(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2011 年 12 月 22 日 (22.12.2011)

PCT

(10) 国际公布号
WO 2011/157156 A2

(51) 国际专利分类号:
G06F 12/00 (2006.01)

(21) 国际申请号: PCT/CN2011/075077

(22) 国际申请日: 2011 年 6 月 1 日 (01.06.2011)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人; 及
(75) 发明人/申请人 (仅对美国): **程实 (CHENG, Shi)** [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: **北京中博世达专利商标代理有限公司 (BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.)**; 中国北京市海淀区大柳树路 17 号富海大厦 B 座 501 室, Beijing 100081 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

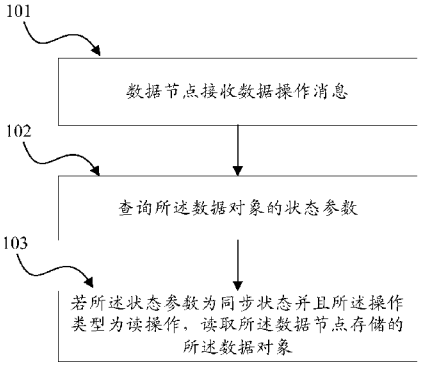
本国际公布:

— 根据申请人的请求，在条约第 21 条(2)(a)所规定的期限届满之前进行。

— 不包括国际检索报告，在收到该报告后将重新公布(细则 48.2(g))。

(54) **Title**: OPERATION METHOD AND DEVICE FOR DATA STORAGE SYSTEM

(54) **发明名称** : 数据存储系统的操作方法和装置



101 A DATA NODE RECEIVING A DATA OPERATION MESSAGE
102 QUERYING A STATUS PARAMETER OF THE DATA OBJECT
103 IF THE STATUS PARAMETER IS A SYNCHRONOUS STATUS AND THE OPERATION TYPE IS A READ OPERATION, READING THE DATA OBJECT STORED BY THE DATA NODE

图 1 / **Fig. 1**

(57) **Abstract**: An operation method and device for a data storage system are provided, which relate to the field of data storage and enhance the efficiency of a read operation to distributed stored data. The method comprises: a data node receiving a data operation message, the data operation message carries a data operation type and a data object of the operation; querying a status parameter of the data object, the status parameter is assigned to the data node by a management node, and is used to indicate an updating status of the data object, the updating status includes synchronous status, semi-synchronous status and asynchronous status; if the status parameter is a synchronous status and the operation type is a read operation, reading the data object stored by the data node. The invention is used to operate the data in a distributed data storage system.

(57) **摘要**: 提供了一种数据存储系统的操作方法和装置，涉及数据存储领域，提高了对分布式存储的数据的读操作效率。该方法包括：数据节点接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象；查询所述数据对象的状态参数，所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述更新状态包括同步状态、半同步状态、异步状态；若所述状态参数为同步状态，并且所述操作类型为读操作，则读取所述数据节点存储的所述数据对象。本发明用于对分布式数据存储系统中的数据进行操作。

WO 2011/157156 A2

数据存储系统的操作方法和装置

技术领域

本发明涉及数据存储领域，尤其涉及一种数据存储系统的操作方法和装置。

背景技术

分布式数据存储系统是由经网络互联的多个存储设备组成的存储系统。在该系统中，数据在多个数据节点上进行备份。传统的分布式数据存储系统的数据节点通常包括多个主节点，各个主节点分别保存全部数据的一部分，每个主节点都连接有一组从节点。当用户读取数据时，直接在主节点执行读数据操作；当用户写入数据时，在主节点执行写数据操作，并由主节点将数据复制到从节点，以使从节点保存与相连的主节点相同的数据副本。在主节点故障时，通过主从节点切换将一个从节点升级为主节点，保证正常的读写操作。传统的主从节点的存储系统中，各个从节点必须配置与主节点性能相近的硬件，以替换故障的主节点进行工作，这导致了过高的硬件成本；另外，在当前网络通常为松散组网的背景下，网络中的节点经常出现连接中断或超时，由此会引发主从节点的频繁切换，影响系统性能。

