(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102968356 A
(43)申请公布日 2013.03.13

(21)申请号 201110456941.2

(22)申请日 2011.12.30

(71)申请人 中国科学院上海光学精密机械研究所
地址 201800 上海市嘉定区 800 － 211 邮政信箱

(72)发明人 刘涛 阮昊

(74)专利代理机构 上海新天专利代理有限公司 31213
代理人 张泽纯

(51)Int.Cl.
G06F 11/14(2006.01)
G06F 11/08(2006.01)

权利要求书 1 页 说明书 3 页

(54)发明名称

云存储系统的数据处理方法

(57)摘要

一种云存储系统的数据处理方法，对云存储系统的数据存入和数据提取分别采用里的所罗门编码数据纠错编码处理和里的所罗门纠错解码处理。本发明提高云存储系统中数据的安全性，可恢复性，降低数据备份数目，可节约数据的存储空间，降低成本。

CN 102968356 A

1. 一种云存储系统的数据处理方法，其特征在于，对云存储系统的数据存入和数据提取分别采用里的所罗门编码数据纠错编码处理和里的所罗门纠错解码处理。

2. 根据权利要求 1 所述的云存储系统的数据处理方法，其特征在于所述的数据存入方法，包括以下具体步骤：

①将待存储的原始数据分成 K 个帧数据，每个帧数据包含相同的固定长度 N 比特的数据，当最后一个原始数据帧的数据长度小于 N 时，对最后一块数据帧填充数据‘0’，以达到长度 N，其中 K 为大于大于 1 的正整数，N 的取值范围为：$200 < N < 2000$；

②对所述的每一个数据帧加上编号，即 ID，得到一个新数据帧，所述的 ID 长为 4 个字节，从 0001 开始递加，故新的数据帧的长度为 (N+4)；

③将所述的新的 K 个数据帧再组合为 W 个数据块，每一个数据块包含 M 个数据帧，形成一个 M*(N+4) 的数据矩阵，当最后一个数据块的数据帧的个数小于 M 时，对该数据块填充‘0’数据帧，以使最后一个数据块达到固定的数据帧 M，其中 M，W 的取值范围分别为：$200 < M < 2000$，$W = K/M$；

④对所述的数据块采用里的所罗门乘积码纠错编码方法进行纠错编码：将第 i 个数据块的行和列分别添加 P_0，P_I 个纠错的冗余数据进行编码，转化成一个包含 $(M+P_0)*(N+4+P_I)$ 个数据的里的所罗门数据块，其中，P_0，P_I 分别为数据块中一行和一列增加的用于纠错的冗余数据个数，且 $0 < P_0 < M/2$，$0 < P_I < M/2)$，$1 < i \leqslant M$；

⑤将所述的每一个里的所罗门数据块按列分解成 $M+P_0$ 个数据片，将同一个数据块的 $M+P_0$ 个数据片的数据分别存储到云存储系统的多个存储器上，且同一存储器中，同一个数据块的片数$\leqslant P_I/2$ 片；

3. 根据权利要求 1 所述云存储系统的数据处理方法，其特征是，所述的数据提取方法，包括下列步骤：

①读取出属于同一数据块的数据片，如果第 i 片中出现 $P_0/2$ 个数据读取错误，根据里的所罗门乘积码的纠错解码算法，对该第 i 片的数据纠错，恢复出原始数据；

②读取完同一数据块的所有的数据片，如果在读取该数据块时有少于 $P_I/2$ 片出现丢失或者无法读取，计算机按里的所罗门乘积码的解码算法，对该少于 $P_I/2$ 片数据纠错，恢复出原始数据；

③重复步骤①、②，读取并处理完属于同一个原始数据的所有数据块，去掉纠错冗余数据，将所有数据块的新数据帧按原编号 ID 顺序排列，再去掉编号，得到原始存储的数据。

云存储系统的数据处理方法

技术领域

[0001] 本发明涉及云存储系统，特点是一种基于里的所罗门的云存储系统的数据处理方法。

背景技术

[0002] 在当今风起云涌的时代，云存储作为云的基础架构和最广泛的应用得到了极大的重视。在云存储系统中，用户数据存放于系统的云端，而构成云端的存储节点是用户不可控的。用户的数据可能被未经授权的第三方挖掘比对或者恶意篡改。

[0003] 同时，在云端单个或者多个存储节点缺失、失效的情况下（随着云端的扩展，存储节点故障的几率将增加），用户数据丢失的可能性极大。上述情况表明，云存储的发展亟需一种能完全保证用户数据完整性、隐私性和可靠性的安全机制。

