doi: 10.7690/bgzdh.2024.07.005

基于逆向运算法的海量大数据安全存储方法

王卓瑜[1]，王　磊[1]，陆　婷[2]，苏　亮[3]，孙　婷[3]
（1. 国网新源控股有限公司，北京 100761；2. 华东宜兴抽水蓄能有限公司，江苏 宜兴 214205；
3. 北京中电飞华通信有限公司，北京 100070）

摘要：为解决传统大数据安全存储方法存在的加解密时间长与存储速率低的问题，提出一种基于逆向运算法的海量大数据安全存储方法。通过 AES 算法中查询表模块、密匙扩展模块及加解密模块的功能设计实现大数据的加密处理；设计一种用于存储加密数据的海量大数据分层存储模型，在加密数据的读取中，通过逆向运算法恢复数据，实现海量大数据的安全存储；搭建 Hadoop 集群测试环境测试设计方法的加密性能与存储性能。测试结果表明：该方法的加密与解密时间均低于 20 s，影像数据存储速度高于 580 MB/s，语音数据存储速度高于 916 MB/s，能有效缩短大数据存储的加解密时间，提高多类资源的存储速度。

关键词：逆向运算法；分布式传感器；海量大数据；安全存储；Rabin 指纹算法
中图分类号：TP309.2　　**文献标志码：**A

Secure Storage Method of Massive Big Data Based on Reverse Operation

Wang Zhuoyu[1], Wang Lei[1], Lu Ting[2], Su Liang[3], Sun Ting[3]
(1. *State Grid Xinyuan Holdings Co*., *Ltd*., *Beijing* 100761, *China*; 2. *East China Yixing Pumped Storage Co*., *Ltd*., *Yixing* 214205, *China*; 3. *Beijing Zhongdian Feihua Communication Co*., *Ltd*., *Beijing* 100070, *China*)

Abstract: In order to solve the problems of long encryption and decryption time and low storage rate of traditional big data security storage methods, a massive big data security storage method based on reverse operation is put forward. Through the functional design of query table module, key expansion module and encryption and decryption module in AES algorithm, the big data encryption processing is realized; A massive big data hierarchical storage model for storing encrypted data is designed. During the reading of encrypted data, data is recovered through reverse operation to achieve the safe storage of massive big data; Build a Hadoop cluster test environment to test the encryption performance and storage performance of the design method. The test results show that the encryption and decryption time of this method is lower than 20 s, the storage speed of image data is higher than 580 MB/s, and the storage speed of voice data is higher than 916 MB/s. This method can effectively shorten the encryption and decryption time of big data storage, and improve the storage speed of multiple resources.

Keywords: reverse operation method; distributed sensors; massive big data; safe storage; Rabin fingerprint algorithm

0　引言

进入 21 世纪后，随着网络应用范围的不断拓展，各行业与领域的数据呈现爆炸式增长状态，形成了海量大数据[1]。在应用海量大数据获得有效信息前，需要解决海量大数据的存储问题。在虚拟的网络环境中，现有数据加密技术能够处理的数据量较小，导致海量数据存在较大的泄露风险[2]。一些不法分子利用这一漏洞获取海量大数据中的重要信息，或者将海量数据中的关键信息篡改，产生不可挽回的损失，对海量大数据的安全存储造成严重威胁[3]；因此，迫切需要研究一种针对海量大数据的安全存储方法。

文献[4]提出基于区块链的海量大数据安全存储方法，首先构建数据调度模型，并通过数据包生成数据存储矩阵。通过计算列不满秩概率完成存储算法研究。以区块链非对称加密技术为基础，获得数据存储的公钥与私钥，完成非关系型数据的安全存储。文献[5]提出基于 Hilbert 曲线与 Cassandra 技术的海量大数据安全存储方法，采用 Hilbert 曲线编码技术将海量数据的存储空间划分为多个存储单元，并映射处理数据的存储轨迹。根据数据存储的时空局部性，设计数据轨迹对应的聚簇键，从而完成海量大数据的安全存储。