为解决传统的主从节点的存储系统的问题，现有技术提供了基于（N,W,R）策略的管理方案。该方案取消了主从节点的概念，每个数据节点保存全部数据中的一部分；并且对于某个数据 X，分配 N 个数据节点作为存储数据 X 的副本节点，即数据 X 在数据存储系统中存有 N 个副本。当对数据 X 进行写操作时，要在 W 个副本节点完成对该数据的写操作后才能结束本次写操作；当对数据 X 进行读操作时，必须在 R 个副本节点中读出。其中，N、W、R 之间满足 W+R>N 的关系，以保证读取的 R 个数据中至少有一个为最新版本。

在实现上述方案的过程中，发明人发现现有技术中至少存在如下问题：首先，基于（N,W,R）策略的管理方案必须经过对 R 个副本节点的读操作后，才能确定一个数据的最新版本，读操作的效率较低。此外，数

据存储系统中通常要支持对数据进行复杂条件查询，即通过遍历数据筛选出符合指定的查询条件的数据，并对筛选出的数据执行计算或写入动作；而基于（N,W,R）策略的方案中，任意数据都在 N 个副本节点中存有副本，这导致在进行复杂条件查询时，对于每个数据必须遍历 R 个副本节点才能确定最新版本的数据，数据遍历的操作量极为巨大，实际应用中难以实现。

发明内容

本发明的实施例提供一种数据存储系统的操作方法和装置，提高了对分布式存储的数据的读操作效率。

为达到上述目的，本发明的实施例采用如下技术方案：

一种数据存储系统的操作方法，包括：

数据节点接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象；

查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述更新状态包括同步状态、半同步状态、异步状态；

若所述状态参数为同步状态，并且所述操作类型为读操作，则读取所述数据节点存储的所述数据对象。

一种数据存储系统的操作装置，包括：

操作消息接收单元，用于接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象；

状态参数查询单元，用于查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述更新状态包括同步状态、半同步状态、异步状态；

数据读取单元，用于在所述状态参数为同步状态并且所述操作类型为读操作时，读取所述数据节点存储的所述数据对象。

本发明实施例提供的数据存储系统的操作方法和装置，对数据节点中保存的数据分配了状态参数，并将状态参数为同步状态的数据作为最新的数据读取出来，避免了现有技术分别从 R 个节点读取数据的操作，

提高了读操作的效率。

附图说明

图 1 为本发明实施例 1 中数据存储系统的操作方法的流程图;

图 2 为本发明实施例 1 中数据存储系统的操作装置的框图;

图 3 为本发明实施例 2 中操作类型为读操作时的操作方法的流程图;

图 4 为本发明实施例 2 中操作类型为写操作时的操作方法的流程图;

图 5 为本发明实施例 2 中数据节点与管理节点进行交互的操作方法的流程图;

图 6 为本发明实施例 3 中数据存储系统的操作装置的框图;

图 7 为本发明实施例 3 中节点状态表更新单元的框图;

图 8 为本发明实施例 2 中一种分布式数据存储系统的结构图;

图 9 为图 8 所示的分布式数据存储系统的数据节点 1 的节点状态表的示意图;

图 10 为基于 account group 的节点状态表的示意图。

具体实施方式

本发明的技术方案中的分布式数据存储系统中没有主从节点的区分，每个数据节点保存全部数据中的一部分。

一般的，分布式数据存储系统中存储的数据项具备 account、key、value、version 四种属性。key 为数据项的唯一标识，value 为数据项的内容，version 为数据项的最新版本，account 为数据项的账户号，并且不同的数据项可分配同一种账户号。

对于每个数据节点仅保存全部数据中的一部分的分布式数据存储系统，在系统建立之初，可将所有的 account 预先分成若干个 account group(账户组)，并且所有数据节点以 account group 为单位来保存数据。举例来说,对于某条数据项 X,若该数据项 X 的 account 为 account_X,则 account_X 必定属于某个 account group。假设 account_X 所属的 account group 为 group_X，则保存数据 X 的数据节点还保存 group_X 中的所有 account 对应的数据项。对于每个 account group，由 N 个数

