信息安全研究 第4卷 第2期 2018年2月
Journal of Information Security Research Vol.4 No.2 Feb. 2018

客户隐私数据流转安全管理系统

艾解清[1] 魏理豪[1] 梁承东[2] 陈 亮[2]
[1](广东电网有限责任公司信息化评测实验室 广州 510080)
[2](广州竞远安全技术股份有限公司 广州 510080)
(dotrai@126.com)

Customer Privacy Data Lifecycle Security Management System

Ai Jieqing[1], Wei Lihao[1], Liang Chengdong[2], and Chen Liang[2]
[1](*Information Testing Laboratory of Guangdong Power Grid Corp Ltd, Guangzhou* 510080)
[2](*Guangzhou Jingyuan Safety Technology Limited by Share Ltd, Guangzhou* 510080)

Abstract Enterprises provide services to the public through web application systems, and they collect personal privacy data at the same time. Associated with enterprises' production data, such personal privacy data has become higher value-added customer privacy data. Multiple applications and network boundaries are involved during the process of data creation, data transmission, data storage, data usage and data destruction. Customers privacy data is faced with a lot of security threats. In the meantime, due to continuous expansions of application systems and external system data exchange interface, customer privacy data is distributed in various nodes of the network, which makes the data hard to manage under unified management and centralized security. This paper analyzes the life cycle of customer privacy data, clarifies the distribution and security status of customer privacy data in the network, and establishes a customer privacy data lifecycle security management system for security protection. The system use DPI technology, document encryption and decryption technology, data desensitization technology and abnormal behavior detection technology, and it provides security technical support for both structured data and unstructured data in the whole life cycle process.

Key words customer privacy; privacy data; DPI; document encryption; abnormal behavior

摘 要 企业通过应用系统对公众提供业务的同时也收集到了个人隐私数据.这些个人隐私数据在与企业生产数据关联后,成为具有更高附加值的客户隐私数据.客户隐私数据面临诸多的安全威胁,在产生、传输、处理、存储、使用、销毁过程中涉及多个应用系统和多个网络边界,同时由于应用系统的不断扩建,与外部系统数据交换的接口不断增加,造成客户隐私数据分布在网络中的各个节点,无法进行统一管理和集中化安全保障.从客户隐私数据在网络流转过程的角度进行分析和设计,建立1个客户隐私数据流转安全管理系统,用于保护客户的隐私数据.该系统使用了深度包检测技术、文档加解密技术、数据脱密技术和异常行为检测技术,为结构化数据和非结构化数据在全生命周期过程中提供了安全技术保障.

收稿日期:2017-11-20
通信作者:陈亮(eddie.chen@hotmail.com)

关键词 客户隐私;隐私数据;深度包检测;文档加密;异常行为
中图法分类号 TP309.2

近年来,随着网络信息化的迅猛发展及其商业应用的广泛普及,推进了企业信息化进程,通过互联网方式为客户办理各种业务,个人信息得到了更快速的融汇.在业务办理过程中,应用系统提供商越来越多地接触到了个人隐私数据(如姓名、身份证号码、联系电话、家庭住址、个人信用等).这些信息有结构化的数据,有非结构化的数据(如图片、文本).个人隐私数据与企业生产数据相关联,成为企业的客户隐私数据(如电网客户的电量、电费,银行客户的交易信息等).客户隐私数据的流转过程涉及到产生、传输、处理、存储、使用、销毁等多个环节,在各个环节中又涉及到众多的合作伙伴、业务办理人员、系统维护人员等,如果不对这些客户隐私数据进行控制,必将会导致数据的泄露,对企业也将造成市场和品牌声誉上的严重损失.Verizon 报告指出最近 10 年以来,在 95 个国家或地区的 6 万多起数据泄露事件中,内部失窃占 19%,排名第 1[1].

企业已经逐渐构建起完整的安全体系,包括:基础安全设备、终端安全设备、网络安全设备、应用安全设备、数据库安全设备及安全管理类产品[2-3],对于外部的攻击起到了一定的防范作用.但是,在数据流转的过程中,更多的开发人员、测试人员、维护人员、管理人员可以接触到客户隐私数据.这些设备的防护往往可以被内部人员找到绕过的途径.在企业内部数据的泄露问题中,有内部人员参与的案例占到了 80%[4].

1 背景知识

客户隐私数据流转安全管理系统紧密结合 DPI 设备、文档加解密系统、数据脱敏系统、异常行为检测系统,构建了 1 套完整的安全管理系统.从客户隐私数据的产生、传输、处理、存储、使用、销毁过程形成全生命周期的客户隐私数据保护.

DPI(deep packet inspection)即深度数据包检测技术,是对流量中的应用层数据进行检测和监控的技术[5].本系统中 DPI 设备主要提供了访问关系管理功能、HTTP(hypertext transfer protocol)流量还原功能[6]和 SQL(structured query language)流量还原功能[7].客户隐私数据在企业业务系统的流转过程中,总是以相对固定的访问关系出现,访问关系管理功能主要监控内部资产间的固定业务、固定访问关系.HTTP 流量还原功能和 SQL 流量还原功能主要为访问关系管理功能和异常行为检测系统提供数据支撑.

文档加解密系统采用"驱动级透明动态加解密技术"对客户隐私数据的文件进行实时、强制、透明的加解密.在正常使用时,计算机内存中的文件是以受保护的明文形式存放,但硬盘上保存的数据却处于加密状态,如果没有合法的使用身份、访问权限和正确的安全通道,所有加密文件都将以密文形式保存.

