(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 111143288 A
(43)申请公布日 2020.05.12

(21)申请号 201911332561.0

(22)申请日 2019.12.22

(71)申请人 北京浪潮数据技术有限公司
地址 100085 北京市海淀区上地信息路2号C栋5层

(72)发明人 岳斌

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 高勇

(51)Int.Cl.
G06F 16/13(2019.01)
G06F 16/16(2019.01)
G06F 16/172(2019.01)

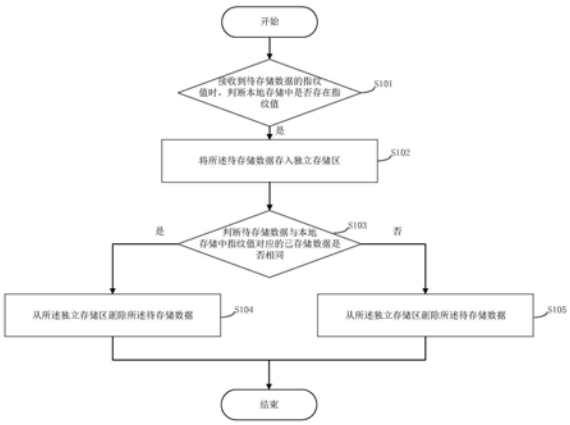
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种数据存储方法、系统及相关装置

(57)摘要

本申请提供一种数据存储方法，包括：接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；若是，将所述待存储数据存入独立存储区；判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；若是，从所述独立存储区删除所述待存储数据；若否，将所述待存储数据从所述独立存储区转存至所述本地存储。本申请利用独立存储区解决由于哈希冲突可能的数据误删问题，保证了数据存储一致性。本申请还提供一种数据存储系统、计算机可读存储介质和服务器，具有上述有益效果。



CN 111143288 A

1.一种数据存储方法,其特征在于,包括:

接收到待存储数据的指纹值时,判断本地存储中是否存在所述指纹值;

若是,将所述待存储数据存入独立存储区;

判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同;

若是,从所述独立存储区删除所述待存储数据;

若否,将所述待存储数据从所述独立存储区转存至所述本地存储。

2.根据权利要求1所述的数据存储方法,其特征在于,判断本地存储中是否存在所述指纹值之前,还包括:

接收所述待存储数据,利用哈希算法计算所述待存储数据的指纹值。

3.根据权利要求1所述的数据存储方法,其特征在于,从所述独立存储区删除所述待存储数据之后,还包括:

将所述待存储数据的LP元数据和PL元数据更新至所述本地存储。

4.根据权利要求1所述的数据存储方法,其特征在于,若所述本地存储中不存在所述指纹值,还包括:

将所述待存储数据写入所述本地存储,并保存所述待存储数据的HP元数据、LP元数据和PL元数据。

5.根据权利要求1所述的数据存储方法,其特征在于,将所述待存储数据存入独立存储区之前,还包括:

在缓存或所述本地存储中划分独立存储区。

6.根据权利要求1所述的数据存储方法,其特征在于,判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同之前,还包括:

删除所述独立存储区中热度值低于预设阈值的待存储数据。

7.根据权利要求6所述的数据存储方法,其特征在于,判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同包括:

按各所述待存储数据的热度值大小顺序判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同。

8.一种数据存储系统,其特征在于,包括:

第一判断模块,用于接收到待存储数据的指纹值时,判断本地存储中是否存在所述指纹值;

独立存储模块,用于所述第一判断模块判断结果为否时,将所述待存储数据存入独立存储区;

第二判断模块,用于判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同;

删除模块,用于所述第二判断模块判断结果为是时,从所述独立存储区删除所述待存储数据;

转存模块,用于所述第二判断模块判断结果为否时,将所述待存储数据从所述独立存储区转存至所述本地存储。

9.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1-4任一项所述的方法的步骤。

10.一种服务器,其特征在于,包括存储器和处理器,所述存储器中存有计算机程序,所述处理器调用所述存储器中的计算机程序时实现如权利要求1-4任一项所述的方法的步骤。

