基于多业务场景的大数据脱敏技术研究及其在电力用户隐私信息保护中的应用

吕 军,杨 超,王跃东,刘 林,王新宁
(国网大连供电公司,辽宁 大连 116011)

摘 要:数据流转过程中,保证数据安全使用的最有效方法即数据脱敏,通过将数据中的敏感信息标识化处理,确保隐私信息在使用过程中的安全可靠。本文首先对数据脱敏技术进行介绍,包括脱敏概念、脱敏原则和脱敏方法等;之后对电力企业数据特点进行分析,梳理形成用户隐私信息类别;进一步对电力系统中用户隐私信息的实际应用场景进行提取划分;在此基础上,应用不可逆的数据脱敏方法,制定形成数据脱敏规则,并根据数据使用的不同业务环境,提出动/静态大数据脱敏技术方案。该方案紧密结合电力企业实际业务需求,在应用过程中,无论是对现有系统进行改造,或是部署新的设备,都最大限度的保证了投入成本与实际产出间的合理配比,使得数据价值得到充分挖掘的同时,保障用户隐私信息安全。

关键词:大数据;数据安全;隐私信息保护;多业务场景;数据脱敏
文章编号:2096 - 4633(2018)07 - 0029 - 07 **中图分类号**:C39 **文献标志码**:B

随着社会网络化的高速发展,数据呈现爆炸式增长,大数据时代已悄然来临。当前,大数据逐渐成为国家基础性战略资源,对国家治理能力、经济运行机制、社会生活方式都将产生深远影响。在大数据应用过程中,应坚持安全与发展并重的原则,在充分发挥数据价值的同时,着力解决数据安全及个人信息保护问题。大数据具有大量、多样、价值、高速、真实等特点,由于大数据系统中普遍存在大量个人信息,在发生数据滥用、网络攻击等安全事件时,时常伴随个人信息泄露等问题。《网络安全法》中对个人信息保护提出了明确要求,突出强调个人信息在采集、使用、管理等各环节的保护。作为关键信息基础设施运营单位,国家电网公司近年来高度重视数据保护工作。目前公司各类信息系统中存储着海量数据,其中包含大量的用户隐私信息,此类数据一旦泄漏,将对公司正常运营造成严重影响[1-5]。通过采用数据脱敏技术,将保障用户隐私信息在数据交换共享、测试开发、对外发布等场景中的安全使用。

本文将以介绍数据脱敏技术为基础,分析用户隐私信息保护重点,结合电力企业生产实际,按场景构建大数据脱敏技术方案,后续章节安排如下:

第 1 章:介绍数据脱敏技术概念,形成用户隐私信息脱敏规则。

第 2 章:提出基于场景的动/静态大数据脱敏方案。

第 3 章:对全文进行总结。

1 数据脱敏技术

1.1 基本概念

数据脱敏(data masking, DM),又称数据漂白、数据去隐私化或数据变形[6]。是指采用既定的脱敏方法,对敏感信息进行数据变形,在不违反系统规则前提下,对数据进行改造并提供使用,保证在访问、开发、测试和其他环境中安全地使用脱敏后真实数据集,实现敏感内容的可靠保护。

1.2 数据脱敏原则

数据脱敏不仅要确保敏感信息被去除,还需要充分考虑脱敏花费、实际业务需求等因素,数据保护与数据挖掘是一对矛盾体,既要使数据潜在价值得到充分应用,又要确保敏感信息不被泄露。因此,在进行数据脱敏时,需要明确脱敏数据范围、脱敏需求以及脱敏后的数据用途[7]。数据脱敏建议遵循以下原则:

