文章编号:1671-4598(2014)08-2629-03 中图分类号:TP393 文献标识码:A

基于 Hadoop 和双密钥的云计算数据安全存储策略设计

涂云杰，白 杨
（呼伦贝尔学院 计算机科学与技术学院，内蒙古 海拉尔 021008）

摘要：针对原有的 Hadoop 平台仅通过 CRC－32 循环冗余校验保证数据存储的安全性，设计了一种基于双密钥和混沌信号的云计算安全存储策略；首先，介绍了原有的 Hadoop 框架下的数据存储对应的文件读写过程，并基于加密机制设计了改进的 Hadoop 数据存储模型，然后根据云存储数据量大和响应要求及时的特点，设计了一种基于双密钥的改进对称密钥算法，在传统的私钥的基础上加入动态公钥，并作为敏感函数的输入获得最终的密钥，从而实现明文的加密和密文的解密，最终定义了具体的基于 Hadoop 和改进双密钥对称加密算法的云计算安全存储算法；通过搭建 Hadoop 仿真实验平台进行实验，结果表明文中方法能有效地实现云计算环境下的安全存储，存储时间与其它方法相比少 15％以上，具有安全性高和存储效率高的优点，具有一定的优越性。

关键词：云计算；数据存储；安全策略；密钥

Design of Cloud Computing Safe Storage Strategy Based on Hadoop and Double Key

Tu Yunjie，Bai Yang
（College of Computer Science and Technology，Hulunbeier College，Hailaer 021008，China）

Abstract：Aiming at given Hadoop platform only considers the safety of CRC cyclic redundancy check can guarantee the safety of data storage，a cloud safe storage strategy based on double keys and chaos signal is proposed. Firstly，the file reading and writing process in the framework of Hadoop is described，then the improved Hadoop model based on encryption mechanism was designed，and according to the big amount of cloud storage and in time response demand，an improved symmetric key algorithm based on double key is designed，the introduced public key and the personal key are as the input of the sensitive function to implement of the text encryption. Finally，the specific algorism based on Hadoop and improved double key symmetric algorithm is defined. The simulation is operated in the Hadoop simulation platform，the result shows the method in this paper can effectively realize the safe storage in cloud environment，and the comparing time has shorted 15％ compared with the other methods，so it has the properties of high safety and high storage efficiency，so it has some priority.

Keywords：cloud computing；data storage；safe strategy；key

0 引言

传统软件企业需要购买新的硬件设备和对软件进行升级[1]，而业务量减少时部分基础设施硬件会闲置而造成浪费。云计算[2-3]（Cloud Computing）应运而生，云存储[4-5]即企业和个人用户将数据和服务交给第三方的云服务提供商，由云服务提供商对数据和服务进行存储、发布和维护，使得用户仅需要花费较小的代价就能弥补软硬件资源的不足，并能有效地防止数据的丢失、设备损坏和移动性弱等方面的问题。

文献[6]设计了一种基于属性加密体制的云存储模型，通过将模型分为加密、存储和解密 3 个阶段以充分保障模型健壮性。文献[7-8]设计了一种分布式纠删码的安全云存储模型，通过引入纠删码技术提高模型的健壮性，能有效地规模云存储的安全风险。文献[9]建立了一种存储安全性好且效率高的存储策略，能综合利用对称加密算和非对称加密算法实现对数据的安全存储。

上述工作均研究了云计算环境下的安全存储问题，往往是独立建立一种模型并用过建立密钥实现安全存储，没有充分利用已有的平台和安全性难以有效实施的问题，为此，本文设计了一种基于 Hadoop 模型和改进对称密钥算法的云存储安全策略，实现证明了文中方法的有效性。

1 Hadoop 模型和 HDFS 数据读写

Hadoop[10]是一个分布式系统的基础构架和开源平台。Hadoop 主要包含两个部分：HDFS（Hadoop Distributed File System）数据存储方式的开源实现和 Map－Reduce 调度方式。

HDFS 是一个云计算环境下的分布式文件系统，具有高度的容错访问性能，可以部署在成本较低的计算上，提供高吞吐量的访问，HDFS 对数据的存储可以分为文件的写入和读出两个阶段。

当客户端有数据要存储到数据节点时，首先计算写入数据的 CRC－32 循环冗余校验和，然后将数据和校验和一起发送到数据节点，数据节点对接收的数据以及其校验码进行存储。当客户端要从数据节点中读取数据时，在读取了数据后，对读取的数据生成 CRC－32 循环冗余校验和，并与从数据节点中读取的校验和进行匹配，以判断数据是否正确。

