一种基于代理重签名的卷烟营销数据细粒度血缘安全分析方法

摘　要：卷烟营销系统非常庞大且复杂，数据来源众多，不仅仅来自于卷烟营销系统内部，也有可能来自第三方应用系统。而且即使同一数据集的数据也有可能来自不同的数据源，这些数据特点给卷烟营销数据的血缘分析带来了巨大挑战。设计了一种基于代理重签名机制的细粒度卷烟营销数据血缘安全分析方法，与现有数据血缘分析方法相比，不仅仅支持数据的高效细粒度血缘分析，也可以确保在外包云数据中心的服务场景下，数据血缘关系的不可篡改性和可验证性。分析了设计的血缘标签数据的安全性，同时基于卷烟营销数据集进行了实验，实验结果表明提出的数据血缘分析方法可以满足细粒度血缘分析需求，查询效率也优于现有数据血缘分析方法。论文提出的基于代理重签名的数据血缘安全分析方法，可以满足复杂卷烟营销数据的血缘分析需求，提供云数据中心环境下卷烟营销数据的安全、高效的细粒度数据血缘分析服务。

关键词：血缘分析；代理重签名；卷烟营销数据

开放科学（资源服务）标识码（OSID）：

DOI：

中图分类号：TP3　　文献标识码：A　　文章编号：1672—7800（2021）001-0001-06

A Fine-grained Bloodline Analysis Method of Cigarette Marketing Data based on Proxy Re-signature Protocol

A**bstract：**The cigarette marketing system usually is very large and complex. And the cigarette marketing data in this system are from many sources which include not only the cigarette marketing system itself but also the third-party systems. Even the data in the same dataset maybe come from different data sources. These data characters always bring great challenge to the bloodline analysis of cigarette marketing data. In this work, we design a fine-grained bloodline analysis method of cigarette marketing data based on the proxy re-signature protocol. Comparing with the existing solutions of data bloodline analysis, the proposed method can support fine-grained bloodline analysis and ensure the data bloodline is not tampered in the cloud outsourcing service scenario. We theoretical analyze the security of our method and carry out experiments to evaluate the performance of the proposed method. The proposed method satisfies all the requirements of bloodline analysis of cigarette marketing data, it provides efficient, secure and fine-grained bloodline analysis services for cigarette marketing data under the scenario of cloud data center.

Key words: bloodline analysis; proxy re-signature; cigarette marketing data

0　引言

随着大数据时代的来临，数据作为重要的企业资产越来越受到重视，面对纷繁复杂的企业数据，如何挖掘和利用其中的信息和知识成为企业大数据分析的关键。数据血缘[1]描述了数据从产生，并随时间推移而演变的整个过程，数据血缘分析的应用领域非常广，包

括数据质量评价、数据核查、数据恢复和数据引用等。本论文以卷烟营销系统为研究对象，研究卷烟营销系统中数据血缘分析问题。

卷烟营销系统庞大且复杂，主要可以分为计划管理、需求预测、货源组织、货源供应、订单管理、客户服务、品牌管理、市场监测、网建管理、综合管理等十大业务环节，各个业务环节之间有着复杂的数据引用关联关系。涉及的主要业务平台系统也包括：省级营销平台、一体化服务平台、市综合业务平台等。营销系统中数据有可能来自于其他的多个数据源，甚至是来自外部的第三方应用。同时即使是同一数据集的数据也有可能来自不同的数据源，这些数据关系的复杂性和不确定性都给卷烟营销数据的血缘分析带来了巨大挑战。因此，当前我国省级卷烟营销系统急需一种高效、支持细粒度的数据血缘追溯分析解决方案。

云计算具有动态可扩展、按需付费、集中管理和算力强大等优点。因此对于企业来说，将传统业务迁移到云端已经成为了一种必然的技术选择。现有云架构大致可以分为公有云、私有云和混合云架构。目前公有云发展较快，但是公有云属于托管性质，云租户缺乏对物理设备的控制权，因此，很多企业和机构出于安全性考虑而选择部署私有云数据中心的方式来集成管理企业内部信息系统。企业选择部署私有云是由于私有云具有更高的安全性和可控性，但是实际运营过程中，私有云数据中心也面临着安全隐患。现有省级卷烟营销平台往往采用混合云的架构，在外包云数据中心环境下如何保障数据血缘分析的安全性和高效性是卷烟营销数据血缘分析面临的又一巨大挑战。