数据脱敏系统是对客户隐私数据的内容进行特殊处理,以达到数据变形的效果,实现客户隐私数据的可靠保护.数据脱密技术的脱敏规则一般分类为可恢复与不可恢复 2 类[8];本系统采用不可恢复类数据脱密技术对关键信息进行模糊化处理.

异常行为检测系统是对每个数据库用户的操作行为进行分析和告警的过程.在数据挖掘算法中,关联规则算法普遍适用于用户行为的分析,多应用于电子购物系统中对用户购买物品的关联分析.在数据库用户的操作过程中,用户针对数据的操作行为与电子购物系统中用户购买物品的行为有很多相似之处,都可以通过关联规则算法来分析用户的行为.数据库用户根据角色岗位的工作要求,在处理数据表、字段的顺序上,在数据处理的操作类型上都具有关联性,因此,异常行为检测系统是在用户行为模型的基础上检测异常行为.

2 客户隐私数据流转安全管理系统整体设计

2.1 整体功能

客户隐私数据流转安全管理系统包括:业务终端、应用服务器、交换机、文件服务器、数据库服

务器. 业务终端包括:移动智能终端和 PC 机. 客户通过业务终端 PC 和智能终端办理开户业务时,产生客户身份证图像、头像、电子签名和文字录入等隐私数据,在业务终端 PC 和智能终端对其进行加密,加密后的隐私数据通过防火墙、交换机、深度包检测设备传送给应用服务器;传送过程中,交换机把访问关系发送给异常用户行为检测系统进行分析和告警,深度包检测设备对数据内容进行细粒度过滤,并把完整数据发送给异常用户行为检测系统进行分析和告警;数据库服务器部署数据脱敏系统对指定数据进行脱敏处理[9],部署方式如图 1 所示,整体功能框架如图 2 所示.

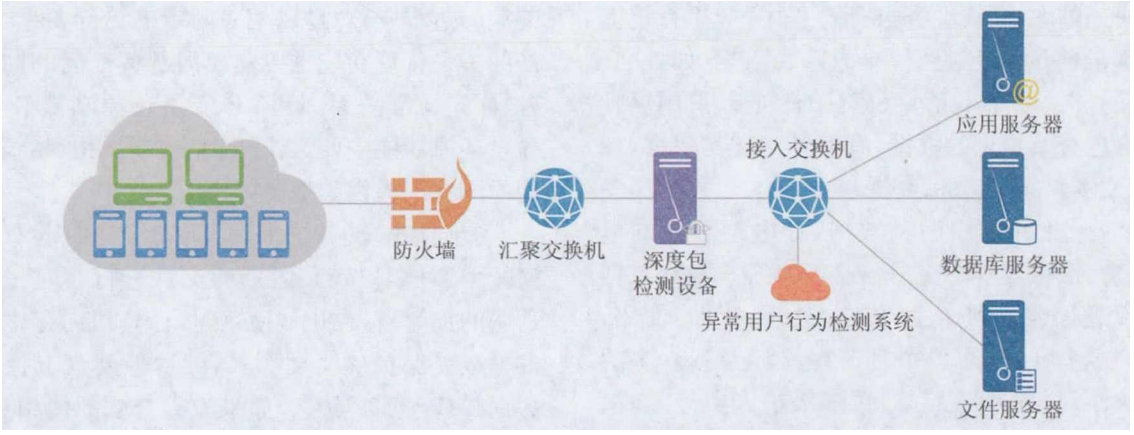

图 1 客户隐私数据流转安全管理系统部署示意图

图 2 客户隐私数据流转安全管理系统功能框架

2.2 基于 DPI 的访问关系管理系统设计

2.2.1 技术架构

DPI 设备采用 Redis 缓存上报日志流,ES 存储全量数据进行计算的数据处理框架,将实时数据与离线数据分离,使用了 Hive,Spark 来处理离线数据分析以及实时数据流,如图 3 所示:

