一种基于数据分割与分级的云存储数据隐私保护机制

徐小龙[1,2] 周静岚[1] 杨 庚[3]
(南京邮电大学计算机学院 南京 210003)[1]
(中国科学院软件研究所信息安全国家重点实验室 北京 100190)[2]
(南京邮电大学计算机技术研究所 南京 210003)[3]

摘 要 云存储系统数据管理权和所有权的分离导致数据安全和隐私保护难题。传统的基于单纯加密技术的云存储数据隐私保障机制在实际的数据操作过程中带来了较大的系统开销。为了以低开销实现云存储系统中异地托管数据的隐私保护机制,提出了一种基于数据分割与分级的云存储数据隐私保护机制。机制首先将数据合理分割为大小数据块;再分别将小块数据和大块数据部署在本地和异地;然后按数据不同的安全级别需求,联合采用数据染色和不同强度的数据加密技术进行数据染色或加密,以在保护云存储用户数据隐私的同时,提高灵活性,降低系统开销。

关键词 云存储,隐私保护,数据分割,数据分级
中图法分类号 TP309 **文献标识码** A

Data Privacy Protection Mechanism for Cloud Storage Based on Data Partition and Classification

XU Xiao-long[1,2] ZHOU Jing-lan[1] YANG Geng[3]
(College of Computer, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)[1]
(State Key Laboratory of Information Security, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)[2]
(Institute of Computer Technology, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)[3]

Abstract The data management and ownership separation in cloud storage systems lead to the difficulties of data security and privacy protection. The traditional cloud storage data privacy protection mechanism based on simple encryption technologies brings larger extra-overhead in the actual process of data manipulation. In order to achieve low overhead of the data privacy protection mechanism for cloud storage systems, this paper proposed a novel data privacy protection mechanism based on data partition and classification. The mechanism reasonably partitions the original data into two block: one is small, which is deployed locally, and the other one is large, which is deployed remotely, and then according to different security requirements of data, the data dyeing and data encryption technologies are adopted jointly in order to protect the cloud data privacy, improve flexibility and reduce overhead at the same time.

Keywords Cloud storage, Privacy protection, Data partition, Data classification

1 引言

云计算(Cloud Computing)[1-5]通过网络有效聚合被虚拟化的计算资源,基于集中构建的数据中心为单一用户或多租客提供动态的、高性价比的、弹性规模扩展的计算、存储和各类信息服务,改变了传统信息技术产业的体系架构和运作模式,目前受到了国内外学术界和产业界的极大关注。云存储(Cloud Storage)[1,6]是在云计算概念上延伸和发展起来的,是指通过集群(Cluster)应用、分布式计算(Distributed Computing)或分布式文件系统等功能,将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作,并通过各种相应的应用软件或应用接口,共同为多租客提供密集数据存储和业务访问功能的复杂存储资源池系统。

目前云服务提供商为用户提供的安全与隐私保护手段还非常有限,由此带来了一系列的安全问题[7,8]。例如,2009年3月 Google 云计算服务系统发生了用户数据泄漏事件;同年,Microsoft、Amazon 等公司的云服务系统均出现了重大故障,导致成千上万的客户的数据存储和信息服务受到影响,进一步加剧了业界对云应用的安全性、可靠性和可信性的担忧;亚马逊云服务平台 AWS(Amazon Web Services)2010年的租

到稿日期:2012-05-15 返修日期:2012-08-25 本文受国家自然科学基金(61202004,61272084),国家教育部高等学校博士学科点专项科研基金(20093223120001,20113223110003),中国博士后科学基金(2011M500095),江苏省自然科学基金(BK2011754,BK2009426),江苏省博士后科研项目(1102103C),江苏省高校自然科学研究计划(12KJB520007),江苏省科技支撑计划(BE2009158)以及信息安全国家重点实验室开放课题(03-01-1)资助。

