学校编码：10384　　分类号＿＿＿＿密级＿＿＿＿

学号：X2013232387　　UDC＿＿＿＿

厦门大学

工　程　硕　士　学　位　论　文

基于 IBM Optim 的证券公司数据脱敏平台的设计与实现

Design and Implementation of Data Desensitization Platform for Securities Companies Based on IBM Optim

张玮希

指 导 教 师：　邱明　助理教授

专 业 名 称：　软　件　工　程

论文提交日期：　2015　年　9　月

论文答辩日期：　2015　年　11　月

学位授予日期：　　年　　月

指　导　教　师：＿＿＿＿

答辩委员会主席：＿＿＿＿

2015 年 9 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

数据脱敏是为了满足非生产环境中对数据的使用要求，而采取的通过一些规则对数据进行变形，从而达到对隐私数据保护的需求。对于证券公司而言，开发和测试环境中多使用真实的用户数据，采取脱敏手段来保护用户的隐私成了必要的环节。然而由于证券公司用户数据不仅量大，而且涉及到用户的敏感、保密数据信息，这些对数据脱敏来说都是一个不小的挑战。对大数据量进行脱敏还会涉及到许多编码、中文字符如何处理等问题。

本文从证券公司的数据脱敏的业务和用户需求出发，深入分析证券公司的数据脱敏的现状及问题，确定了需要实现对生产环境数据源进行有效的变形处理，屏蔽涉及客户关键信息的敏感数据，保证测试数据维护生产数据的基本属性和依赖关系；灵活的数据变形策略和变形方案，能够支持生产数据中的中文字符变形，满足中文数据的变形和处理要求。因此，本文采用了基于 IBM Optim 系统设计了数据脱敏平台，主要设计了数据脱敏可配置、任务运行、脱敏内容、数据一致性及效率的功能点；由于 Optim 的特性，脱敏内容设计有名称脱敏函数、客户号脱敏函数、电话号码脱敏函数、地址脱敏函数、电子邮箱脱敏函数及网址脱敏函数等，通过了代码的调用及引用图展示了关键技术的实现过程。文章最后采用流程测试、数据脱敏测试、安全性测试及可用性测试对系统进行测试，其中数据脱敏测试和可用性测试是功能测试的主要部分，旨在验证脱敏的内容的正确性及脱敏后系统交易的可用性，最终测试结果也证明了系统的可用性。

本系统实现了证券公司的开发测试中对于真实数据的要求，同时保证了生产数据的安全性，保护了客户的隐私，具有极高的价值与意义。

关键词：证券公司；数据脱敏；IBM Optim；

Abstract

Data desensitization is to meet the non-production environments require data, taken by some of the rules for data modification, so as to achieve the demand for privacy data protection. For securities companies, development and test environments using real multi-user data, desensitization to take measures to protect users' privacy has become a necessary part. However, since the user data is not only a large amount of securities companies, but also related to the user's sensitive and confidential data, these data are desensitization is no small challenge. Large amount of data desensitization will issue involves many aspects, such as how to better the encoded data in order to save resources, issues such as how to deal with Chinese characters, we need to choose the right tools and platform approach to solve.

In this dissertation, we analyze the current situation and problems of data desensitization in securities companies, and determine the sensitive data, which is required to realize the data source of production environment, and protect the sensitive data, which is the basic attribute and dependency relationship. Therefore, this paper uses the Optim IBM system to design the data desensitization platform, which mainly designs the function of data desensitization, task operation, desensitization, data consistency and efficiency. Because of the characteristics of Optim, we design the function of desensitization, customer number, phone number, address desensitization, electronic mail, and web site. At the end of this paper, flow testing, data desensitization, safety testing and usability testing are used to test the system, and the testing results are the main part of functional tests.

This system realizes the requirements of real data in the development and testing of securities companies, and ensures the safety of production data, protecting the privacy of customers, and has a very high value and significance.