(1)不可逆原则。数据经过脱敏处理后,敏感信息已被移除,且无法通过技术手段还原敏感内容。

(2)主动防御原则。数据脱敏时,要充分考虑同质属性攻击、概率攻击、知识推断攻击和相似性攻击等攻击手段,提升脱敏处理的主动防范能力。

(3)可用性原则。保证脱敏后数据在非原始环境中的可用性,保障数据的真实性。

(4)自主可控原则。脱敏工具开发与脱敏工作分离,脱敏规则可灵活配置,确保脱敏工作的自主可控。

1.3 数据脱敏方法

脱敏方法是实现数据脱敏的关键所在,一般可分为两类:可逆类与不可逆类[8-9]。

可逆类方法。脱敏后数据可通过一定方式恢复敏感信息。可逆类方法会造成脱敏数据的不当使用。实际业务中,应尽量避免使用可逆类脱敏方法。

不可逆类方法。脱敏后数据不能恢复敏感信息,但不可逆方法的安全性也不是绝对的,实际业务中,要对脱敏后数据的安全性进行评估,避免对不可逆方法的不当使用。

目前针对个人隐私信息保护,常用的脱敏方法主要包括:

(1)时间偏移取整。对时间随机进行向上(向下)偏移取整,即保证时间数据满足一定的分布特征,同时隐藏原始时间。例如:将时间20170528 02:04:11,采用日期按2天、时间按5秒的粒度进行向下取整,得到20170525 02:04:06。

(2)不可逆替换。使用随机数据对原始数据进行不可逆替换。例如:将34567替换为cdefg。

(3)不可逆轮询。将原始数据排成一个序列,指针向前(向后)移动n位得到新数据。例如:对姓氏做百家姓不可逆轮询。例如姓氏"李"按百家姓顺序向后移4位,转换为"王"。

(4)截断:舍弃关键信息,仅保留部分信息,以保证数据的模糊性。例如:将地址"北京市西城区宣武门东大街4号楼2单元611"截断为:北京市西城区宣武门东大街。

(5)掩码:对敏感数据的部分内容用通用字符(如"X、*"等)进行统一置换,使敏感数据仅部分内容公开,对信息持有者易于辨别。该方法在实现脱敏的同时,保证信息的长度不变,是当前使用最为广泛的脱敏方法。例如:手机号码13803412597经掩码得到138****2597。身份证号230154197703284115经掩码得到230154XXXXXXXX4115。

(6)随机化:参考原始数据的特征,重新随机生成数据,部分情况下进行加盐(随机盐)处理,提升安全性。这种随机生成数据与原始数据间没有映射关系,因此具有不可逆性。

此外,面对大数据环境下,多源数据分析带来的安全威胁,部分脱敏方法无法保证脱敏的有效性(敏感信息可能被还原),需要研究适用于此的脱敏方法,目前较为有效的脱敏方法是K-匿名。

K-匿名模型(K-Anonymity model):K-匿名模型是在发布数据时保护个人信息安全的一种模型。要求发布的数据中,指定标识符(直接标识符或准标识符)属性值相同的每一等价类至少包含K个记录,使攻击者不能判别出个人信息所属的具体个体。K-匿名模型还包括一些增强概念,如L-多样性和T-接近性。

①L-多样性(L-diversity)。L-多样性要求在K-匿名的基础上,实现每一等价类在每一敏感属性上存在至少L个不同值。

②T-接近性(T-closeness)。T-接近性是L-多样性的增强概念,要求任何等价类中敏感属性的分布与整个数据集中相应属性的分布之间的距离小于阈值T。

差分隐私(differential privacy):对于数据集A,设定一个算法扰动机制,扰动后得到A′,再从原数据集A里随意拿掉一行记录得到B,对这个数据集B做扰动得到B′,如果得到的A和B几乎在数据分析的结果上是一致的。即A里面任何单独一行数据存在或不存在都几乎不影响结果,简单理解"攻击者无法确定我的信息在不在这个数据集里",此方法具有较强的隐私保护能力,但是在大数据环境下实现较为复杂。

1.4 用户隐私信息划分

用户隐私信息的划分需要以数据资产的梳理为基础。通过确定本单位包含的数据类型、数据分布存储情况、数据使用情况、数据流向等,进一步分析出敏感数据并进行分级分类管控,在此基础上,识别划分用户隐私信息[10]。