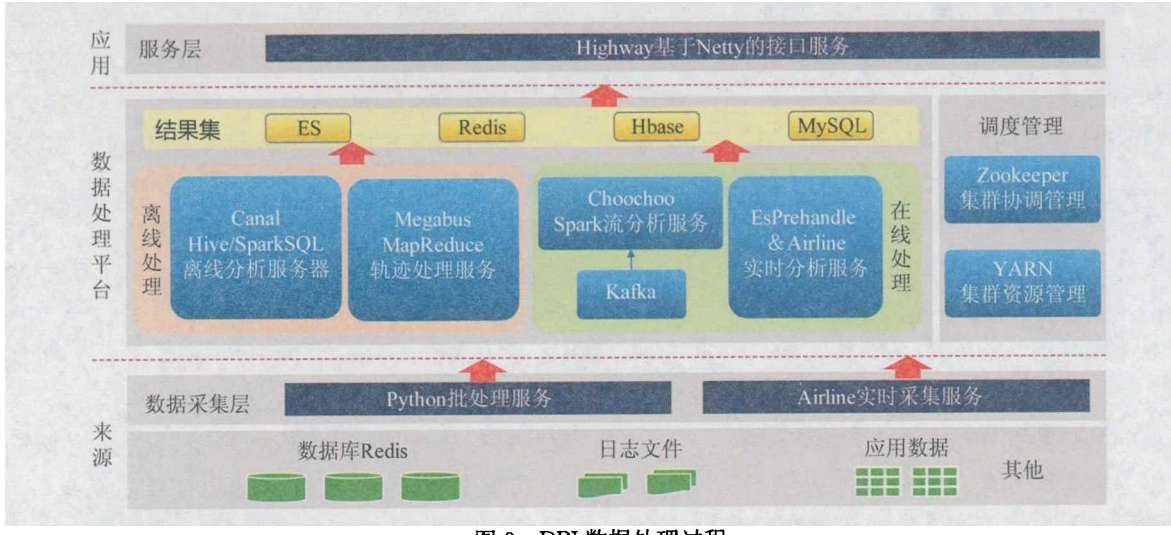

图 3　DPI 数据处理过程

2.2.2　HTTP 流量还原

DPI 设备提供了 HTTP 流量还原功能. HTTP 流量还原功能采用层次结构，分为 4 层：数据采集层、数据处理层、协议解析层和数据还原层，如图 4 所示：

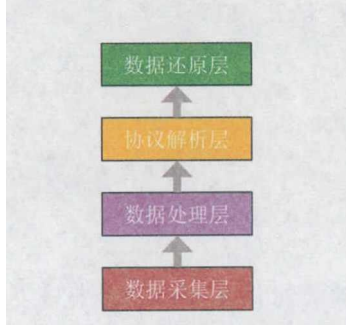

图 4　DPI 数据处理过程层次结构

数据采集层：从交换机流量镜像中获取数据包；数据处理层：根据数据包信息，先进行 IP (Internet protocol) 分片重组处理，然后将 TCP (transmission control protocol) 包重组为完整 TCP 会话数据流；协议解析层：根据 HTTP 协议特征从完整的 TCP 会话数据流中进行筛选；数据还原层：解析 HTTP 协议数据包的内容，并存入数据库.

HTTP 内容还原的具体实现：根据 HTTP 协议特征从 TCP 流中提取数据内容，对 options，head，put，delete，trace，connect 数据内容直接进行解析还原，对 get，post 方式传输的数据进行二次筛选，当 get，post 方式的 HTTP 请求中包含 jsp，php，asp，html，action 关键字的内容时进行解析还原. 使用 zlib 库对 gzip 压缩请求进行解压. 将还原后的数据内容存入数据库，提供给访问关系监控模块进行分析和告警.

2.2.3　SQL 流量还原

SQL 流量还原功能的处理流程同 HTTP 流量还原功能的处理流程一致，包括数据采集层、数据处理层、协议解析层和数据还原层. 二者的区别在于协议解析层和数据还原层. 协议解析层需要通过数据库的协议特征进行筛选，不同的数据库类型有不同的协议特征；数据还原层也会根据不同的数据库类型建立不同的数据库语法还原方法.

经过协议解析层和数据还原层处理后的流量数据已经可以阅读和理解了，如图 5 所示，但是有些数据的顺序是混乱的，通过排序、格式化处理等过程，形成完整的 IP、数据库账号、操作时间、操作内容.

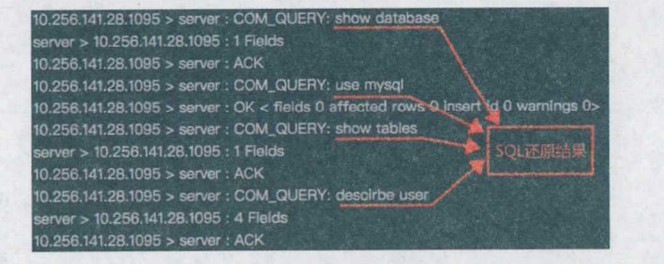

图 5　SQL 语句还原结果

2.2.4　基于资产的访问关系监控

资产间的访问动态监控：根据客户隐私数据流转过程明确 DPI 设备的部署位置，对资产间的访问关系进行监控.

2.2.5 基于内容的访问关系监控规则

基于内容的访问关系监控规则包括静态访问规则和动态访问规则.

HTTP 静态访问规则:建立角色、源 IP、业务 URL 这 3 者关系的白名单. 比如,办理用户开户业务的业务人员,其应用系统角色是固定的,访问应用系统的终端 IP 是固定的,业务 URL(uniform resource locator)的集合是固定.

SQL 静态访问规则:建立角色、源 IP、业务数据表这 3 者关系的白名单. 比如:数据库运维人员经常访问的是数据库维护相关的系统表.

因为数据库的角色划分不如应用系统的角色划分细致,数据库不同角色的账号存在数据表的交叉访问. 如业务运维人员处理数据内容错误时也会查看到系统表,静态规则显然不适用于这类访问关系的监控和告警.

动态访问规则是根据 SQL 数据的访问特征进行建立的. 我们采用了 K-means 算法用于动态规则库的创建[10],如图 6 所示,其包括 2 个步骤:1)建立关联关系,由账号为关键条件,建立与账号相关的 2 阶和 3 阶关系,如账号和源 IP 的关系、账号、源 IP 和业务数据表等,需要用穷举法罗列所有与账号相关的 2 阶和 3 阶关系. 2)用 K-means 算法进行分析,发现偏离数据,分析数据的业务含义,排除数据干扰. 分析 K-means 算法结果,验证合规性,形成新的违规访问关系规则库.

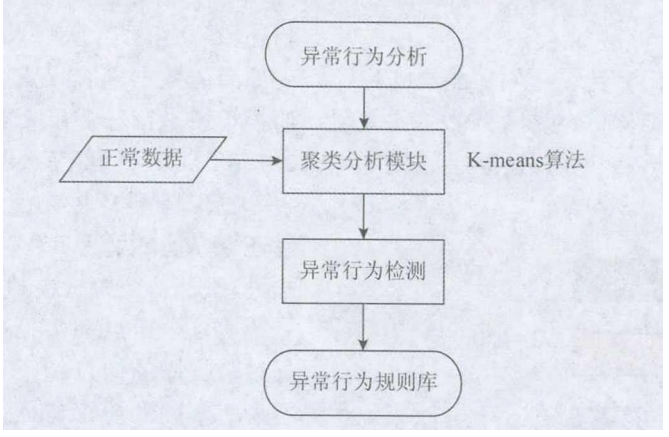

图 6 异常访问行为规则库

2.3 文档加解密系统设计

用户对客户隐私数据使用的方式有 2 种:1)在本地终端使用 Web 应用系统访问后台的主机/数据库;2)在本地终端使用主机/数据库维护工具软件对存储有客户隐私数据的主机/数据库进行直接操作. 由于这些用户拥有较高的权限,因此可以访问主机/数据库中的客户隐私数据信息,如:客户资料、交易信息等,并可以在本地终端将数据导出,造成敏感信息的泄密.