据节点保存每个 account group 的数据的副本，其中 N 称为副本阈值并且 N 小于数据节点的总数。因此，该分布式数据存储系统中的每条数据项都在 N 个数据节点中存有副本。

下面结合本发明实施例的附图对本发明实施例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。

实施例 1:

本发明实施例提供了一种数据存储系统的操作方法，如图 1 所示，所述方法包括以下步骤:

101、数据节点接收数据操作消息。

所述数据操作消息由客户端发送到数据节点，并且所述数据操作消息中携带了数据操作类型以及操作的数据对象。

102、查询所述数据对象的状态参数。

当所述数据节点收到所述数据操作消息后，执行查询存储的所述数据对象的状态参数的操作。

所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述更新状态包括同步状态、半同步状态、异步状态。对于更新为最新版本的数据，其状态参数为同步状态；对于还未更新至最新版本的数据，其状态参数为半同步状态或异步状态。特别的，还可以在状态参数中设置一种未指定状态，用来表示可能出现的既不是同步状态、也不是半同步或异步状态的异常情况。

103、若所述状态参数为同步状态并且所述操作类型为读操作，读取所述数据节点存储的所述数据对象。

当所述数据节点获取到所述状态参数为同步状态时，进而查询所述数据操作消息中的数据操作类型。如果所述数据操作类型为读操作，则所述数据节点对所述状态参数为同步状态的数据对象执行读操作，并将读取的结果返回给客户端。

本发明实施例还提供了一种数据存储系统的操作装置，如图 2 所示，所述装置包括：操作消息接收单元 21、状态参数查询单元 22、数据读取单元 23。

操作消息接收单元 21 用于接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象。状态参数查询单元 22 用于查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述更新状态包括同步状态、半同步状态、异步状态。数据读取单元 23 用于在所述状态参数为同步状态并且所述操作类型为读操作时，读取所述数据节点存储的所述数据对象。

本发明实施例提供的数据存储系统的操作方法和装置，对数据节点中保存的数据分配了状态参数，并将状态参数为同步状态的数据作为最新的数据读取出来，避免了现有技术分别从 R 个节点读取数据的操作，提高了读操作的效率。

实施例 2:

本发明实施例提供了一种数据存储系统的操作方法，如图 3 所示，所述方法包括以下步骤:

301、数据节点接收数据操作消息。

所述数据操作消息由客户端发送到数据节点，并且所述数据操作消息中携带了数据操作类型以及操作的数据对象。

如果所述数据节点未存储所述数据操作消息中携带的所述数据对象，则所述数据节点向客户端返回一个异常消息，以使客户端向其他的数据节点重发该数据操作消息。

302、查询所述数据节点当前的管理状态参数。

在所述数据节点的管理状态参数为正常状态时，执行步骤 303；当所述数据节点的管理状态参数为中断状态时，终止当前的数据操作。

303、查询所述数据对象的状态参数。

所述状态参数由所述管理节点为所述数据节点分配，包括同步状态、半同步状态、异步状态。对于更新为最新版本的数据，其状态参数为同步状态；对于还未更新至最新版本的数据，其状态参数为半同步状态或异步状态。特别的，还可以在状态参数中设置一种未指定状态，用来表示可能出现的既不是同步状态、也不是半同步或异步状态的异常情况。

查询所述数据对象的状态参数后，如果所述状态参数为同步状态，则转向步骤 304；如果所述状态参数为半同步状态或异步状态，则转向

步骤 305。

304、若所述数据操作类型为读操作，读取所述数据节点存储的所述数据对象。

所述数据节点读取自身存储的所述数据对象，并将读取的结果返回客户端。

305、若所述数据操作类型为读操作，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点。

每个数据节点都保存有一份节点状态表，用于记录各个数据节点保存的数据以及所述保存的数据对应的状态参数。所述数据节点通过查询节点状态表来获取数据的状态参数为同步状态的其他数据节点，从而可以通知所述其他数据节点进行读操作。