上述研究方法虽然取得了一定成果，但在隐私保护方面仍有所欠缺；因此，笔者设计一种基于逆向运算法的海量大数据安全存储方法，实现更加安全的海量大数据存储。

收稿日期：2024-03-23；**修回日期：**2024-04-25
第一作者：王卓瑜(1980—)，女，河北人。

1 基于逆向运算法的海量大数据安全存储

1.1 海量大数据加密处理

为确保海量大数据的安全性，通过 AES 算法实施海量大数据的加密处理。在 AES 算法中有查询表、密匙扩展和加解密 3 个模块[6]。

通过查询表模块生成基础查询表，基础查询表包括 S 及逆 S 盒表，其中构造 S 盒表的过程具体如下：

1) 对于任意元素 x，获取 GF(2)^8 上 x 的逆元。具体公式如下：

$$a(x)=a_7x^7+a_6x^6+a_5x^5+a_4x^4+a_3x^3+a_2x^2+a_1x。\quad(1)$$

式中 a_7、a_6、a_5、a_4、a_3、a_2、a_1 均为字节运算阈值[7-10]。

2) 对逆元实施简化仿射变换，获得经过 S 转换后元素 x 的值，并存储在 1 维数组 S[256]中[11]。其中简化仿射变换的公式如下：

$$\begin{bmatrix}b_7\\b_6\\b_5\\b_4\\b_3\\b_2\\b_1\end{bmatrix}=\left(\begin{bmatrix}a_7\\a_6\\a_5\\a_4\\a_3\\a_2\\a_1\end{bmatrix}\oplus\begin{bmatrix}a_6\\a_5\\a_4\\a_3\\a_2\\a_1\\a_7\end{bmatrix}\oplus\begin{bmatrix}a_5\\a_4\\a_3\\a_2\\a_1\\a_7\\a_6\end{bmatrix}\oplus\begin{bmatrix}a_4\\a_3\\a_2\\a_1\\a_7\\a_6\\a_5\end{bmatrix}\oplus\begin{bmatrix}a_3\\a_2\\a_1\\a_7\\a_6\\a_5\\a_4\end{bmatrix}\oplus\begin{bmatrix}a_2\\a_1\\a_7\\a_6\\a_5\\a_4\\a_3\end{bmatrix}\right)\bullet a(x)。\quad(2)$$

式中 b_7、b_6、b_5、b_4、b_3、b_2、b_1 为简化仿射变换后的逆元值[12]。

3) 构造 S 盒表，其表达式具体如下式：

$$y='05'b_7^{-1}+'09'b_6^{-2}+'05'b_5^{-4}+'9'b_4^{-8}+'4'b_3^{-16}+'01'b_2^{-32}+'5'b_1^{-191}。\quad(3)$$

接着通过同样的方式构造逆 S 盒表，具体表达式为：

$$y=\left[\left('05'b_7^{-1}+'09'b_6^{-2}+'05'b_5^{-4}+'9'b_4^{-8}+'4'b_3^{-16}+'01'b_2^{-32}+'5'b_1^{-191}\right)^{-1}\right]^{256}。\quad(4)$$

通过 S 盒表与逆 S 盒表的表达式构建轮函数的 32 位 T 真值。

密匙扩展模块由解密子密钥产生单元和加密子密钥产生单元构成。解密子密钥单元能够实施逆列混合运算，以产生加密子密钥；加密子密钥单元能够实施 S 盒表与逆 S 盒表的置换操作以及循环移位操作，以产生加密子密钥[13]。

通过加密模块实现轮变换操作，具体过程如下：

定义 $a_{i,j}$ 为轮变换的实际输入状态阵列，$b_{i,j}$ 为其对应的 SubBytes 变换输出，则下式成立：

$$\left.\begin{aligned}&b_{i,j}=S_{\mathrm{RD}}\left[a_{i,j}\right]^2\\&0\leqslant i\leqslant 4\\&0\leqslant j\leqslant N_b\end{aligned}\right\}。\quad(5)$$