文档加解密系统是对从数据库导出的客户隐私数据进行保护的一种安全技术手段. 文档加解密系统包括:敏感数据管控客户端、密钥服务器、管理门户和受控网关. 文档加解密系统和原有的 4A(authentication account authorization audit)系统、受控网关进行了紧密结合;当运维人员通过 4A 访问数据库,并把数据导出到受控网关形成数据文件时,文档加解密系统对数据文件进行了加密;当运维人员把数据文件拷贝到终端时,敏感数据管控客户端负责对文件的透明加解密和权限控制,如图 7 所示.

敏感数据管控客户端被安装在用户的本地终端,提供对文档加解密、文档授权、访问控制、截屏控制、剪贴板控制等功能,同时支持离线授权、用户行为日志记录等功能. 敏感数据管控客户端保证和第三方软件的兼容,不会影响用户终端的正常使用.

密钥服务器用于管理用户私钥及系统公钥,管理用户个人信息及功能授权信息,同时使用安全的通道与敏感数据管控客户端进行通信,包括密钥下发和各类授权信息下发.

管理门户对敏感信息管控系统的各个组件进行管理,对文档访问、文档授权、用户行为等信息

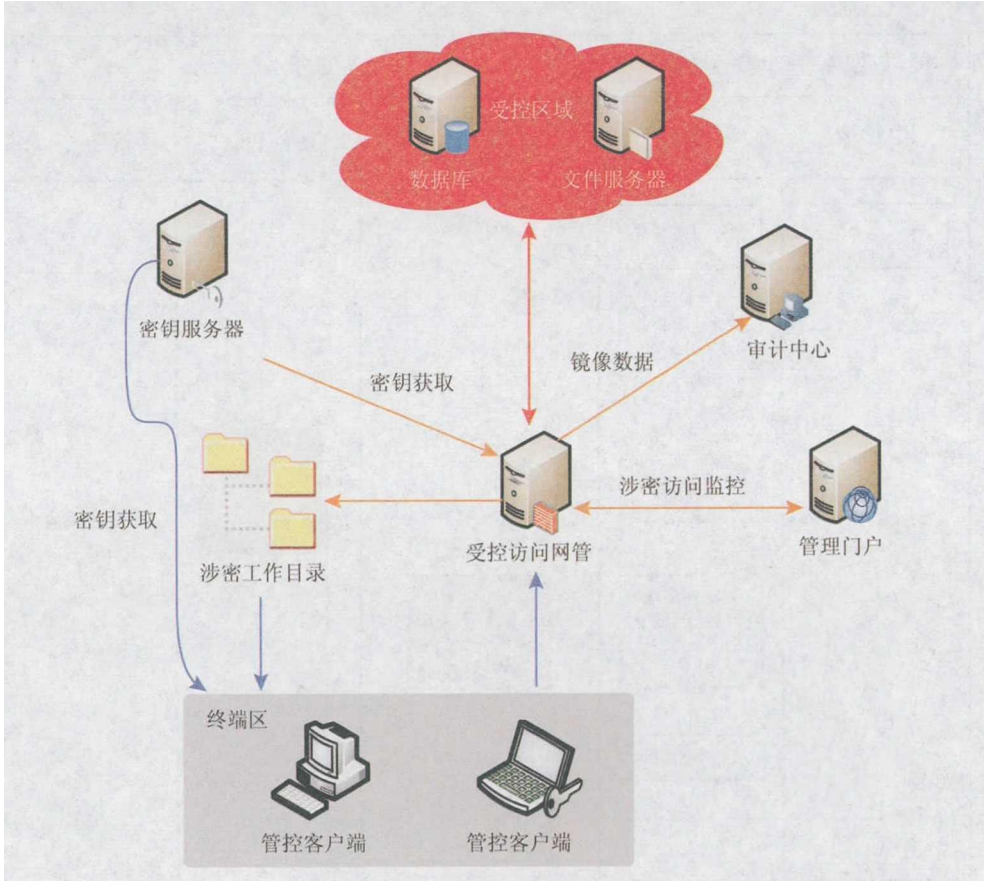

图 7　文档加解密部署方式

进行统计分析，并提供详细的数据访问报告，同时提供相应的控制接口，与外部系统进行协作，如：4A、SOC（security operations center）、第三方审计等.

受控访问网关对受控区域（例如：存储客户隐私数据的数据库、存有机密文件的服务器）的访问路径进行统一控制，同时对从受控区域内导出的客户隐私数据文件进行加密，保证受控区域内的敏感数据不被非法窃取.

文档加解密处理流程如图 8 所示.

2.4　数据脱敏系统设计

数据脱敏是一种数据失真处理的技术，目的是为了通过缩小用户权限而降低敏感数据的暴露风险. 在 Web 应用系统和第三方接口使用过程中会接触到客户的隐私数据，如姓名、身份证号、地址、收入、电话号码等；通过采用数据脱敏技术对敏感数据进行变形处理，实现客户隐私数据的可靠保护. 这样，在开发、测试和其他非生产环境以及外包环境中就可以安全地使用脱敏后的真实数据集.