徐小龙(1977-),男,博士,副教授,主要研究方向为计算机软件、分布式计算、信息安全、Agent 技术等,E-mail:xuxl@njupt.edu.cn;**周静岚**(1988-),女,硕士生,主要研究方向为云存储、信息安全技术、基于网络的计算机软件和信息安全技术等;**杨 庚**(1961-),男,博士,教授,博士生导师,主要研究方向为计算机应用、网络计算技术、信息安全技术等。

户协议就明确指出 AWS 不能保证租户数据的安全性。

云存储系统中的数据安全问题[6-8]的核心根源是数据管理权和所有权的分离。用户所属的数据外包给云服务提供商，云服务提供商就获得了该数据或应用的优先访问权。事实证明，由于存在内部人员失职、黑客攻击及系统故障导致安全机制失效等多种风险，云服务商没有充足的证据让用户确信其数据被正确地存储和使用。例如，用户数据没有被盗卖给其竞争对手，用户使用习惯等数据隐私没有被提取或分析，用户数据被正确存储在其指定的国家或区域，数据严格按用户要求被彻底地销毁、删除等。

为了保障云存储系统中的数据安全，特别是隐私，目前常见的方法仍然是基于传统的数据加密技术，即简单地用某种加密技术将加密后的数据托管到云存储系统中。然而这些方法在实际的数据操作过程中都带来了较大的开销。针对这种情况，为了以低开销实现云存储系统中异地托管数据的隐私保护机制，本文提出了一种基于数据分割与分级的云存储数据隐私保护机制。机制首先对数据进行合理分割，接着分别在本地和异地进行部署；然后按数据不同的安全级别需求，联合采用数据染色和不同强度的数据加密技术进行染色或加密，从而在保护云存储用户数据隐私的同时，降低系统开销。

2 基于数据分割与分级的云存储数据隐私保护机制

2.1 工作流程

云存储的用户端系统首先自动将待托管的数据分割为小数据块和大数据块，小数据块存储于用户本地，大数据块则按照用户指定的安全级别需求进行加密后，由云端文件系统分块存储。

整个系统按其部署地点分为 3 个部分。

(1)用户端系统：用户端系统由内核态的文件系统过滤驱动和用户态的控制程序构成。其负责监控和捕获本地的系统操作，对获得的数据进行大、小块的分割，并进行哈希运算(Hash)，然后在本地保存小数据块和哈希值，采用数据染色和不同强度的数据加密技术进行染色或加密。

(2)主控服务器系统：主控服务器系统主要进行数据分块和元数据处理服务，并对客户端的数据请求进行认证和访问控制。

(3)存储服务器系统：存储服务器系统主要负责实际的托管数据存储。用户上传数据时，根据主控服务器的分配，将数据块存入对应的物理设备；用户提取数据时，通过查询主控服务器的元数据信息，提取相应的数据块。

当用户请求系统进行数据托管时，系统工作过程如图 1 所示。

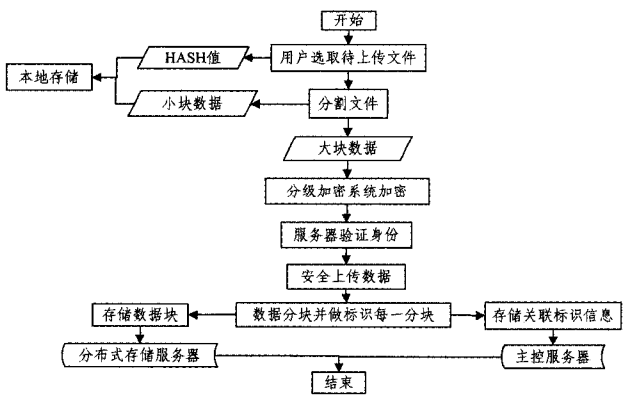

图 1 系统工作过程(数据托管时)