[0004] 目前基于所有的云存储技术都是采用一种类似 Hadoop（云计算的一种开源软件）系统中的 HDFS(Hadoop Distributed File System，即 Hadoop 分布式文件系统）。该技术主要是将一个数据文件按照设定的大小分成若干块，再通过对每一个数据完整备份（例如 Hadoop 中的 HDFS 设置有 3 份相同备份）来提高可靠性，该技术的缺点是浪费存储空间。

[0005] 里的所罗门纠错编码方法：其原理是，计算信息码符多项式除以校验码生成多项式之后的余数，具体公式是：

[0006] $F \bmod D = C$；

[0007] 其中 F 为原始数据，D 是生成多项式，C 为生成的冗余纠错数据。mod 为求余运算。

[0008] 而在解码时，为简单起见，假定存入原始的信息符号为 m_3、m_2、m_1、m_0 和由此产生的检验符号 Q_1、Q_0，而读出的符号为 m_3'、m_2'、m_1'、m_0'、Q_1' 和 Q_0'，如果计算由此得到的校正子 s_0 和 s_1 不全为 0，则说明有差错，再通过计算错误多项式以及错误值，将错误纠正过来。

[0009] 该方法广泛应用于 DVD 光盘数据的处理，该编码方法能很好的提高对原始数据的纠错能力，能将数据的随机错误率从 2×10^{-2} 降低到 1×10^{-15}。在本发明中我们将该编码方法将一个数组形式的数据块，在横向和纵向分别进行里的所罗门编码，获得横向和纵向纠错冗余数据，这样就对数据进行了双重的纠错保护，提高了纠错能力，而且这些冗余数据只占原来数据量的 13%。

[0010] 正是由于在较低的数据冗余的情况下能如此高效的进行数据纠错，而一般的云存储系统均未采取此种纠错方法，只是通过数据备份来达到数据恢复的能力，一般云存储系统均须备份 3 份及以上，显然将极大地浪费数据的存储空间，提高成本，

发明内容

[0011] 本发明要解决的技术问题在于，提供了一种云存储系统的数据处理方法，该方法将提高云存储系统中数据的安全性，可恢复性，降低数据备份数目，极大的节约数据的存储空间，降低成本。

[0012] 本发明的技术解决方案如下：

[0013] 一种云存储系统的数据处理方法，其特点在于，对云存储系统的数据存入和数据提取分别采用里的所罗门编码数据纠错编码处理和里的所罗门纠错解码处理。

[0014] 所述的数据存入方法，包括以下具体步骤：

[0015] ①将待存储的原始数据分成 K 个帧数据，每个帧数据包含相同的固定长度 N 比特的数据，当最后一个原始数据帧的数据长度小于 N 时，对最后一块数据帧填充数据‘0’，以达到长度 N，其中 K 为大于大于 1 的正整数，N 的取值范围为：200 ＜ N ＜ 2000；

[0016] ②对所述的每一个数据帧加上编号，即 ID，得到一个新数据帧，所述的 ID 长为 4 个字节，从 0001 开始递加，故新的数据帧的长度为 (N+4)；

[0017] ③将所述的新的 K 个数据帧再组合为 W 个数据块，每一个数据块包含 M 个数据帧，形成一个 M*(N+4) 的数据矩阵，当最后一个数据块的数据帧的个数小于 M 时，对该数据块填充‘0’数据帧，以使最后一个数据块达到固定的数据帧 M，其中 M，W 的取值范围分别为：200 ＜ M ＜ 2000，W ＝ K/M；

[0018] ④对所述的数据块采用里的所罗门乘积码纠错编码方法进行纠错编码：将第 i 个数据块的行和列分别添加 P_0，P_I 个纠错的冗余数据进行编码，转化成一个包含 $(M+P_0)*(N+4+P_I)$ 个数据的里的所罗门数据块，其中，P_0，P_I 分别为数据块中一行和一列增加的用于纠错的冗余数据个数，且 $0 < P_0 < M/2$，$0 < P_I < M/2$，$1 < i \leqslant M$；

[0019] ⑤将所述的每一个里的所罗门数据块按列分解成 $M+P_0$ 个数据片，将同一个数据块的 $M+P_0$ 个数据片的数据分别存储到云存储系统的多个存储器上，且同一存储器中，同一个数据块的片数≤ $P_I/2$ 片；

[0020] 所述的数据提取方法，包括下列步骤：

[0021] ①读取出属于同一数据块的数据片，如果第 i 片中出现 $P_0/2$ 个数据读取错误，根据里的所罗门乘积码的纠错解码算法，对该第 i 片的数据纠错，恢复出原始数据；