一种数据存储方法、系统及相关装置

技术领域

[0001] 本申请涉及数据存储领域，特别涉及一种数据存储方法、系统及相关装置。

背景技术

[0002] 在存储领域，海量数据查询和存储需要占用超大的资源，严重影响了数据存储的性能。在这些数据中，存在大量重复数据，为了降低存储数据占用的资源，提高数据存储性能，这种重复数据完全可以只保存一份在存储介质中，在不影响数据一致性的前提下，减少盘上数据存放量。判断数据是否重复的方法是通过哈希算法计算数据指纹值，通过比对指纹值作为数据是否相同的依据。但是哈希算法存在哈希冲突，即不同数据通过哈希运算得到相同的哈希值。

[0003] 因此，如何解决由于哈希冲突对数据存储造成的影响是本领域技术人员亟需解决的技术问题。

发明内容

[0004] 本申请的目的是提供一种数据存储方法、系统、计算机可读存储介质和服务器，能够解决数据存储时的哈希冲突。

[0005] 为解决上述技术问题，本申请提供一种数据存储方法，具体技术方案如下：

[0006] 接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；

[0007] 若是，将所述待存储数据存入独立存储区；

[0008] 判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；

[0009] 若是，从所述独立存储区删除所述待存储数据；

[0010] 若否，将所述待存储数据从所述独立存储区转存至所述本地存储。

[0011] 其中，判断本地存储中是否存在所述指纹值之前，还包括：

[0012] 接收所述待存储数据，利用哈希算法计算所述待存储数据的指纹值。

[0013] 其中，从所述独立存储区删除所述待存储数据之后，还包括：

[0014] 将所述待存储数据的LP元数据和PL元数据更新至所述本地存储。

[0015] 其中，若所述本地存储中不存在所述指纹值，还包括：

[0016] 将所述待存储数据写入所述本地存储，并保存所述待存储数据的HP元数据、LP元数据和PL元数据。

[0017] 其中，将所述待存储数据存入独立存储区之前，还包括：

[0018] 在缓存或所述本地存储中划分独立存储区。

[0019] 其中，判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同之前，还包括：

[0020] 删除所述独立存储区中热度值低于预设阈值的待存储数据。

[0021] 其中，判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同包括：

[0022] 按各所述待存储数据的热度值大小顺序判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同。

[0023] 本申请还提供一种数据存储系统，包括：

[0024] 第一判断模块，用于接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；

[0025] 独立存储模块，用于所述第一判断模块判断结果为否时，将所述待存储数据存入独立存储区；

[0026] 第二判断模块，用于判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；

[0027] 删除模块，用于所述第二判断模块判断结果为是时，从所述独立存储区删除所述待存储数据；

[0028] 转存模块，用于所述第二判断模块判断结果为否时，将所述待存储数据从所述独立存储区转存至所述本地存储。

[0029] 其中，还包括：

[0030] 指纹值计算模块，用于接收所述待存储数据，利用哈希算法计算所述待存储数据的指纹值。

[0031] 其中，所述删除模块还包括：

[0032] 元数据更新单元，用于将所述待存储数据的LP元数据和PL元数据更新至所述本地存储。

[0033] 其中，还包括：

[0034] 直接存储模块，用于所述第一判断模块判断结果为是时，将所述待存储数据写入所述本地存储，并保存所述待存储数据的HP元数据、LP元数据和PL元数据。

[0035] 其中，还包括：

[0036] 独立存储区划分模块，用于在将所述待存储数据存入独立存储区之前，在缓存或所述本地存储中划分独立存储区。

[0037] 其中，还包括：

[0038] 数据删除模块，用于判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同之前，删除所述独立存储区中热度值低于预设阈值的待存储数据。

[0039] 其中，所述第二判断模块具体为按各所述待存储数据的热度值大小顺序判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同的模块。