式中：S_{RD} 为 SubBytes 变换；N_b 为输入状态阵列的最大数量[14]。

用 $c_{i,j}$ 表示经 ShifRows 变换后的输出，则下式成立：

$$\begin{bmatrix}c_{0,j}\\c_{1,j}\\c_{2,j}\\c_{3,j}\\c_{4,j}\end{bmatrix}=\begin{bmatrix}b_{0,\ j+c_0}\\b_{1,\ j+c_1}\\b_{2,\ j+c_2}\\b_{3,\ j+c_3}\\b_{4,\ j+c_4}\end{bmatrix}。\quad(6)$$

合并式(5)与(6)，实现轮变换操作。

算法的实现方法：将 AES 算法的最优效率作为算法的实现目标，通过空间换取时间，高效率实现海量大数据的加密处理。

1.2 实现海量大数据安全存储

设计一个海量大数据分层存储模型，用于存储加密后的海量大数据，在加密数据读取中，通过逆向运算法将加密后的海量大数据转换为明文数据，实现海量大数据的安全存储。

加密后海量大数据逆向运算过程如下：

1) 获得海量大数据的公钥 Q_A，计算 $(x_1, y_1)=d_BQ_A$。

2) 计算 $m_1=x_2x_1^{-1}\bmod p$，$m_2=y_2y_1^{-1}\bmod p$。

逆向运算如图 1 所示。

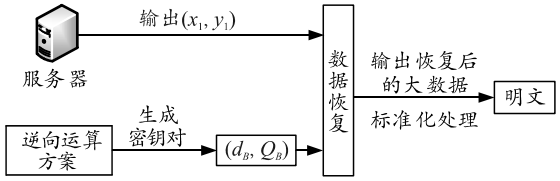


图 1 逆向运算

设计的分层存储模型具体由 3 个模块构成，分别为云服务层模块、边缘节点层模块以及用户应用层模块。模型结构具体如图 2 所示。

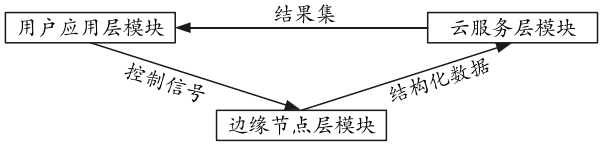


图 2 分层存储模型结构

上图中，云服务层模块主要由多个云服务器构

成，负责对海量大数据进行分析，其配置与数量可根据运算要求与用户需求来调整，从而满足用户对于海量大数据的存储要求。

在设计海量大数据分层存储模型中，加密数据的读取需要使用逆向运算法对加密数据进行恢复[15]。具体来说，当需要对数据进行读取时，在云服务器端获取加密数据块，根据拆分的逆运算对原始数据进行恢复，具体步骤如下：

1) 根据各拆分块所对应的平面坐标，恢复 3 维空间中拆分的长方体结构，并用 $C[m][n]\left[L'/(m\times n)\right]$ 来表示。

2) 以 z 轴对应坐标为依据，对 $C[m][n][L'/(m\times n)]$ 进行拆分，使其成为多个独立的拆分平面。根据 z 轴对应坐标对这些拆分平面实施分组，具体为：

$$C=\left(C[m][n][L'/(m\times n)]_1, C[m][n][L'/(m\times n)]_2, \cdots, C[m][n][L'/(m\times n)]_l\right)。\tag{7}$$

式中 l 为 z 轴对应的坐标个数[16]。

3) 对于各拆分面，根据对应的拆分规则 G_i 恢复其中的字节数据，并依据恢复的顺序依次对字节数据进行读出。以 z 轴对应坐标为依据拼接各拆分面恢复的字节，获取原始数据。

通过该方式拆分数据获取的拆分块中不包括连续字节。当用户合法获取全部拆分块时，才能够依据之前的拆分规则恢复数据，实现数据隐私的有效保护。

2　海量大数据安全存储测试

2.1　搭建 Hadoop 集群测试环境

对于设计基于逆向运算法的海量大数据安全存储方法，为测试其性能，搭建一个 Hadoop 集群测试环境。共设置 1 个 NameNode 节点和 5 个 DataNode 节点，其中，NameNode 节点的主机名是 Master，DataNode 节点的主机名是 Slave1、Slave2、Slave3、Slave4、Slave5。在搭建的 Hadoop 集群测试环境中对设计方法实施仿真测试。