比如，在如图 8 所示的分布式数据存储系统中，数据 1 至数据 4 在数据节点 1 至数据节点 4 中有着不同的存储分布。数据节点 1 的节点状态表可以是如图 9 所示的情况。在图 9 的节点状态表中记录了数据节点 1 至数据节点 4 的存储情况。其中，考虑到整个分布式数据存储系统的数据更新可能不是完全实时，数据节点 1 的节点状态表记录的其他数据节点（数据节点 2 至数据节点 4）的情况可能会与实际情况不符。图 8、图 9 描述的是一种极简化的情况，由于分布式数据存储系统中的数据量通常很庞大，像图 9 一样对每条数据进行记录较难实现。实际的分布式数据存储系统中的数据节点以 account group 为单位保存数据，因此，数据节点的节点状态表通常是以 account group 为单位来进行数据的状态参数的记录，如图 10 所示。这样，一个数据节点所保存的一个 account group 下的所有数据都具有相同的状态参数。本发明实施例下文中提及的数据节点的节点状态表中的状态参数，可以理解为以每条数据项为单位分配的状态参数，也可以理解为以每个 account group 为单位分配的状态参数。通常，在数据量很大的分布式数据存储系统中，以 account group 为单位分配状态参数是优选的方案。

306、向所述其他数据节点中的其中一个数据节点发送读操作消息。

所述读操作消息中包括所述数据对象，所述其他数据节点接收所述读操作消息后读取自身存储的所述数据对象，并将读取的结果返回客户端。

此外，对于所述数据操作类型为写操作的情况，如图 4 所示，所述

方法还包括如下步骤:

303、查询所述数据对象的状态参数。

查询所述数据对象的状态参数后，如果所述状态参数为同步状态，则转向步骤 307；如果所述状态参数为半同步状态或异步状态，则转向步骤 308。

307、若所述状态参数为同步状态且所述数据操作类型为写操作，对所述数据节点存储的所述数据对象进行写入。

308、若所述状态参数为半同步状态或异步状态且所述数据操作类型为写操作，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点。

309、向所述状态参数为同步状态的其他数据节点中的其中一个数据节点发送第一写操作消息。

所述第一写操作消息中包括所述数据对象，所述其中一个数据节点接收所述第一写操作消息后对自身存储的所述数据对象进行写入操作。

310、向保存所述数据对象且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息。

在步骤 307 完成写入动作后，以及在步骤 309 完成写入动作后，完成写入动作的所述数据节点或所述状态参数为同步状态的其他数据节点中的其中一个数据节点将向保存有所述数据对象并且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息，以便所述所有其他数据节点都进行对所述数据对象的写入，实现整个分布式数据存储系统对所述数据对象的更新。

下面举例来对读操作和写操作进行说明。假设在分布式数据存储系统中， 数据节点 A、B、C、D 存储了数据 a。

如果数据节点 A 收到了客户端发送的数据操作消息并且该数据操作消息要求对数据 a 进行读操作，则当数据节点 A 查询自身的节点状态表并获取到自身存储的数据 a 的状态参数为同步状态时，数据节点 A 直接读取自身存储的数据 a；当数据节点 A 查询自身的节点状态表并获取到自身存储的数据 a 的状态参数为半同步或异步状态时，数据节点 A 将向自身的节点状态表中记录的存有数据 a 且其状态参数为同步状态的其他数据节点中的其中一个数据节点（不妨假设所述其中一个数据节点为数据节点 C）发送读操作消息，从而由数据节点 C 进行对数据 a 的