1.4.1 数据分类

电网企业包含不同业务部门及单位,其中涉及个人信息最多的主要为客户服务中心。根据对公司客服中心数据进行的调研,结合实际业务应用、内部安全管理和对外开放共享等特点,将相关数据整合划分为"客户身份相关数据"、"业务权属数据"、"业务辅助数据"以及"服务衍生数据"四大类。

1.4.2 数据分级

数据分级依据如下原则:一是界限明确。数据

分级需按照数据敏感程度进行划分。二是就高不就低。如果同一批数据中各个属性或字段的分级不同,则需要参照定级最高的属性或字段级别,一并实施安全管控。

数据分级方法。基于对数据分类及上述分级原则,按照数据敏感程度进行敏感性分级。同时,为了便于进一步对数据进行敏感等级标记,每一级别的数据再进一步细分。

(1)高敏感级数据。高敏感级数据是指其泄露会对信息基础设施和客户造成极其重大影响的数据。将涉及信息基础设施和 VIP 级别客户的数据定为“H1 级”,该级数据的泄露会造成严重网络安全事件;将涉及客户的身份鉴权和金融类信息的数据定为“H2 级”,该级数据的泄露会对客户造成严重的声誉和财产损失。根据数据特性,对高敏感级数据,需实施最严格的技术和管理措施,建立最严格的数据安全管理规范以及数据实时监控机制。严禁该级别数据输出。

(2)中敏感级数据。敏感级数据是指其泄露会对客户隐私和本单位生产运营造成重大影响的数据。将客户标识与资产数据定为“M1 级”,该级数据泄露,会对客户声誉和财产造成一定损失;将涉及本单位生产运营的数据定为“M2 级”,该级数据泄露,会对企业形象和财产造成一定的损失。根据数据特性,对中敏感级数据应实施严格管理措施与审批机制。针对特定数据项实施严格的数据脱敏和客户隐私保护措施。该级别数据经过脱敏处理后可以对内共享,若未进行信息泄露评估,应避免对外输出。

(3)低敏感级数据。低敏感级数据是指其泄露对本单位的业务和客户造成较小影响的数据。对于低敏感级数据,将基本业务相关的数据定为“L1 级”,将与客户相关的低敏感数据定为“L2 级”。低敏感级数据在经过必要的数据脱敏处理并满足相关管理审核条件的前提下,可对内共享,减少对外输出。需做好数据共享和输出的详细记录。

1.4.3　用户隐私信息

结合公司业务系统数据分析结果,目前用户隐私信息可划分为用户身份标识信息、用户身份鉴权信息和用户身份辅助信息三类:

(1)用户身份标识信息,主要包括 VIP 用户身份标识、自然人身份标识、网络身份标识、机构客户身份标识等;

(2)用户身份鉴权信息,主要包括自然人身份实体证明信息、虚拟身份鉴权信息、机构身份实体证明信息等;

