

基于差分隐私模型的云数据副本安全控制方案

任君^{1,2}, 熊金波^{1,2}, 姚志强^{1,2}

(1. 福建师范大学软件学院, 福建 福州 350108;

2. 福建省公共服务大数据挖掘与应用工程技术研究中心, 福建 福州 350108)

摘要: 针对敌手利用重复数据删除作为侧信道攻击用户隐私信息的问题, 提出一种安全的文件副本数量控制方案。核心思想是构建混合云架构, 结合差分隐私中的隐私保护参数及收敛加密算法计算文件密钥值并对其进行二次加密, 保证每个文件在服务器中只存在一个副本。与已有的通过设置阈值来控制副本数量的方案进行比较分析, 所提方案可抵抗侧信道攻击、相关文件的攻击, 以及针对加密算法的特定攻击, 并节约了网络带宽及磁盘存储空间。

关键词: 重复数据删除; 差分隐私; 收敛加密; 混合云; 侧信道攻击

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00163

Security control scheme for cloud data copy based on differential privacy model

REN Jun^{1,2}, XIONG Jin-bo^{1,2}, YAO Zhi-qiang^{1,2}

(1. Faculty of Software, Fujian Normal University, Fuzhou 350108, China;

2. Fujian Engineering Research Center of Public Service Big Data Mining and Application, Fuzhou 350108, China)

Abstract: In cloud computing environment, adversary may use data deduplication as a side channel to eavesdrop users' privacy. In order to tackle this serious issue, a secure copy number control scheme was proposed. The highlights of the proposed scheme lie in constructing a hybrid cloud framework. Then file was encrypted for second times by using the privacy protection parameters of differential privacy as well as the key of the file computed by convergent encryption algorithm. After the file is uploaded, each file has only one copy in the server. Meanwhile, compared with the scheme that use setting threshold value to control the number of copies, the proposed scheme can resist against the side channel attack, related files attack, and the specific attack of encryption algorithm and save the network bandwidth and disk storage space.

Key words: data deduplication, differential privacy, convergent encryption, hybrid cloud, side channel attack

1 引言

随着云计算、大数据技术不断进步, 由社交网站、健康应用、医疗卫生以及一些其他网络公司所产生的数据呈爆炸式增长, 且随着网络社会的

信息化, 这些数据来源多样, 形式多样^[1]。据统计, 平均每秒有 200 万用户在使用谷歌搜索, Twitter 每天处理的推特数量超过 3.4 亿^[2], 同时, 科学计算、医疗卫生等各大行业也有大量数据产生。2012 年, 全球信息总量已经达到 7 ZB, 而到 2016 年

收稿日期: 2017-01-06; 修回日期: 2017-02-17。通信作者: 姚志强, yzq@fjnu.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61370078, No.61402109)

Foundation Item: The National Natural Science Foundation of China (No.61370078, No.61402109)

这一数值达到 8 ZB 以上^[3]，且在新型信息传播方式和个性化服务模式下，越来越多的用户选择将数据外包给云端进行存储和管理，据 Gartner 数据统计，预计到 2020 年，全球数据量将达到 44 ZB^[4]。庞大的数据量不仅会造成空间浪费，还会导致存储效率降低，因此先后出现了一些解决此问题的技术，最直接的方法是采用数据压缩技术对原始数据压缩处理后再上传到云端，但对于同一份数据文件，不同用户可能采用不同的压缩技术，生成不同的压缩文件，造成多副本共存的现象，反而使云端存储压力更大。因此，迫切需要数据副本重复性检测与删除机制，通过仅保留一份数据副本来提高云服务效率与服务质量，即重复数据删除（deduplication）机制。研究发现，通过重复数据删除技术，可以为备份系统节约 83% 的存储空间，为主存系统节约 68% 的存储空间，为云虚拟机中通用数据的存储节约 80% 的空间，直接影响用户使用云服务功能上传和下载文件的效率^[5]，作为目前节约网络带宽和磁盘存储空间的重要手段，重复数据删除技术现已被广大云存储服务提供商接纳和认可^[6]，但同时，攻击者会利用重复数据删除对特定的用户进行隐私攻击，或收集大量用户数据中的隐私信息进行针对性的决策分析^[7]，因此，如何能够在云服务中进行重复数据删除时确保用户数据的隐私安全，已成为隐私保护领域一个亟待解决的问题。

针对上述问题，本文结合差分隐私及收敛加密算法提出一个安全的混合云架构的重复数据删除方案来保护重复数据删除过程中用户的隐私信息，证明其安全性，并且能够抵抗侧信道攻击以及相关文件的攻击；再引入收敛加密算法，并结合差分隐私中的隐私保护参数计算出新的密钥值来代替文件 F ，作为判断重复数据删除的条件以及与服务端沟通的信息，减少通信过程中的带宽消耗，并且可抵抗针对加密算法的特定攻击；最后根据不同文件内容结合隐私保护参数设置不同文件的安全等级，实现不同安全等级的文件上传。