当用户需要下载已被托管云存储系统中的数据时，系统工作过程如图 2 所示。

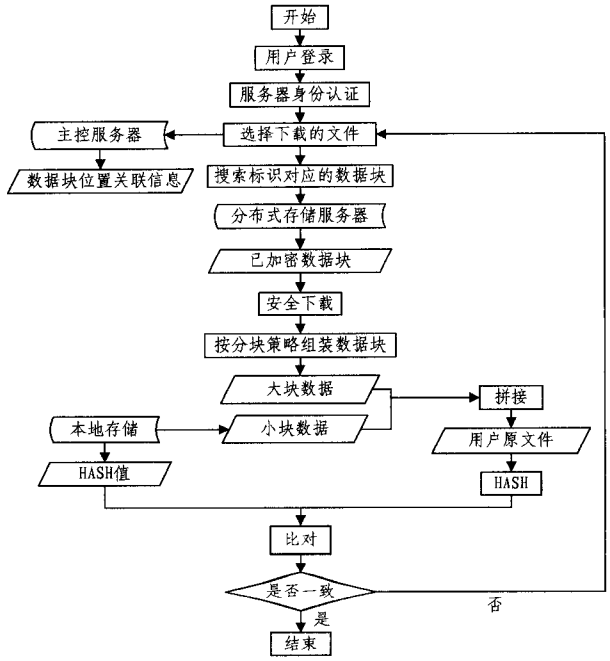

图 2 系统工作过程(数据下载时)

(1)通过身份认证后，用户登录服务器，选择需要下载的文件。

(2)存储服务器系统根据标识向主控服务器查询数据分块信息及数据块存储位置。

(3)根据数据块分块信息和存储位置，从存储服务器取出所有数据块，根据分块策略将数据块组装成大块数据。

(4)将所需数据下载到本地，根据数据对应的分级策略对数据进行解密或去色。

(5)从本地取出保存的小块数据，与大块数据进行拼接，得到原文件，并验证数据完整性。

2.2 数据分割

LPCA(Linear Partition-Combination Algorithm)算法[9]提出利用线性的方法对数据进行分割，并提供了安全恢复数据的策略，可应用于实现大规模存储数据的分布式系统，但是对于隐私数据的保护过于薄弱。另一种数据分割机制是将真实的数据在客户端分块后加密传输到网络上的文件存储服务器，并将数据文件的目录信息存储于本地[10]，这种方式实现了文件数据与其元数据的分离，使得服务器端系统无法获得元数据，解决了来自内部人员的安全问题。

基于数据分割的安全机制的实现以密钥分解理论为基础：将一个文件分解成 n 个分块，完全具备其中任意至少 k($k \leqslant n$)个分块时，才能恢复原文件。这种设计方法使得任意 $n-k$ 个分块丢失或损坏时仍能恢复原文件，从而提高了可靠性和可用性；且任意不足 k 个分块被窃取时，不能还原成原文件，从而提高了安全性。但若所有数据块均存储于服务器端，则系统仍能获取数据隐私信息。

本文的数据分割机制是将数据文件分割成大、小数据块后异地存储，设计了以下两种分割方案。

方案 1 抽取固定大小的小块数据

将小块数据大小固定为 1k，则抽取 1024 个字节，工作流程如图 3 所示。

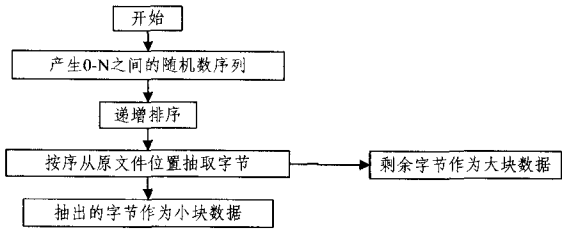

图 3　抽取固定大小的小块数据流程

(1)产生 0～N(N 为文件大小)之间的随机数序列。序列长度等于小块数据的大小(1k)。

(2)将随机数序列从小到大排列,就得到了要从文件中抽取字节的位置。

(3)将对应位置的字节从原文件中分割出来,与顺序排列的随机数序列一起保存作为小块数据;被分割后的文件作为大块数据。

该数据分割方案的优点是小块数据大小固定,不会带来用户端存储或者读写处理的压力,便于本地存储管理;缺点是产生的随机数可能不均匀,使得可能有大段的连续数据存在,被解密后可以从中分析出一些信息。