2.4.1　明确数据模糊化实施原则

1）目的明确原则：处理个人信息具有特定、明确、合理的目的，不扩大使用范围，不在个人信息主体不知情的情况下改变处理个人信息的目的.

2）最少够用原则：只提供与业务需求有关的最少信息，达到业务需求后，在最短时间内删除个人信息.

3）最大化原则：在不影响使用的情况下，尽可能多的模糊化客户信息，保证敏感数据不会在流转过程中泄露.

4）处理源头化原则：尽可能选择在客户信息服务端进行模糊化处理，特别是在数据导出端或数据存储端进行模糊化处理.

5）健壮性原则：对于无需还原原始信息的数据，尽可能采取不可逆算法进行模糊化处理；对于需要还原原始信息的数据，可以适当采用具有足够健壮性的可逆算法.

6）参照业内所遵循的数据管理办法确定数据

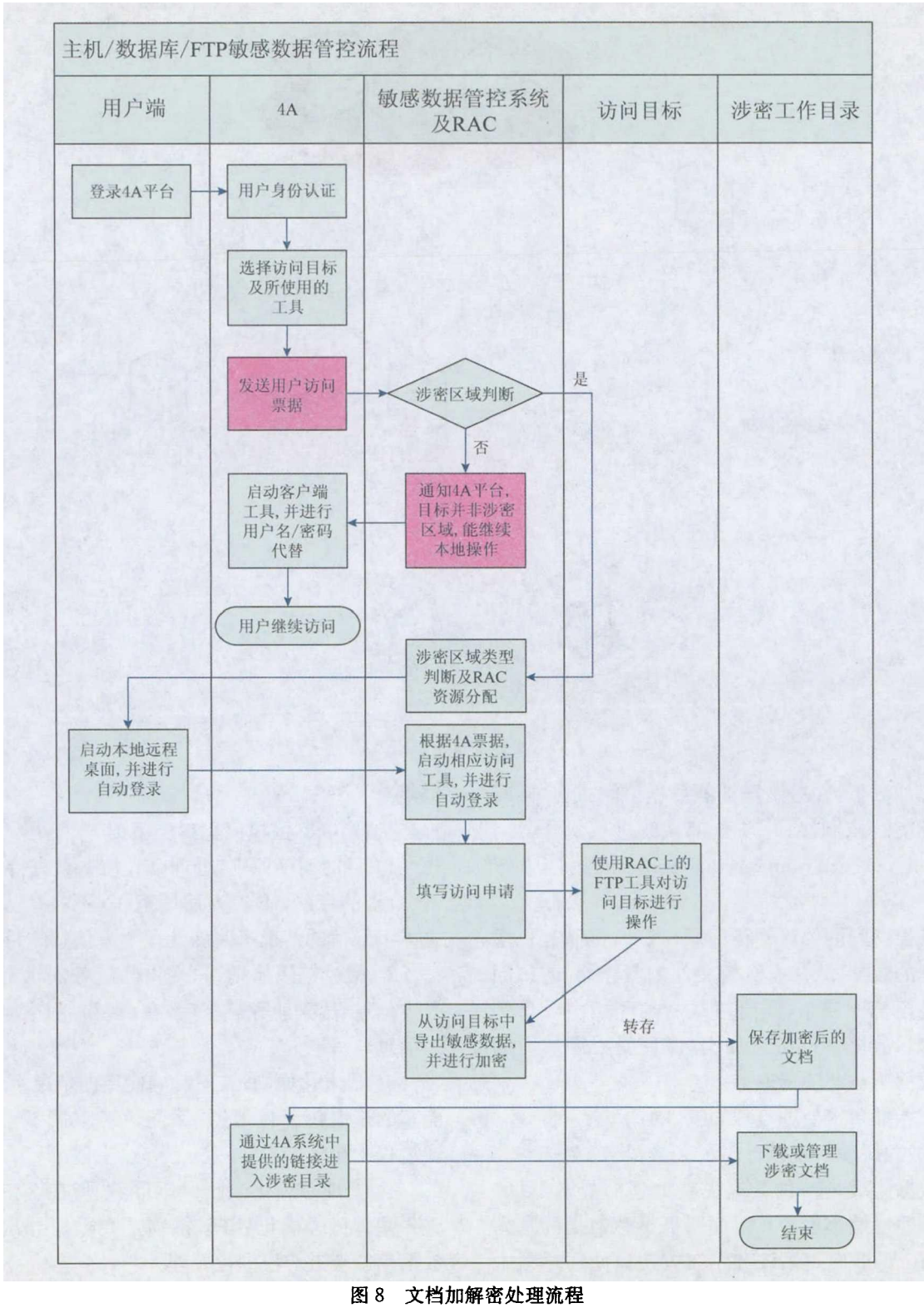

图 8　文档加解密处理流程

的价值和安全等级，并结合数据的使用场景完成数据模糊化. 根据数据的价值和安全等级，如：1 级敏感数据需要进行模糊化处理；2 级敏感数据需要进行评估后，再进行部分模糊化处理；非敏感数据不需要进行模糊化处理.

