文章编号:1671-4598(2015)07-2475-03 DOI:10.16526/j.cnki.11—4762/tp.2015.07.069 中图分类号:TP393 文献标识码:A

基于双混沌系统的大数据环境并行加密算法设计

司红伟，钟国韵
(东华理工大学 理学院，南昌 330013)

摘要：为了克服大数据在采用串行加密方式时具有的加密效率低的问题，设计了一种基于双混沌系统的大数据环境的并行加密算法；首先，设计了基于 Map—Reduce 的大数据环境的并行加密模型；然后，引入了改进的 Logistic 映射和 Tent 映射构成双混沌系统，并设计了 Map 函数、Sort 函数和 Reduce 函数实现并行加密，在 Map 函数中通过 Logistic 映射和 Tent 映射的不断迭代计算加密密钥或解密密钥，在 Sort 函数对由 Map 函数输出的键值对进行排序并剔除重复的数据块，在 Reduce 函数中对加密后的密文数据块或解密后的明文数据块进一步合并构成输出数据；仿真实验表明：文中设计的基于双混沌系统的 Map—Reduce 并行加密模型能高效地进行数据加密或解密，能提高数据安全性和加密效率，具有较强的可行性。

关键词：混沌系统；并行加密；大数据环境；Map—Reduce

Design of Parallel Encryption Algorithm for Big Data Environment Based on Double Chaos System

Si Hongwei, Zhong Guoyun
(School of Science, East China Institute of Technology, Nanchang 330013, China)

Abstract: In order to solve the problem of using the serial encryption having the problem of low efficiency, a parallel encryption algorithm is proposed based on double chaos system. Firstly, the parallel encryption model based on Map—Reduce for big data environment is designed. Then the parallel chaos system based on Logistic mapping and Tent mapping is designed, and the Map function, sort function and Reduce function are all designed, the Logistic mapping and Tent mapping is iterated to compute the encryption and decryption key in the Map function, and the data from Map function is merged to obtain the output data in the Reduce function, and the iteration initial value is computed and stored in the history data information for Logistic mapping and Tent mapping. The simulation experiment shows that the parallel model Map—Reduce model designed in this paper can achieve the encryption and decryption for big data, and it can improve the safety for data and enlarge the encryption efficiency with strong feasibility.

Keywords: chaos system; parallel encryption; big data environment; Map—Reduce

0 引言

随着信息技术、物联网和云计算技术的发展，大数据(Big Data)[1-3]越来越吸引人们的视线，IBM 预言：到 2020 年，全球产生的数据总量将超过今天的 44 倍[4-5]。大数据包含了海量数据，同时这些数据是复杂类型的数据。

大数据的特征可以表示为[6-7]：1) 数据体量大；2) 数据种类多；3) 信息价值大；4) 数据处理速度快。具有以上 4 个特征的大数据往往含有较多的敏感信息，为了保证大数据的安全性，应从多方面进行安全防护。

传统的串行数据加密算法难以实现海量数据的加密，要对大数据进行加密需要进行并行化操作。由于云计算中心具有大量计算节点，因此，可以用于存储和处理海量数据，Hadoop[8]是云计算中心的一个并行处理框架，将数据加密算法与 Hadoop 结合在一起能实现大数据的并行加密。

为了实现大数据并行加密，文中设计了一种基于 Hadoop 和双混沌系统的并行加密算法，并通过实验证明了其有效性。

收稿日期:2015-02-06； **修回日期**:2015-05-20。
基金项目:国家自然基金项目(61402102)。
作者简介:司红伟(1980-)，男，甘肃渭源人，硕士，讲师，研究方向：计算机网络与分布式数据库。

1 并行加密模型

本文通过采用 Map—Reduce 模型来采用将基于双混沌映射的加密和解密过程分解为对 Map 函数、Sort 函数和 Reduce 函数的设计，从而利用 Hadoop 平台进行数据并行加密和解密过程。