方案 2:抽取非固定大小的小块数据

分割后的小块数据大小不是固定的,而是由原文件大小决定,流程图如图 4 所示。

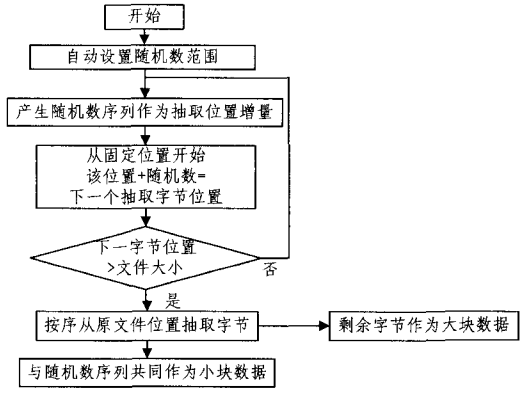

图 4　抽取非固定大小的小块数据流程

(1)根据文件大小自动设置随机数范围(如 1～10)。

(2)系统产生一组在设置的范围内的随机数作为抽取位置增量。

(3)从固定位置开始抽取字节,该位置加上随机数得到下一个抽取字节位置,直到下一个字节位置大于待上传数据大小为止。

(4)将对应位置的字节从原文件中抽取出来,与顺序排列的随机数序列一起保存作为小块数据;剩余文件作为大块数据。

该数据分割方案的优点是抽取的字节位置比较均匀;缺点是随机数位置增量范围是根据文件大小自动设置的,其设置策略较难制定。若范围分得过于细致会使算法繁琐,若分得过于粗略则使同一位置增量范围内的文件得到的小块数据大小差距很大,加重本地存储管理小块数据的负担。

2.3　分级加密

经抽取分割后,被托管云存储系统中的大块数据按其安全级别需求被分级加密系统进行加密。本文提供 3 种不同安全级别的加密策略:

(1)高强度加密:安全程度最高,但处理速度较慢。适合保护对隐私要求极高的数据。

(2)中强度加密:安全程度适中,计算复杂度低于高强度加密,适合保护对隐私安全要求一般的数据。

(2)低强度加密:安全程度低,但处理速度很快,适合保护对隐私安全要求不高的数据。

用户托管数据时,分级加密系统负责绑定用户数据与选择的安全策略,根据用户的选择用相应的算法对待上传数据进行处理,并负责维护用户文件与选择的安全策略的映射表,将加密相关参数保存于本地,然后上传文件;用户使用数据时,由分级加密系统负责查找文件、加密策略映射表,并提取加密算法相关参数,然后解密数据。其中,加密算法的相关参数可由用户端本地保存的小块数据生成。

1. 高强度加密

高强度加密采用基于椭圆曲线的加密算法(ECC, Elliptic Curve Cryptography)[11]。Koblitz 和 Miller 于 1985 年分别提出将椭圆曲线系统首次应用于密码学。椭圆曲线密码方案属于公钥机制,它的安全性依赖于解决椭圆曲线离散对数问题(ECDLP, Elliptic Curve Discrete Logarithm Problem)的困难性。ECDLP 定义如下:给定素数 p 和椭圆曲线 E,对 $Q=kP$,在已知 P、Q 的情况下求小于 p 的正整数 k。将椭圆曲线中的加法运算、乘法运算分别与离散对数中的模乘运算、模幂运算相对应,建立相应的基于椭圆曲线的密码体制。