(3)用户身份辅助信息,主要包括相关联系人信息、用户偏好信息、用户资产相关信息、用户金融信息、用户档案信息等。

1.5　脱敏规则

综合上述分析后,遵照数据脱敏原则,基于典型的数据脱敏方法,形成公司用户隐私信息保护脱敏规则,如表 1 所示。

表 1　用户隐私信息脱敏规则

Tab. 1　Masking rules of user privacy information

序号	字段	示例数据	脱敏方法	脱敏效果
1	标识类数据	ID50384153	对数字做随机盐	ID21895502
2	身份证号码	230154197703284115	保留前 6 位,对其后 8 位做掩码	230154＊＊＊＊＊＊＊＊4115
3	客户姓名	张三丰	对姓做百家姓不可逆轮询,名做掩码	赵＊＊
4	联系电话	13803412597	最后 4 位用 0000 替换 中间 4 位做掩码	13803410000 138＊＊＊＊2597
5	联系地址	北京市海淀区 学院南路 15 号	根据地址中的关键字 (如省、市、区等)做截断	北京市海淀区
6	车牌号码	京 N5678L	保留地域信息,对其中的字母做不可逆轮询, 数字做不可逆随机替换	京 N7682S
7	开票日期	2017/03/12 12:43:27	日期按 2 天粒度向下偏移,时间按小时取整	2017/03/10 13:00:00
8	IP 地址	192. 168. 11. 21	掩码后四个字节	192. 168. 11. ＊＊
9	电量数据	35104. 68	偏移取整	35000
10	电表用户号	1001786249	中间四位做掩码	100＊＊＊＊49

2 基于场景的大数据脱敏方案研究

2.1 脱敏场景分析

(1)交换与共享场景:数据在公司内部的交换和共享主要用于分析评估、审计、培训等环节,是其价值的重要体现。在数据交换和共享过程中存在隐私信息泄露的风险,需要对相关数据进行脱敏处理,保证用户隐私信息在内部使用过程中的安全可靠。

(2)开发与测试场景:为保证功能开发和集成测试的顺利进行,公司各类系统在研发和测试环节需要导入大量原始数据,直接使用势必引发信息泄露事件。因此,需要对原始数据进行脱敏处理后在相关环境中使用。

(3)对外发布场景:数据外部流转是公司数据业务的重要应用形式之一,无论是用电客户的访问请求,还是外部单位的数据交易,涉及到用户个人信息在外发过程中不应存在隐私泄露风险,需要考虑在大数据环境下对用户隐私信息进行分析和脱敏处理。

2.2 数据脱敏实现方法

按使用敏感数据时是否进行脱敏操作进行划分,数据脱敏可分为静态数据脱敏(SDM)和动态数据脱敏(DDM)两种主要方法[11-13]。

2.2.1 动态数据脱敏

动态数据脱敏一般用于生产环境,不改变生产数据库中的原始数据,只对"输入请求"和"输出数据"进行实时脱敏处理,防止敏感数据外泄。这种脱敏形式适用于对生产数据的动态访问和检索,通常与访问权限结合使用。如图 1 所示。

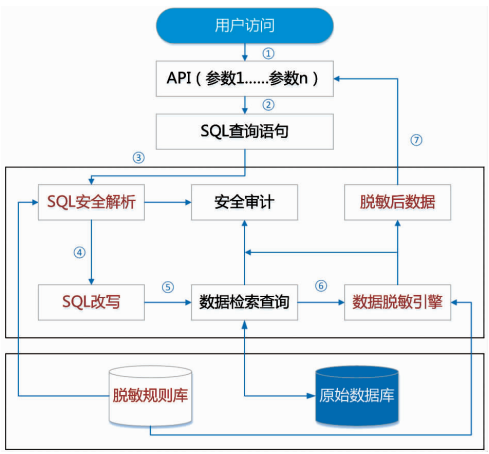

图 1 动态数据脱敏流程图

Fig. 1 Flow chart of dynamic data masking

动态数据脱敏主要流程包括:

①用户端输入数据访问请求至 API(参数 1……参数 n);

②系统将数据访问请求转换为 SQL 查询语句;

③系统对 SQL 语句进行安全性、合规性检查(对于不安全的 SQL 语句直接抛弃并反馈警告信息);

④对于安全的 SQL 查询语句,结合脱敏规则库,进行 SQL 语句改写;

⑤基于改写的 SQL 语句进行数据检索查询;

⑥查询后数据经脱敏引擎,进行实时脱敏处理;