读取。

如果数据节点 A 收到了客户端发送的数据操作消息并且该数据操作消息要求对数据 a 进行写操作，则当数据节点 A 查询自身的节点状态表并获取到自身存储的数据 a 的状态参数为同步状态时，对自身存储的数据 a 进行写入；当数据节点 A 查询自身的节点状态表并获取到自身存储的数据 a 的状态参数为半同步或异步状态时，数据节点 A 将向自身的节点状态表中记录的存有数据 a 且其状态参数为同步状态的其他数据节点中的其中一个数据节点（不妨假设所述其中一个数据节点为数据节点 C）发送第一写操作消息，从而由数据节点 C 进行对数据 a 的写入。在完成上述的对数据 a 的写入后，还要向保存数据 a 且其状态参数为同步、半同步或异步状态的所有其他数据节点（比如数据节点 B、D）发送第二写操作消息，以使数据节点 B、D 也进行对自身存储的数据 a 的写入。这样，系统中所有保存有数据 a 的数据节点都进行了写入动作，从而完成对数据 a 的更新。

在执行上述的读操作、写操作的过程中，所述数据节点还要对异常情况进行分析处理。

对于数据的读操作，所述数据节点发出所述读操作消息后，在所述第一等待时间段中等待所述其中一个数据节点的响应。

若在所述第一等待时间段内收到所述其中一个数据节点因管理状态参数为中断状态而返回的管理状态异常消息，则将所述读操作消息发送到所述其他数据节点中的另一个数据节点；若所述第一等待时间段内收到所述其中一个数据节点因数据对象的状态参数不是同步状态而返回的更新状态异常消息，则对所述数据节点的节点状态表中记录的所述其中一个数据节点的数据对象的状态参数进行更新；在完成所述状态参数的更新后，将所述读操作消息发送到所述其他数据节点中的另一个数据节点；若在所述第一等待时间段内未接收所述其中一个数据节点的响应，则在所述数据节点存储的超时记录表中将所述其中一个数据节点的超时次数增加一次。

举例来说，数据节点 N 向数据节点 M 发送了读操作消息，以使数据节点 M 对数据 n 进行读操作。数据节点 N 在所述第一等待时间段内等待数据节点 M 返回操作结果，包括以下几种情况：

S1、数据节点 M 完成了读操作，并返回了操作结果。

S2、如果数据节点M的自身保存的管理状态参数为中断状态，则数据节点M向数据节点N发送管理状态异常消息。数据节点N接收所述管理状态异常消息后从自身的节点状态表查找到另一个存有数据 n 且其状态参数为同步状态的数据节点，并向该数据节点发送读操作消息。

S3、如果数据节点M接收所述读操作消息后，从自身的节点状态表查找到自身存储的数据n不处于同步状态，则数据节点M向数据节点N返回更新状态异常消息，所述更新状态异常消息携带了数据节点M存储的数据n的状态参数。数据节点N接收所述更新状态异常消息后，将对自身的节点状态表记录的数据节点M存储的数据n的状态参数进行更新动作。所述更新动作包括：

S301、数据节点N从所述更新状态异常消息中获取所述数据节点M存储的数据n的状态参数。

S302、将节点状态表记录的数据节点M的数据n的状态参数更新为从所述更新状态异常消息中获取的状态参数。数据节点N的节点状态表中记录的数据节点M的数据n的状态参数原本为同步状态，在接收所述更新状态异常消息并进行状态参数的更新后，数据节点N的节点状态表记录的数据节点 M 的数据 n 的状态参数将变更为半同步状态或异步状态。

如果在所述第一等待时间段内，数据节点 N 没有收到数据节点 M 的任何响应，则在数据节点 N 存储的超时记录表中将所述数据节点 M 的超时次数增加一次。

对于数据的写操作，所述数据节点发出所述第一写操作消息后同样将在所述第一等待时间段中等待接收了所述第一写操作消息的数据节点的响应。对不同响应的处理可参照上述读操作所举的例子。

对于发送所述第二写操作消息的情况，所述数据节点发出所述第二写操作消息后，将在另一个预设的时间段中等待响应。假设所述另一个预设的时间段为第二等待时间段。举例来说，数据节点P完成对自身存储的数据 p 的写入操作后，向所有存储由数据 p 的数据节点发送第二写操作消息。其中，对于保存数据 p 且其状态参数为同步或半同步的数据节点，数据节点P将发送同步型第二写操作消息；对于保存数据 p 且其状态参数为异步的数据节点，数据节点 P 将发送异步型第二写操作消息。不妨假设数据节点Q接收到所述同步型第二写操作消息并且数据节

