极跟进,相继建设了大数据平台,实现了用户数据、话单数据、

互联网 DPI 数据、网络信令数据等数据汇聚,并在此基础上开始逐步尝试数据应用,使信息价值显性化。

面向运营商内部,可以通过分析用户的上网行为偏好,推出类似面向特定内容的、有针对性的优惠流量包,在提升运营网络流量的同时,为用户节省上网费用;还可以通过分析基站等网络设备的状态和流量特性,促进网络规划、布局、配置的优化,进而提升管道的智能化水平。面向运营商外部,可以对运营商数据进行包装,形成可售卖的数据产品。

积极探索大数据对内、对外应用的同时,必须特别注意运营商的很多数据都涉及用户敏感信息。笔者注意



到,一些互联网公司追踪用户 cookie,通过 cookie 盗取和收集用户信息并用于盈利目的,这引发了广泛的舆论关注。因此,敏感信息保护如果处理不当,一方面会导致个人隐私和企业机密的泄露,损害用户利益,另一方面也会给运营商自身带来一系列严重的社会责任和法律风险问题。

2 敏感信息保护概述

2.1 敏感信息的定义

敏感信息,是指其丢失、不当使用、未经授权被人接触或修改会不利于国家利益或政府计划的实行、不利于个人依法享有个人隐私权的所有信息^[2]。企业都拥有敏感数据,包括商业秘密、知识产权、关键业务信息、业务合作伙伴信息或用户信息等。其中涉及个人隐私的用户个人信息是运营商大数据平台中最重要最广泛的敏感信息。

2013年7月19日,工业和信息化部依据《全国人民代表大会常务委员会关于加强网络信息保护的決定》、《中华人民共和国电信条例》和《互联网信息服务管理办法》等法律、行政法规,制定颁布了《电信和互联网用户个人信息保护规定》。

在该规定中,用户个人信息是指电信业务经营者和互联网信息服务提供者在提供服务的过程中收集的用户姓名、出生日期、身份证件号码、住址、电话号码、账号和密码等能够单独或者与其他信息结合识别用户的信息以及用户使用服务的时间、地点等信息^[3]。

参照这一规定,运营商大数据平台中涉及敏感的用户个人信息主要包括用户资料信息、用户密码信息、用户通话行为记录、用户上网行为记录、用户位置信息等。须依据国家法律规定、行业规范标准、公司政策要求等对所有此类数据进行严格保护。

2.2 敏感信息保护的方法

企业实施敏感信息保护,一般遵循以下的方法步骤。

(1) 敏感信息的分类分级

敏感信息保护的第一步,是要依据对敏感信息管理的相关规定,对企业的敏感信息进行分类分级。数据分类的数量和定义,需要兼顾数据治理、风险管理、合规性和业务要求等方面。

(2) 基于信息生命周期的信息安全风险识别

敏感信息的保护,需要从信息生命周期各个环节综合考虑采取的安全管理或技术措施。在此基础上,识别造成

信息资产泄露、篡改、破坏的安全风险,基于风险对业务造成的影响开发有针对性的信息安全规划和风险处理计划。

(3) 针对敏感信息的数据脱敏

数据脱敏,指对某些敏感信息通过脱敏规则进行数据的变形,实现敏感隐私数据的可靠保护。在开发、测试等非生产环境中安全地使用脱敏后的数据集;对于数据限制发布的应用,生产环境中也需要执行数据脱敏操作。

(4) 涉及敏感信息行为的审计监控

通过安全审计及监控,使敏感信息的授权使用、权限管理等活动处于控制当中。在一旦出现风险迹象时,能够及时识别并且采取相应控制措施规避风险。需要对涉及用户敏感信息的访问行为进行实时处理,发生敏感信息泄露行为必须进行报警、阻断、跟踪和追溯。

3 敏感数据保护的关键技术

学术界投入精力进行大数据敏感信息保护的关键技术研究,不断优化技术方法及算法,力求达到数据保密性和可用性之间更好的平衡。依据技术方法原理和适用场景的差异,敏感数据保护技术可分为3类。

3.1 数据失真技术

数据失真技术,是使敏感数据失真但同时保持某些数据或数据属性不变的方法,包括随机化、凝聚、阻塞等技术。

数据随机化,是对原始数据加入随机噪声,然后发布扰动后数据的方法。对外界而言,只可见扰动后的数据,从而实现了真实数据值的隐藏^[4]。由于重构后的数据可与原始数据在分布上保持一致,因此对于更关注群体特征规律的统计分析和数据挖掘应用依然有效。