1 相关工作

数据血缘技术对多源数据集成、演化过程进行分析、研究，获取原始数据到目标数据的具体生成、转换流程。数据血缘包括静态的源数据信息和动态的数据演化过程。

针对数据血缘分析，Cui 等人从数据库关系运算符操作出发，定义了数据血缘的具体流程[2]，例如 SPJ 段是通过查询、投影和选择操作构成的标准形式查询。在后续研究[3-4]中，Cui 等人数据血缘分析的定义，构建了数据血缘分析的完整体系，从操作对应的元组起源，数据起源追踪查询，数据分割操作，视图元组起源到数据集合起源等，并给出了数据血缘查询的具体系统实现。Buneman 等人利用辅助数据库对数据血缘信息进行管理[6]，根据使用者对辅助数据库的操作来追溯使用者的操作。Ruan 等人针对区块链设计了一种细粒度、安全、高效的区块链数据溯源系统 LineageChain[9]。朱运磊等人利用布尔公式、逻辑蕴含和图模型的性质及等价转换机制，将不确定性数据的世系表达式等价地转换为贝叶斯网络，并基于贝叶斯网络的概率推理回答查询[10]。Porkodi 等人设计了一种基于混合属性加密的数据溯源方法[11]，该溯源方法可以实现基于区块链针对物联网数据流的高效、安全溯源操作。Simon 等人针对数据存储优化场景，提出了一种基于图数据库的医疗应用数据溯源方法[12]，保障在云平台环境下医疗数据的高效溯源。Marchetti 等人提出了一种针对网络中数据泄露的溯源追踪方法[22]，利用有向无环图及 K-means 聚类等方法对网络流量监控识别，从而确定大型网络中恶意活动的特定主机。Priebe 等人利用水印技术添加加密安全标签和安装监视器的方式使得云租户能够实时监控其数据流[14]。Bertino 等人将数据溯源技术、机密访问

控制以及可信计算相结合，提出了一种构建安全数据来源路线链图的方式[15]，确保不影响人员隐私的情况下，实现数据来源的高度保证。

以上方法都不能兼顾在外包云数据服务中心场景下数据血缘分析的高效性和安全性，近年来大数据技术普遍应用于我国烟草数据平台和信息系统建设[7-8]，我国卷烟营销系统大多基于云服务平台建设，但目前缺乏有效的烟草营销数据血缘分析解决方案，本文借鉴代理重加密方法，研究云环境下的卷烟营销数据血缘分析方法，实现了混合云模式下安全、高效的细粒度数据血缘分析服务。

2 基于代理重签名的卷烟营销数据细粒度血缘分析方法

为了实现数据血缘的细粒度、高效查询，设计基于代理重签名的卷烟营销数据血缘分析方法，首先介绍用到的密码学工具，包括代理重签名算法和双线性映射，然后介绍卷烟营销数据，最后介绍卷烟营销数据的血缘查询分析方法。

2.1 代理重签名算法

代理重加密[5]由一组多项式时间算法：**KeyGen**，**ReKey**，**Sign**，**ReSign**，**Verify** 组成。代理重加密算法允许一个半可信的机构（云数据中心）将用户 Alice 的加密数据安全地转换成另一个用户 Bob 的加密数据。对于安全的代理重加密算法来说，半可信机构不能通过代理重加密算法本身或分析存储的加密数据、收到的通信消息获取任何参与者（Alice，Bob）的信息。论文中假设云数据中心是半可信的，即云数据中心会正确执行协议，但是云数据中心存在猜测平台用户秘密信息的可能。

2.2 双线性映射

设 G_1和 G_2是两个阶为素数 p 的循环群，g 是 G_1的一个生成元，若映射 e: $G_1\times G_1\rightarrow G_2$为一个双线性映射，则映射 e 满足以下条件：

（1）双线性：$\forall a,b\in Z_p^*$，满足 $e(g^a, g^b)=e(g, g)^{ab}$。

（2）非退化性：$e(g, g)\neq 1\in G_2$，其中1代表 G_2群的单位元。

（3）可计算性：$\forall g_1, g_2\in G_1$，存在一个有效的算法，可以在多项式时间内计算 $e(g_1, g_2)$。

2.3 卷烟营销数据

卷烟营销系统数据复杂，有可能来自烟草公司的基础数据，或者来自卷烟营销系统内部。为了进行卷烟营销数据血缘分析，将卷烟营销系统平台数据分成三类：B 表、K 表、R 表。

