(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109032536 B
(45)授权公告日 2021.08.10

(21)申请号 201811011203.5

(22)申请日 2018.08.31

(65)同一申请的已公布的文献号
申请公布号 CN 109032536 A

(43)申请公布日 2018.12.18

(73)专利权人 郑州云海信息技术有限公司
地址 450018 河南省郑州市郑东新区心怡路278号16层1601室

(72)发明人 刘洪栋

(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 罗满

(51)Int.Cl.
G06F 3/06(2006.01)

(56)对比文件
CN 103809147 A,2014.05.21
CN 105141685 A,2015.12.09
CN 103902479 A,2014.07.02
CN 105025106 A,2015.11.04
CN 101888395 A,2010.11.17
CN 106527958 A,2017.03.22
CN 102223382 A,2011.10.19
CN 104246767 A,2014.12.24
CN 104932953 A,2015.09.23
US 2013073522 A1,2013.03.21

审查员 刘婷婷

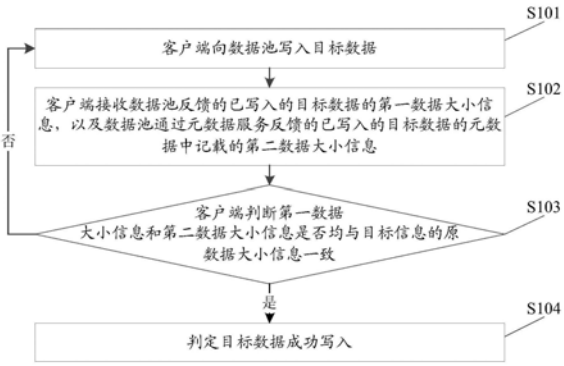
权利要求书2页 说明书6页 附图2页

(54)发明名称

基于分布式集群系统的数据存储方法、装置、系统及设备

(57)摘要

本发明公开了一种基于分布式集群系统的数据存储方法，包括：客户端向数据池写入目标数据；接收已写入的目标数据的第一数据大小信息，以及已写入的目标数据的元数据中记载的第二数据大小信息；判断第一数据大小信息和第二数据大小信息是否与原数据大小信息一致；若是，目标数据成功写入；若否，重新写入目标数据。可见，本方案将数据写入数据池后，需要确定写入数据池的数据大小是否正确、写入数据池的数据的元数据记载的数据大小是否正确；若均正确，则判定数据写入成功，从而通过这种方式确保数据的一致性，提高了分布式集群系统的可靠性；本发明还公开了一种基于分布式集群系统的数据存储装置、系统及设备，同样能实现上述技术效果。



CN 109032536 B

1.一种基于分布式集群系统的数据存储方法,其特征在于,包括:

客户端向数据池写入目标数据;

所述客户端接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;

所述客户端判断所述第一数据大小信息和第二数据大小信息是否均与所述目标数据的原数据大小信息一致;

若一致,则判定所述目标数据成功写入;若不一致,则重新向所述数据池写入所述目标数据。

2.根据权利要求1所述的数据存储方法,其特征在于,所述客户端向数据池写入目标数据之前,还包括:

所述客户端向元数据服务发送元数据请求,以使所述元数据服务根据所述元数据请求生成对应的日志事件,并落盘至所述数据池。

3.根据权利要求2所述的数据存储方法,其特征在于,还包括:

所述元数据服务检测所述日志事件是否成功落盘至所述数据池;

若否,则所述元数据服务根据日志事件回放机制,利用所述日志事件回放对应的元数据请求,并生成日志事件落盘至所述数据池。

4.根据权利要求1至3中任意一项所述的数据存储方法,其特征在于,所述客户端向数据池写入目标数据,包括:

所述客户端通过分段方式依次向所述数据池写入数据;其中,所述客户端判定数据成功写入后,才会向所述数据池写入下一段数据。

5.一种基于分布式集群系统的数据存储装置,其特征在于,基于客户端,所述数据存储装置包括:

数据写入模块,用于向数据池写入目标数据;

信息接收模块,用于接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;