Key Words: Securities Companies; Data Desensitization; Optim IBM;

目 录

CONTENTS

第一章 绪论

1.1 研究背景及意义

数据信息安全对于证券等金融服务行业是致关重要的，往往客户和企业的机密数据都需要证券金融服务行业重点保护，防止因数据管理不当造成泄露，无论是对企业的信誉、企业的经济收益、法律形象以及社会对企业的评价等方面都会造成极大的影响，甚至会带来社会危害。而当前的证券行业做项目中仍然需要用到生产数据，因此建立一个脱敏平台就具有极高的意义。

在一项最新公布的面向金融服务行业进行的数据调研显示，多数金融服务行业数据在开发期处于极高的风险之中，数据脱敏[1]必不可少。数据脱敏是为了满足非生产环境中对数据的使用要求，而采取的通过一些规则对数据进行变形，从而达到对隐私数据保护的需求。非生产环境是除了真实的使用场景外的其他场景，例如测试环境、培训环境、研究等等。数据脱敏要求对用户的隐私数据进行保护，包括用户的名称、联系方式、证件号以及住址等私密信息[2-3]。对此银监会先后发布了两份针对商业银行以及金融机构的风险管理文件。其中，针对信息安全做出了明确的规定。例如：针对客户信息的保密，包括信息输入/输出的保密，还包括生产系统、开发系统、测试系统的分离等。原则上禁止服务供应商进入安全区域。而生产数据是最好的数据和开发数据来源，它最能够反映系统实际情况的数据。因此如何在屏蔽开发和测试环境中，一方面能够用敏感信息进行测试，另一方面又能够保证数据具有完整性和安全性有效性，这是目前亟需解决的问题[4-6]。调查结果表明，超过百分之 30 的用户表示，如果他们在证券公司的基本信息被泄露，那么他们将会选择换一家金融机构。

报告显示，使用真实的客户数据作为开发测试，不仅会给金融机构带来违章的可能，而且会导致客户资源的严重流失。因此，报告中给出了一些能有效降低这两个风险的手段和指导性原则，研究发现：第一，多数金融机构对敏感数据都没有进行有效的隐私处理，而是对一些敏感数据进行曝光，其中百分之 80 以上的金融机构都在使用真实的用户数据进行开发测试，其中有百分之 70 的数据是消费者数据；第二，金融机构没有采取有效的敏感数据保护措施，对于使用真实数据的开发和测试过程，没有对数据进行加密或者保护；第三，超过一半的金融

机构存在违规，在实际的研发环境中使用用户的实际数据，而许多机构都不清楚是否有违规；第四，出现违规的机构要不就是运营中断，要不就是客户流失，损失代价很大；第五，多数金融机构在数据流失的情况下仍旧一头雾水，不确定数据是如何泄露的，也不知道是否真的存在泄露问题；第六，在实际的开发过程中，金融机构多采用外包或者云计算的形式进行开发，无形中增加了数据丢失的概率。与外包公司进行合作过程中，金融机构的真实数据往往会提供给开发公司。而使用云换进资源时，四分之一的机构都相信该环境是安全可信的。

针对目前金融机构在真实开发和测试环境中多使用真实的用户数据的情况，Ponemon 机构给出了一下两点建议：

1.保护客户的隐私，制定确定的规章制度和流程，建立单一的职责管理机制；

2.购买一些生产工具技术，能够有效地保证真实数据的安全性，提高保密程度的同时又能有效进行开发测试。

虽然数据脱敏技术能够有效地对真实数据中的敏感数据进行保密，例如支付卡行业中对支付者个人信息的保护等，但是目前的数据脱敏技术仍然存在以下问题：

数据脱敏能够有效地处理小规模数据，但是对于大型企业而言，数据量庞大，数据脱敏的效率便有所下降[7-8]。比如说，大型数据存在多个数据分块，数据块之间存在调度与被调用的关系，脱敏技术需要保证脱敏后数据之间不再具有关联，且仍然能够实现原有的调度处理任务。另外，在数据脱敏中会遇到不同类型的数据，需要对数据进行编码以及中文字符等各种问题，因此，在选择数据脱敏技术的时候，需要设计一种通用的脱敏技术。对于大规模的数据而言，设计这种技术便大大增加了复杂程度。因此，对于不同类型的需求目标，需要采取不同的脱敏方案，从而实现具体的目标。