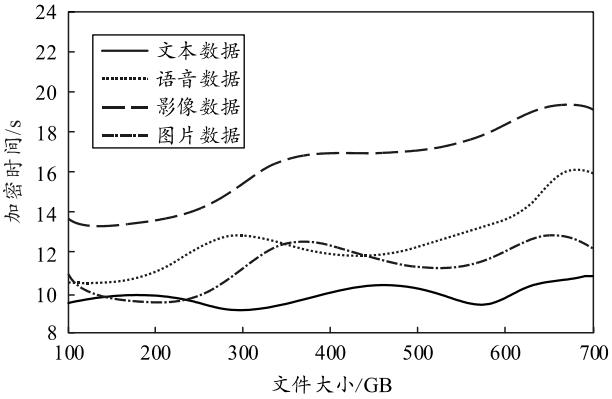
为各节点配置 Intel Core i3 2310M 的 CPU，1 GB 的内存与 100 GB 的硬盘。

2.2　实验数据选用

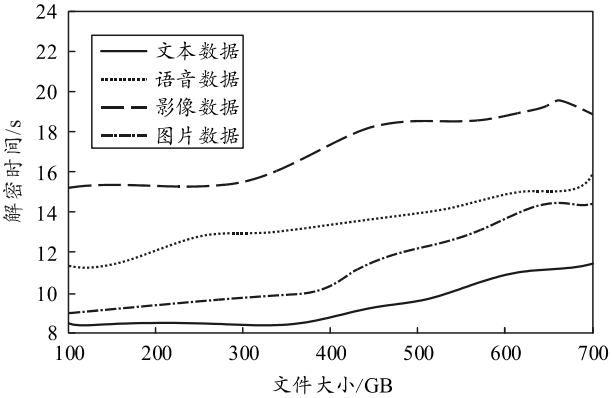
在测试中选用部分文本数据、图片数据、语音数据以及影像数据作为实验所用的大数据，以测试设计方法的存储性能。

2.3　性能测试结果分析

首先对 4 种不同类型大数据的加密时间与解密时间进行测试，测试结果如图 3 所示。



(a) 加密时间测试结果



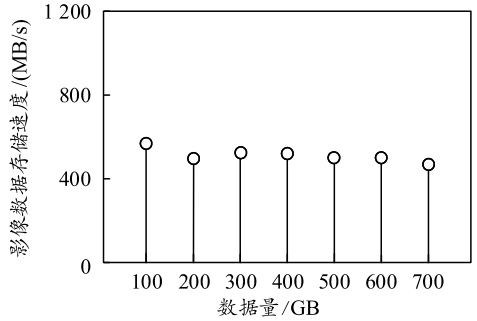
(b) 解密时间测试结果

图 3　大数据加密时间与解密时间测试结果

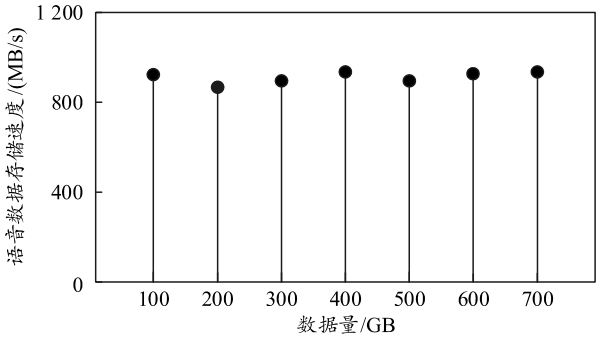
从上图所示各种大数据的加密时间与解密时间测试结果中可看出，影像数据与语音数据的加密与解密处理需耗费较长时间，图像数据与文本数据的加密与解密所耗费的时间相对较短。

综合所有测试数据可发现，设计方法的加密时间与解密时间均较短，所耗费的时间均低于 20 s，证明了设计方法的加解密效率较高。

接着对设计方法的存储速度进行测试，分别测试影像数据与语音数据的存储速度，观察设计方法的存储性能。测试结果如图 4 所示。



(a) 影像数据



(b) 语音数据

图 4 存储速度测试数据

上图存储速度测试结果表明，设计方法的影像数据存储速度高于 580 MB/s，语音数据存储速度高于 916 MB/s，表明其存储速度较快，能够满足海量大数据的存储要求。