判断模块,用于判断所述第一数据大小信息和第二数据大小信息是否均与所述目标数据的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则触发所述数据写入模块重新向所述数据池写入所述目标数据。

6.根据权利要求5所述的数据存储装置,其特征在于,还包括:

数据请求发送模块,用于向数据池写入目标数据之前,向元数据服务发送元数据请求,以使所述元数据服务根据所述元数据请求生成对应的日志事件,并落盘至所述数据池。

7.根据权利要求5或6所述的数据存储装置,其特征在于,所述数据写入模块具体用于通过分段方式依次向所述数据池写入数据;其中,所述客户端判定数据成功写入后,才会向所述数据池写入下一段数据。

8.一种基于分布式集群系统的数据存储系统,其特征在于,包括客户端、元数据服务和数据池;

所述客户端,用于向所述数据池写入目标数据;接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元

数据中记载的第二数据大小信息;判断所述第一数据大小信息和第二数据大小信息是否均与所述目标数据的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则重新向所述数据池写入所述目标数据。

9.根据权利要求8所述的数据存储系统,其特征在于,所述元数据服务还用于:检测日志事件是否成功落盘至所述数据池;若否,则所述元数据服务根据日志事件回放机制,利用所述日志事件回放对应的元数据请求,并生成日志事件落盘至所述数据池。

10.一种基于分布式集群系统的数据存储设备,其特征在于,包括:

存储器,用于存储计算机程序;

处理器,用于执行所述计算机程序时实现如权利要求1至4任一项所述基于分布式集群系统的数据存储方法的步骤。

基于分布式集群系统的数据存储方法、装置、系统及设备

技术领域

[0001] 本发明涉及分布式存储技术领域，更具体地说，涉及一种基于分布式集群系统的数据存储方法、装置、系统及设备。

背景技术

[0002] 分布式存储系统是将数据分散存储在多台独立的设备上。传统的网络存储系统采用集中的存储服务器存放所有数据，存储服务器成为系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要。分布式网络存储系统采用可扩展的系统结构，利用多台存储服务器分担存储负荷，利用位置服务器定位存储信息，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

[0003] 目前，分布式集群存储系统在当前海量数据的多种场景下广泛应用，例如：高性能场景、视频监控场景、广电媒资等。在进行分布式集群存储时，若在数据存储过程中出现数据不一致的情况，则会降低分布式集群系统的数据可靠性。

[0004] 因此，如何提高分布式集群系统的可靠性，是本领域技术人员需要解决的问题。

发明内容

[0005] 本发明的目的在于提供一种基于分布式集群系统的数据存储方法、装置、系统及设备，以提高分布式集群系统的可靠性。

[0006] 为实现上述目的，本发明实施例提供了如下技术方案：

[0007] 一种基于分布式集群系统的数据存储方法，包括：

[0008] 客户端向数据池写入目标数据；

[0009] 所述客户端接收所述数据池反馈的已写入的目标数据的第一数据大小信息，以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息；

[0010] 所述客户端判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致；

[0011] 若一致，则判定所述目标数据成功写入；若不一致，则重新向所述数据池写入所述目标数据。

[0012] 其中，所述客户端向数据池写入目标数据之前，还包括：

[0013] 所述客户端向元数据服务发送元数据请求，以使所述元数据服务根据所述元数据请求生成对应的日志事件，并落盘至所述数据池。

[0014] 其中，本方案还包括：

[0015] 所述元数据服务检测所述日志事件是否成功落盘至所述数据池；

[0016] 若否，则所述元数据服务根据日志事件回放机制，利用所述日志事件回放对应的元数据请求，并生成日志事件落盘至所述数据池。

[0017] 其中，所述客户端向数据池写入目标数据，包括：

[0018] 所述客户端通过分段方式依次向所述数据池写入数据;其中,所述客户端判定数据成功写入后,才会向所述数据池写入下一段数据。

[0019] 一种基于分布式集群系统的数据存储装置,基于客户端,所述数据存储装置包括:

[0020] 数据写入模块,用于向数据池写入目标数据;

[0021] 信息接收模块,用于接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;