⑦脱敏后数据作为结果反馈发送给用户端。

系统内部各关键环节的操作,都需进行安全审计,并保存相关网络安全日志。动态数据脱敏装备模块示意图如图 2 所示。

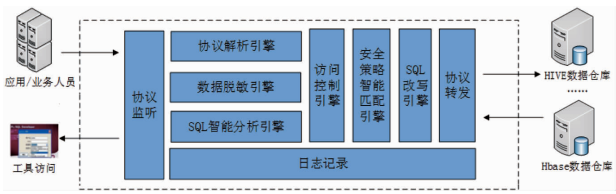

图 2 动态数据脱敏装备模块图

Fig. 2 Module diagram of dynamic data masking equipment

2.2.2 静态数据脱敏

静态数据脱敏一般用于非生产环境,或是对在线数据进行离线脱敏处理。脱敏完毕后在非生产环境中使用,主要用于数据批量外发共享、系统开发测试等。如图 3 所示。

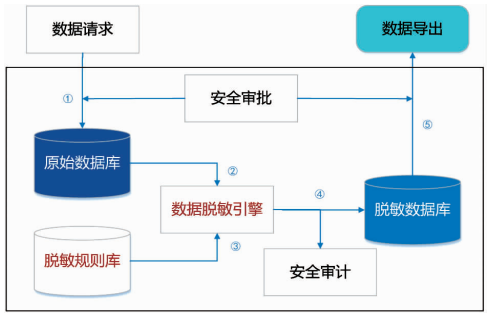

图 3 静态数据脱敏流程图

Fig. 3 Flow chart of static data masking

静态数据脱敏主要流程包括:

①根据实际需求场景的不同,提出数据使用请求;

②、③依托数据脱敏引擎,结合脱敏规则库(脱敏算法),对原始数据库中的重要数据及客户隐私信息进行脱敏处理;

④脱敏后数据存储于专用的脱敏数据库;

⑤根据数据使用请求,将脱敏数据批量导出;

在数据请求及数据导出阶段,需要经过安全审批流程,确保相关行为的安全合规。

静态数据脱敏在离线场景下的应用[14-15],如图 4、图 5 所示,构建的脱敏系统可以实现数据的离线脱敏批量处理,可满足大数据导出共享、内部分析试验、研发测试等典型业务需求。