ECC 的优点之一是用很短的数字就可以表示一个可观的存储,与比其他方法(例如 RSA)相比,它使用更小的密钥就能提供相当的或更高的安全级,被广泛认为是在给定密钥长度时最强大的非对称加密算法。ECC 的另一个优点是可以定义群之间的基于 Weil 对或 Tate 对的双线性映射,双线性映射在密码学领域中被大量应用,如基于身份的加密、密钥协商协议等。

但是双线性对的计算非常费时,导致了 ECC 加密和解密操作的实现比采用其他机制需要花费更长的时间。在提高双线性对计算效率的成果中,最重要的是 Miller 在 2004 年提出的 Miller 算法[12],它使得双线性对的计算可在多项式时间内完成。在此基础上,Barreto 提出了改进的方案[13]。随后又有一些加速双线性对计算的算法被提出[14-16]。

2. 中强度加密

中强度加密采用基于数据染色(Data Dyeing)的加密方案。数据染色是将数据经过若干函数变换后,其表现形态发生很大改变的一种方法。这种方法能保证非授权用户在没有获得函数的模糊化参数情况下,即使得到模糊化后的数据,也无法在多项式时间内还原得到原始数据。

李德毅院士提出了正态云模型的概念[17,18]。云是用来将定性概念转换为定量的不确定性的转换模型,其数字特征由期望值 Ex、熵 En、超熵 He 3 个值表征,把随机性和模糊性完全结合起来。其中,期望值 Ex 指空间内最能代表某个定性概念的点;熵 En 反映定性概念的不确定性;超熵 He 用来度量熵的不确定性。在云模型中,"模糊"的概念被定义为一个边界弹性不同、收敛于正态分布函数的云。云是由一系列的云滴构成的,每个云滴是定性概念映射到一维、二维或多维

空间的一个点；云模型同时给出了这个点代表此定性概念的确定度，以反映其不确定性。基于正态云模型通过数据染色实现隐私保护主要通过使用正态云模型的 3 个特征值生成颜色[8,18]：期望 Ex 的值决定于用户数据的内容，熵 En 和超熵 He 则是独立于数据内容的，只有数据拥有者知道的随机值，可作为数据拥有者的私钥。使用这 3 个特征值经过云发生器生成一组云滴，这组云滴是云存储服务提供商以及其他用户无法获得的。然后使用这组云滴对数据进行模糊化后再上传，可降低用户数据隐私被非法用户窃取后泄露的风险。数据染色(模糊化)过程如图 5 所示。

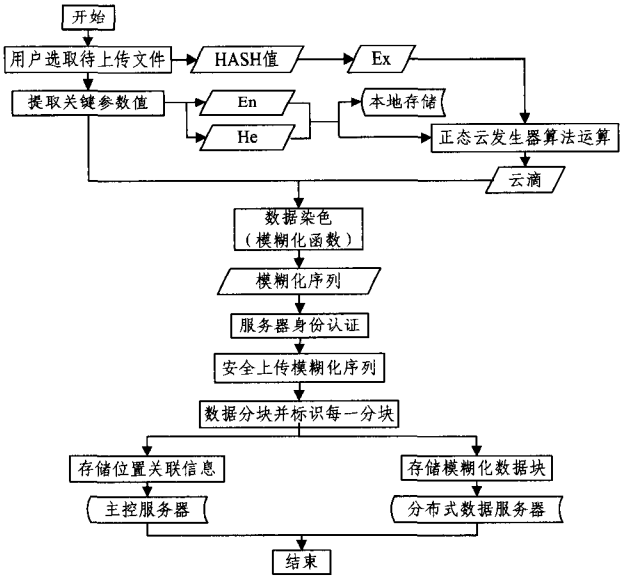

图 5　数据染色过程

(1)用户端上传的文件通过哈希后得到参数 Ex 的值，生成独立于数据内容、只有数据拥有者才知道的关键参数值 En、He，并将哈希值和 En、He 的值保存在本地。