[0022] 判断模块,用于判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则触发所述数据写入模块重新向所述数据池写入所述目标数据。

[0023] 其中,本方案还包括:

[0024] 数据请求发送模块,用于向数据池写入目标数据之前,向元数据服务发送元数据请求,以使所述元数据服务根据所述元数据请求生成对应的日志事件,并落盘至所述数据池。

[0025] 其中,所述数据写入模块具体用于通过分段方式依次向所述数据池写入数据;其中,所述客户端判定数据成功写入后,才会向所述数据池写入下一段数据。

[0026] 一种基于分布式集群系统的数据存储系统,包括客户端、元数据服务和数据池;

[0027] 所述客户端,用于向所述数据池写入目标数据;接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则重新向所述数据池写入所述目标数据。

[0028] 其中,所述元数据服务还用于:检测所述日志事件是否成功落盘至所述数据池;若否,则所述元数据服务根据日志事件回放机制,利用所述日志事件回放对应的元数据请求,并生成日志事件落盘至所述数据池。

[0029] 一种基于分布式集群系统的数据存储设备,包括:

[0030] 存储器,用于存储计算机程序;

[0031] 处理器,用于执行所述计算机程序时实现上述基于分布式集群系统的数据存储方法的步骤。

[0032] 通过以上方案可知,本发明实施例提供的一种基于分布式集群系统的数据存储方法,包括:客户端向数据池写入目标数据;所述客户端接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;所述客户端判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则重新向所述数据池写入所述目标数据。

[0033] 可见,在本方案中,在向数据池写入数据后,需要利用元数据同步机制和数据同步机制,确定写入数据池的数据大小是否正确,以及写入数据池的数据的元数据记载的数据大小是否正确;若均正确,则判定数据写入成功,从而通过这种方式保证数据的一致性,提高了分布式集群系统的可靠性;

[0034] 本发明还公开了一种基于分布式集群系统的数据存储装置、系统及设备，同样能实现上述技术效果。

附图说明

[0035] 为了更清楚地说明本发明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发明的一些实施例，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，还可以根据这些附图获得其他的附图。

[0036] 图1为本发明实施例公开的一种基于分布式集群系统的数据存储方法流程示意图；

[0037] 图2为本发明实施例公开的一种基于分布式集群系统的数据存储装置结构示意图；

[0038] 图3为本发明实施例公开的一种基于分布式集群系统的数据存储系统结构示意图。

具体实施方式

[0039] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0040] 本发明实施例公开了一种基于分布式集群系统的数据存储方法、装置、系统及设备，以提高分布式集群系统的可靠性。

[0041] 参见图1，本发明实施例提供的一种基于分布式集群系统的数据存储方法，包括：

[0042] S101、客户端向数据池写入目标数据；

[0043] 其中，所述客户端向数据池写入目标数据，包括：

[0044] 所述客户端通过分段方式依次向所述数据池写入数据；其中，所述客户端判定数据成功写入后，才会向所述数据池写入下一段数据。

[0045] 具体的，在本实施例中，向数据池写入的目标数据可以是一段完整的数据，也可以是将一段完整的数据通过分段方式分为多个数据段，其中的一个数据段便是本方案中的目标数据；这种将完整的数据分段处理的方式，可以分别确认每一段数据的一致性，如果出现错误，只需要重新写入出现错误的数据段便可以；而如果直接将完整的数据写入数据池，若该完整的数据出现问题，则需要重新将所有数据重新写入，会浪费大量的时间。

[0046] 在本方案中，将数据写入数据池后，如果通过S102-S103确定该数据成功写入，则继续将下一段数据写入数据池，并继续执行后续步骤；而如果未成功写入，则需要重新写入该数据，直至判定该数据成功写入后，方可写入下一段数据。

[0047] S102、所述客户端接收所述数据池反馈的已写入的目标数据的第一数据大小信息，以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息；