1.1 Hadoop 模型

Hadoop 是一个分布式计算平台，具有可靠性高和容错性强的优点，用户可以在不了解平台细节的前提下，利用平台提供的接口实现分布地并行存储和计算，其框架如图 1 所示。

从图 1 中可以看出，Hadoop 平台主要包括 HDFS 文件系统（Hadoop Distributed File system，HDFS）、Map—Reduce 分布式计算模型、列式数据库（HBase）、Hive（SQL 解析引擎）和分布式应用程序协调系统（Zookeeper）组合。HDFS 主要负责集群中各节点 DataNode 数据的存取。

1.2 大数据加密的 Map—Reduce 并行模型

一个典型的 Map—Reduce 模型可以表示为如图 2 所示。

从图 2 可以看出，大数据加密/解密作业的处理过程可以描述如下。

1) 大数据分割：将需要处理的加密或解密作业大数据集分割成若干独立的数据块，每个数据块用键值对＜key,$value$＞的形式进行存储。

2) Input 输入阶段：通过在 Map 任务中指定输入位置、

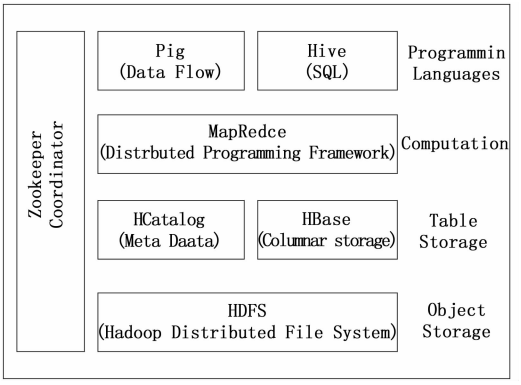

图 1 Hadoop 框架模型

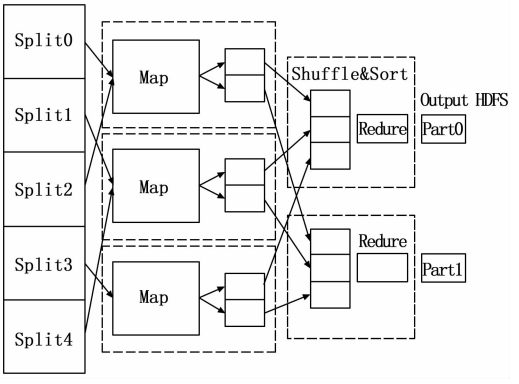

图 2 Map－Reduce 计算模型

输出位置和一些运行参数，将输入目录下的各分割加密或解密数据块读入，读入的格式为＜*key*，*value*＞。

3）Map 阶段：根据自定义的加密 Map 函数和解密 Map 函数进行 Map 操作，生成中间结果对应的键值对集，即＜*key*，*value*＞集，这组键值对的类型与输入键值对具有不同的类型。

4）Sort 阶段：此过程的主要任务就是对 Map 操作的输出结果的＜*key*，*value*＞集进行排序，将具有相同的键值 *key* 的中间结果尽可能地交由同一个 Reduce 函数处理。在 Shuffle 阶段，主要完成混排交换数据；在 Sort 阶段，模型根据键值 *key* 对＜*key*，*value*＞集进行排序，将相同 *key* 的中间结果尽可能地汇集到同一个节点上。

5）Reduce 阶段：此过程将遍历由 Shuffle & Sort 阶段产生的中间数据，对每个不同的 *key* 执行用户自定义的 Reduce 函数，Reduce 函数的输入为＜*key*，(list of values)＞，输出为＜*key*，*value*＞。