[0040] 本申请还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如上所述的方法的步骤。

[0041] 本申请还提供一种服务器，包括存储器和处理器，所述存储器中存有计算机程序，所述处理器调用所述存储器中的计算机程序时实现如上所述的方法的步骤。

[0042] 本申请提供一种数据存储方法，包括：接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；若是，将所述待存储数据存入独立存储区；判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；若是，从所述独立存储区删除所述待存储数据；若否，将所述待存储数据从所述独立存储区转存至所述本地存储。

[0043] 本申请通过判断待存储数据的指纹值，确定待存储数据在本地存储中是否已经存

储，若已经存在该指纹值，先将待存储数据存入独立存储区，然后通过比对独立存储区中的待存储数据和本地存储中相同指纹值的已存储数据，若二者一致，说明该待存储数据重复；若二者不一致，则确认发生哈希冲突，此时将待存储数据从独立存储区转存至本地存储，利用独立存储区解决由于哈希冲突可能的数据误删问题，保证了数据存储一致性。本申请还提供一种数据存储系统、计算机可读存储介质和服务器，具有上述有益效果，此处不再赘述。

附图说明

[0044] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

[0045] 图1为本申请实施例所提供的一种数据存储方法的流程图；

[0046] 图2为本申请实施例所提供的一种数据存储系统结构示意图。

具体实施方式

[0047] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0048] 请参考图1，图1为本申请实施例所提供的一种数据存储方法的流程图，该方法包括：

[0049] S101：接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；若是，进入S102；

[0050] 本步骤需要在接收到指纹值时，判断本地存储中是否包含该指纹值。

[0051] 通常，指纹值为根据数据的特征信息利用哈希算法计算得到的数据内容，在此对于指纹值的具体格式等不作限定。采用不同的哈希算法得到的指纹值位数并不相同，例如利用SHA-256算法即可将任意长度的待存储数据转为256位的指纹值。而由于哈希算法存在哈希冲突的问题，即不同的数据经过哈希算法计算能得到相同的指纹值。而此时相同指纹值的两份数据实际上可能并不相同。

[0052] 在接收到指纹值时，先判断本地存储中是否存在该指纹值，若本地存储中不存在该指纹值，意味着该待存储数据并未在本地存储中存储，可以直接将所述待存储数据写入所述本地存储。容易理解的是，在数据存储时，通常还需要保存相关的元数据。在本地存储中不存在该指纹值时，可以保存所述待存储数据的HP元数据(hash-pba modedata)、LP元数据(lba-pba modedata)和PL元数据(pba-lba modedata)。LP元数据和PL元数据均为与逻辑区块地址和物理区块地址相关的元数据。

[0053] S102：将所述待存储数据存入独立存储区；

[0054] 若本地存储中包含该指纹值，说明待存储数据可能在本地存储中已经存储了相同数据。但由于此时尚不能排除是哈希冲突造成的指纹值相同的情况，因此本步骤需要将待

存储数据存入独立存储区。

[0055] 独立存储区是与本地存储相区分的存储区块，其可以为本地存储中独立划分出的临时存储区，专门用于存储与已存储数据疑似相同的待存储数据。也可以利用缓存区域作为独立存储区。

[0056] 容易理解的是，在执行本步骤之前，还需要在缓存或所述本地存储中划分独立存储区。而划分独立存储区的过程只需在发现本地存储存在相同指纹值之前即可，与步骤S101并无既定的执行顺序限定。

[0057] 在此对于具体的划分规则不作限定，本领域技术人员可以根据经验值，或者根据本地存储量或内存大小的预设百分比确定。例如无论采用缓存还是本地存储，均可以选取20％作为独立存储区。

[0058] S103：判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；若是，进入S104；若否，进入S105；

[0059] 在待存储数据存入独立存储区后，将待存储数据与本地存储中指纹值相同的已存储数据进行比对，判断两份数据是否为同一份数据。

[0060] S104：从所述独立存储区删除所述待存储数据；

[0061] 若待存储数据与本地存储中指纹值对应的已存储数据相同，则此时无需存储该待存储数据，可以从独立存储区中删除该待存储数据。

[0062] 此外，优选的，虽然待存储数据已经存在于本地存储，但还可以将待存储数据的LP元数据和PL元数据更新至本地存储。即在本地存储中更新该数据元数据，便于后续数据的遍历。

