美创科技

数据脱敏原理及方法简析

数据爆炸式增长， 大数据成为国家基础性战略资源。 大数据中蕴藏的巨大商业价值被认可，但也带来了一个问题：“ 大数据对人们来说， 真的只有好处吗？ ”。

十三五规划纲要中明确提出： “ 实施国家大数据战略， 推进数据资源开放共享。 ”然而，各行业数据中包含大量的个人隐私数据与敏感、重要数据，一旦泄露或遭到非法利用， 将会给个人甚至国家带来无法弥补的损失。 同时， 随着大数据分析的成熟和价值挖掘的深入， 利用大数据学习技术从大量相关联的普通数据中还原出用户的敏感、隐私信息已不再困难。

如何在数据交换、 共享及使用等过程中实现对敏感数据的定向、 精准和彻底脱敏，达到数据安全、可信、受控使用的目标，是数据产生者和管理者亟待解决的技术问题。因此，数据安全技术和数据隐私相关技术成为安全技术热门。

当前，数据安全技术包括数据加密、数据脱敏、访问控制、安全审计、备份恢复、运维管理等。本文主要从数据脱敏这一安全控制手段入手。

一、 数据脱敏与安全控制

数据脱敏又称数据去隐私化，或数据变形，是在给定的规则、策略下对敏感数据进行变换、修改的技术机制，能够在很大程度上解决敏感数据在不可控环境中使用的问题。国内银行、通信运营商等是最早开始使用数据脱敏工具的单位。多以静态脱敏为主。

在各行业中以金融、政府和医疗行业涉及敏感信息最多，都有明确的数据脱敏需求，特别是在应用开发、测试、培训等环节。因为开发、测试、培训等环境的安全风险较大，如果在这种情况下使用真实数据，恐将面临严重泄露。