B 表为基础表，接口表入库后被命名为 B 表。B 表来源于源端业务系统，即基础业务层的原始粒度数据。

K 表为加工过程表，其数据由 B 表加工生成，可能来自一个或多个 B 表。B 表加工后形成共享程度高，业务含义也丰富的 K 表。

R 表为业务指标表，其数据有 B 表和 K 表数据加工生成，将数据按业务单元、分析主题进行加工整合，用于对外提供数据服务。

卷烟营销数据中心内 B 表、K 表、R 表须严格遵守数据分层存储、层级间加工转换的规则：

（1）K 层数据表可由 B 层、K 层数据表加工而成；

（2）R 层数据表可由 B 层、K 层数据表加工而成；

（3）R 层数据表不能生成 R 层数据表。

卷烟营销数据表分层存储样例如图1所示，卷烟营销数据平台中每个数据表拥有唯一表标识（如 B_id, K_id, R_id）。

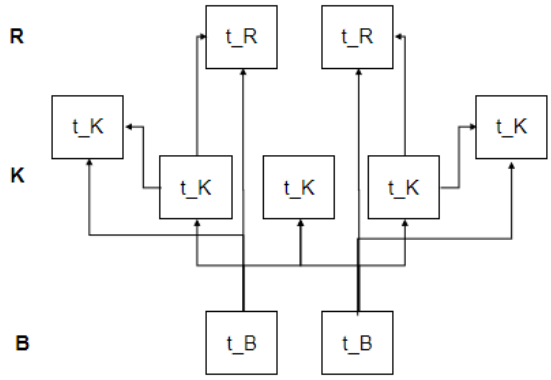

图1 卷烟营销数据表分层存储示例

卷烟营销数据在源端业务系统产生后，按业务域和业务环节采集进入数据中心。数据中心对数据进行清洗转换、汇聚加工、分级存储。源层数据清洗转换后形成业务数据基础表（B 表层）和主数据表。基础表（B 层）数据经过加工汇聚，形成加工过程表（K 层）和业务指标表（R 层）。建立数据服务目录，对外提供数据服务，支撑源端业务系统运行及各类数据服务应用。烟草营销数据总体分布存储及处理框架如图2。

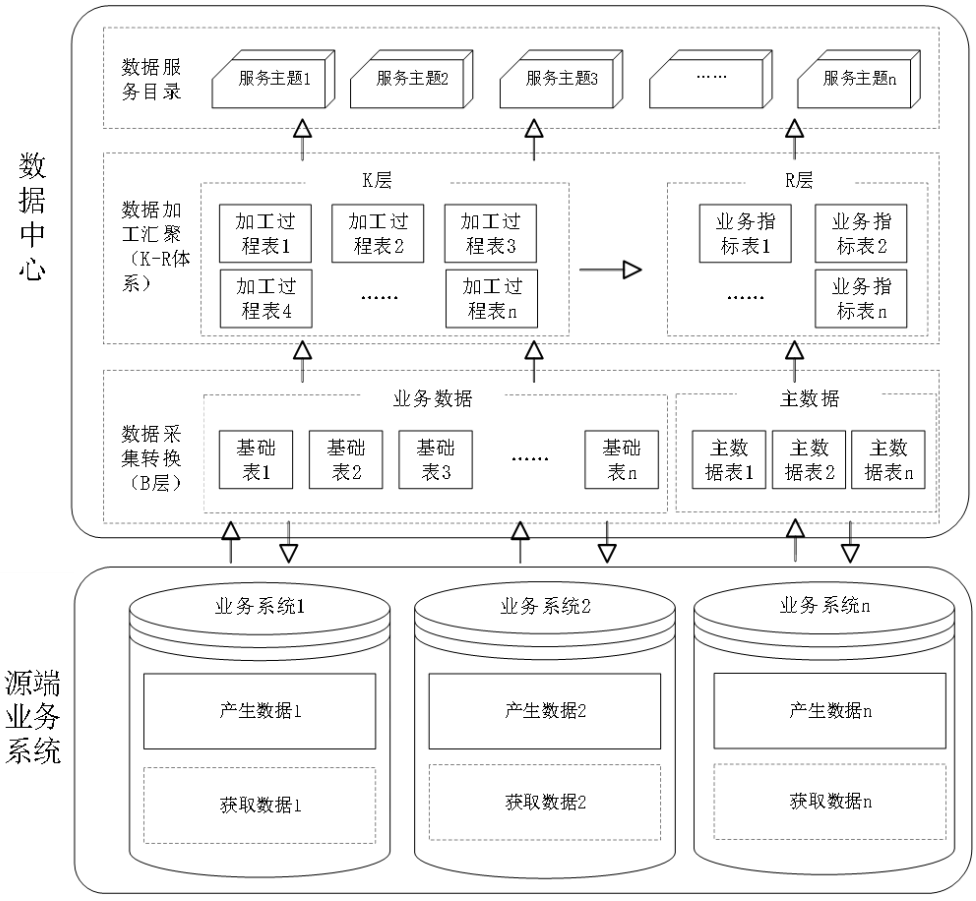

图2 烟草营销数据总体分布框架

卷烟营销数据安全、高效血缘分析针对 K 表和 R 表，可以查询数据来源，演化途径，针对卷烟营销数据实际情况，即使同一个 K 表或 R 表中数据也可能来自不同数据源（B 表或 K 表），论文设计细粒度的血缘分析机制，可以针对元组进行细粒度的数据血缘追溯。考虑到云数据中心并不完全可信，数据库中血缘溯源数据具有不可抵赖性，可以有效抵抗来自云服务器段的伪造和替换攻击等攻击方式。