6）Output 阶段：将 Reduce 函数输出的结果写入到输出目录文件中。

2 基于 Logistic 和 Tent 映射的双混沌系统

2.1 双混沌加密原理

混沌[9-10]是一种无规则的运动，是一种非线性系统中的确定和抽象的随机过程，其具有参数敏感、有界和遍历性等特征，非常适合于数据安全领域的加密操作。

传统的单混沌系统具有较好的加密速度，但其密钥空间小，算法简单且安全性不高，因此，设计一种基于一维 Logistic 映射和 Tent 映射的双混沌系统方法，在指定初始值和参数的情况下进行不断迭代，并通过相互反馈产生子密钥序列对明文进行加密，双混沌系统不仅可以有效增加密钥空间，增强算法的鲁棒性，同时还能利用反馈生成的子密钥，由于其与明文和密文都相关，因此可以用于增加信息加密的安全性。

2.2 改进的 Logistic 映射

Logistic 映射作为一种非常简单和典型的一维映射，其动力学方程可以描述为：

$$x_{n+1} = u * x_n * (1 - x_n) \tag{1}$$

在公式（1）中，u 为控制参数，对于任意的初值 $x_0 \in (0, 1)$，当确定了 u 的值可以迭代出一个确定时间序列：

1）当 $0 < u \leqslant 1$ 时，除了不动点 $x_0 = 0$ 外，无其他周期点，$x_0 = 0$ 也被称为吸引不动点。

2）当 $1 < u \leqslant 3.5699456\cdots$ 时，系统的动力学形态较为简单，周期点有两个为：0 和 $1 - 1/u$，0 为排斥的不动点。

3）当 $3 \leqslant u \leqslant 4$ 时，系统由倍周期向混沌状态转换，系统动力学形态较为复杂。

4）当 $u > 4$ 时，系统动力学形态更为复杂。

由于函数在固定点处的斜率大于或等于 1 时，固定点具有不稳定性，同时当函数在不可忽略的固定点上的梯度值越大，其混沌性能越好，因此，设计改进的 Logistic 映射。

$$x_{n+1} = \beta\left[1 + \frac{1}{\beta}\right]^{\beta} * x_n * (1 - x_n)^{\beta} \tag{2}$$

在公式（2）中，β 表示［1，4］之间的一个常数，x_0 的值满足 $x_0 \in (0,1)$。

2.3 Tent 映射

Tent 映射可以被称作帐篷映射，其动力学方程可以描述为：

$$y_{n+1} = \begin{cases} \lambda * y_n, & 0 < y_n \leqslant 0.5 \\ \lambda * (1 - y_n), & 0.5 < y_n < 1 \end{cases} \tag{3}$$

由于 Tent 映射具有混沌特性，因此，可以通过参数 λ 取不同的值可以得到当 λ 的值为 1.4 和 2 之间时，Tent 映射能进入完全混沌状态。

3 基于双混沌的大数据加密并行算法设计

为了实现基于双混沌的大数据并行加密算法，在每次对大数据进行加密和解密前，将大数据分为若干数据块，每次数据块加密和解密的迭代过程可以通过 Map－Reduce 的并行化操作实现。

3.1 加密和解密 Map 函数设计

在 Map 阶段函数中，其输入为数据集中键值对集组成的键值对集合，即＜key1，value1＞集，输出为＜key2，value2＞，加密和解过程的 Map 函数中均包含下列两个函数：

$$key2 = RepresentIndividual(key1) \tag{4}$$

$$value2 = Encyl(value1) \tag{5}$$

$$value2 = Dncyl(value1) \tag{6}$$

其中，$RepresentIndividual(key1)$ 对应了数据块编号为 $key1$ 的数据块经过映射函数 $RepresentIndividual$，将其映射到键值为 $key2$ 的数据块，当 $value1$ 为明文时，调用公式（5）所示的加密函数 $Encyl$ 进行加密操作，当 $value1$ 为密文时，调用公式（6）所示的解密函数 $Dncyl$ 进行解密操作。

加密函数 $Encyl$ 和解密函数 $Dncyl$ 依次调用下面几个函数

进行加密操作：

1）调用 $justify(\beta,\lambda)$ 对 Logistic 映射和 Tent 映射的参数 β 和 λ 是否满足规定的要求进行判断，当不满足时，重新从历史 Logistic 映射的初始迭代次数集合和 Tent 映射的初始迭代次数集合中选择两个参数直到满足条件为止；