(2)由 Ex、En、He 通过正态云发生器算法得到一组云服务提供者和其他用户不能获得的云滴。

(3)使用生成的云滴对待上传数据进行数据染色。

(4)通过服务器身份认证后，将模糊化序列送到云端，分块后数据块存储到存储服务器，将数据分块信息和数据块存储位置信息存储到主控服务器。

数据去色(去模糊化)过程如图 6 所示。

(1)通过身份认证后，用户登录服务器，透明地选择需要下载的文件。

(2)云存储文件系统根据标识查询对应的模糊化文件，并向主控服务器查询数据分块信息及数据块存储位置。

(3)根据数据块分块信息和存储位置，从存储服务器中取出所有数据块，并组装成模糊化数据序列。

(4)将模糊化数据序列安全下载到本地，根据模糊数据和参数 En、He 得到去模糊化函数。

(5)用去模糊化函数对模糊数据序列进行去模糊化，得到原始数据，对该数据序列进行哈希运算，与本地存储的哈希值比对，若相同，则说明数据未被篡改。

数据染色方法保护图像、软件、视频、文档以及其他类型数据的隐私性，其开销远远低于传统加密解密计算。

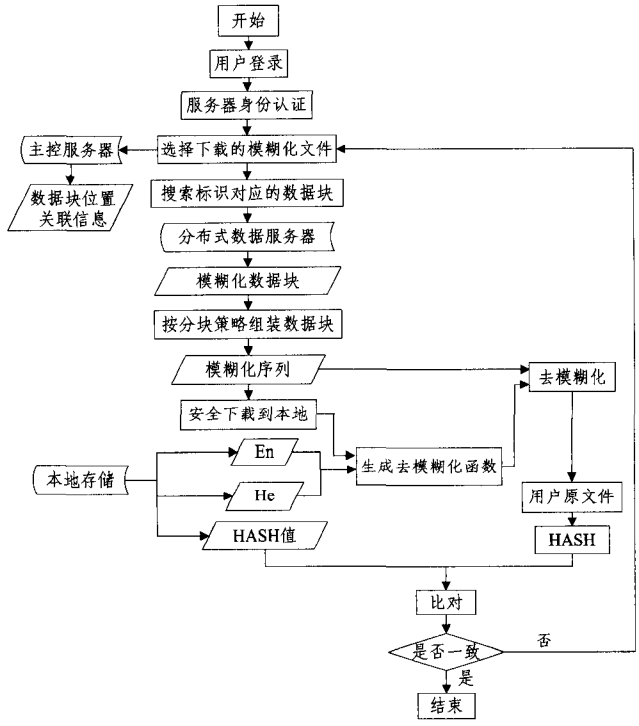

图 6　数据染色过程

3. 低强度加密

低强度加密采用基于 TEA 算法的加密方案。TEA 是一种小型的分组对称加密算法，最初由 David Wheeler 和 Roger Needham 于 1994 年设计。TEA 的特点是速度快、效率高、实现简单，用 C 语言实现仅需 26 行代码。尽管算法十分简单，但它具有很强的抗差分分析能力[19]。相比其他算法，它的安全性相当好，可靠性不是通过算法的复杂度而是通过加密轮数保证。加密过程中，密钥不变，主要的运算是移位和异或。TEA 使用长度为 64 位的明文分组和 128 位的密钥，进行 64 轮迭代，每轮数据经过 Feistel 结构模块进行处理。它使用一个来源于黄金比率的神秘常数 δ 作为倍数，来保证每轮加密都不同。

在 TEA 的基础上，QQ 建立了 QQTEA 算法，其使用 16 轮加密，并采用了一些填充和交织的算法来满足需要加密不定长数据的需要。

随后，针对 TEA 的攻击不断出现。TEA 被发现存在一些缺陷，几个升级的版本被提出，分别是 XTEA、Block TEA 和 XXTEA。这些算法降低了加密过程中密钥混合的规律性，提高了安全性，但降低了处理速度[20]。