[0063] 但需要注意的是，由于此时已经确定待存储数据与本地存储中已存储的数据相同，则此时更新的元数据只包括LP元数据和PL元数据，而HP元数据则由于两份数据实际为同一份数据，其对应的哈希算法得到的指纹值也均相同，因此无需存储至本地存储。

[0064] S105：将所述待存储数据从所述独立存储区转存至所述本地存储。

[0065] 若待存储数据与本地存储中指纹值对应的已存储数据不同，则待存储数据需要存至本地存储，即从独立存储区中转存至本地存储。

[0066] 由此可以看出，独立存储区实际上作为一个临时存储区，在其中执行判断发生哈希冲突时冲突双方数据是否一致的过程。一旦冲突双方数据不一致，即说明是由于哈希算法导致的不同数据得到了相同的指纹值，则此时需要将待存储数据从独立存储区转存至本地存储。一旦冲突双方数据一致，说明待存储数据已经存于本地存储中，则可以直接将待存储数据从独立存储区删除。

[0067] 本申请实施例通过判断待存储数据的指纹值，确定待存储数据在本地存储中是否已经存储，若已经存在该指纹值，先将待存储数据存入独立存储区，然后通过比对独立存储区中的待存储数据和本地存储中相同指纹值的已存储数据，若二者一致，说明该待存储数据重复；若二者不一致，则确认发生哈希冲突，此时将待存储数据从独立存储区转存至本地存储，利用独立存储区解决由于哈希冲突可能的数据误删问题，保证了数据存储一致性。

[0068] 进一步的，在上述实施例的基础上，作为优选的实施例，在步骤S103之前，还可以对独立存储区中的待存储数据进行热度划分，同时删除热度值较低的待存储数据。具体的，可以先根据各待存储数据的热度值从高到低排序，将热度值低于预设阈值的待存储数据删

除。由于热度值较低时说明其使用量和检索率均较低，即并非常用数据，因此在该待存储数据的指纹值已经存在于本地存储中时，将其直接删除。

[0069] 此后，将热度值低于预设阈值的待存储数据删除后，将剩余待存储数据按照热度值大小排序，对热度值较高的待存储数据优先执行步骤S103中的判断过程，使得热度值较高的待存储数据在于本地存储中同一指纹值对应的已存储数据不同时可以优先落盘，提高数据存储的高价值数据存储率。

[0070] 则此时整个执行过程可以如下：

[0071] S201：接收到待存储数据的指纹值时，判断本地存储中是否存在指纹值；若是，进入S202；

[0072] S202：将待存储数据存入独立存储区；

[0073] S203：删除所述独立存储区中热度值低于预设阈值的待存储数据；

[0074] S204：按各所述待存储数据的热度值大小顺序判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；若是，进入S205；若否，进入S206；

[0075] S205：从独立存储区删除待存储数据；

[0076] S206：将待存储数据从独立存储区转存至本地存储。

[0077] 本申请实施例在上述实施例的基础上，删除了热度值低于预设阈值的待存储数据，同时比对待存储数据和已存储数据时优先比对热度值较高的待存储数据，使得本地存储可以优先存储高热度数据，能够保证高价值数据尽快落盘，同时剔除低热度数据，避免本地存储由于存储过多低热度数据导致的磁盘有效利用率降低等问题。