2 相关工作

2002 年，Douceur 等^[8]提出收敛加密算法（CE，convergent encryption）来实现数据的保密性，该

方案的密钥生成算法为确定性算法，通常由原数据经过散列运算得到，因此可以确保相同的数据得到相同的密钥。之后的很多研究成果都利用其来实现对密文的重复数据删除，Li 等^[9]提出 CDStore 方案，采用收敛扩散（CD，convergent dispersal）机制，将秘密共享 AONT-RS（all-or-nothing transform reed-solomon）与 CE 结合，把扩散算法中的随机信息替换为数据的散列指纹，保证扩散算法的确定性以实现数据的重复删除；Stanek 等^[10]基于 CE 提出为数据提供不同安全等级加密保护的重复数据删除方案，该方案将文件分为流行文件和非流行文件，不同的文件类别采用不同的重复数据删除方法；此外，Bellare 等^[11]又在 CE 的基础上提出一种新的消息锁加密（MLE，message locked encryption）机制，该机制使用 MLE 密钥对信息进行对称加密；DupLESS^[12]提出使用 KS 来产生与管理密钥，即当 MLE 密钥生成时，由数据本身和一个系统层面的密钥作为输入，增加了 MLE 密钥的安全性。以上这些重复数据删除方案皆可在一定程度上节省云存储空间，部分方案也加强了对密钥的安全保护，但仍有较多的隐私泄露是攻击者利用云服务中的重复数据删除进行的攻击，如蛮力攻击、侧信道攻击、字典攻击及伪造攻击等。

目前，已有很多学者致力于研究客户端重复数据删除的隐私保护工作。Harnik 等^[13]通过利用随机事件的发生概率来混淆重复数据删除事件的发生，即在服务器设置阈值，当单个文件副本数量达到此阈值时，执行重复数据删除以删除多余的副本文件，从而达到释放存储空间以及保护隐私数据的目的，但同时这种方法无法抵抗穷举攻击；Lee 等^[14]在 Harnik 的方案上改进，在阈值的基础上增加 2~3 个随机数，设置阈值和其进行相减，从而增大随机事件的发生概率，这在一定程度上防止敌手利用穷举法进行攻击，增加了隐私数据的安全性，但这 2 种方法都是基于所有文件是相互独立的假设下，所以无法抵抗相关文件的攻击，并且在重复数据删除的过程中都会由于不必要的文件上传而导致巨大的网络带宽消耗。此外，Olivier 等^[15]提出通过一个互联网服务提供商和混合云存储网络通信与其他业务通信，在家用

路由器上运行重复数据删除协议。然而,该解决方案缺乏灵活性,因为它需要在家通过 ISP 使用路由器,限制其应用到一个特定的服务模型。本文所提的基于差分隐私的重复数据删除方案,可以保证云服务器中单个文件只有一个单一副本,且利用密钥值代替文件本身进行通信,在一定程度上节省了通信带宽,同时在保证用户隐私安全上较以前方案有较大提高。

3 基础知识

3.1 收敛加密

在加密文件时,如果采用传统的随机加密算法,当密钥不同时,即使相同的文件加密后也会得到不同的密文,从而无法进行重复性检测,使相同文件在云端同时存储多份副本,严重浪费存储空间。针对此问题,方案引入引言中提到的收敛加密算法^[8],主要操作如下。

加密文件 F : 对文件进行一个散列运算,计算密钥 $k = H(F)$, 这样,相同的 F 得到的 k 相同,同时, k 还可以作为文件 F 的身份 ID,再用密钥对文件进行加密得到密文

$$C = E_k(F)$$

解密时依旧利用 k 进行解密

$$F = D_k(C)$$

由于 CE 加密算法具有密文可校验的特性,因此迅速在云服务器中的重复数据删除领域得到广泛应用,许多研究成果都将 CE 结合各种不同机制实现对密文的重复数据删除。

3.2 差分隐私

差分隐私的概念最早在文献[16,17]中提出,是针对隐私泄露问题提出的一种新的隐私数据保护方法。例如,有一个数据集 D ,其中包含 Alice 的隐私信息,针对 D 进行任意的查询操作 f ,输出结果为 $f(D)$,如果将 Alice 的隐私信息从数据集 D 中删除,查询的结果依然为 $f(D)$,则可以认为 Alice 的隐私信息并没有因为被包含在数据集 D 中而产生泄露风险,具体的定义如下。