[0048] 具体的，将目标数据写入数据池后，需要向客户端反馈第一数据大小信息和第二

数据大小信息，该第一数据大小信息具体为将目标数据写入数据池后，占用存储的空间大小，可以通过算法确定，如果数据在写入过程中出现问题，那么第一数据大小信息便与目标数据的原数据大小信息不同，例如比原数据大小信息大或者比原数据大小信息小。

[0049] 该第二数据大小信息，是将目标数据写入数据池后，数据池中的目标数据中的元数据中记载的目标数据的大小信息，如果元数据中成功记录了目标数据的大小信息，那么该第二数据大小信息应该也与目标数据的原数据大小信息相同，如果不同，则说明元数据错误。

[0050] S103、所述客户端判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致；若一致，则执行S104；若不一致，则重新执行S101；

[0051] S104、判定所述目标数据成功写入。

[0052] 具体的，本方案中的原数据大小信息是指：目标数据还没有存入数据池之前，在客户端中的数据大小信息，该原数据大小信息是对比基准，是正确的信息。需要说明的是，本实施例其实公开了两个机制，一个是数据同步机制，另一个是元数据同步机制，在本方案中，数据同步机制用来确定数据是否正确的写入数据池，即通过写入数据池中的目标数据的大小信息来确定是否正确写入，如果正确写入，则第一数据大小信息与原数据大小信息相同。元数据同步机制用来确定数据池中的元数据是否正确统计写入的目标数据的大小，如果第二数据大小信息与原数据大小信息相同，则说明元数据正确。

[0053] 在将目标数据成功写入数据池时，第一数据大小信息、第二数据大小信息和原数据大小信息三者应该相同，如果存在任意两者不同，则说明写入失败，则需要重新写入目标数据，即重新执行S101，直至重新写入的数据的第一数据大小信息、第二数据大小信息和原数据大小信息相同为止。

[0054] 具体来说，如果客户端发起文件写操作，需要通过数据同步机制及元数据同步机制分别确认数据及元数据的正确性；例如：应该写入的数据为1GB文件，通过分段方式先写入4MB大小数据，这里的4MB即为原数据大小信息；将4MB大小数据写入数据池后，接收返回的第一数据大小信息和第二数据大小信息，如果第一数据大小信息为4MB，则写入数据池的数据正确，否则不正确；如果第二数据大小信息为4MB，则与写入的数据对应的元数据正切，否则不正确；如果存在任意一者不正确，则客户端重新发送该写操作。

[0055] 综上可以看出，本方案通过客户端的元数据同步机制保证元数据的一致性，通过客户端的数据同步机制确保数据的一致性落盘，从而可以避免故障场景下数据丢失的风险，从根本上保障数据的可靠性。当然，本方案中的数据存储方式同样可以适用于统一存储，在此并不具体限定。

[0056] 基于上述任意实施例，在本实施例中，所述客户端向数据池写入目标数据之前，还包括：

[0057] 所述客户端向元数据服务发送元数据请求，以使所述元数据服务根据所述元数据请求生成对应的日志事件，并落盘至所述数据池。

[0058] 所述元数据服务检测所述日志事件是否成功落盘至所述数据池；

[0059] 若否，则所述元数据服务根据日志事件回放机制，利用所述日志事件回放对应的元数据请求，并生成日志事件落盘至所述数据池。

[0060] 具体的，客户端对数据池进行创建、修改等操作时，首先需要向元数据服务(Meta

Data Server,MDS)发起(创建、修改等)元数据请求;MDS收到该元数据请求后,会生成相应日志事件,并将日志事件落盘;该日志事件仅仅用来标识该客户端对数据池的操作类型,该日志事件可以理解为记录接下来向数据池执行操作的事件,例如上述实施例中的将目标数据写入数据池的事件,同样的,本方案中的日志事件落盘即在上述实施例中客户端将目标数据落盘至数据池。

[0061] 进一步,客户端向MDS发送该元数据请求后,MDS会在目标数据成功落盘至数据池后,向客户端发送一个日志事件落盘成功的应答,如果未落盘成功,则不向客户端发送应答信息,说明在将日志事件落盘时出现故障,这时MDS可以通过元数据日志事件回放机制,将元数据请求回放出来,如果该请求未执行完成,即与该请求对应的数据写入操作或者数据修改操作未执行完成,则继续处理。