[0078] 下面对本申请实施例提供的一种数据存储系统进行介绍，下文描述的数据存储系统与上文描述的数据存储方法可相互对应参照。

[0079] 参见图2，图2为本申请实施例所提供的一种数据存储系统结构示意图，本申请还提供一种数据存储系统，包括：

[0080] 第一判断模块100，用于接收到待存储数据的指纹值时，判断本地存储中是否存在所述指纹值；

[0081] 独立存储模块200，用于所述第一判断模块判断结果为否时，将所述待存储数据存入独立存储区；

[0082] 第二判断模块300，用于判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同；

[0083] 删除模块400，用于所述第二判断模块判断结果为是时，从所述独立存储区删除所述待存储数据；

[0084] 转存模块500，用于所述第二判断模块判断结果为否时，将所述待存储数据从所述独立存储区转存至所述本地存储。

[0085] 基于上述实施例，作为优选的实施例，还包括：

[0086] 指纹值计算模块，用于接收所述待存储数据，利用哈希算法计算所述待存储数据的指纹值。

[0087] 基于上述实施例，作为优选的实施例，所述删除模块400还可以包括：

[0088] 元数据更新单元，用于将所述待存储数据的LP元数据和PL元数据更新至所述本地存储。

[0089] 基于上述实施例，作为优选的实施例，还可以包括：

[0090] 直接存储模块，用于所述第一判断模块判断结果为是时，将所述待存储数据写入所述本地存储，并保存所述待存储数据的HP元数据、LP元数据和PL元数据。

[0091] 基于上述实施例，作为优选的实施例，还包括：

[0092] 独立存储区划分模块，用于在将所述待存储数据存入独立存储区之前，在缓存或所述本地存储中划分独立存储区。

[0093] 基于上述实施例，作为优选的实施例，还包括：

[0094] 数据删除模块，用于判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同之前，删除所述独立存储区中热度值低于预设阈值的待存储数据。

[0095] 基于上述实施例，作为优选的实施例，所述第二判断模块具体为按各所述待存储数据的热度值大小顺序判断所述待存储数据与所述本地存储中所述指纹值对应的已存储数据是否相同的模块。

[0096] 本申请还提供了一种计算机可读存储介质，其上存有计算机程序，该计算机程序被执行时可以实现上述实施例所提供的步骤。该存储介质可以包括：U盘、移动硬盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0097] 本申请还提供了一种服务器，可以包括存储器和处理器，所述存储器中存有计算机程序，所述处理器调用所述存储器中的计算机程序时，可以实现上述实施例所提供的步骤。当然所述服务器还可以包括各种网络接口，电源等组件。

[0098] 说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例提供的方法而言，由于其与实施例提供的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0099] 本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想。应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