凝聚技术,是将原始数据记录分组,每一组内存储由 k 条记录产生的统计信息,包括每个属性的均值、协方差等,保证同一组内的 k 条记录是两两不可区分的^[5]。由于凝聚技术在同一组内利用统计值替代原始值,消除了记录的个体差异,从而隐藏了原始的敏感信息。

阻塞技术采用的是不发布某些特定数据的方法。具体反映到数据表中,就是将某些特定的值用一个不确定符号代替。

数据经过失真技术的处理,内容上发生了变化,但处理后的数据仍然保持了原始数据的某些统计特性,并不影响数据分析和挖掘使用。

3.2 数据加密技术

数据加密技术,是指将一个信息(或称明文)经过加密钥匙及加密函数转换,变成无意义的密文,而接收方则将此密文经过解密函数、解密钥匙还原成明文。

数据加密技术要求只有在指定的用户或网络下,才能解除密码而获得原来的数据,这就需要给数据发送方和接收方一些特殊的信息用于加解密,即密钥。按加密算法的不同,可分为专用密钥(对称密钥)和公开密钥(非对称密钥)。

数据加密技术能够满足分布式环境下的通信安全性需求,基于数据加密的敏感信息保护主要应用包括分布式数据挖掘、分布式查询、安全多方计算(SMC)等。

3.3 匿名化限制发布技术

限制发布即有选择地发布、不发布,或者发布精度较低的敏感数据,以实现敏感信息的保护。有两类不同级别的数据匿名化处理,一是抑制,即不发布某个敏感数据项;二是泛化,即对敏感数据进行更概括、抽象的描述。

常见的数据匿名化模型方法主要有以下 3 种。

(1) k -匿名

k -匿名(k -anonymity)是要求所发布的数据表中的每一条记录不能区分于其他 $k-1$ 条记录^[6]。不能相互区分的 k 条记录称为一个等价类。由于 k -anonymity 是针对所有属性集合而言,对于具体的某个属性则未加限制,容易出现某个属性匿名处理不足的情况。若某等价类中某个敏感属性上取值一致,则攻击者可以有效地获得该敏感属性与个体的联系。

(2) l -多样化

l -多样化(l -diversity)要求在每一个匿名属性组里敏感数据的多样性满足大于或等于 l 。 l -diversity 保证攻击者最多以 $1/l$ 的概率确认某个体的敏感信息。 l -diversity 尽量使敏感数据出现的频率平均化。但当同一等价类中数据范围很小时,攻击者可猜测其值^[7]。

(3) t -贴近

t -贴近(t -closeness)在 l -diversity 基础上,考虑了敏感属性的分布问题,它要求所有等价类中敏感属性值的分布尽量接近该属性的全局分布^[8]。

4 运营商大数据平台的敏感信息保护

4.1 运营商大数据平台的架构

运营商的大数据平台架构通常在逻辑上可以分为几个层次,包括数据采集层、数据存储处理层、数据服务层和数据应用层,如图 1 所示。

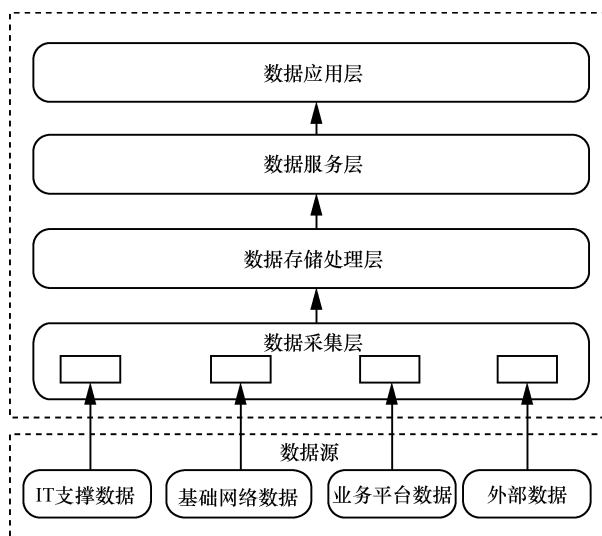


图 1 运营商大数据平台架构

数据采集层负责全业务数据的采集、清洗和安全传输,支持通道加密、单向传输、接入节点控制、弹性采集等。采集的数据源包括 IT 支撑数据(BOSS)、基础网络数据、业务平台数据、外部数据等。

数据存储处理层提供数据的整合和存储以及数据处理功能,原始数据只进不出,输出分析结果数据。物理节点采用 x86 分布式集群,技术上采用混搭架构,不同类型数据采用不同的技术进行处理,包括分布式的批量处理、流式处理、内存计算等。