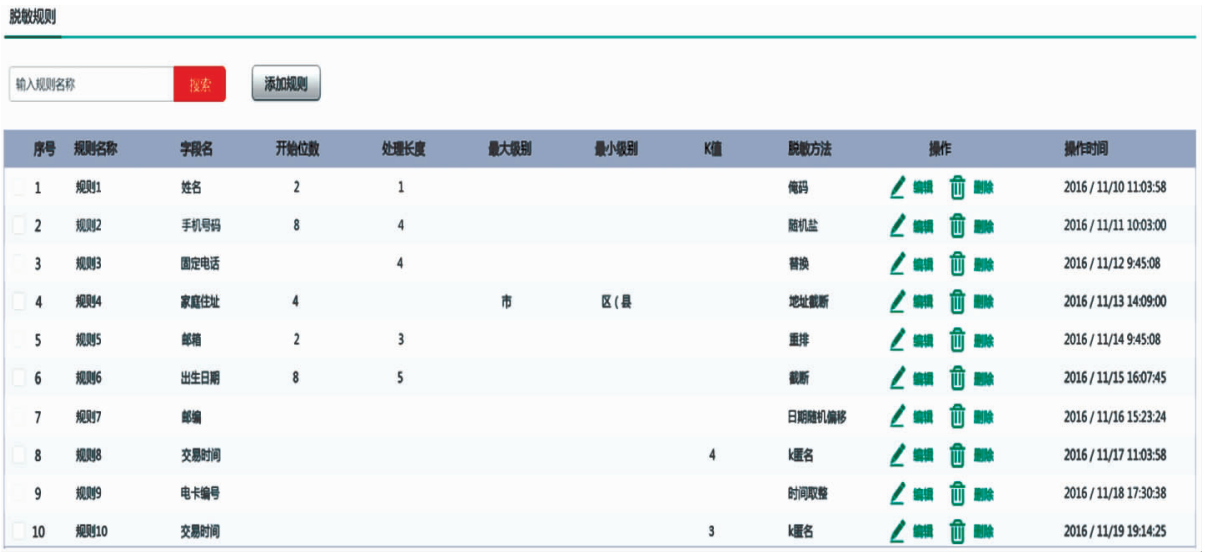

图 4　数据脱敏规则管理

Fig. 4　Data masking rule management

图 5　数据脱敏任务管理

Fig. 5　Data masking task management

3　结语

本文在综合分析数据脱敏技术的基础上,结合电力企业实际,对用户隐私信息制定合理的数据脱敏规则,同时结合不同业务场景,提出动/静态大数据脱敏方案。该方案可以很好的保障用户隐私信息在各类典型应用环境下的安全。在应用过程中,无论是对现有系统进行改造,或是部署新的设备,该方案都最大限度的保证了投入成本与实际产出间的合理配比。

针对个人信息保护,数据脱敏是目前最行之有效的方法之一,需要注意的时,数据脱敏离不开两个前提:一是敏感数据的识别,除个人隐私信息外,涉及企业正常生产经营的数据,按不同专业划分,都应明确哪些为敏感数据;二是数据的分类分级,在识别数据敏感程度基础上,应建立数据资产目录以及分类分级规范,从而分层次有依据的对不同数据实施保护[16]。

数据安全作为网络与信息安全的一个重要组成,有其独有的侧重点,但不应与传统的安全防护措施割离开来,诸如身份认证、访问控制、入侵防范、鉴权授权、加解密、审计备份、追踪溯源等,都应是数据

安全所必须的技术手段。在当前大数据环境下,更应充分考虑大数据海量、多源、异构、动态的特性,不断研究适应新形势下的数据安全防护技术,保证数据使用的安全可靠[17-19]。

参考文献:

[1] 傅拥钢. 利用大数据提高电力企业社会保险管理水平[J]. 电力大数据,2018,21(01):35-38.
FU Yonggang. Using big data to improve the social insurance management level of power enterprises[J]. Power System and Big Data,2018,21(01):35-38.

[2] E KENNED,C MILLARD. Data security and multi-factor authentication: Analysis of requirements under EU law and in selected EU member states[J]. Computer Law & Security Review the International Journal of Technology Law Practice,2016,32(01):91-110.

[3] D TALIA. Clouds for scalable big data analytics[J]. Computer,2013,46(05):98-101.

[4] Z TAN,UT NAGAR. Enhancing big data security with collaborative intrusion detection[J]. IEEE Cloud Computing,2015,1(03):27-33.

[5] ROB J HYNDMAN. Visualizing big energy data:solutions for this crucial component of data analysis[J]. IEEE Power and Energy Magazine,2018,16(03):181-182.

[6] 王丹,赵文兵. 大数据安全保障关键技术分析综述[J]. 北京工业大学学报,2017,43(03):335-349.
WANG Dan,ZHAO Wenbing. review of big data security critical technologies[J]. Journal of Beijing University of Technology,2017,43(03):335-349.

[7] 吕欣,韩晓露. 大数据安全和隐私保护技术架构研究[J]. 信息安全研究,2016,2(03):244-250.
LV Xin,HAN Xiaolu. Research on the technology architecture of big data security and privacy system[J]. Journal of Information Security Research,2016,2(03):244-250.

[8] 李一平,王晨. 大数据平台的敏感数据保护研究[J]. 电信工程技术与标准化,2017,30(11):35-38.
LI Yiping,WANG Chen. Research on sensitive data protection of big data platform[J]. Telecom Engineering Technics and Standardization,2017,30(11):35-38.

[9] 乔宏明,梁奂. 运营商面向大数据应用的数据脱敏方法探讨[J]. 移动通信,2015,39(13):17-20.
QIAO Hongming,LIANG Huan. Discussion on data masking oriented to big data application for operators[J]. Mobile Communications,2015,39(13):17-20.