定义 1 设有随机函数 K , r 为函数 K 所有可能输出构成的集合,对于任意 2 个相邻数据集 D_1 和 D_2 (至多相差一个数据),若满足式(1),则

称函数 K 提供 ε 差分隐私保护。

$$\Pr\{K(D_1) = r\} \leq e^\varepsilon \Pr\{K(D_2) = r\} \quad (1)$$

定义 1 表明随机函数 K 通过对输出结果的随机化操作达到隐私保护的,且操作结果的概率是独立的,目前主要的方法是对输出结果添加噪声,增加其不确定性从而降低数据真实性以达到保护隐私数据的目的,在此基础上, Ebadi 等^[18]增加更加复杂的个性化隐私机制,即用户可根据不同文件的隐私程度设置保护级别;另外, Geng 等^[19]设计出更加优化的阶梯机制来添加噪声。

4 方案构造

方案中的符号与相应的描述如表 1 所示。

表 1 符号含义

符号	含义
F	用户需要上传的文件
f	文件 F 的大小
k	第一次散列运算所得到的散列值
Key	第二次散列运算所得到的散列值
C'	利用 k 值对文件 F 加密得到的加密文件
C	利用 Key 值对文件 C' 加密得到的加密文件
D	虚拟数据
$Hash_1 : \{0,1\}^* \rightarrow \{0,1\}^a$	计算 k 的散列函数
$Hash_2 : \{0,1\}^b \rightarrow \{0,1\}^7$	计算 Key 的散列函数
α, β, γ	任意整数
$H_{pr}(h_1, h_2, \dots, h_n)$	私有云服务器存储的 Key 值列表
$H_{pu}(h_1, h_2, \dots, h_n)$	公有云服务器存储的 Key 值列表
N	已存储文件的数量
ID	用户身份的标识符
T	传送队列 T
R	泊松分布生成的值
ε	隐私保护参数

4.1 系统模型

系统模型如图 1 所示。

1) 用户 (user): 选择需要上传的文件或数据发送给私有云服务器,或通过私有云服务器向公有云服务器请求需要下载的文件或数据。

2) 私有云服务器 (private cloud server): 这是一个诚实且可信的服务器, 当用户将需要上传的文件发送给私有云服务器后, 私有云服务器将会计算文件密钥值并对其进行加密, 用密钥值判断该文件是否已经存在于公有云服务器, 若不存在, 则上传该文件; 若存在, 则基于差分隐私执行重复数据删除。私有云服务器会存储一段时间内已上传的文件以方便用户下载。

3) 公有云服务器 (public cloud server): 这是一个半可信的服务器, 提供给用户上传和下载的服务功能, 并在文件上传成功之后返回给用户文件的密钥值。

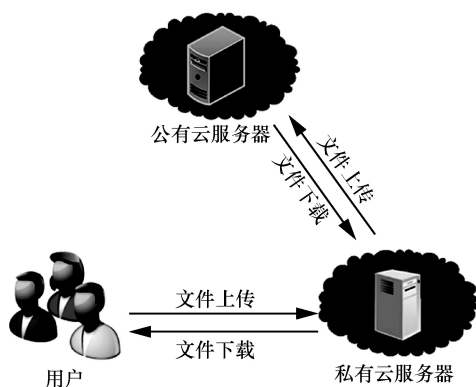


图1 系统模型

4.2 威胁模型

云服务提供商在采用重复数据删除技术控制单个文件副本数量的同时, 有很多攻击者会利用重复数据删除对用户的隐私信息发起攻击, 或针对加密算法进行攻击, 从而获取用户的隐私信息, 主要包括以下几种攻击类型。

4.2.1 侧信道攻击

侧信道攻击是攻击者通过不断上传文件 F 获取用户文件中的隐私信息, 主要有 2 种攻击方式。

1) 文件识别^[13]: 攻击者上传特定的文件到服务器, 根据重复数据删除是否发生来判断服务器是否拥有该文件。例如, Alice 想确定 Bob 是否有一个敏感文件, 该文件不太可能被其他用户拥有, 此时她只需要备份一份完全相同的敏感文件并上传, 然后验证服务器是否执行重复数据删除即可。

2) 学习文件内容: 识别一个特定文件是否存在于服务器后, 攻击者可能为了确定文件的内容而进行穷举攻击。例如, Alice 已经确定只含有 Bob

身份证号码的文件存储于服务器, 她已知前 5 位, 那么为了确定完整的身份证号码, 她便可以采取穷举的办法, 创建所有只包含 Bob 可能身份证号码的副本文件, 根据重复数据删除可以确定包含 Bob 身份证号码的文件是哪一个副本文件。

4.2.2 相关文件攻击

相关文件攻击利用文件之间的相关性, 不仅通过上传 F , 还通过上传与 F 相关的其他文件, 敌手有很大概率可以推断出 F 是否存在, 因此在重复数据删除的过程中, 防止相关文件的攻击尤为重要。