2.4.2　创建数据模糊化规则

根据实际的业务需求，创建明确的模糊化规则，如表 1 所示：

表 1 SQL 操作数据项目集

分类	模糊化字段	模糊化规则	举例
名称信息	客户名称	2 个字或 3 个字的至少 1 个字用*代替，大于 3 个字的至少 2 个字用*代替，模糊化名称前后、中间均可，也可以全部模糊化.	例如： 张三→张*或*三 李二宝→李**、**宝或李*宝 欧阳正华→欧阳**、**正华或欧**华 广西地税局→广**税局
	银行账户名称		
	工作单位名称	全部用 8 个*代替	
地址信息	证件地址	全部用 8 个*代替	
	联系地址		
	家庭地址		
	单位地址		
	邮件地址	@前面的用 3 个*代替，或全部使用 8 个*代替	例如： 13901234567@139. com→***@139. com
证件信息	身份证号码	出生年月日用*代替，最后一位用*代替，或全部使用 8 个*代替	例如： 330101197701014237→330101********423*
	护照号码	末 4 位用*代替，或全部使用 8 个*代替	例如： G12345678→G1234****
	军官证号码	末 4 位用*代替，或全部使用 8 个*代替	例如： 空字第 12345678→空字第 1234****
	其他证件号码	参照上述 3 种方式进行模糊化处理	
银行信息	银行账号	保留前 5 位和末 4 位，中间用*代替，或全部使用 8 个*代替	例如： 9558801202106562334→95588**********2334

2.5 异常行为检测系统设计

异常行为检测系统是客户隐私数据流转安全管理系统的核心部分；它明确了客户信息流转的各个环节，保证了 DPI 设备、文档加解密系统、数据脱敏系统的部署位置，并涵盖了客户隐私数据从产生、传输、处理、存储、使用、销毁的整个过程. 从而降低了没有纳入到管理系统而产生数据泄露的风险.

同时，异常行为检测系统通过 DPI 设备解析、存储过来的 HTTP 数据和 SQL 数据，根据用户异常行为检测方法对数据内容进行分析和告警.

2.5.1 客户信息流转梳理

通过人工访谈、业务系统梳理、4A 系统梳理、数据库字典分析的方法明确客户信息的流转过程，如图 9 所示.

1) 人工访谈：主要了解人员的角色划分和业务职能.

2) 业务系统梳理：了解业务系统的任意查询功能、批量查询功能、导出功能的角色和数据内容. 了解业务系统在客户隐私数据产生、处理、传输和存储的流转情况，其中传输包括本身系统内的传输和与第三方系统的传输.

3) 4A 系统梳理：了解 4A 管理资产中涉及客户隐私数据的资产信息的使用和分布情况；了解这些资产信息的网络结构，与业务系统梳理的资产进行查漏补缺.

4) 数据库字典分析：分析数据库中涉及客户隐私数据的字段在数据表的分布情况.

2.5.2 异常用户行为分析

用户行为数据，来源于 DPI 设备，由 DPI 设备进行 HTTP 流量还原和 SQL 流量还原后，将格式化的数据存入数据库，通过关联分析规则对数据内容进行分析，如图 10 所示.

关联分析是利用 Apriori 算法对用户行为进行分析，并将结果建立为用户正常行为规则库(对每一类用户即角色使用关联分析算法建立行为规则库). 下面给出用户行为规则库的分析和建立过程.

角色 A 的 SQL 操作数据项目集中有 $T_1, T_2, \cdots, T_{10}$ 共 10 个事务，有 sysuser1，sysuser2，sysuser3