数据脱敏就是通过变形屏蔽数据中的敏感信息，便于数据在非生产环境中使用，减少数据使用中的限制。变形后的数据主要用于：测试，开发，培训，外包，数据挖掘/研究等。不同于数据加密，数据脱敏是不可逆的过程，并且需要要保持数据的完整性[17]。数据脱敏是为了满足非生产环境中对数据的使用要求，而采取的通过一些规则对数据进行变形，从而达到对隐私数据保护的需求。在系统测

试中，真实生产环境中获得的数据更能够反映系统所面临的真实问题；而数据挖掘的目的是从数据中发现数据的规律。非生产环境是除了真实的使用场景外的其他场景，例如测试环境、培训环境、研究等等。数据脱敏要求对用户的隐私数据进行保护，包括用户的名称、联系方式、证件号以及住址等私密信息以及证券公司柜员的基本信息等。数据漂白（data Masking）[18]说的也是数据脱敏，数据脱敏一般是指对敏感信息的保护，数据漂白的范围不限于敏感信息，可能包括数据的分布也会予以改变。数据变形说的是数据脱敏或数据漂白的具体方法。例如：将“A”变成“B”，即通过数据变形，实现了数据脱敏。

目前采用的数据脱敏方式是半手工方式，即技术人员编写简单的数据操作程序直接对数据库操作的方式实现数据变换，例如：针对用户密码进行变换。可能使数据的一致性、完整性遭到破坏，导致测试的软件系统某些功能无法实现。例如：身份证号如果替换成一个固定值，则通过身份证号连接的功能就无法实现。一旦规模扩大，问题会变得复杂。例如：数据量大，需要对多个系统的数据进行脱敏，并且保持数据一致性。例如：针对不同系统的姓名，还要求变换成姓名。对于半手工操作的技术人员，无论是效率还是质量都是一个挑战。可能会破坏数据的可用性。例如：把姓名替换成一个固定值，测试人员看起来就很不舒服，想找一个测试过的案例，无法通过最容易记忆的方式获得。同时受限于技术手段，脱敏不彻底。未经彻底脱敏的数据，依然存在泄密的风险[9]。

1.2 国内外研究现状

目前国内外已经存在部分的数据脱敏平台，例如国内的华瑞数据脱敏平台，结合业务对象知识库针对各业务系统，能有效地混淆、加密或屏蔽测试数据库中的敏感数据，并能确保用于测试的数据格式以及业务关联有效性。华瑞数据脱敏平台预装了丰富的脱敏算法来处理测试环境中的敏感信息，同时确保个人敏感信息的有效性[10]。同时基于完整性的业务对象进行脱敏操作，确保不破坏数据的二义性以及业务关联性；内置多种脱敏算法包括替换、加密和解密、随机、模糊、数据格式化、用户自定义算法等；同时支持抽取式脱敏与就地脱敏 2 种模式：业务唯一一款同时支持抽取式不落地脱敏以及就地脱敏 2 种模式的系统；通过调度器自动执行测试数据抽取以及脱敏工作，减少人工干扰；通过多任务、多线程、

分批处理等技术实现脱敏的最高性能；具备完善的用户权限管理策略，可以针对不同角色、不同用户、不同系统进行权限设置。而博尔信公司支持多种数据脱敏转换规则以及规则的组合，同时提供了脱敏转换的预览功能，从而可以帮助用户更好的创建和测试转换规则。系统设计了一套标准化的数据保护流程，从数据的申请到审核，再到数据规则的设计与审核等，最后实现数据的监控，不仅规范了数据的应用，而且为用户提供了方便的管理，能够实时对数据进行监控。