4.2.3 针对加密算法攻击

传统的利用文件加密进行重复数据删除多采用在基础知识中提到的收敛加密算法, 计算文件的散列值, 相同文件会得到相同的散列值, 通过判断散列值是否相同控制文件的副本数量, 但在信道传输的过程中, 攻击者易获取到散列值, 若其已知计算散列值的散列函数和原始文件的范围, 便可在有限的计算开销内, 计算范围内每一个文件的散列值, 并与截取到的散列值进行对比, 若相同, 则判断此文件为原始文件。

4.3 协议描述

根据系统模型图设计方案交互的总体框架, 如图 2 所示。本文所设计方案的具体执行步骤分为文件上传和文件下载这 2 个阶段, 文件上传主要利用收敛加密和差分隐私保护文件中的隐私信息, 而文件下载则利用私有云服务器中一定时间内的文件缓存减少与公有云服务器之间的通信带宽消耗。

4.3.1 文件上传

根据系统模型, 利用算法 1 设计上传步骤如下。

- 1) 用户选择文件 F 上传到私有云服务器。
- 2) 私有云服务器依据文件内容的等级选择相应的 ε 。
- 3) 利用散列函数 $Hash_1$ 计算文件 F 的散列值 k , 并对文件 F 进行加密得到 C' 。
- 4) 利用散列函数 $Hash_2$ 计算 Key , 并对加密文件 C' 进行二次加密得到文件 C 。
- 5) 私有云服务器利用 Key 值查找散列列表 H_{pr} , 判断公有云服务器是否已存在文件 F 。
- 6) 根据差分隐私的定义, 若存在, 利用算法

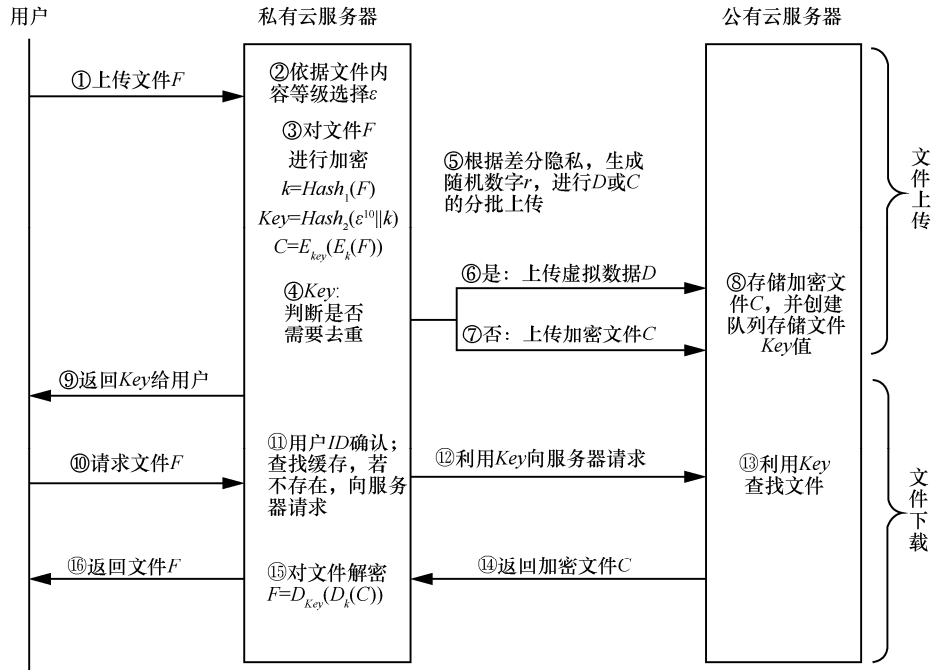


图2 副本数量控制步骤

2 随机生成一个 $[0,1]$ 之间的数字 r , 根据加密文件 C 的大小, 每次创建大小为 r 的虚拟数据放入队列 T 传给公有云服务器, 执行 $\frac{f}{r}+1$ 次。

7) 若不存在, 同样利用随机函数生成一个 $[0,1]$ 之间的数字 r , 将加密文件 C 分 $\frac{f}{r}+1$ 次上传给公有云服务器。

算法1 文件上传

输入: A File

输出: Transfer(T) to Public Cloud Server

- 1) $k=Hash_1(F)$;
- 2) $C=E_k(F)$;
- 3) $Key=Hash_2(\epsilon^{10} || k)$
- 4) $C=E_{Key}(C)$;
- 5) FileExists=check(Key);
- 6) $n=n+1$;
- 7) $h_n=Key$;
- 8) GetTransferSize(FileExists, ϵ);
- 9) if FileExists is false then
- 10) For (int $x=1$; $x \leq \left(\frac{f}{r}+1\right)$; $x++$)
- 11) {
- 12) Enqueue(T , r megabytes of C);
- 13) Transfer(T)→ Public Cloud Server;
- 14) }

15) else

16) For(int $y=1$; $y \leq \left(\frac{f}{r}+1\right)$; $y++$)

17) {

18) Create dummy data of f ;

19) Enqueue(T , r megabytes of dummy data);

20) Transfer(T)→ Public Cloud Server;