[0022] ②读取完同一数据块的所有的数据片，如果在读取该数据块时有少于 $P_I/2$ 片出现丢失或者无法读取，计算机按里的所罗门乘积码的解码算法，对该少于 $P_I/2$ 片数据纠错，恢复出原始数据；

[0023] ③重复步骤①、②，读取并处理完属于同一个原始数据的所有数据块，去掉纠错冗余数据，将所有数据块的新数据帧按原编号 ID 顺序排列，再去掉编号，得到原始存储的数据。

[0024] 本发明的技术效果：

[0025] 1、本发明的最显著效果为，对原始数据块的行和列均进行了里的所罗门纠错编码，达到了双重纠错的能力，并且将这些经纠错后的数据块重新分片存储到云系统中的不同存储器上，这样不仅能对各个存储器上的片数据能进行纠错，并且当云系统中有一定数目的存储器出现故障造成一些片数据不能读取时，依然可得到完全的恢复，这极大的提高了系统的可靠性。

[0026] 2、一般的云存储系统需要将原始数据复制多份来确保数据的安全（一般云存储系统均须备份 3 份及以上），本发明可以减少在确保存储数据安全时使用的冗余存储量（可以只需备份两份或者一份），这将极大的节约的数据的存储空间，降低成本，使云存储系统空间得到更加充分的利用。

[0027] 3、本发明的另一显著特点是，由于数据是被分散存储在多个存储器上的，所以当外部有人非法入侵云系统中某单一存储器时，所获得的数据都是不完整的，也就提高了在面对系统外部的非法入侵时，数据的安全性。

具体实施方式

[0028] 下面结合实例对本发明做进一步说明，但不应以此限制本发明的保护范围。

[0029] 本实例用于对一个 100MB 的原始数据进行云存储，具体实施步骤如下

[0030] 步骤一，将一个 100MB 的原始待存储的数据分成 102401 帧数据，每一帧数据为固定长度 1020 个比特的数据。

[0031] 所述将待存储数据分为固定长度 1020 的 102401 帧数据，最后一块原始数据帧长度小于 255 时，对最后一块数据帧填充数据‘0’，以达到长度 1020。

[0032] 步骤二，对所述的每一个数据帧加上编号，即 ID，得到新的数据帧。

[0033] 所述将每一帧数据加上 ID，ID 长为 4 个字节，从 1 开始递加，实例中 ID 号从 1 到 102401。

[0034] 步骤三，将所述的数据帧再组合为若干个数据块，每一个数据块包含 1024 个数据帧，共得到 100 个这样的数据块。

[0035] 所述将数据帧再组合为数据块，这些数据帧将组成一个 1024*1024 的数据矩阵，形成一个数据块，如果最后一个数据块数据帧个数少于 1024，对最后一块数据块填充数据‘0’，以达到固定数据块数 1024。

[0036] 步骤四，通过里的所罗门纠错编码，将第 i 个包含 1024*1024 个数据的数据块的行和列分别进行编码，转化成一个包含 (1024+PO)*(1024+PI) 个数据的数据块，其中 0 ＜ i ＜＝ M。PO ＝ PI ＝ 100 为数据块中一行和一列增加的用于纠错的冗余数据个数。具体公式是：

[0037] F mod D ＝ C；

[0038] 本实施例中 F 为 1024 位数据，D 是生成多项式，C 为生成的 100 个冗余纠错数据。mod 为求余运算。

[0039] 步骤五，将上述得到的每一个数据块按列分解成 1024+100 片，将同一个数据块得到的若干片数据分别存储到若干个云存储系统的存储器上，且同一存储器上中同一个数据块的片数不能多于 50 块。

[0040] 步骤六，实例中当从云存储系统中读取数据时，通过 RS-PC 中的解码算法，解码出需要提取的数据。

[0041] 所述从云存储系统中读取出数据，是指：

[0042] 1) 读取出属于同一数据块的不同片，第 i 片中出现少于 50 个数据读取错误，根据里的所罗门纠错解码算法，可以将整片数据纠错，恢复成原始数据；

[0043] 2) 读取完同一数据块的不同片，如果在读取这些片时有少于 50 片出现丢失或者无法读取，根据里的所罗门纠错解码算法，可以将该段数据纠错，恢复出来。

[0044] 3) 读取完所有属于同一个原始数据的不同数据块，去掉纠错冗余数据，将这些数据块按编号 (1 到 102401) 顺序排列，再去掉编号，最终得到原始存储的数据。