2.4 细粒度卷烟营销数据溯源标签

卷烟营销数据平台系统模型有平台管理中心、平台用户（平台应用子系统）、云数据中心三部分组成，其中平台管理中心主要负责对平台用户身份管理，并为平台用户生成密钥信息，平台管理中心不需要强大的计算能力和存储能力，构建在烟草公司私有云环境下，可以认为是完全可信的。云数据中心负责存储管理卷烟营销数据平台中的全部数据，构建在公有云环境被认为是半可信的，有可能受到外部攻击者的攻击，也有可能处于好奇的原因而探测平台数据。平台用户或平台应用子系统为卷烟营销平台的授权用户，拥有访问、修改平台数据的权限，同时可以查询平台营销数据的血缘信息。卷烟营销数据平台系统架构如图3所示。

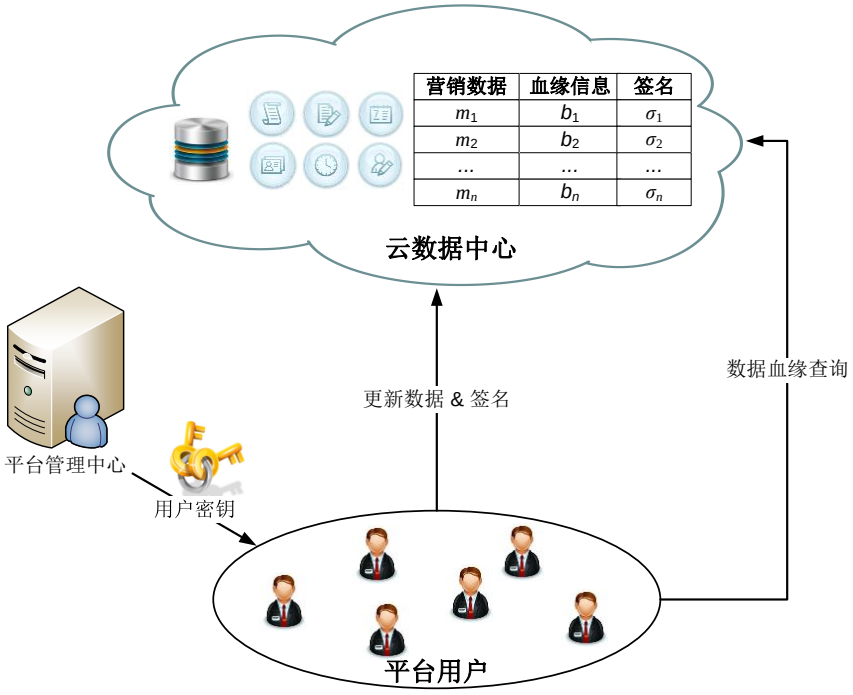

图3 卷烟营销数据平台系统模型

卷烟营销数据平台中，用户更新营销数据时，写入数据血缘信息，并对数据血缘信息进行签名。基于卷烟营销数据分层存储结构，本系统中平台用户只能更新 K 表和 R 表数据，B 表数据作为基础数据不能更新。平台设计代理重加密算法$\prod$={**Setup**, **KeyGen**, **Sign**, **ReKey**, **ReSign**, **Verify**}实现平台营销数据血缘分析管理。

Setup(1^{λ})→(e, g, G_1, G_2, H)为系统初始化函数，生成全局安全参数，由平台管理中心调用，算法以系统安全参数 λ 作为输入，输出系统全局安全参数(e, g, h, G_1, G_2, H)，其中 e: $G_1 \times G_1 \rightarrow G_2$为双线性映射，$g$, h（$g \neq h$）为 G_1的生成元，H:$\{0,1\}^{*} \rightarrow Z_p$ 为安全哈希函数。

KeyGen()→(pk_i, sk_i)由平台管理中心调用，算法由系统安全参数 λ 作为输入，管理中心选择随机数 $sk_i=\alpha\in Z_p$ 作为用户 u_i 的私钥，$pk_i = g^{sk_i} = g^{\alpha}$ 作为 u_i 的公钥，并存储用户公钥 pk_i 和 $\frac{1}{sk_i}$。