2）调用 $logistic(100,b)$ 进行第一次混沌运算：首先采用 Logistic 映射迭代 100 次，使其进入完全混沌的状态，然后再此基础上，再迭代 $b = b * 11$ 次，此时可以得到其输出为 x_b；

3）调用 $k1 = key1(x_b)$ 求取加密密钥或解密密钥 $k1$：$k1 = key1(x_b)$ 将 Logistic 映射输出的 x_b 的后三位赋值给 $key1$，然后将其对 256 取模从而得加密密钥或解密密钥 $k1$；

4）调用 $k2 = key2(x_b)$ 计算 $k2$：$k2 = key2(x_b)$ 表示对 x_b 的小数点的后四、五和六位赋值给 $key2$，然后对其将其对 256 取模从而得密钥 $k2$；

5）调用 $Tent(x_b,b)$ 进行第二次混沌运算：将 Logistic 映射的输出作为 $Tent(x_b)$ 映射的输入，从而进入 Tent 映射，对 Tent 映射共迭代 b 次，得到输出值为 y_b；

6）调用 $k3 = key3(y_b)$ 计算 $k3$：$k3 = key3(x_b)$ 表示对 y_b 的小数点的后四、五和六位赋值给 $key3$，然后对其将其对 256 取模从而得密钥 $k3$；

7）调用 $value2 = f1(k1)$ 计算密文或调用 $value2 = f2(k1)$ 计算明文：将明文与读取的加密密钥 $k1$ 进行异或操作得到密文，将密文与读取的解密密钥 $k1$ 进行异或操作得到明文。

8）调用 $value2 = attach(k2,k3)$ 将密钥 $k2$ 和 $k3$ 附加在明文或密文 $value2$ 后，作为新的 $value2$。

经过上述过程，可以对划分的数据块进行并行的加密和解密，得到对应的加密数据块或解密数据块，然后进入 Sort 函数进行相应的处理。

3.2 Sort 函数设计

Sort 函数的输入为由 Map 函数输出得到的无序键值对集＜key2，value2＞，输出为有序的键值对集＜key3，value3＞，其具体的操作过程为：

1）调用键值映射函数生成新键值 key3 即根据 $RepresentIndividual(key2_1,...)$ 生成多个键值对应的唯一新键值 key3；

2）收集由 Map 函数输出得到的无序键值对 list {＜key2，value2＞}，构成有序的键值对列表 list1 {＜key2，value2＞}；

3）对有序键值对列表 list1 {＜key2，value2＞} 中的重复的键值进行删除，得到无重复元素的序列键值对列表 list2 {＜key2，value2＞}；

4）将 list2 {＜key2，value2＞} 赋给 value3；

3.3 Reduce 函数设计

Reduce 函数的输入为由 Sort 函数输出得到的键值对＜key3，value3＞，然后输出＜key4，value4 ＞，输出为明文或密文大数据，其具体的操作过程为：

Reduce 函数的具体操作过程如下：

1）提取由 Sort 函数输出得到的键值对＜key3，value3＞，将 value3 这个有序列表中的所有密钥数据块和明文数据块提取出来；

2）将每个密钥数据块或明文数据块中的明文或密文与对应的密钥 $k2$ 和 $k3$ 分离；

3）将密文与 $k2$ 进行异或得到新的 Logistic 映射初始迭代次数 b，将明文与 $k3$ 进行异或得到 Tent 映射对应的新的迭代次数 b'；

4）将 b 和 b' 添加到历史 Logistic 映射的初始迭代次数集合和 Tent 映射的初始迭代次数集合中。

5）将分离了密钥 $k2$ 和 $k3$ 的明文或密文前后连接组成新的大数据集，作为最终加密或解密的大数据，从而实现整个大数据的并行加密或解密。

4　仿真实验

为了对文中方法进行验证，构建实验环境，硬件服务器平台为 Dell PowerEdge R910，软件平台为 Hadoop，服务器中包含 24 核 Intel Xeon 处理器，内存 200G，以加密算法为例，随机输入一组由于明文构成的数据集，明文的类型包含文本、图片、音频和视频数据，采用文中设计的基于 Map－Reduce 并行框架的模型进行明文的加密，基于统计学方法得到的类型为文本的明文和加密后的密文数据分别如图 3 和图 4 所示。