3　安全性分析

3.1　攻击方式分析

结合云存储系统多用户的特点可总结出 2 个级别的云内部攻击方式：第一级(L1)是非授权恶意用户，第二级(L2)是系统管理员甚至是云存储服务提供商本身。而云系统外部攻击者可通过系统入侵等操作，达到与 L1 级攻击者相近程度的威胁。攻击者可通过以下方式破坏用户数据的隐私与安全：

(1)L1 级攻击者能够通过网络远程攻击系统内部漏洞绕过访问控制权限限制，获得更高级别的权限；或运行恶意程序获得加密密钥及其他运行时信息。

(2)L2 级攻击者具有对于数据处理的最高权限，他们如果对用户数据好奇，则可以从主控服务器系统得到用户的元数据信息，直接访问存储于服务器系统中的任意用户的数据，并进行窃取、篡改等操作。

3.2 数据保密性

安全的云存储系统要求云存储服务提供商无法得到与用户数据有关的任何信息。用户的敏感数据在云端是以密文形式存储的，因而具有密码学上的安全性。L1 级攻击者由于无法直接接触硬件，因此无法直接得到用户的数据块，必须采取非直接物理攻击的网络攻击方式来获取云存储用户的数据。由于云存储系统中数据是分块存放的，因此 L1 级攻击者并不能获取完整的元数据信息从而得到所有的数据块以恢复出云端存储的大块数据。而数据文件分割后上传进一步保证了云端数据隐私安全，这样，即使是拥有管理员权限的 L2 级云服务提供者也无法恢复原始数据，只能获取被分块的加密后的不完整数据；本方案使用分级加密系统，加密策略选择信息保存于本地，处于云端的攻击者并不知道数据用何种方式进行加密，故无法从加密后的不完整数据获取用户的私密信息。

3.3 数据完整性

本方案中，用户上传文件前计算其 HASH 值并保存于本地，且将小块数据保存于本地，故每次用户下载大块数据后都要与本地小块数据拼接后才能得到完整数据，拼接后都要进行 HASH 计算，并与保存在本地的原文件的 HASH 值进行比对。这样既使 L1 级攻击者对云端部分数据块进行篡改，或 L2 级攻击者对存储在云端的恢复出的大块数据进行篡改，或是文件本身部分丢失，都无法通过完整性验证，从而有效保证了数据完整性。

4 性能开销

与传统的云存储隐私保护方案相比，本方案增加了数据分割的操作，产生了一定的额外时间开销。

若采用分割方案 1，由于小块数据大小为定值，产生随机数序列及从文件中抽取对应的字节的时间复杂度为 O(1)；若采用分割方案 2，数据大小块分割的时间开销为 $O(n)$，其中 n 为小块数据的大小，与待上传文件大小及设置的随机数范围有关，不是定值，但可通过设置策略将其控制在合理的范围。

本方案将待上传数据分割为大、小数据块后上传，将用户数据处理的过程分为两个阶段，这确实带来了一定的系统开销，但分割数据的操作大大增强了数据隐私的安全性，并且分级加密系统的引入又平衡了安全性与性能开销的矛盾。用户使用云存储应用时的数据隐私要求大多处于中强度和低强度，本方案应用的中、低强度加密算法的时间开销远低于传统公钥加密算法，而其经过数据大、小块分割后上传，安全性可达到甚至超过公钥加密算法。而对于选择高强度加密策略的用户，他们本身对于高私密性的要求高于低数据开销的要求，本方案可有效防止不可信服务器的安全威胁，因此可认为在这样的安全性前提下，上传前数据分割至多造成 $O(n)$ 的时间开销是可以接受的。