上述 HDFS 提供了一种简单有效的数据存储安全性访问机制，但由于未对文件进行加密，所以仍存在着安全隐患，如用户

收稿日期：2014-01-19； **修回日期**：2014-03-26。
基金项目：内蒙古自治区高等学校科学技术研究一般项目(NJZY14308)。
作者简介：涂云杰(1975-)，女，河北沧州人，硕士，副教授，主要从事数据安全及应用方向的研究。

认证问题、DataNode 认证问题以及文件存储和传输的问题。

因此，下面设计一种基于加密和 Hadoop 的文件安全存储模型。

2 基于加密的 HDFS 读写访问

2.1 HDFS 写文件

客户端需要写入数据到数据节点中时，可以通过 DistributedFileSystem 对象的 Create（）方法来创建文件，并在命名节点中创建一个以文件名作为关键字的新记录，此时由于文件还没有加入到对应的数据节点，因此，该新记录的内容为空。为了对 HDFS 的文件写入文件进行加密，此时运行加密算法对文件进行 encrypt 操作，从而获得存储明文对应的密文，然后计算密文对应的校验和，此时，从 DistributedFileSystem 对象获取文件输出流 FSdataOutputStream，通过文件输出流将数据进行分包，并加入等待队列，等待队列中的数据包按照先后顺序发送到数据节点集中的数据节点中，其中校验包发送到最后的一个数据节点中，并将数据包的存储位置，其文件写入过程如图 1 所示。

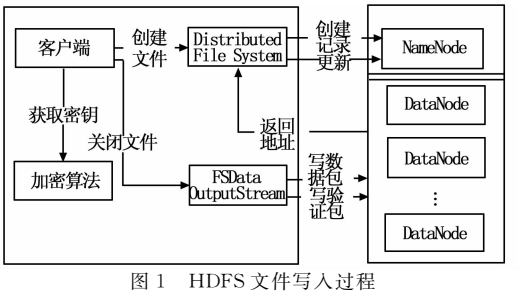

图 1 HDFS 文件写入过程

2.2 HDFS 读文件

HDFS 读取文件的过程与写文件过程类似，客户端通过调用 DistributedFileSystem 对象的 open（）方法打开文件，并通过 DistributedFileSystem 对象调用命名节点，以获取文件的存储位置。客户端通过反复地读取数据节点中的数据，从而完成对数据节点中数据的读取，当客户端获取了全部的密文和校验和后，重新对密文生成校验文件，并与读取的校验和进行比较，同时通过解密算法对密文进行解密，获取明文。

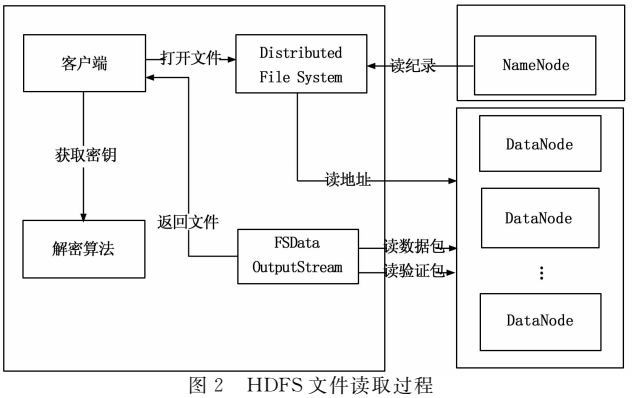

图 2 HDFS 文件读取过程

3 基于双密钥和混沌的对称加密算法

传统的加密技术可以分为两类：对称加密算法和非对称加密算法，对称加密算法中通信双方采用相同的密钥，具有加密速度快、硬件实现简单和安全程度较高等优点，而非对称的加密算法则在加密和解密阶段采用不同的密钥，密钥长度较对称密钥大，同时加密效率不高，主要用于身份认证和数据签名等领域。

由于云计算中心的数据量大，同时数据加密要求效率高，因此，设计了一种基于双密钥的对称密钥算法。

3.1 基于双密钥的文件加密

双密钥的文件加密过程在传统的对称加密算法的仅有一个私钥的基础上，加入了一个公开的动态公钥，并将私钥和公钥均作为加密敏感函数的输入，产生一个随机变化的密钥流，从而使得明文和密文之间的相关性随着时间不断变化。私钥可以通过专用通道或人工协商方式来进行交换，而公钥则由客户端在对其指定后存入命名节点。