Sign(t_id, b, sk)→σ 由平台用户调用，为元组的血缘数据生成签名。系统中为了实现细粒度的数据血缘追溯，每个元组都由最后的修改者根据数据血缘信息进行签名。卷烟营销系统中 B 表为基础表，其数据来源于源端业务系统，可以认为在卷烟营销业务中是未经加工数据，因此 B 表数据并不进行数据血缘签名。平台用户更新 K 表或 R 表中数据时，对数据血缘信息进行签名。用户 u（密钥为 sk）修改 K 表数据项 t_id（$t\_id\in Z_p$ 是元组的唯一标识）时，t_id 元组数据源来自 B 表集合{B_id1, B_id2,…, B_idn}，而当用户 u 修改 R 表数据项 t_id 时，t_id 元组数据源来自 B 表集合{B_id1, B_id2,…, B_idn}和 K 表集合{K_$id1$, K_$id2$,…, K_idm}，u 对 t_id 元组血缘数据签名如下：

（1）u 选择随机数 $k\in Z_p$，并计算 $r=h^k$；

（2）u 计算 H(B_id1||B_id2||…||B_idn||K_$id1$||K_$id2$||…||K_idm||r)，输出 $s=sk\times$(H(B_id1||B_id2||…||B_idn||K_$id1$||K_$id2$||…||K_idm||r)+$k\times t\_id$) **mod** p

（3）u 输出签名 σ 如公式（1）：

$$\sigma=(r,\ s)=(h^k,\ sk\times(\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)+k\times t\_id)\ \textbf{mod}\ p) \quad (1)$$

ReKey(pk, sk')→$rk_{u\to u'}$由平台管理中心和平台用户 u'调用，算法输入用户 u 的公钥 pk 以及用户 u' 的私钥 sk'，输出代理重加密密钥 $rk_{u\to u'}$。

ReSign(σ, $rk_{u\to u'}$)→σ'由云数据中心执行，算法输入血缘数据签名 σ（由数据的最后访问者 u 签名），由用户 u 到 u'的代理重加密密钥 $rk_{u\to u'}$（u'为数据血缘关系的验证者）。

Verify(t_id, b_id, σ)→{1, ⊥}为验证算法，由平台用户 u 执行，当血缘数据签名正确时输出1，否则输出⊥。

将在基于代理重签名的卷烟营销数据血缘分析算法中详细介绍 **ReKey**，**ReSign** 和 **Verify** 等算法的实现细节。

2.5 基于代理重签名的卷烟营销数据血缘分析算法

卷烟营销数据平台中，平台用户 u 修改 K 表或 R 表数据时，根据新数据来源记录数据血缘信息，并调用 Sign 签名算法对数据血缘信息进行签名，上传存储在云数据中心，表结构如图4所示。

K table t_K_1 bloodline data

tuple_id	bloodline info of B table	signature
t_1	t_B_1, t_B_3	σ_1
t_2	t_B_3	σ_2
t_3	t_B_1, t_B_2	σ_3
…	…	…

R table t_R_1 bloodline data

tuple_id	bloodline info of B table	bloodline info of K table	signature
t_1	t_B_2	t_K_2, t_K_5	σ_1
t_2	t_B_3	t_K_1, t_K_4	σ_2
t_3	t_B_4, t_B_7	t_K_1	σ_3
…	…	…	…

图4 卷烟营销数据表血缘信息存储结构

卷烟营销平台中所有的平台用户可以查询 K 表和 R 表中元组的数据血缘信息，如 K 表 t_K_1的元组 t_1中的数据来源于 B 表 t_B_1和 t_B_3，而 R 表 t_R_1的元组 t_3中的数据来源于 B 表 t_B_4和 K 表 t_K_1。用户 u 利用个人私钥 sk 对元组血缘数据进行签名 $\sigma(r, s)$如公式1所示。

基于云数据中心的安全性需求，设计了两种数据血缘查询验证方法：1）修改者身份公开血缘查询机制；2）修改者隐私保护血缘查询机制。在修改者身份公开血缘查询模式下，数据的最终修改者身份信息公开，血缘查询者可以查询数据血缘信息，并基于血缘签名利用数据修改者的公钥验证数据血缘信息。而考虑到卷烟数据平台安全性需求，部分数据的修改者信息不能公开，则血缘查询者不能利用修改者公钥对数据血缘进行验证，论文利用代理重加密机制实现修改者身份信息隐私保护前提下的数据血缘高效查询及安全验证。

2.5.1 修改者公开血缘查询机制

在修改者公开血缘查询机制下，平台用户 u 访问卷烟营销平台中 K 表或 R 表数据元组 t_id，u 可以知道 t_id 的最终修改用户 u'（公钥为 $pk'=g^{sk'}$），以及 u'记录的元组血缘信息。u 可以基于 pk'调用 **Verify** 算法对元组血缘数据进行查询验证，过程如下：

（1）u 访问元组 t_id，获取元组的数据血缘信息$\{B_id1, B_id2, \ldots, B_idn, K_id1, K_id2, \ldots, K_idm\}$；