21) }

22) end if;

23) end procedure

算法2 随机数生成算法

输入: FileExists, ϵ

输出: Random number

1) if FileExists is true

2) $r=poi\left(\frac{1+\epsilon}{2}\right)$;

3) else

4) $r=poi\left(\frac{1-\epsilon}{2}\right)$;

5) return r ;

6) end procedure

4.3.2 文件下载

根据系统模型中对私有云服务器的设计, 其可存储一定时间内已上传的文件, 在这段时间内, 用户若需要相同文件可以直接从私有云服务器下

载，否则再通过私有云服务器向公有云服务器请求，具体步骤如下。

- 1) 用户向私有云服务器请求下载文件 F 。
- 2) 私有云服务器判断用户 ID 是否为上传过文件的用户，若不是，拒绝查询文件 F ；若是，执行下一步。
- 3) 在缓存区查询是否存在该文件。
- 4) 若存在，则将该文件返回给用户。
- 5) 若不存在，则利用 Key 值向公有云服务器请求。
- 6) 公有云服务器查找并返回文件 C 给私有云。
- 7) 私有云服务器利用 Key 值对加密文件 C 进行解密得到 C' ，再利用 k 值对 C' 进行解密得到文件 F 继而返回给用户。

5 方案分析

5.1 安全性分析

5.1.1 单一文件隐私

1) 在本文方案的安全假设中，数据拥有者、私有云服务器以及两者之间的通信信道是可信且安全的，不会泄露相关的密钥数据或传播已解析出的原始明文，而公有云服务器为半可信的服务器，攻击者可能通过非正常手段进行攻击获取到文件从而得到原始明文，因此本文方案的安全性可以规约为算法安全性。根据差分隐私的定义，本文衍生出以下推论：设 S_1 和 S_2 表示服务器是否包含文件 F ， $S_2 = S_1 \cup \{F\}$ ，若有随机函数 $QS(F)$ 满足式(2)，则称随机函数 $QS(F)$ 对文件 F 进行了 ϵ 差分隐私保护。

$$\Pr\{QS_2(F) = r\} \leq e^\epsilon \Pr\{QS_1(F) = r\} \quad (2)$$

由此可知，本文在算法 2 中所定义的 r 即为 $QS(F)$ 函数所得，而本文的随机函数 $QS(F)$ 采用泊松分布，通过定义及如下推导表明本文方案满足差分隐私的定义。

$$\begin{aligned} \frac{\Pr\{QS_2(F) = r\}}{\Pr\{QS_1(F) = r\}} &= \frac{\Pr\left\{Poi\left(\frac{1-\epsilon}{2}\right) = r\right\}}{\Pr\left\{Poi\left(\frac{1+\epsilon}{2}\right) = r\right\}} \\ &= \frac{\left(\frac{1-\epsilon}{2}\right)^r e^{\frac{-1+\epsilon}{2}}}{\left(\frac{1+\epsilon}{2}\right)^r e^{\frac{-1-\epsilon}{2}}} = \left(\frac{1-\epsilon}{1+\epsilon}\right)^r e^\epsilon \leq e^\epsilon \end{aligned}$$

因此，2 个相差一个文件的存储设备，会有近似同样大小的网络传输量，此属性保证单一文件的存在或不存在将不会显著影响本文所提方案的输出。因此，它可以防止对手使用重复数据删除作为侧信道来判断某个文件存在与否。

2) 文件的 Key 值会作为文件 F 的标识符，发送给服务器用来判断是否需要执行重复数据删除，若需要，则基于差分隐私随机生成虚拟数据分队列上传，否则加密后的文件 C 分批次上传给公有云服务器。因为重复数据删除是否发生都执行了数据上传操作，所以在此过程中攻击者无法判断是否执行了重复数据删除，从而无法判断服务器中是否包含文件 F ，因此无法进行文件内容的识别。

5.1.2 加密算法的特性

本文方案在 k 值的基础上引入和原始文件安全等级有关的参数 ϵ^{10} ，在对计算开销影响较小的前提下，计算一个新的 Key 值，再利用 Key 值对 C' 进行加密得到 C ，这样，在传输的过程中即使攻击者获取了 Key 值和计算 Key 值的散列函数，也无法得知 k ，从而无法得到加密文件 C' ，继而无法判断哪个文件是原始文件 F 。

5.1.3 相关文件隐私

根据攻击者会利用文件之间的相关性进行攻击，首先定义服务器的 3 种存储状态： S_1 ， $S_2 = S_1 \cup \{F\}$ ， $S_3 = S_1 \cup \{F_1, F_2, \dots, F_n\}$ ，并且 $F \in \{F_1, F_2, \dots, F_n\}$ ，根据差分隐私的定义，很明显 $\Pr\{QS_2(F) = r\} \leq \Pr\{QS_3(F) = r\}^{[13]}$ ，因此概率的比值为

$$\frac{\Pr\{QS_2(F) = r\}}{\Pr\{QS_1(F) = r\}} = \frac{\Pr\{QS_3(F) = r\}}{\Pr\{QS_1(F) = r\}} \quad (3)$$