最后测试设计方法在存储时的通信开销，分别为测试图片数据、语音数据以及影像数据不同数据量下的通信开销，测试结果如表 1 所示。

表 1 3 种数据不同数据量下的通信开销测试结果

数据量/TB	通信开销/(bits•10^2)		
	图片数据	语音数据	影像数据
5	2.3	3.2	4.9
10	2.5	3.5	5.6
15	2.8	3.8	6.0
20	2.9	4.1	6.5
25	3.2	4.5	7.0
30	3.4	4.8	7.4
35	3.7	5.0	7.8
40	3.9	5.2	8.2

根据上表可知，影像数据的通信开销最大，语音数据的通信开销居中，图片数据的通信开销最低，与各种大数据的特点相符合。综合全部测试数据，设计方法的整体通信开销较低，说明设计方法在较高的存储效率下有着较低的通信开支，证明了设计方法的节能性。

3 结束语

笔者设计一种基于逆向运算法的海量大数据安全存储方法，实现了海量大数据的安全、高效存储。今后，将对该方法进行扩展，争取取得更系统化的研究成果。

参考文献：

[1] 张青凤. 基于大数据的高密度信息安全存储系统设计[J]. 现代电子技术, 2020, 43(14): 83-85.

[2] 谷保平, 马建红. 可撤销属性加密结合快速密度聚类算法的非结构化大数据安全存储方法[J]. 计算机应用与软件, 2021, 38(5): 337-343.

[3] 陈宇翔, 郝尧, 赵越, 等. 面向制造大数据的安全存储交换技术[J]. 电子技术应用, 2019, 45(12): 38-41, 46.

[4] 段平. 基于区块链的非结构化大数据动态安全存储[J]. 吉林大学学报(信息科学版), 2020, 38(5): 595-600.

[5] 曹布阳, 冯华森, 梁峻浩, 等. 利用 Hilbert 曲线与 Cassandra 技术实现时空大数据存储与索引[J]. 武汉大学学报(信息科学版), 2021, 46(5): 620-629.

[6] 申东凡, 杨庚. 面向隐私保护的异构数据库集成中间件系统[J]. 计算机技术与发展, 2020, 30(1): 99-105.

[7] 赵英豪, 吕亮, 徐青, 等. 一种面向海量时空数据的多维检索策略[J]. 测绘科学, 2020, 45(6): 199-204.

[8] 李劭, 黄诚. 云计算下分布式大数据智能存储算法仿真[J]. 计算机仿真, 2020, 37(5): 443-447.

[9] 徐敏, 胡聪, 王萍, 等. 基于云计算技术的大规模数据存储策略研究[J]. 微型电脑应用, 2022, 38(4): 80-83, 92.

[10] 王向华, 刘颖. 基于改进索引散列树的动态大数据审核方法[J]. 计算机应用与软件, 2019, 36(1): 302-307, 333.

[11] 吴林, 吴超, 吴娥. 大数据视域下安全信息资源管理模式研究[J]. 科技管理研究, 2020, 40(9): 156-162.

[12] 王黎, 吕殿基. 基于 Spark 框架的大数据局部频繁项集挖掘算法设计[J]. 微型电脑应用, 2021, 37(4): 130-132, 136.

[13] 郎为民, 马卫国, 张寅, 等. 一种支持数据所有权动态管理的数据去重方案[J]. 信息网络安全, 2020, 20(6): 1-9.

[14] 唐桂文, 韩嘉福, 李洪省. 面向空间大数据的分布式存储策略[J]. 计算机技术与发展, 2019, 29(3): 194-197.

[15] 郭威, 谢光伟, 张帆, 等. 一种分布式存储系统拟态化架构设计与实现[J]. 计算机工程, 2020, 46(6): 12-19.

[16] 张明辉, 张劲波. 无线通信链路数据的分布式融合存储系统设计[J]. 微型电脑应用, 2021, 37(5): 106-109, 112.