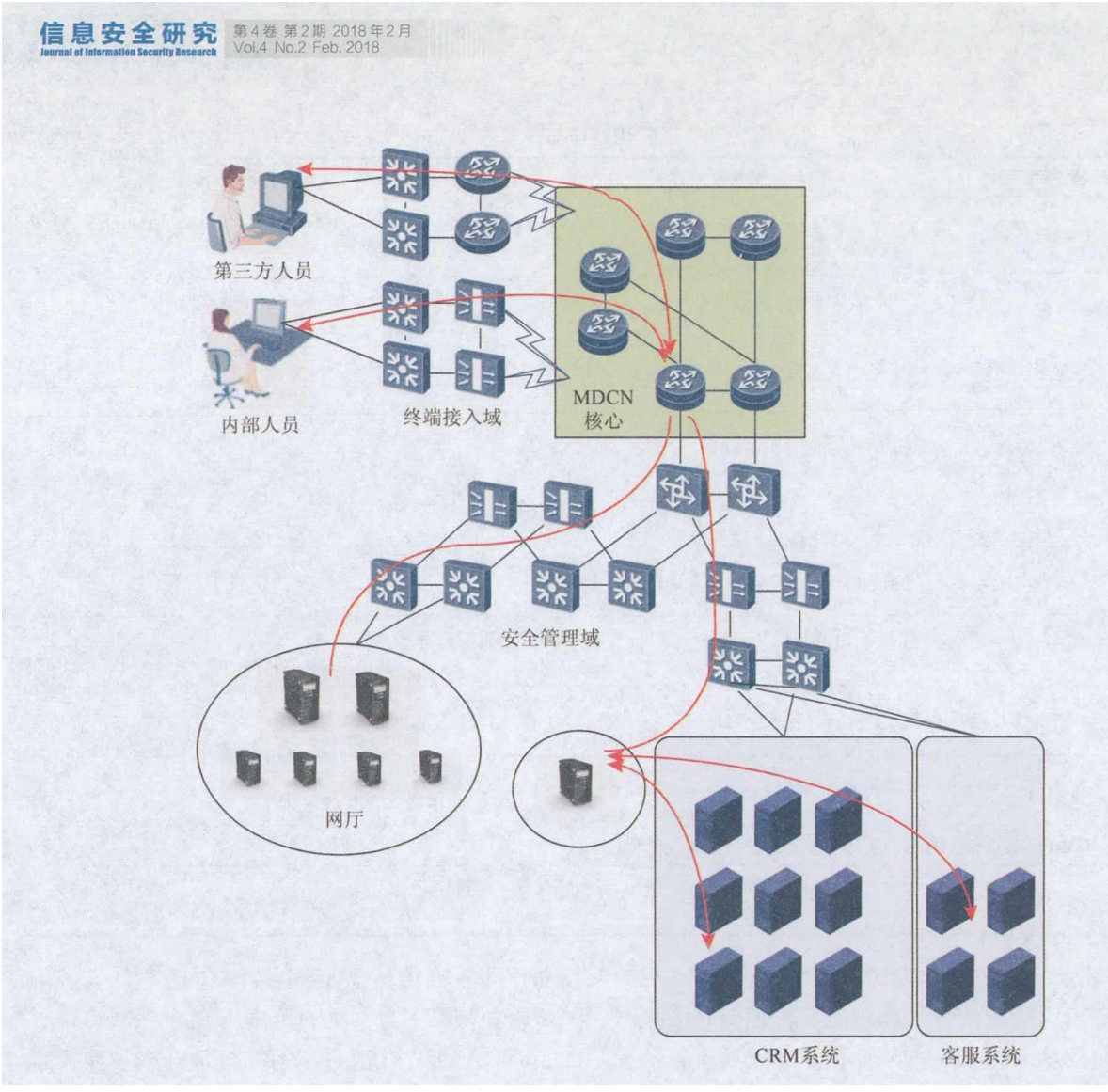

图 9　客户隐私数据流转过程

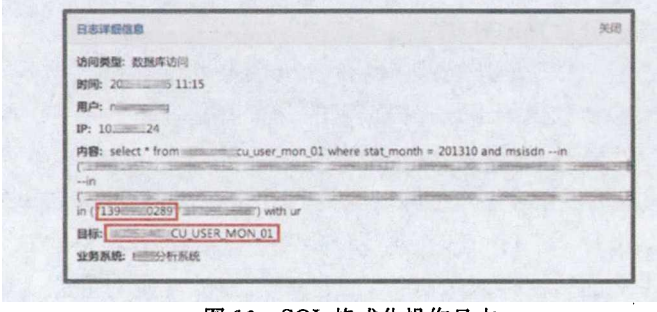

图 10　SQL 格式化操作日志

这 3 个用户使用行为对象表 DBA_SYNONYMS，DBA_USERS，进行的操作有 select，insert，delete，如表 2 所示.

将上面的数据转换为布尔型，如表 3 所示.

在数据处理过程中，有些属性值是互斥的，如 sysuser1 和 sysuser2 是互斥的，这些在候选集中进行了去除. 设最小支持事务计数为 2，则 $minS=20\%$，最小可信度 $minC=60\%$.

表 2　SQL 操作数据项目集

事务	用户名	操作对象	操作类型
T_1	sysuser1	DBA_SYNONYMS	insert
T_2	sysuser1	DBA_USERS	select
T_3	sysuser1	DBA_USERS	delete
T_4	sysuser2	DBA_SYNONYMS	select
T_5	sysuser2	DBA_SYNONYMS	select
T_6	sysuser2	DBA_SYNONYMS	insert
T_7	sysuser3	DBA_USERS	select
T_8	sysuser3	DBA_USERS	insert
T_9	sysuser3	DBA_USERS	select
T_{10}	sysuser3	DBA_USERS	select

最后由频繁 3-项集可以得到强关联规则：

sysuser2∩DBA_SYNONYMS⇒select 和 sysuser3∩DBA_USERS⇒select 且

confidence(sysuser2∩DBA_SYNONYMS⇒select)=2/3>$minC$,

confidence(sysuser3∩DBA_USERS⇒select)=2/3>$minC$,

⋮

上述公式既满足最小支持度,也满足最小可信度,所以它们为强关联规则,存入数据库中的规则库.

表 3 转换为布尔型数据

事务	sysuser1	sysuser2	sysuser3	DBA_SYNONYMS	DBA_USERS	select	insert	delete
T_1	1	0	0	1	0	0	1	0
T_2	1	0	0	0	1	1	0	0
T_3	1	0	0	0	1	0	0	1
T_4	0	1	0	1	0	1	0	0
T_5	0	1	0	1	0	1	0	0
T_6	0	1	0	1	0	0	1	0
T_7	0	0	1	1	0	1	0	0
T_8	0	0	1	0	1	0	1	0
T_9	0	0	1	0	1	1	0	0
T_{10}	0	0	1	0	1	1	0	0

根据表 3,采用 Apriori 算法进行分析,过程如图 11 所示:

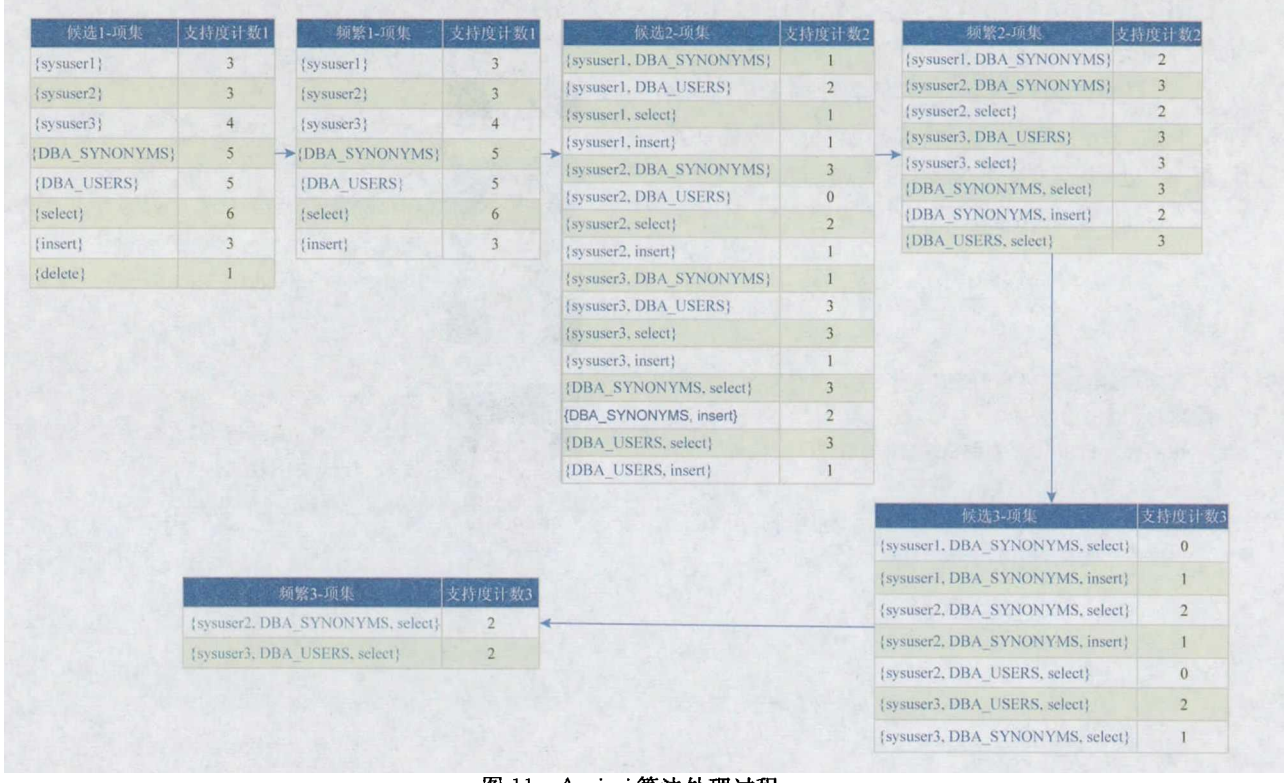

图 11 Apriori 算法处理过程

然后使用异常检测算法,将新的用户操作行为与正常行为规则库进行匹配,如果产生较大偏差,则认为用户操作行为异常,检测流程如图 12 所示:

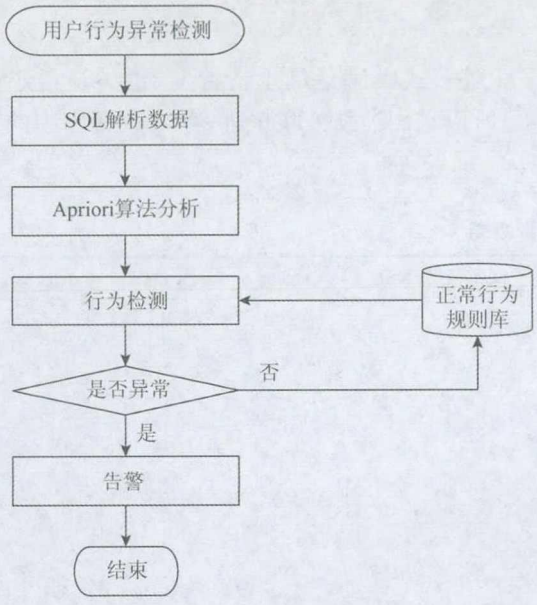

图 12 用户行为异常检测流程

3 结束语

DPI 技术、文档加解密技术、数据脱敏技术、异常行为分析技术都是信息安全领域比较领先的技术. 我们通过客户隐私流转安全管理系统，将各种技术紧密结合在一起，形成一套客户隐私在产生、传输、处理、存储、使用、销毁完整流程的安全管理系统，从而有效地提高了系统的安全保障能力.

参 考 文 献

[1] 郭瑞. 数据泄露风险的趋势分析[J]. 信息安全与技术，2014，5(10)：18-21

[2] 张勇进，张知恒. 信息安全产品体系概述[J]. 网络安全技术与应用，2001，1(1)：54-55

[3] 刘丽娜. 国内信息安全产品应用技术及市场现状分析[J]. 网络安全技术与应用，2016，16(12)：25-25

[4] 陈小文. 网络安全审计系统中数据采集的研究与实现[D]. 哈尔滨：哈尔滨工程大学，2009

[5] 武光达，蒋朝惠. 基于 DPI 的流量识别系统的研究[J]. 信息网络安全，2014，14(10)：44-48

[6] 向宇. HTTP 协议还原系统的设计与实现[D]. 武汉：华中科技大学，2011

[7] 陈炜. 基于网络的数据库审计和风险控制研究[D]. 武汉：武汉理工大学，2013

[8] 卞超轶，朱少敏，周涛. 一种基于保形加密的大数据脱敏系统实现及评估[J]. 电信科学，2017，33(3)：119-125

[9] 艾解清，魏理豪，王建永，等. 一种基于 DPI 的敏感文件流转监控方法：中国，CN106713067A [P]. 2017-05-24

[10] 刘亮，王雷，陈亮. 一种基于聚类分析的敏感数据异常访问检测方法：中国，CN106570131A [P]. 2017-04-19

艾解清
博士研究生，主要研究方向为信息化评测技术.
dotrai@126.com

魏理豪
硕士研究生，主要研究方向为信息化评测技术.
wlh_wind@126.com

梁承东
硕士研究生，主要研究方向为信息化评测技术.
liangcd@chinagdn.com

陈 亮
硕士研究生，主要研究方向为安全检测技术、数据安全技术.
eddie.chen@hotmail.com