[0062] 下面对本发明实施例提供的数据存储装置进行介绍,下文描述的数据存储装置与上文描述的数据存储方法可以相互参照。

[0063] 参见图2,本发明实施例提供的一种基于分布式集群系统的数据存储装置,基于客户端,所述数据存储装置包括:

[0064] 数据写入模块101,用于向数据池写入目标数据;

[0065] 信息接收模块102,用于接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;

[0066] 判断模块103,用于判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则触发所述数据写入模块重新向所述数据池写入所述目标数据。

[0067] 其中,本方案还包括:

[0068] 数据请求发送模块,用于向数据池写入目标数据之前,向元数据服务发送元数据请求,以使所述元数据服务根据所述元数据请求生成对应的日志事件,并落盘至所述数据池。

[0069] 其中,所述数据写入模块具体用于通过分段方式依次向所述数据池写入数据;所述客户端判定数据成功写入后,才会向所述数据池写入下一段数据。

[0070] 下面对本发明实施例提供的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以相互参照。

[0071] 参见图3,本发明实施例提供的一种基于分布式集群系统的数据存储系统,包括客户端100、元数据服务200和数据池300;

[0072] 所述客户端100,用于向所述数据池写入目标数据;接收所述数据池反馈的已写入的目标数据的第一数据大小信息,以及所述数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息;判断所述第一数据大小信息和第二数据大小信息是否均与所述目标信息的原数据大小信息一致;若一致,则判定所述目标数据成功写入;若不一致,则重新向所述数据池写入所述目标数据。

[0073] 其中,所述元数据服务还用于:检测所述日志事件是否成功落盘至所述数据池;若否,则所述元数据服务根据日志事件回放机制,利用所述日志事件回放对应的元数据请求,并生成日志事件落盘至所述数据池。

[0074] 其中，所述客户端还用于：向数据池写入目标数据之前，向元数据服务发送元数据请求，以使所述元数据服务根据所述元数据请求生成对应的日志事件，并落盘至所述数据池。

[0075] 其中，所述客户端具体用于通过分段方式依次向所述数据池写入数据；其中，所述客户端判定数据成功写入后，才会向所述数据池写入下一段数据。

[0076] 本发明实施例还公开了一种基于分布式集群系统的数据存储设备，包括：

[0077] 存储器，用于存储计算机程序；

[0078] 处理器，用于执行所述计算机程序时实现上述任意方法实施例中基于分布式集群系统的数据存储方法的步骤。

[0079] 本发明实施例还公开了一种计算机可读存储介质，所述计算机可读存储介质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上述任意方法实施例中基于分布式集群系统的数据存储方法的步骤。