式(3)表明 $QS(F)$ 在 S_1 和 S_3 之间的统计差异总是等同于 S_1 和 S_2 。因此，攻击者在区分 S_1 和 S_3 这 2 种存储状态时并没有比区分 S_1 和 S_2 这 2 种存储状态容易，因此本文方案可以抵抗相关文件的攻击。

5.2 性能分析

5.2.1 理论分析

在重复数据安全删除的过程中，不仅要关注隐私保护程度的高低，更要考虑安全重复数据删除的效率、带宽消耗等问题。因为 Harnik^[13]及 Lee^[14]等的方案都是利用在服务器设置阈值的方

法来防止边信道攻击,与本文方案有可比性,所以表2给出了几个方案的性能对比分析,其中, y 表示上传文件的次数, f 表示文件 F 的长度。

表2 不同方案的性能评估

方案	隐私保护程度	与服务器通信次数	服务器内存消耗
Harnik 方案	低	$2y$	$O(yf)$
Lee 方案	中	$2y$	$O(yf)$
本文方案	高	1	$O(f)$

1) 隐私保护程度。Harnik 与 Lee 等的方案隐私保护程度不如本文方案高,原因在于他们的方案无法抵抗侧信道攻击,即在攻击者连续不断尝试上传文件 F 时,其数量达到方案为服务器所设定的阈值后便会重复数据删除,这样会导致攻击者判断出文件 F 的存在与否或识别文件内容。

2) 通信带宽消耗。已有方案皆是用户和服务器直接通信,在服务器设置文件 F 可上传的最大数量 y ,则上传一个文件 F ,要使重复数据删除得以执行,至少需要通信 $y+1$ 次,导致带宽消耗大。在本文方案中,当用户选择上传文件 F 到私有云服务器后,私有云服务器会根据文件 F 的密钥值在列表进行查找是否公有云服务器中已存在文件 F ,用户和私有云服务器都无需直接与公有云服务器通信,减少一定的通信开销。

3) 服务器内存消耗。设定阈值的方法会导致服务器内存消耗因为阈值的增大而增大,而在本文所设计的方案中,若服务器接收到一组虚拟数据,会自行删除,因此服务器只需要存储文件 F 的单一副本即可,所以服务器内存消耗极少,减轻了服务器的负担。

5.2.2 实验评估

为了具体分析方案的时间效率,对方案的运行时间进行监测。实验采用2台处理器为Intel Core i5-4590 CPU, 3.30 GHz, 64 bit, 8 GB RAM的计算机分别担任私有云服务器和公有云服务器,两者之间通过百兆带宽相连,操作系统均为Microsoft Windows 7,算法采用AES-256进行文件加密及解密,散列函数 $Hash_1$ 采用SHA-1、 $Hash_2$ 采用SHA-256分别进行散列运算,数据主要来自本机的文本文件及视频音频等,且以上算法均由Java语言实现。

1) 方案不同阶段时间开销

根据方案设计,无论重复数据删除事件是否

发生,都会有相同文件大小数据上传到公有云服务器,所以在上传文件时,重复数据删除对时间消耗并未有太大影响。

从图3可以看出,在方案中,将文件传输给公有云服务器所需要的时间最大,除去文件上传下载中传输的时间消耗,主要开销在文件密钥值的计算以及加密解密上。

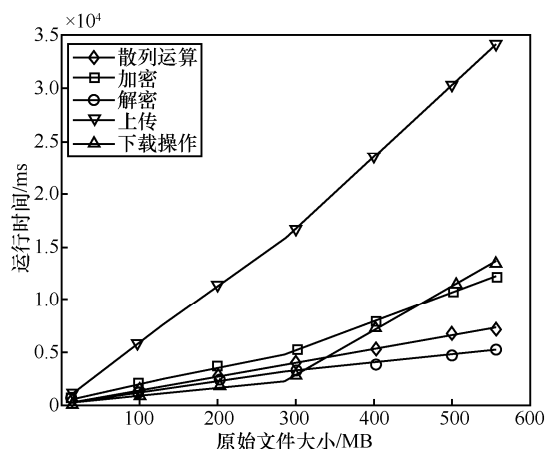


图3 不同阶段的时间消耗

2) 不同阶段时间消耗分析

对每个阶段进行细化分析,图4是2次散列值的运算时间对比分析,从图4中可以看出,计算Key值所用时间小于1 ms,相比获得的实用性和安全性而言,该时间消耗可忽略不计。