Optim 本身提供了一系列数据变形函数，可以通过这些变形函数对数据进行脱敏。例如：电子邮箱，可以通过 Optim 自带函数实现的脱敏。例如，LOOKUP 方法是先建立一张表，里面放置的事先做好的一些与敏感信息无关的数据。例如：“假地址”。然后通过 LOOK_UP 随机替换的方式，用表中的数据替换掉目标表中的相应数据 ，从而实现数据脱敏。自带的 BASIC 方法，可以对电话号码，包括电话、传真、手机实现脱敏。而姓名、身份证号码的脱敏，采用的是出口函数，可通过 C 语言编制一个出口函数，嵌入到 Optim 中执行，实行数据的变换。

在这些应用平台上，大连银行通过采用“Informatica 数据脱敏”解决方案，帮助其管理对最敏感数据的访问，建立了企业内部完善统一的脱敏机制与管理流程[11]。Informatica Persistent Data Masking 对数据创建了内外部的安全共享机制，不仅能够保证数据的真实性，而且使得外部无法识别数据的归属，能够有效地防止数据的意外泄露。而农行在应用平台的同时，还进行了特殊的脱敏算法设计；例如在客户姓名脱敏上，建立一个人为构建的百万级的中文名字码表，然后将原来的名字进行哈希映射替换。这种方法好处是客户的多样性和分布性得以保留，但是在处理上需要增加系统的开销和处理时间，而且构建的客户数量是有限的，依然无法做到真正的分布特征的自由维度保留[12]。

1.3 研究内容

本文根据国内外的脱敏需求的发展趋势及不同平台的优劣性，设计了基于 IBM Optim 的脱敏平台，部署于证券公司内部网中。因此设计的功能内容主要有：

1、该脱敏平台能够支持从主流的操作系统，包括商业的和开放的进行数据抽取、数据脱敏和数据加载。主流的操作系统例如：OS400，SUN SOLARIS，HP-UX, IBMAIX, LINUX， WINDOWS 等；

2、针对目前的主流数据库，具有数据抽取、数据脱敏、数据装载能力，包括 ORACLE，DB2, INFORMIX，SYBASE，SQLSERVER，TERADATA 等各种主流商业数据库。

3、数据脱敏的数据抽取、数据漂白、数据加载、任务定制等需能够实现可配置功能。

4、任务的执行可以通过人工发起执行或定时发起执行，实现自动化的脱敏处理过程。

5、针对证券公司的需求设计脱敏函数，通过脱敏的算法，保持系统间的数据一致性。

本文从证券业的数据使用管理入手，在充分了解数据应用上的问题，进行了数据脱敏的需求分析，进而进行系统设计与实现，最终进行系统测试。

1.4 结构安排

本研究的主要内容包括以下 6 章：

第一章绪论，开展整个论文研究课题内容的相关信息阐述，包括背景，意义，研究现状，技术路线，以及研究内容等等。

第二章数据脱敏平台的需求分析，针对数据脱敏平台的业务需求内容，分析与确定支持的操作系统、数据库、数据脱敏配置、任务运行及脱敏内容的功能节点需求内容，以及非功能性需求内容和业务流程内容。

第三章数据脱敏平台的设计，研究分析当前需求设计的原则性内容，设计了系统的物理结构、逻辑结构及数据流程，进行了数据脱敏可配置设计、任务运行设计、脱敏内容设计、数据一致性设计及针对效率的设计。

第四章数据脱敏平台实现，针对于系统的实现与代码运行需求进行展示，完成脱敏函数的实现过程及 Optim 的完整流程过程。

第五章数据脱敏平台测试，采纳软件设计与软件流程测试理论基础，实现流程测试、数据脱敏测试、可用性测试及安全性测试。

第六章是总结及展望，对此次开发工作进行了总结，客观评价了系统的优点及不足，并对系统将来的发展方向提出了展望。