明文加密变为密文的过程为可以描述为客户端在指定动态密钥后，向命名节点发送动态密钥，然后将动态密钥 $k2$ 和私钥 $k1$ 输入加密敏感函数，获得最终的密钥，并将该密钥对明文进行加密，获得密文，如图 3 所示。

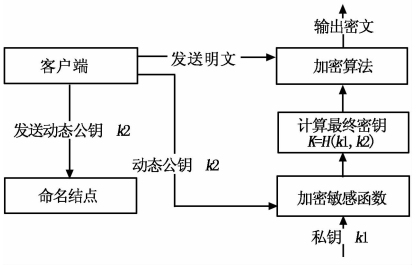

图 3 明文加密过程

3.2 基于双密钥的文件解密

客户端从数据节点中获取密文，并从命名节点中读取记录获得动态公钥 $k2$，然后根据动态公钥 $k2$ 和私钥 $k1$，采用解密敏感函数获得最终的密钥 k，再通过解密算法对密文进行解密，从而得到明文，其过程如图 4 所示。

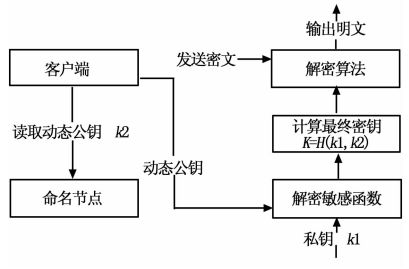

图 4 密文解密过程

3.3 基于 Lorenz 混沌的敏感函数选取

Lorenz 混沌系统是由美国 Lorenz 于 1963 年首次提出，该系统对初始条件敏感，当初始条件变化为 10^{-15} 时，系统运行轨迹偏差为 -30^{-30} 即为初始条件误差的 10^{16}，加解密敏感函数可以设计为：

$$H(k1,k2) = \mathrm{mod}(abs(round(F_{k1,\tau(t)} \times k2)),256) \quad (1)$$

在式（1）中，$k1$ 为混沌系统初始条件，采用私钥对其初始化，$k2$ 为动态密钥，$F_{k1,\tau(t)}$ 为采样后的混沌信号，mod 为求余操作，*abs* 为求绝对值运算，*round* 是四舍五入运算。

从式（1）可以看出，加解密敏感函数 H 的值域为［0，255］的整数。

加密算法可以通过加解密敏感函数 H 获得的最终值与明文进行异或操作得到，即：

$$C = K \oplus M \quad (2)$$

其中，在式（2）中，K 为经过加解密敏感函数 H 获得的

最终密钥，M 为密文。

4　基于 Hadoop 和混沌双密钥的加解密算法描述

基于 Hadoop 和混沌双密钥的加解密算法可以描述为：

当客户端写入文件和读入过程是相互逆过程，因此仅对客户端写入文件的过程进行描述如下：

1）客户端创建 DistributedFileSystem 对象，并通过 Create() 方法来创建文件，在命名节点中创建以文件名作为关键字的记录，记录格式如表 1 所示。

表 1　文件记录

关键字(文件名)	文件地址	动态公钥 k2

2）客户端采用私钥 $k1$ 和 Lorenz 混沌系统产生长度足以对明文加密的混沌信号 $F_{k1}(t)$；

3）客户端指定动态公钥 $k2$，采用规则对混沌信号 $F_{k1}(t)$ 进行采样处理，获得用于最终加密的混沌信号 $F_{k1,\tau(t)}$；

4）客户端以动态公钥 $k2$ 和混沌信号 $F_{k1,\tau(t)}$ 作为输入，通过加密敏感函数求得密钥序列 K；

5）客户端通过加密算法即异或操作来得到密文；

6）将密文写入由 DistributedFileSystem 对象创建的文件中，并对其计算 CRC－32 循环冗余校验和；

7）FSdataOutputStream 流将密文和校验和进行分包处理，加入等待队列，等待队列中的数据包按照先后顺序发送到数据节点集中的数据节点中，其中校验包发送到最后的一个数据节点中；

8）客户端调用 DistributedFileSystem 对象将密文和校验和对应的文件地址发送到命名节点中。

5　仿真实验

为了对文中方法进行验证，搭建 Hadoop 集群，集群中设置 5 台计算机作为存储节点，每个存储节点配置相同，设置如下：内存为 4 G，硬盘为 250 G，CPU 采用 Intel 双核 2.8 Ghz，操作系统为 Windows7.0，采用 Cygwin 来虚拟 Linux 环境，并搭建 Hadoop 开源平台，采用 HBASE 数据库，以某公司的数据为例进行仿真实验。