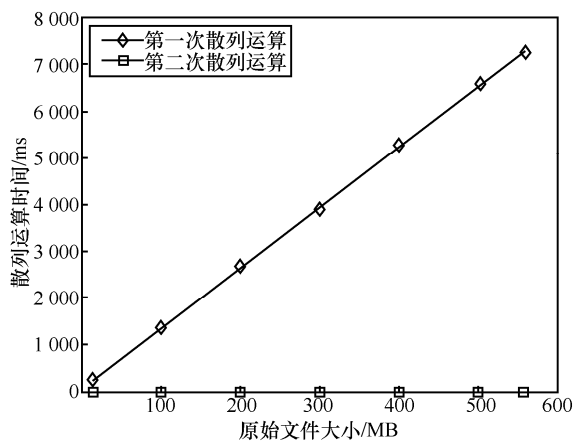


图4 2次散列运算时间消耗

图5和图6分别是方案中2次加密及2次解密的时间消耗对比分析,可以很明显看出,文件越大,第一次加密所用的时间会大于第二次,且第一次解密所用的时间在通常情况下也会大于第二次。

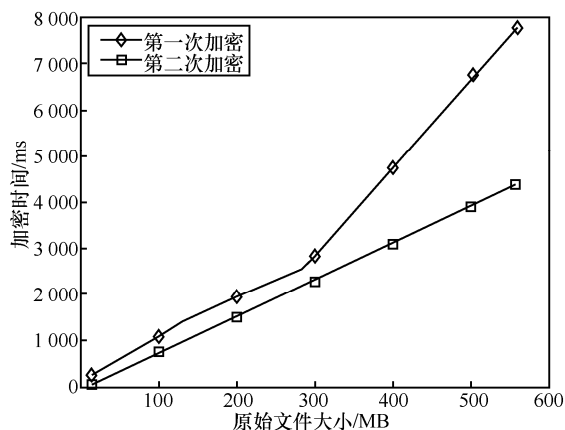


图5 2次加密时间消耗

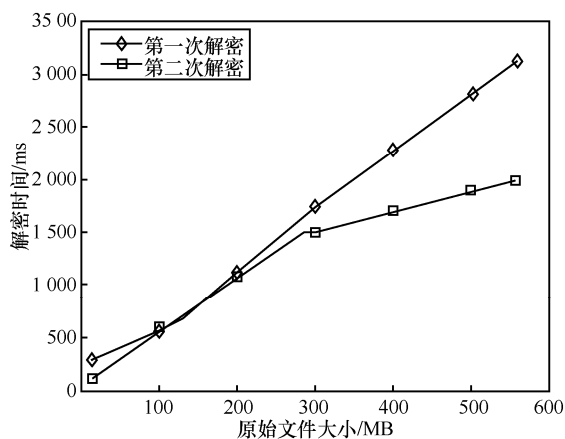


图6 2次解密时间消耗

3) 不同 ϵ 值对方案的影响

图7表示同一个文件下,当 ϵ 取值分别为0.1、0.5、0.9时,散列运算、加密及解密的时间消耗前后不会有太大改变,说明不同 ϵ 值对计算消耗没有影响。

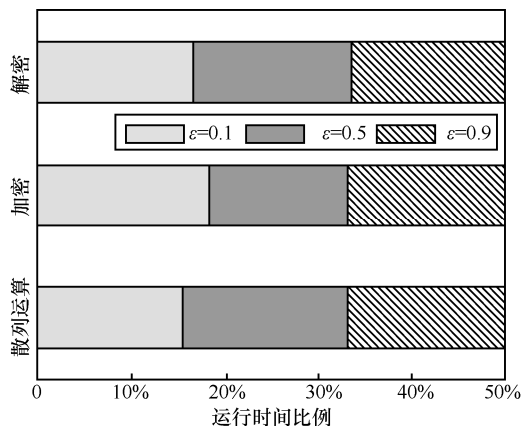
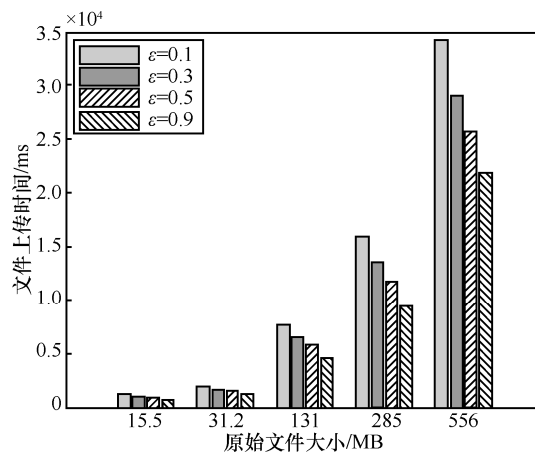
图7 不同 ϵ 值下不同阶段的时间消耗对比