数据服务层负责提供大数据平台服务的数据封装和统一接口,避免数据泄露风险。通过数据封装,建立统一的标准化输出接口,开放数据服务,促进信息共享和应用重用。

数据应用层负责提供对内对外的数据应用,实现大数据平台的数据提供和数据展现,并提供开放工具,支撑定制数据应用的敏捷化和标准化开发。

4.2 大数据平台的敏感信息保护方法

参考本文第 2.2 节所述的敏感信息保护一般方法,对于运营商大数据平台的敏感信息保护,可以遵循以下方法。

(1)对运营商大数据平台的信息进行分类分级

根据数据机密性把运营商大数据平台数据分为敏感数据和非敏感数据两大类,对照相关的标准和规范对敏感数据按机密程度进行分类分级。

针对不同等级的敏感信息,制定差异化的数据保护策略。



(2)结合运营商大数据平台分层架构,识别信息生命周期的安全风险

大数据平台的分层架构,反映了信息的生命周期阶段特征。

数据采集层对应于信息的收集和传输,需重点关注数据的可靠性保证和数据通信安全;数据存储处理层对应于信息的存储、挖掘和归档,侧重于数据的保密性和完整性保证;数据服务层对应于信息的共享和发布,强调数据的保密性和可用性保证;数据应用层对应于信息的使用,重点是数据访问安全。

(3)基于运营商大数据平台分层架构,实现对敏感信息的保护

针对每种敏感信息保护技术的特点和适用场景,设计基于运营商大数据平台分层架构的实现方案。本文第4.3节将具体展开描述。

(4)针对大数据平台的敏感信息访问行为实施

审计监控主要包括实时监控和日志审计。

- 实时监控主要用于事前防范,提供实时监测大数据平台的数据访问操作,并通过预定义的规则,实时地对涉及敏感信息的风险进行识别,对泄露敏感信息的行为进行处理。
- 日志审计主要用于事后追溯,包括对于用户的敏感信息访问行为,记录下完整的数据访问操作信息;进行日志格式化、日志信息校验和日志敏感级别管理,以保证日志的有效性;提供日志查询和导出功能以供审计需要。

4.3 大数据平台的敏感信息保护技术实现

本文第4.1节描述了运营商大数据平台的分层架构,每层对应数据生命周期不同阶段的功能实现。而不同的敏感信息保护技术,有各自所适应的应用场景。在大数据平台的分层架构和不同的敏感信息保护技术之间,存在一定的关联关系,这决定了大数据平台的敏感信息保护的技术实现,具体如图2所示。

(1)数据采集层的传输数据加密/解密

数据采集层的采集点前置部署于数据源一侧,采用分布式ETL架构模式实现多数据源的分布式采集。采集点完成数据采集和清洗之后,要将数据传输到大数据平台进行加载。为保障敏感信息的数据传输安全,需要在数据源端对数据进行加密操作,加密数据分组传输到大数据平台之后,再利用密钥对数据进行解密操作,完成解密后再由大

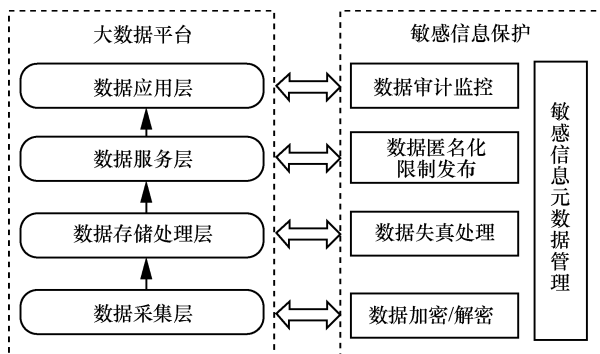


图2 运营商大数据平台的敏感信息保护技术实现

数据平台进行数据加载。

(2)数据存储处理层的数据失真处理

数据存储处理层负责对采集进来的数据进行持久化存储,有些敏感信息在数据加载时就需要进行数据脱敏,这类敏感信息属性具有以下特征:后续应用环节对于个体数据的失真不关注,仅关注群体的分布特征,用于统计和数据挖掘应用,对于这一类敏感信息,在持久化存储之前就要依据数据特征使用随机化、阻塞、凝聚等技术方法对敏感信息进行数据失真处理。

(3)数据服务层的数据匿名化限制发布

数据服务层负责对大数据平台的数据进行统一封装和对外发布提供。数据限制发布在数据服务层进行配置和管控。不对应用层发布的敏感信息数据项,应当采用抑制方式做数据匿名化处理;对应用层进行“裁剪”发布的敏感信息数据项,应当采用泛化方式做数据匿名化处理。而数据匿名化的计算处理过程,则可以下沉到数据存储处理层运行,充分利用分布式处理平台的强大的计算能力。