为了进一步说明文中方法的总体性能，以存储耗时为标准对文中方法进行验证，在不同的数据文件大小情况下进行测试，即：2 M，4 M，15 M，30 M，80 M，200 M，500 M，800 M，客户端在写入文件时，通过用文中方法对明文进行双密钥加密，然后再进行 RC－32 循环冗余校验，对密文和校验码写入后，再通过逆向的解密过程对其解密，得到了完全相同的数据，同时在加入各类恶意攻击，得到的实验结果如表 2 所示。

从表 2 中可以看出，文中设计的存储安全策略中的数据包在解密后数据仍然能完整地获取，这说明了文中方法不仅能保证数据的完整性和正确性，同时文中方法在显著地改善了存储安全性的同时，没有显著地增加时间开销，因此，具有很强可行性。

将文中方法的存储时间与文献［7］和文献［8］进行比较，结果如图 4 所示。

从图 4 中可以看出，文中方法在存储不同大小的数据包时对应的存储时间均远低于另外两种方法，较其分别高 15.7% 和 30.04%，显然，文中方法更优。

表 2　存储耗时和完整性仿真

序号	文件大小/M	直接存储时间/ms	安全存储时间/ms	完整性/(%)
1	2	321	356	100%
2	4	612	664	100%
3	15	1 355	1 456	100%
4	30	3 567	3 895	100%
5	80	7 831	7 972	100%
6	200	11 043	11 546	100%
7	500	13 535	14 867	100%
8	800	67 123	76 465	100%

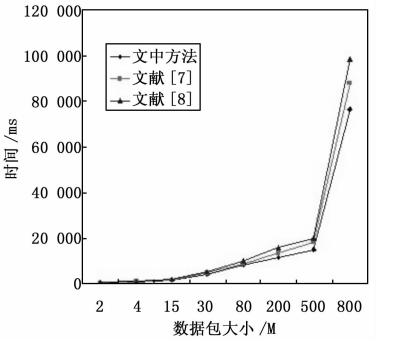

图 4　存储时间对比

6　结论

本文研究了基于云计算 Hadoop 开源平台和双密钥的存储安全策略，在传统的 Hadoop 开源平台 CRC－32 循环冗余校验方式的基础上，建立了一种双密钥和混沌信号的明文加解密算法，并对算法的总体流程进行了描述。仿真实验表明了文中方法不仅能保证文件读写的完整性，同时还能保证文件存储的高效性，是一种具有很强可行性的方法。

参考文献：

[1] 张建华，吴　恒，张文博. 云计算核心技术研究综述［J］. 小型微型计算机系统，2013，11（34）：2417－2424.

[2] Vaquero L，Rodero Marino L，Cacerce J，et al. A break in the clouds：towards a cloud definition［J］. SIGCOMM Computer Communication Review，2009，39（1）：50－55.

[3] 王意洁，孙伟东，周　松，等. 云计算环境下的分布存储关键技术［J］. 软件学报，2012，23（4）：962－986.

[4] Armbrust M，Fox A，Griffith R，et al. Above the clouds：A Berkeley view of cloud computing，USB－EECS－2009－28［R］. Berkeley：University of California，2009.

[5] 程芳权，彭智勇，宋　伟，等. 可信云存储环境下支持访问控制的密钥管理［J］. 计算机研究与发展，2013，50（8）：1613－1627.

[6] 吴胜艳，许　力，林昌露. 基于门限属性加密的安全分布式云存储模型［J］. 计算机应用，2013，33（7）：1880－1884.

[7] Lin H Y，Tzeng W G. A secure decentralized erasure code fordistributed network storage［J］. IEEE Transactions on Parallel and Distributed Systems，2010，21（11）：1586－1594.

[8] 肖长水，姒茂新，傅颖丽，等. 自认证可信云存储框架与算法设计［J］. 计算机工程与设计，2013，10（34）：3459－3464.

[9] 胡光永. 基于云计算的数据安全存储策略研究［J］. 计算机测量与控制，2011，19（10）：2539－2542.

[10] Shvachko K，Kuang Hairong，Rania S，et al. The Hadoop distributed file System［A］. Proc. of Mass Storage Systems and Technologies［C］. 2010：1－10.