（2）u 访问平台管理中心，获取元组 t_id 修改者 u'的公钥 $pk'=g^{sk'}$；

（3）u 选择随机数 $\lambda \in Z_p$，并发送挑战信息$\{t_id, \lambda\}$给云数据中心；

（4）云数据中心计算 $h^{\lambda s}$，r^{t_id}，生成验证消息$\{r, h^{\lambda s}, r^{t_id}\}$并返回给 u；

（5）u 计算 $H(B\_id1||B\_id2||\ldots||B\_idn||K\_id1||K\_id2||\ldots||K\_idm||r)$ **mod** p，并验证 t_id 的数据血缘信息如式（2）所示：

$$e(g, h^{\lambda s})=e(pk'^{\lambda}, h^{H(B\_id1||B\_id2||\ldots||B\_idn||K\_id1||K\_id2||\ldots||K\_idm||r)\times r^{t\_id}}) \quad (2)$$

基于双线性映射性质，式（2）的正确性可以验证如下：

$$e(pk'^{\lambda}, h^{H(B\_id1||B\_id2||\ldots||B\_idn||K\_id1||K\_id2||\ldots||K\_idm||r)\times r^{t\_id}})=e(g^{sk'\lambda}, h^{H(B\_id1||B\_id2||\ldots||B\_idn||K\_id1||K\_id2||\ldots||K\_idm||r)\times r^{t\_id}})$$

$$=e(g, h^{sk'\lambda(H(B\_id1||B\_id2||\ldots||B\_idn||K\_id1||K\_id2||\ldots||K\_idm||r)\times r^{t\_id})}) =e(g, h^{\lambda s}) \quad (3)$$

修改者公开血缘查询模式中，用户无需下载血缘签名即可实现血缘数据完整性验证，可以实现云环境下卷烟营销数据安全、高效血缘查询。

2.5.2 修改者隐私保护血缘查询机制

修改者公开血缘查询模式中，用户可以高效的实现数据血缘数据查询及验证，但是在卷烟营销平台中很多数据并不能公开修改者信息，用户无法利用签名者公钥进行验证。基于代理重签名机制，设计修改者隐私保护的数据血缘查询服务。

当用户 u 在修改者用户隐私保护模式下查询数据元组 t_id 的数据血缘信息时，验证过程如下所示：

（1）用户 u（私钥为 sk）选择随机数 $\beta \in Z_p$ 并将 $\beta \times sk$ 发送给平台管理中心；

（2）平台管理中心选取数据更新用户的密钥信息 $\frac{1}{sk'}$（sk'为更新用户的私钥），生成 $\frac{\beta \times sk}{sk'}$ 并返回用户 u；

（3）u 选择随机数 $\lambda \in Z_p$，并发送挑战信息$\{t_id, \lambda, \frac{\beta \times sk}{sk'}\}$给云数据中心；

（4）云计算中心针对 t_id 的数据签名 $\sigma(r, s)$计算

$$s=s^{\wedge}\frac{\beta\times sk}{sk'}=\beta\times sk\times(\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)+k\times t\_id)\ \textbf{mod}\ p \quad (4)$$

（5）云数据中心计算 $h^{\lambda s}$，r^{t_id}，生成验证消息$\{r, h^{\lambda s}, r^{t_id}\}$并返回给 u；

（6）u 计算 $\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)$ **mod** p，并验证 t_id 的数据血缘信息如式（5）所示：

$$e(g, h^{\lambda s})=e(pk^{\lambda\beta}, h^{\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)}\times r^{t\_id}) \quad (5)$$

基于双线性映射性质，式（5）的正确性可验证如下：

$$e(pk^{\lambda\beta}, h^{\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)}\times r^{t\_id})=e(g^{sk\lambda\beta}, h^{\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)}\times r^{t\_id})$$

$$=e(g, h^{\lambda\beta sk\ (\mathrm{H}(B\_id1\|B\_id2\|\ldots\|\mathrm{B}\_idn\|\mathrm{K}\_id1\|\mathrm{K}\_id2\|\ldots\|\mathrm{K}\_idm\|r)}\times r^{t\_id})}=e(g, h^{\lambda s}) \quad (6)$$

2.6 安全性分析

假设攻击者 $\mathscr{A}$ 修改元组 t_id 的血缘信息$\{B_id1, B_id2, \ldots, \mathrm{B}_idn, \mathrm{K}_id1, \mathrm{K}_id2, \ldots, \mathrm{K}_idm\}$，显然基于公式（2），（5）平台用户可以发现这一伪造事实，进一步假设攻击者 $\mathscr{A}$ 可以伪造签名 σ’，并通过后续验证。那么攻击者 $\mathscr{A}$ 可以找到哈希函数 H(·)的一个有效碰撞，假设 H(·)的输出宽度为 l（$l\geq 64$），那么 $\mathscr{A}$ 找到有效碰撞的概率不大于 $\frac{1}{2^l}=2^{-64}=5.42\times10^{-20}$，显然这个可能性是可忽略的，因此攻击者 $\mathscr{A}$ 无法伪造签名，破坏系统安全性。