例如在例行拷贝敏感数据或者常规生产数据到非生产环境中时不经意的泄露信息。具体表现有：

1. 大部分公司将生产数据拷贝到测试和开发环境中，允许系统管理员来测试、升级、更新和修复。

2. 为在商业上保持竞争力，需要新的和改进后的功能。而应用程序的开发者需要一个环境仿真来测试新功能，以确保已经存在的功能没有被破坏。

3. 零售商将各个销售点的销售数据与市场调查员分享，从而分析顾客们的购物模式。

4. 医药组织向调查员分享病人的数据，来评估诊断疗效。

这些被拷贝到非生产环境中的真实数据，变成了黑客们或内部心怀不轨人员的目标。一旦被窃取或者泄露，可能会造成难以挽回的损失。

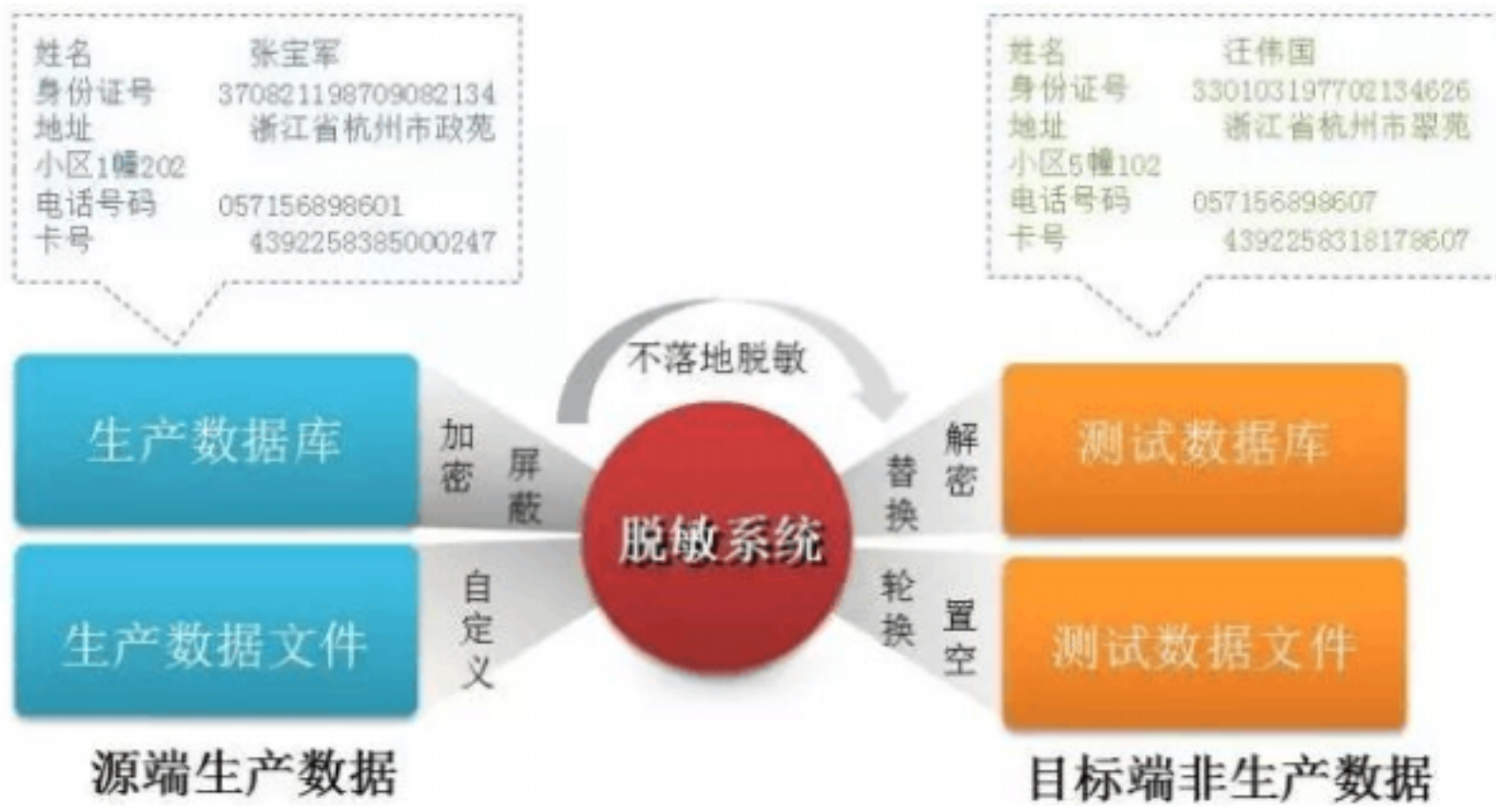
二、 数据脱敏的原理

数据脱敏在保留数据原始特征的条件下， 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。 在不违反系统规则条件下， 对真实数据进行改造并提供测试使用，如身份证号、手机号、卡号、客户号等个人信息都需要进行数据脱敏。 只有授权的管理员或用户， 在必须知晓的情况下， 才可通

过特定应用程序与工具访问数据的真实值，　从而降低重要数据在共享、　移动时的风险。

数据脱敏在不降低安全性的前提下，　使原有数据的使用范围和共享对象得以拓展，因而，成为大数据环境下最有效的敏感数据保护方法之一。

在数据脱敏系统的帮助下，　单位企业能够按照数据使用目标，　通过定义精确、灵活的脱敏策略，按照用户的权限等级，针对不同类别的数据以不同方式脱敏，实现跨工具、应用程序和环境的迅速、一致性的访问限制。



数据脱敏通常遵循的几条原则包括：

(1)　数据脱敏算法通常应当是不可逆的，必须防止使用非敏感数据推断、　重建敏感原始数据。但在一些特定场合，也存在可恢复式数据复敏需求。

(2)　脱敏后的数据应具有原数据的特征，　因为它们仍将用于开发或测试场合。带有数值分布范围、具有指定格式　　(如信用卡号前四位指代银行名称　　)的数据，脱敏后应与原始信息相似。　姓名和地址等字段应符合基本的语言认知，　而不是无意义的字符串。在要求较高的情形下，还要求具有与原始数据一致的频率分布、字段唯一性等。

(3) 数据的引用完整性应予保留，如果被脱敏的字段是数据表主键， 那么相关的引用记录必须同步更改。

(4) 对所有可能生成敏感数据的非敏感字段同样进行脱敏处理。 例如，在病人诊治记录中为隐藏姓名与病情的对应关系，将“ 姓名 ”作为敏感字段进行变换。但是，如果能够凭借某“ 住址 ”的唯一性推导出“ 姓名 ” ，则需要将“ 住址 ”一并变换。

(5) 脱敏过程应是自动化、可重复的。 数据处于不停的变化中， 期望对所需数据进行一劳永逸式的脱敏并不现实。 生产环境中数据的生成速度极快， 脱敏过程必须能够在规则的引导下自动化进行， 才能达到可用性要求， 更多强调的是不同环境的控制功能； 另一种意义上的可重复性， 是指脱敏结果的稳定性。 在某些场景下，对同一字段脱敏的每轮计算结果都相同或者都不同， 以满足数据使用方可测性、模型正确性、安全性等指标的要求。

三、 脱敏的方法

替换： 以虚构的数据代替真实的数据，如建立一较大的字典数据表，对每一真实值记录产生随机因子，对原始数据内容进行字典表内容的替换。这种方法得到的数据与真实数据非常相似。

无效化： 以特殊符号代替真值或真值的一部分，如遮盖身份证号码前 6-14 位。

乱序： 对敏感数据列的值进行重新随机分布，混淆原有值和其他字段的联系，这种方法不影响原有数据的统计特性，如该列总金额与原数据无异。

平均取值： 针对数值型数据，首先计算它们的均值，然后使脱敏后的值在均值附近随机分布，从而保持数据的总和不变。通常用于成本表、工资　　表等场合。

反关联： 查找可能由某些字段推断出另一敏感字段的映射，并对这些字段进行脱敏，如从出生日期可推断出身份证号、性别、地区的场景。

偏移： 通过随机移位改变数字数据。

对称加密： 这种加密是一种特殊的可逆脱敏方法。通过加密密钥和算法对原始数据进行加密，密文格式与原始数据在逻辑规则上一致，通过解密密钥可以恢复原始数据。

动态环境控制： 根据预定义规则，仅改变部分回应数据，如不在约定情况下访问业务数据时，控制数据内容，屏蔽特定字段内容。如不给　　DBA 账号显示重要客户信息，仅对业务模块的关键用户显示。（在生产环境中使用较多）。