图8表示 ϵ 的取值对文件上传时间的影响。针对同一个文件,要使文件安全等级增大,根据

差分隐私的定义, ϵ 的取值需越小。在本文方案中,为了控制方案在重复数据删除过程中的文件安全,笔者控制 ϵ 取值在 $[0,1]$,根据泊松分布的特点, ϵ 取值越小,生成的结果值也就越大,从而每次上传队列中需要负荷的数据也就越多,根据图8显示,文件上传所需时间也就越长。

图8 不同 ϵ 取值下文件大小与上传时间的关系

综上所述,本文提出的方案在时间效率上仍有很大的改进空间。

6 结束语

本文针对云服务中重复数据删除下的用户隐私泄露问题,构建混合云架构,引入差分隐私,针对不同文件的安全等级对原始文件进行加密,并在用户上传中保证每个文件在云服务器中只存在一个副本。通过实验和安全性分析表明,该方案安全性较高,保护单一文件的隐私,抵抗侧信道攻击、相关文件的攻击以及针对加密算法的特定攻击。另外,相比已有方案,本文方案所占用服务器内存较少,且通信次数大大减少,节约了带宽及存储空间。接下来,将针对文件加密算法及隐私保护参数的设置做进一步研究。

参考文献:

- [1] MEHMOOD A, NATGUNANATHAN I, XIANG Y, et al. Protection of big data privacy[J]. IEEE Access, 2016, 4:1821-1834.
- [2] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1):246-258.
- [3] FENG D G, ZHANG M, LI H. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.
- [3] MAI J E. Big data privacy: the datafication of personal informa-

- tion[J]. Information Society, 2016, 32(3):192-199.
- [4] 熊金波, 张媛媛, 李凤华, 等. 云环境中数据安全去重研究进展[J]. 通信学报, 2016, 37(11): 169-180.
XIONG J B, ZHANG Y Y, LI F H, et al. Research progress on secure data deduplication in cloud[J]. Journal on Communications, 2016, 37(11): 169-180.
- [5] PAULO J, PEREIRA J. A survey and classification of storage deduplication systems[J]. ACM Computing Surveys (CSUR), 2014, 47(1): 11.
- [6] RUSSELL D. Data deduplication will be even bigger in 2010[J]. Gartner, 2010.
- [7] SORIA-COMAS J, DOMINGO-FERRER J. Big data privacy: challenges to privacy principles and models[J]. Data Science and Engineering, 2016, 1(1): 21-28.
- [8] DOUCEUR J R, ADYA A, BOLOSKEY W J, et al. Reclaiming space from duplicate files in a serverless distributed file system[C]//The International Conference on Distributed Computing Systems. 2002: 617-624.
- [9] LI M, QIN C, LEE P P C. CDStore: toward reliable, secure, and cost-efficient cloud storage via convergent dispersal[C]//2015 USENIX Annual Technical Conference (USENIX ATC 15). 2015: 111-124.
- [10] STANEK J, SORNIOTTI A, ANDROULAKI E, et al. A secure data deduplication scheme for cloud storage[C]//Financial Cryptography and Data Security. 2014: 99-118.
- [11] BELLARE M, KEELVEEDHI S, RISTENPART T. Message-locked encryption and secure deduplication[C]//Advances in Cryptology-EUROCRYPT. 2013: 296-312.
- [12] KEELVEEDHI S, BELLARE M, RISTENPART T. DupLESS: server-aided encryption for deduplicated storage[C]//The 22nd USENIX Security Symposium. 2013: 179-194.
- [13] HARNIK D, PINKAS B, SHULMAN-PELEG A. Side channels in cloud services: deduplication in cloud storage[J]. IEEE Security & Privacy, 2010, 8(6):40-47.
- [14] LEE S, CHOI D. Privacy-preserving cross-user sourcebased data deduplication in cloud storage[C]//The International Conference on ICT Convergence. 2012: 329-330.
- [15] OLIVIER, NEUMANN C, MONTALVO L, et al. Improving the resistance to side-channel attacks on cloud storage services[C]//The International Conference on New Technologies, Mobility and Security (NTMS'12). 2012: 1-5.
- [16] DWORK C. Differential privacy[J]. Lecture Notes in Computer Science, 2006, (2):1-12.
- [17] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1):101-122.
XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1): 101-122.
- [18] EBADI H, SANDS D, SCHNEIDER G. Differential privacy: now it's getting personal[J]. ACM Special Interest Group on Programming Languages, 2015, 50(1): 69-81.
- [19] GENG Q, KAIROUZ P, OH S, et al. The staircase mechanism in differential privacy[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(7):1.

作者简介:



任君(1993-), 女, 山西临汾人, 福建师范大学硕士生, 主要研究方向为云计算与安全服务。



熊金波(1981-), 男, 湖南益阳人, 福建师范大学副教授、硕士生导师, 主要研究方向为云数据安全与隐私保护技术。



姚志强(1967-), 男, 福建莆田人, 博士, 福建师范大学教授、硕士生导师, 主要研究方向为信息安全。