[0100] 还需要说明的是，在本说明书中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

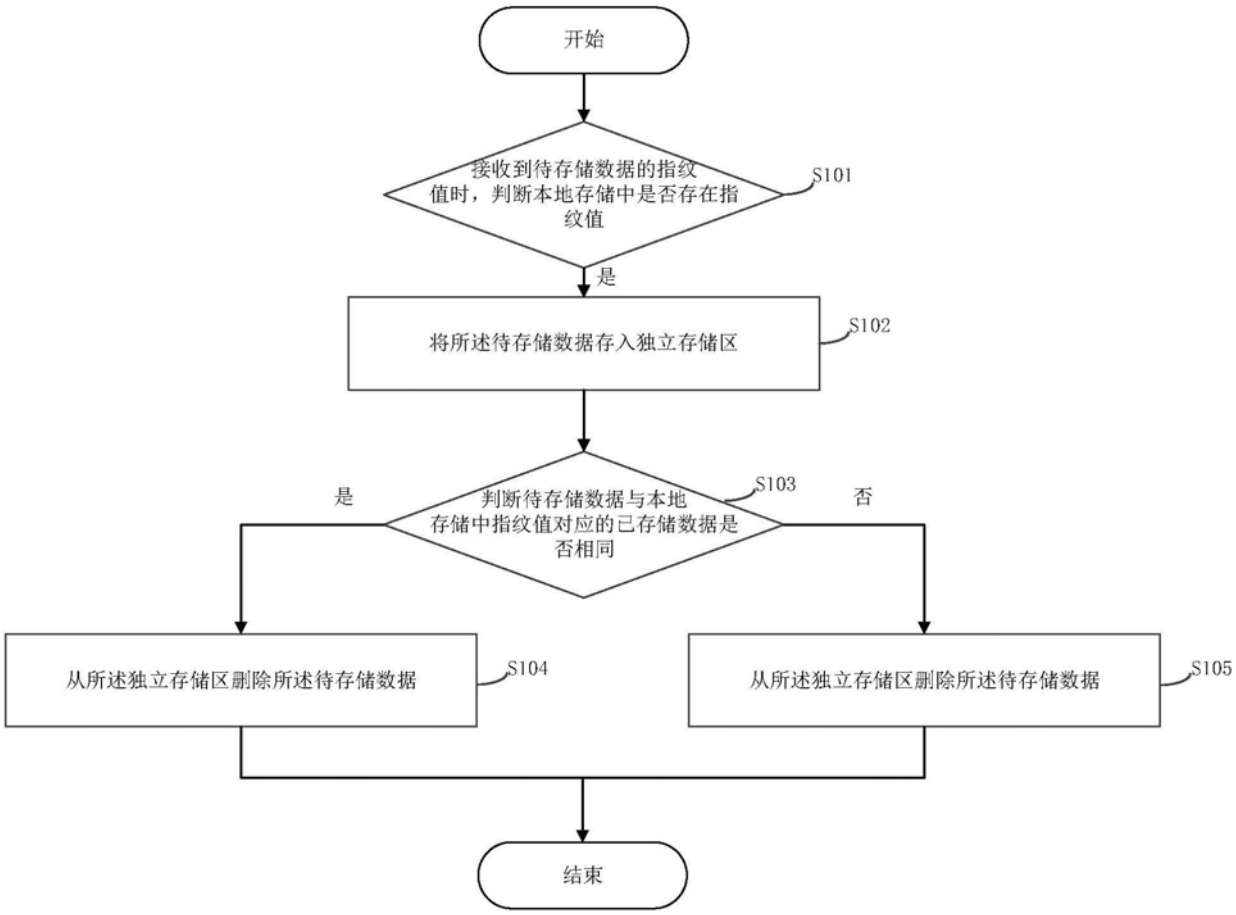

开始
接收到待存储数据的指纹值时，判断本地存储中是否存在指纹值
S101
是
将所述待存储数据存入独立存储区
S102
判断待存储数据与本地存储中指纹值对应的已存储数据是否相同
S103
是
否
从所述独立存储区删除所述待存储数据
S104
从所述独立存储区删除所述待存储数据
S105
结束

图1

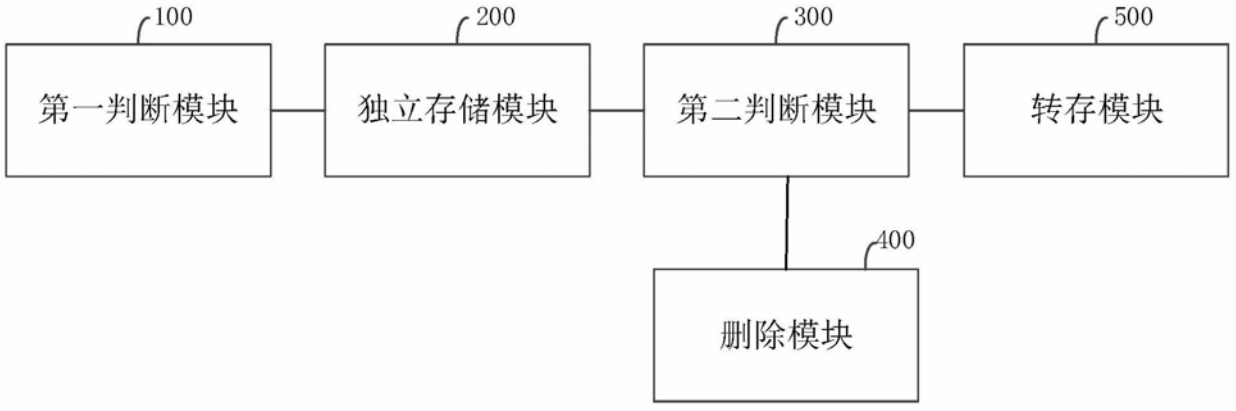

100
第一判断模块
200
独立存储模块
300
第二判断模块
500
转存模块
400
删除模块

图2