结束语 云存储系统中的隐私保护问题已经成为了云存储应用发展的瓶颈。在数据所有权与管理权分离的情况下，既要保证数据拥有者的隐私，又要兼顾客户端的性能开销。本文提出的基于数据分割与分级的云存储数据隐私保护机制综合应用了数据分割和分级加密保护方案。机制在服务器不可信的云存储环境下，防止恶意用户和拥有管理员权限的系统管理员非法窃取、篡改用户隐私数据，同时便于用户根据偏好设置数据在云端存储时的安全等级，兼顾了安全性与性能开销，增加了云存储应用的灵活性。但将原文件切割成大、小块数据时，切分策略还有待改进，如何合理、巧妙地切割原文件，使得从大数据块无法恢复出原文件的任何隐私信息，以及选择更适合的算法加入分级加密系统，还需做进一步研究，使其更好地适用于云存储系统。

参 考 文 献

[1] Miller M. Cloud computing [M]. Beijing: Machinery Industry Press, 2009

[2] Boss G, Malladi P, Quan D, et al. Cloud computing[EB/OL]. http://download. boulder. ibm. com/ibmdl/pub/software/dw/wes/hipods/Cloud_computing_wp_final_8Oct. pdf, 2012-03-09

[3] 陈康，郑纬民. 云计算：系统实例与研究现状[J]. 软件学报，2009，20(5)：1337-1348

[4] 金海，吴松，廖小飞，等. 云计算的发展与挑战[R]. 2009 中国计算机科学技术发展报告. 北京：机械工业出版社，2010：21-51

[5] 刘鹏. 云计算(第二版) [M]. 北京：电子工业出版社，2011

[6] 马玮骏，吴海佳，刘鹏. MassCloud 云存储系统构架及可靠性机制[J]. 河海大学学报：自然科学版，2011，39(3)：348-352

[7] 冯登国，张敏，张妍，等. 云计算安全研究[J]. 软件学报，2011，22(1)：71-83

[8] 邹德清，金海，羌卫中，等. 云计算安全挑战与实践[J]. 中国计算机学会通讯，2011，7(12)：55-61

[9] 张薇，马建峰. LPCA——分布式存储中的数据分离算法[J]. 系统工程与电子技术，2007，29(3)：453-458

[10] 王保兵. 电子数据分离存储与安全恢复系统的研究及实现[D]. 南京：南京邮电大学，2009

[11] Hankerson D, Menezes A, Vanstone S. Guide to Elliptic Curve Cryptography[M]. New York: Springer-Verlag, 2004: 1-15

[12] Miller V. Short Programs for Functions on Curves [EB/OL]. http://tcs. uj. edu. pl/~mistar/pdf/Miller1986ShortPrograms. pdf, 2012-03-12

[13] Barreto P, Galbraith S, Eigeartaigh C, et al. Efficient Pairing Computation on Supersingular Abelian Varieties[J]. Des Codes Cryptograp, 2007, 42(3): 239-271

[14] Koblitz N, Menezes A. Pairing-based Cryptography at High Security Levels[C] // Proceeding of Cryptography and Coding. LNCS 3796, Berlin: Springer-Verlag, 2005: 13-36

[15] Barreto P, Kim H, Lynn B, et al. Efficient Algorithms for Pairing-Based Cryptosystems[C] // CRYPT0'02 Proceedings of 22nd Annual International Cryptology Conference on Advances in Cryptology. LNCS 2442, Berlin: Springer-Verlag, 2002: 354-368

[16] Hess F, Smart N, Vercauteren F. The Eta Pairing Revisited[J]. IEEE Transactions on Information Theory, 2006, 52(10): 4595-4602

[17] 李德毅，孟海军，史雪梅. 隶属云和隶属云发生器[J]. 计算机研究和发展，1995，32(6)：16-21

[18] 刘常昱，李德毅，潘莉莉. 基于云模型的不确定性知识表示[J]. 计算机工程与应用，2004，6(8)：28-34

[19] Andem V. A Cryptanalysis of the Tiny Encryption Algorithm [D]. Tuscaloosa: The University of Alabama, 2003

[20] Russell M D. Tinyness: An Overview of TEA and Related Ciphers[EB/OL]. http://www-users. cs. york. ac. uk/~matthew/TEA/, 2012-03-13