3 应用实验

将论文设计的数据血缘安全分析方法用于烟草营销系统中的数据血缘分析，设计实验分析算法的查询效率和网络消息量，并将论文提出方法的查询效率和 Hybrid Attribute[11]，CloudSafetyNet[14]方法进行比较。三种方法都可以对血缘数据进行安全性验证，因此也设计实验对比三种方法的血缘数据验证效率。论文实验数据选择来自营销系统的20个表，分别隶属于卷烟营销和物流业务域，包括10个 B 表，7个 K 表和3个 R 表。

3.1 数据血缘查询效率

设计实验统计三种方法查询不同规模数据血缘（50-300个元组）的时间开销，实验结果如图5所示。

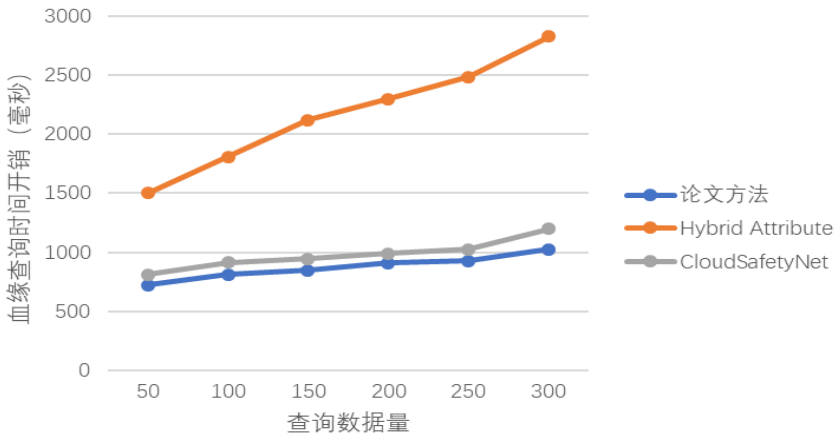

图5 血缘查询效率实验

实验结果显示论文方法的数据血缘查询效率优于 Hybrid Attribute[11]和 CloudSafetyNet[14]两种对比方法。实验中选择查询50-300个元组数据的血缘信息，论文方法的查询效率从723毫秒到1023毫秒，可以有效支持卷烟营销平台系统对于数据血缘查询的效率需求。

3.2 数据血缘验证效率

为确保云平台环境下血缘数据的安全性和可验证性，论文设计数据签名机制保障血缘数据的不可抵赖性和防篡改特性。针对卷烟营销平台的数据安全需求，论文分别设计了“修改者公开血缘查询机制”和“修改者隐私保护血缘查询机制”两种服务模式，设计模拟实验统计不同规模查询数据元组量条件下血缘数据验证效率，实验统计结果如图6所示。

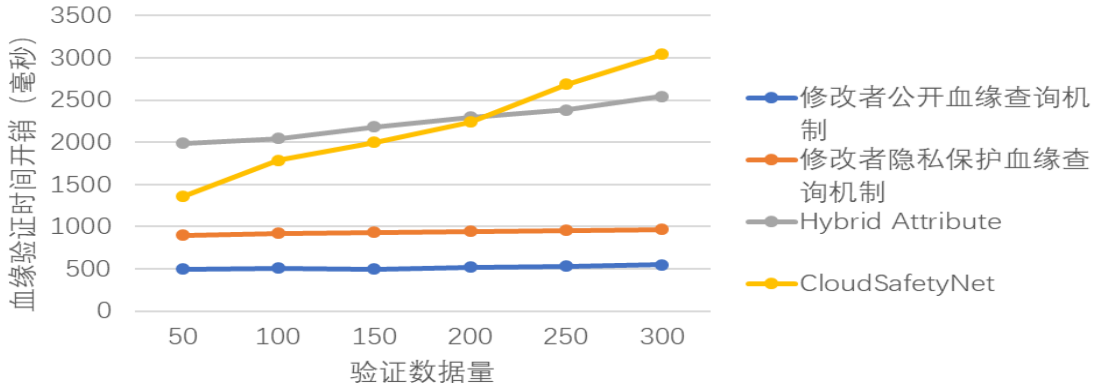

图6 血缘验证效率实验

论文设计方法大量的数据计算操作由云平台完成，充分发挥了云服务外包模式的优势，实验结果显示论文方法的数据血缘验证效率优于 Hybrid Attribute[11]和 CloudSafetyNet[14]两种对比方法。实验中选择查询验证50-300个元组数据的血缘信息，“修改者公开血缘查询机制”验证时间开销从498毫秒到547毫秒，“修改者隐私保护血缘查询机制”验证时间开销从897毫秒到970毫秒，可以满足卷烟营销平台系统对于数据血缘验证的效率需求。