(4)数据应用层的数据审计监控

在数据应用层,各类对内对外数据应用都可能涉及对大数据平台敏感信息的访问操作。数据审计监控保障应用访问行为在授权范围的约束之下活动,对于涉及敏感信息泄露风险的用户动作,能够及时识别并且及时阻断。对所有使用人员的所有操作包括查看、赋权、修改、删除、导出等提供日志功能,记录访问操作的处理参数和返回的处理结果,对异常行为进行事后跟踪和追溯。

(5)贯穿各层的敏感信息元数据管理

敏感信息元数据管理贯穿大数据平台各层,保存运营商敏感信息分类分级的资料数据以及对各类敏感信息的数据保护策略的资料数据,包括对于某个敏感信息数据项

拟采取何种数据脱敏技术及具体技术方法、是否纳入审计监控范围及审计监控级别等。敏感信息元数据是运营商大数据平台进行敏感信息保护管理的“知识库”。

4.4 大数据平台的敏感信息保护方案验证

依据上述技术实现框架,笔者在中国电信股份有限公司某省公司的大数据平台现网环境中,结合该省大数据安全管理的需求,构建了大数据敏感信息保护平台,进行了方案验证。

大数据敏感信息保护平台与原大数据平台之间保持相对独立,侧重提供业务无关的安全管理基础能力,支撑大数据平台敏感信息的安全保护。平台的功能架构如图 3 所示。

从功能架构上看,大数据敏感信息保护平台主要由 3 部分功能组成。

敏感信息配置管理,主要面向数据安全管理员,提供敏感信息管理规则相关的配置功能,包括敏感信息分类分级定义、敏感信息保护策略管理、敏感信息监控规则定义、日志审计级别管理等。这些功能会产生与之相关的敏感信息元数据,这些配置数据保存在敏感信息元数据库中。

敏感信息数据转换,主要面向大数据平台的分层架构,提供各类数据转换处理引擎实现对应的敏感信息保护技术,包括数据匿名化处理引擎、数据失真处理引擎、数据加密处理引擎、数据解密处理引擎等。处理引擎从敏感信

息元数据库中读取已配置的各敏感属性保护策略,调用大数据平台分布式处理计算能力,实现对敏感信息的脱敏转换处理。处理引擎还要与大数据平台的各层实现接口,完成处理前后敏感数据的双向交互。

敏感信息审计监控,主要面向大数据平台数据应用层的数据访问,提供实时监控和日志审计功能,包括敏感信息访问风险识别、敏感信息访问风险管理、日志分析与审计、敏感信息风险模型管理等。该功能组从敏感信息元数据库中读取监控规则定义和日志级别等配置数据,将动态获取的用户访问操作信息与读入的规则数据进行匹配,实时识别敏感信息泄露风险。此外还支持对离线的日志数据进行分析 and 建模,实现对风险跟踪、追溯和预警的需求。

大数据敏感信息保护平台的引入,对于大数据平台的信息安全起到了预期的加固作用。根据上线试运行以来的效果观察,大数据平台 4 层架构的 10 类原有风险点被有效控制。具体见表 1。