[0080] 其中，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。

[0081] 本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。

[0082] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

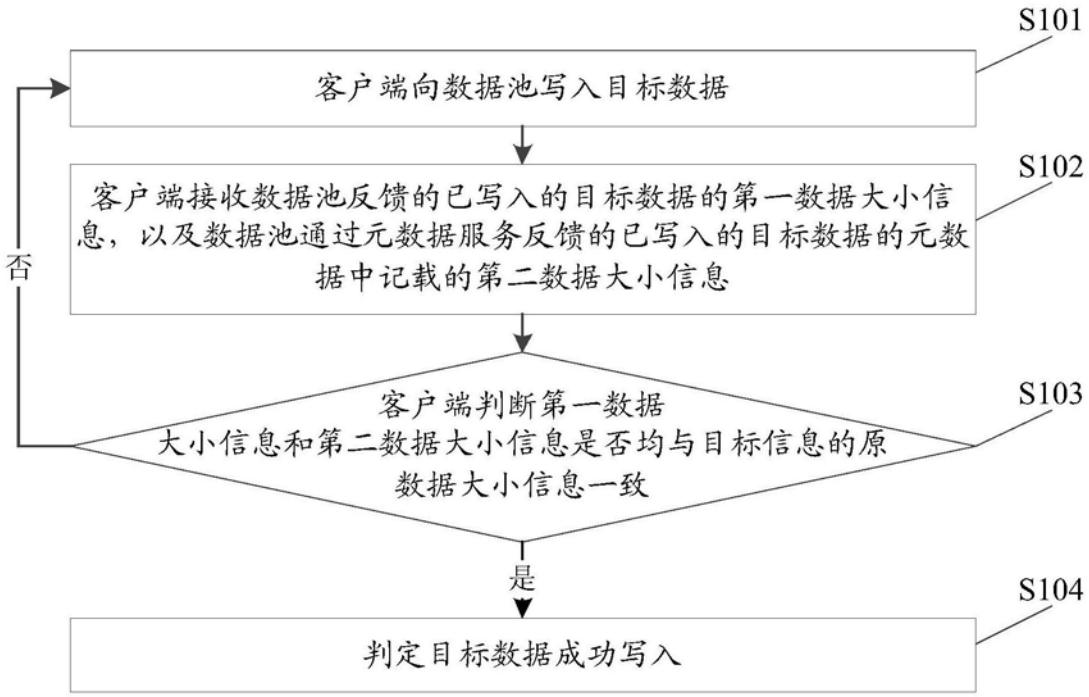

S101
客户端向数据池写入目标数据
S102
客户端接收数据池反馈的已写入的目标数据的第一数据大小信息，以及数据池通过元数据服务反馈的已写入的目标数据的元数据中记载的第二数据大小信息
S103
客户端判断第一数据大小信息和第二数据大小信息是否均与目标信息的原数据大小信息一致
否
是
S104
判定目标数据成功写入

图1

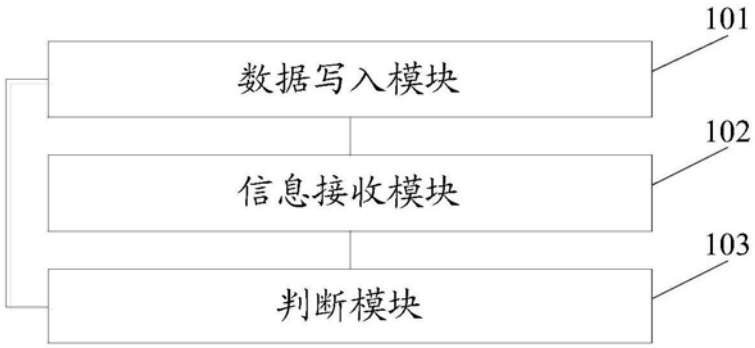

101
数据写入模块
102
信息接收模块
103
判断模块

图2

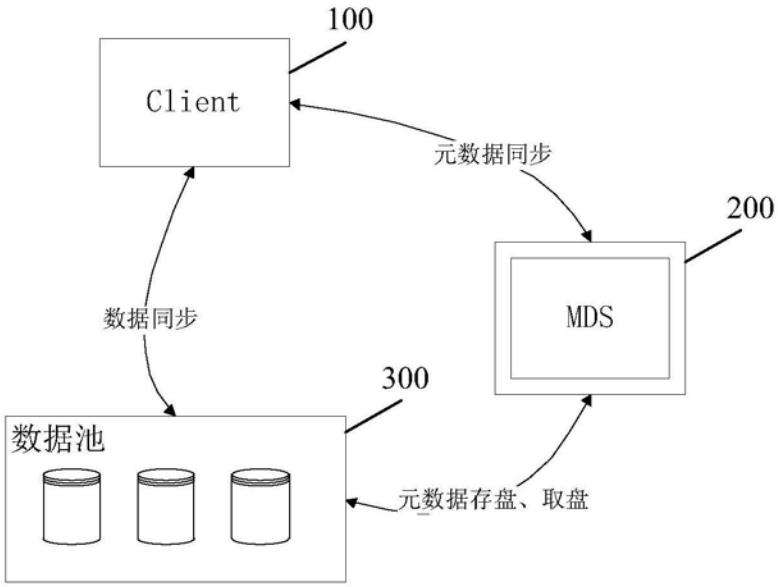


图3