3.3 网络消息量

网络通讯开销也是卷烟营销平台需要重点关注的，设计仿真实验统计在“修改者公开血缘查询机制”和“修改者隐私保护血缘查询机制”两种服务模式下的网络流量开销，实验结果如表1所示。

Table 1 Network traffic overheads

表1 网络流量开销

查询验证元组规模	修改者公开血缘查询机制（KB）	修改者隐私保护血缘查询机制（KB）
50	3.91	5.47
100	7.81	10.94
150	11.72	16.41
200	15.63	21.88
250	19.53	27.34
300	23.44	32.81

如图1所示，在修改者公开血缘查询机制下，网络流量开销从3.91KB 到23.44KB，在修改者隐私保护血缘机制下，网络流量开销从5.47KB 到32.81KB。通过实验结果可以显示论文提出方法的网络流量开销较小不会对卷烟营销平台带来性能影响。

4 结语

本文设计并实现了一种外包云数据中心环境下的卷烟营销数据血缘安全分析方法。实现卷烟营销数据修改者隐私保护前提下的安全、高效数据血缘分析。与现有方法相比，论文方法在查询效率，网络开销方面都具有明显优势。论文血缘分析算法应用于卷烟营销系统，实现卷烟营销数据的细粒度血缘追溯查询并保障卷烟营销数据在外包云服务中心的数据安全，是大型卷烟营销系统数据血缘分析的理想解决方案。

参考文献：

[1] GAO M, JIN C Q, WANG X L, TIAN X X, ZHOU A Y. A survey on management of data provenance[J]. Chinese Journal of Computers, 2010, 33(3): 373-389.

高明，金澈清，王晓玲，田秀霞，周傲英. 数据世系管理技术研究综述[J]. 计算机学报, 2010, 33(3): 373-389.

[2] CUI Y W, WIDOM J. Lineage tracing in a data warehousing system[C]. International Conference on Data Engineering, 2000.

[3] CUI Y W, WIDOM J. Tracing the lineage of view data in a warehouse environment[J]. ACM Transactions on Database Systems, 2000, 179-227.

[4] CUI Y W, WIDOM J. Lineage tracing for general data warehouse transformations[C]. VLDB, 2003, 41-58.

[5] Ateniese G, Hohenberger S. Proxy re-signature: new definitions, algorithms, and applications[C]. CCS, 2005, 310-319.

[6] Buneman P, Cheney J, Vansummeren S. On the expressiveness of implicit provenance in query and update languages[J]. ACM Transactions on Database Systems, 2007, 33(4): 209-223

[7] 杨华龙，王东，李元威. 大数据建设在烟草执法方面的研究和应用[J]. 科技创新与应用，2018, (34): 182-184.

[8] DENG C, SUN R, CHEN Z, et al. Visual analysis of tobacco market big data based on spatial-temporal grid[J], Tobacco Science & Technology, 2018, 51(6): 106-112.

邓超，孙瑞志，陈志斌. 基于时空网格的烟草市场大数据可视分析[J]. 烟草科技, 2018, 51(6): 106-112.

[9] RUAN P, DINH T T A, LIN Q, et al. LineageChain : a fine-grained, secure and efficient data provenance system for blockchains[J]. The VLDB Journal, 2021, pp. 1-22.

[10] ZHU Y, YUE K, WANG C, et al. LBNS: a Bayesian-Network-Based system for representation and query processing of lineages over uncertain data[J]. Journal of Computer Research and Development, 2012, 49(Suppl.): 344-348.

朱运磊，岳昆，王朝禄等. LBNS: 基于贝叶斯网的不确定性数据世系表示和查询处理系统. 计算机研究与发展, 2012, 49(Suppl.): 344-348.

[11] POKODI S, KESAVARAJA D. Secure data provenance in Internet of Things using hybrid attribute based crypt technique[J], Wireless Personal Communications, 2021, pp. 1-22.

[12] SIMON W, HUGO H, PAUL W. Applications of provenance in performance prediction and data storage optimization[J].

[13] MARCHETTI M, PIERAZZI F, COLAHANNI M, et al. Analysis of high volumes of traffic for advances persistent threat detection[J]. Computer networks, 2016, 109(2): 127-141.

[14] PRIEBE C. CloudSafetyNet: Detecting data leakage between cloud tenants[C]. In: Proceedings of the 6th Edition of the ACM Workshop on Cloud Computing Security. ACM, 2014: 117–128.

[15] BERTINO E, GHINITA G, KANTARCIOGLU M. A roadmap for privacy-enhanced secure data provenance[J]. Journal of Intelligent Information Systems, 2014, 43(3): 481–501.