5 结束语

运营商大数据应用方兴未艾,为积极稳妥地推进大数据平台的服务能力提升,避免探索创新过程中可能遇到的舆论和法律风险,需要高度重视大数据平台的敏感信息保护问题。

本文探讨了运营商大数据平台的敏感信息范围,给出

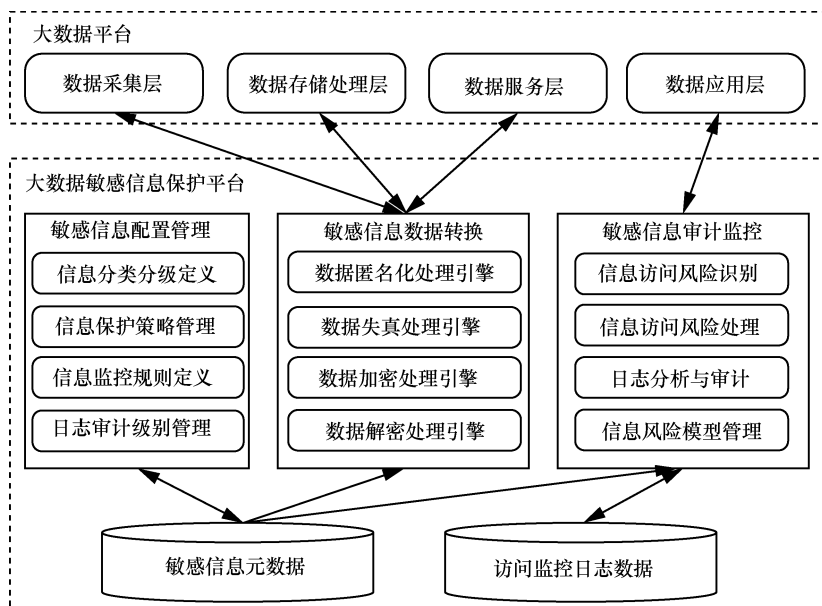


图 3 运营商大数据敏感信息保护平台功能架构



表 1 运营商大数据敏感信息保护平台风险控制效果

	大数据平台原有风险	敏感信息保护平台加固效果
数据采集层	信息源伪造	利用私钥对信息源加密完成数字签名,确保数据来源可靠
	信息传输泄密	利用加密技术对敏感信息的传输提供安全保障
数据存储处理层	高度敏感信息明文存储	利用数据失真技术对高度敏感信息进行脱敏处理,再做持久化存储
	使用未脱敏的数据进行数据分析挖掘	利用随机化等技术对敏感数据进行失真处理,在保证敏感数据保密性的同时,保持数据的统计特性和分析可用性
	挖掘结果产生新的敏感信息明文	利用数据失真技术对挖掘结果中新产生的敏感信息进行脱敏处理,再做持久化存储
数据服务层	不需泄露应用层发布的非法/违规/敏感信息	利用抑制的方式对此类敏感信息数据项进行限制发布,即不发布这类数据项
	需要泄露应用层进行模糊处理的敏感信息	利用泛化的方式对此类敏感信息进行限制发布,即发布匿名化处理后的数据项
数据应用层	用户对于敏感信息的增/删/改/查等操作	利用实时监控功能对用户数据访问中涉及敏感信息的风险进行识别和阻断
	日志分散不完整,不能进行有效的审计追踪	利用增强的日志分析功能,对异常的用户数据访问行为进行跟踪和追溯
	对于敏感信息泄露风险缺乏预判	利用日志进行分析建模,发现风险出现的规律前兆,对敏感信息泄露提前预警

了敏感信息保护的方法步骤,总结了该研究领域内的关键技术及每类技术的主流技术方法,并针对实际的运营商大数据平台的分层架构,设计了不同敏感信息保护技术的实现方案。

需要指出,学术界所研究的敏感信息保护技术,一般是基于单一静态数据集的环境给出的理想模型。而实际的运营商大数据生产环境要复杂得多,数据集是无时无刻不在变化的,而且这种变化一般都不是完全随机、独立的,而是相互关联的。如何在实际的复杂环境下同时实现对动态数据的利用和敏感信息保护,还有待于更深入的理论研究和实践探索。

参考文献

- 1 国内首个运营商自主的 RTB 平台在浙江电信上线. <http://it.21cn.com/tel/a/2013/1206/07/25384164.shtml>, 2013
- 2 敏感信息. <http://baike.baidu.com/view/3061484.htm?fr=aladdin>, 2014
- 3 电信和互联网用户个人信息保护规定. <http://www.miit.gov.cn/n11293472/n11294912/n11296542/15514014.html>, 2014
- 4 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. 计算机学报, 2009, 32(5):1~18
- 5 Aggarwal C C, Yu P S. A condensation approach to privacy preserving data mining. Proceedings of the 9th International Conference on Extending Database Technology (EDBT). Heraklion, Greece, 2004: 183~199

- 6 Sweeney L. k -anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5):557~570
- 7 冯登国,张敏,李昊. 大数据安全与隐私保护. 计算机学报, 2014, 37(1):246~258
- 8 Li N, Li T. T -closeness: privacy beyond k -anonymity and l -diversity. Proceedings of the 23rd International Conference on Data Engineering (ICDE), Istanbul, Turkey, 2007: 106~115

[作者简介]



潘文字,男,中国电信股份有限公司上海研究院高级工程师,主要研究方向为信息管理、信息技术。

蔡鑫,男,中国电信股份有限公司上海研究院高级工程师,主要研究方向为数据规划、信息管理、数据标准。

阮宜龙,男,中国电信股份有限公司安徽分公司工程师,主要研究方向为数据架构、信息管理、数据应用。

姚晓辉,男,中国电信股份有限公司上海研究院工程师,主要研究方向为数据规划、数据挖掘、信息管理。

(收稿日期:2014-07-22)