[10] 彭小圣. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报,2015,35(03):503-511.
PENG Xiaosheng. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE,2015,35(03):503-511.

[11] 卞超轶,朱少敏. 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(03):119-125.
BIAN Chaoyi,ZHU Shaomin. Implementation and evaluation of big data desensitization system based on format-preserving encryption[J]. Telecommunications Science,2017,33(03):119-125.

[12] 陈克非,翁健. 云计算环境下数据安全与隐私保护[J]. 杭州师范大学学报(自然科学版),2014,13(06):561-570.
CHEN Kefei,WENG Jian. Data security and privacy protection in cloud computing[J]. Journal of Hangzhou Normal University (Natural Sciences Edition),2014,13(06):561-570.

[13] S ALDOSSARY,W ALLEN. Data security,privacy,availability and integrity in cloud computing:issues and current solutions[J]. International Journal of Advanced Computer Science & Applications,2016,7(04):485-498.

[14] ZHIYUAN TAN,UPASANA T NAGAR. Enhancing big data security with collaborative intrusion detection[J]. IEEE Cloud Computing,2014,1(03):27-33.

[15] 张少敏,李晓强,王保义. 基于 Hadoop 的智能电网数据安全存储设计[J]. 电力系统保护与控制,2013,41(14):136-140.
ZHANG Shaomin,LI Xiaoqiang,WANG Baoyi. Design of data security storage in smart grid based on Hadoop[J]. Power System Protection and Control,2013,41(14):136-140.

[16] 郭仁超,徐玉韬. 内外网数据安全交换技术在电网企业的应用研究[J]. 电力大数据,2018,21(02):61-66.
GUO Renchao,XU Yutao. Research onb the application of data security exchange technology of Internal and external network in power grid enterprises[J]. Power System and Big Data,2018,21(02):61-66.

[17] 崔敏龙. 商业秘密保护中数据脱敏技术研究[D]. 陕西:西安电子科技大学,2015.

[18] 乔亚男. 云计算服务下信息网络传播权侵权责任研究[J],电力大数据 2017,20(12):72-73.
QIAO Yanan. Research on tort liability of information network dissemination right under cloud computing services[J]. Power systems and big data. 2017,20(12):72-73.

[19] 赵晓明,张学强,曹岚. 基于关键词的电力系统"大数据"与"云计算"专题文献分析[J]. 浙江电力,2016,35(02):27-30.
ZHAO Xiaoming,ZHANG Xueqiang,CAO Lan. Thematic analysis of "big data" and "cloud computing" in power system based on key words[J]. Zhejiang Electric Power,2016,35(02):27-30.

收稿日期:2018-06-16

作者简介:

吕 军(1987),男,硕士,工程师,主要从事电力通信、网络与信息安全方面的研究工作。

(本文责任编辑:范 斌)

Research on big data masking technology based on multi service scenarios and its application in privacy information protection of power users

LV Jun, YANG Chao, WANG Yuedong, LIU Lin, WANG Xinning

(State Grid Dalian Power Supply Company, Dalian 116011 Liaoning, China)

Abstract: Data masking is the most effective method to ensure the safe use of data in the process of data transfer. By hide identify the sensitive information in the data, it ensures the security and reliability of the privacy information in the process of use. This paper first introduces the data masking technology, including masking concept, masking principle and masking method, and then analyzes the data characteristics of the power enterprise, combs and forms the user privacy information category, and further extracts and divides the actual application scene of the user privacy information in the power system. On this basis, by using the irreversible data masking method, the data masking rules are formed, and the dynamic / static large data masking scheme is proposed according to the different business environment. This scheme closely combines the actual business needs of the electric power enterprises. During the application process, whether the existing system is reformed or the new equipment is deployed, the reasonable ratio between the input cost and the actual output is guaranteed to the maximum extent, so that the data value is fully excavated and the security of the user's privacy information is ensured.

Key words: big data; data security; privacy information protection; multi service scenarios; data masking