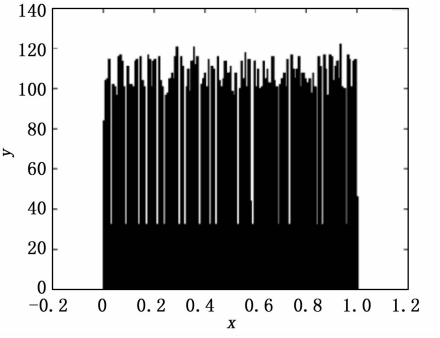

图 3　文本类型的明文数据

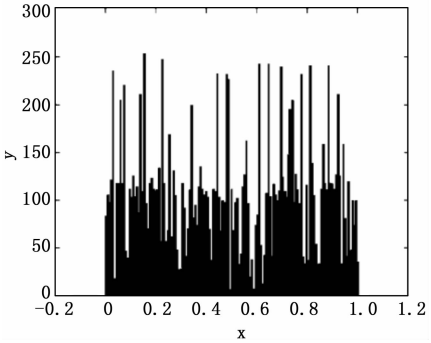

图 4　文本类型的密文数据

从图 3 和图 4 中可以发现，类型为文本的明文数据在进行了加密后，其统计学特性发生了较大的变化，具有较好的加密效果，同时，由于文中的加密算法或解密算法仅选用两个控制参数和初始化两个混沌迭代次数，因此，具有较大的密钥空间，即使攻击者在已知算法中应用了 Logistic 映射和 Tent 映射，也无法通过穷举来获得这 4 个初始参数，因此，该加密算法具有较好的安全性。

为了验证文中基于双混沌系统的并行加密算法的加密性能，将文中方法与串行的数据加密程序进行对比，得到的结果如表 1 所示。

（下转第 2481 页）

统中，分析人员可以立即下载进行分析处理并将前一晚的数据传输给 NOAA 总部以及其他的异地分析人员。我国在试验数据管理网络建设的远程同步试验数据应用模式上还很不完善。

4　总结

随着空间环境模拟设备试验数据管理要求的不断增高，以及 IT 技术的发展，新的测试技术和测试方法的不断涌现，空间环境模拟设备试验数据管理必将朝着数字化、网络化、综合化方向发展[1]，从解决大量试验数据、试验文档的保存和检索问题转变为到对试验计划内容、参数数据、相关文档、试验报告、测试人员和设备信息等一系列相关数据的管理。本文提出的一些建议，为空间环境模拟器试验系统数据管理提供借鉴与参考。

参考文献：

[1] 王素丽. 基于 Web 技术的试验数据管理系统的研究与应用 [D]. 洛阳：河南科技大学，2008.

[2] Duprat，Raymond；Mouton，Andre. INTESPACE's new thermal－vacuum test facility：SIMMER，NASA/N93－15613 [R].

[3] http：//www.ipo.nasa.gov. NPP Pre－Launch Test Data Collection and Archive [EB/OL].

[4] Guijt H，Popovitch A. STAMP：A new data acquisition system for ESA' S large space simulator [A] . Proceedings of the 5th International Symposium on Environmental Testing for Space Programmes [C]. 2004.

[5] 李　娜，刘劲松，顾　苗. 基于吉时利 3706 的真空热试验数据采集系统 [A]. 空间环境与材料科学论坛论文集 [C] . 2009.

（上接第 2477 页）

表 1　串行和并行运行时比较

数据总量/G	处理节点个数/个	Map /s	Sort /s	Reduce /s	串行时间 /s	文中方法 /s
0.2	10	0.04	0.01	0.02	0.02	0.09
1.6	15	0.25	0.07	0.13	0.2	0.45
2.9	20	1.34	0.15	0.45	0.43	3.56
4	25	2.56	0.53	2.14	1.13	5.64
6.4	32	5.32	0.74	3.35	3.53	8.75
16.7	61	5.93	0.85	4.13	9.35	9.13
26.2	83	6.43	0.94	4.24	14.53	10.34
45.2	94	7.32	1.45	5.03	25.25	15.59
81.3	104	7.93	2.53	5.34	46.32	16.52
105.2	117	8.47	3.24	6.27	64.32	20.05
156.4	127	14.43	3.57	9.60	153.59	29.61
210.3	143	17.73	3.86	15.42	253.25	39.10
302.1	158	26.35	6.43	24	454.38	58.56
345.1	210	37.26	7.64	31.56	590.67	74.25
421.4	240	73.24	8.54	61.35	934.29	157.34
531.3	250	125.33	9.43	115.43	1 345.24	264.22
612.3	250	167.35	11.34	156.35	超出内存	345.22
724.2	250	185.24	13.20	164.63		407.43
843.7	250	201.53	15.64	178.33		426.34
943.5	250	232.53	17.53	187.64		562.06

从表 1 中可以看出，文中方法在算法处理数据的过程中，具有很好的自适应性，在数据量小于 16.7 之前，文中方法虽然落后于传统的串行方法，但是在随着数据量的进一步增加，文中方法在处理的大数据方面的并行性能的优势逐渐显现出来，而串行的执行方式虽然在初期具有较少的加密或解密时间，但是在数量大于 16.7G 后，并行方法所需时间较文中方法显著增大，且在数据量为 612.3 时发生了溢出。而文中并行算法增幅较缓慢，在所有数据均已加密执行完毕时花费的总时间也仅为 562.06 s。

5　结论

为了提高云计算数据中心中大数据的安全和并行处理能力，利用云计算 Hadoop 平台提供的 Map－Reduce 模型，实现大数据加密的并行化。设计了改进的 Logistic 映射和 Tent 映射，通过初始化迭代来寻求加密密钥和解密密钥，将双混沌加密系统分解为 Map 函数、Sort 函数和 Reduce 函数，从而实现了大数据的并行加密和解密。仿真实验证明文中设计的基于并行框架的加密和解密算法是一种高效，可靠和通用的模型。

参考文献：

[1] 刘智慧，张泉灵. 大数据技术研究综述 [J]. 浙江大学学报，2014，6 (48)：957－970.

[2] Goldston D. Big data：data wrangling [J/OL]. Nature，2008，455：15.

[3] 宫夏屹，李伯虎，柴旭东，等. 大数据平台技术综述 [J]. 系统仿真学报，2014，3 (26)：489－496.

[4] Executive Office of the president. Designing a future：Federally funded research and development in network and information technology [R]. New York：Executive Office of the President，2010，10.

[5] 涂新莉，刘　波，林伟伟. 大数据研究综述 [J]. 计算机应用研究，2014，6 (31)：1612－1623.

[6] Lu Lin，Liang Yi－wen，Yang He，et al. Danger theory：a new approach in big data analysis [A]. Proc of International conference on automatic control and artificial intelligence [C]. 2012：739－742.

[7] Hsinchun Chen，Roger H L Chiang，Veda C Storey. Business intelligence and analytics：From big data to big impact [J]. MIS Quarterly，2012，36 (11)：1－24.

[8] 郭其标，吕春峰. 基于云计算 Hadoop 异构集群的并行作业调度算法 [J]. 计算机测量与控制，2014，22 (6)：1846－1849.

[9] 王小龙，赵庶旭. 基于分段线性混沌映射的算术编码与加密 [J]. 计算机研究与发展，2014，5 (31)：14811487.

[10] 郎　讯，魏立线，王绪安，等. 基于代理重加密的云存储密文访问控制方案 [J]. 计算机应用，2014，34 (3)：724－727.