点 R 接收到所述异步型第二写操作消息。数据节点 P 在所述第二等待时间段中等待返回的操作结果，包括以下几种情况:

S4、数据节点 P 接收到写入成功的消息。

S5、数据节点 P 接收到更新状态异常消息。

以数据节点 Q 为例说明 S5 的情况。数据节点 Q 接收所述同步型第二写操作消息后，在自身的节点状态表中查询存储的数据 p 的状态参数。如果查询到所述数据 p 的状态参数不是同步状态或半同步状态，则向数据节点 P 返回更新状态异常消息，所述更新状态异常消息携带了数据节点 Q 存储的数据 p 的状态参数(所述更新状态异常消息携带的状态参数为异步状态)。数据节点 P 接收所述更新状态异常消息后，将进行更新动作。所述更新动作包括: 数据节点 P 根据接收到的所述更新状态异常消息中携带的数据 p 的状态参数，将自身的节点状态表记录的数据 p 的状态参数更新为异步状态。更新完成后，向数据节点 Q 发送异步型第二写操作消息，并等待返回结果。

如果在所述第二等待时间段内，数据节点 P 没有收到数据节点 Q 的任何响应，则在数据节点 P 存储的超时记录表中将所述数据节点 Q 的超时次数增加一次。

对于接受所述异步型第二写操作消息的数据节点 R 的处理过程与上述对数据节点 Q 的处理过程类似。此外，为提升整个系统的处理性能，数据节点 P 可以在收到数据节点 Q 返回的写入成功的消息后结束在所述第二等待时间段中的等待。也就是说，数据节点 P 向所有存储数据 p 的数据节点发送所述第二写操作消息后，在接收到所有保存数据 p 且其状态参数为同步或半同步状态的数据节点返回的写入成功的消息后就可以结束当前在所述第二等待时间段中的等待，并继续进行其他操作，而不必等待保存数据 p 且状态参数为异常状态的数据节点返回的写入结果。这样，分布式数据存储系统需要等待的数据节点越少，可能出现的操作异常的情况也越少，从而有利于提升系统整体的处理效率。

所述数据节点除了进行读写操作及异常处理，还需与所述管理节点进行连通和信息交互，如图 5 所示，包括如下步骤:

501、所述数据节点以预设的时间段为周期，向管理节点发起管理连接请求。

所述管理节点接收到所述管理连接请求后，将所述数据节点的管理

状态参数设置为正常状态。当所述管理连接请求的连续连接失败的次数达到预设的连接失败计数阈值时，所述数据节点将自身的管理状态参数设置为中断状态。

502、所述数据节点向所述管理节点发送所述数据节点的超时记录表。

在所述管理连接请求被所述管理节点接收后，所述数据节点向所述管理节点发送所述数据节点的超时记录表。所述管理节点接收所述数据节点的超时记录表，从所述超时记录表获取所述数据节点的超时次数，并将达到超时阈值的数据节点的所有数据的状态参数修改为异步状态并记录到管理节点状态表。

作为实际应用中的一种实现方式，所述数据节点发送的超时记录表中可以还包括：在预设的超时记录时间段内的超时比值，所述超时比值为是指在预设的超时记录时间段内所述数据节点的总超时次数与所述数据节点向其他数据节点进行连接请求的次数的比值。所述管理节点获取到所述超时比值后，与所述超时阈值进行比对，对于所述超时比值大于等于所述超时阈值的数据节点，所述管理节点将其所有数据的状态参数变更为异步状态。此外，所述管理节点还根据预设的节点数阈值 S 对状态参数进行动态调整。比如，如果存储数据 h 且状态参数为同步和半同步状态的所有数据节点的数量不足所述节点数阈值 S 时，所述管理节点从所有的存有数据 h 且状态参数为异步状态的数据节点中，获取超时比值最小(即在预设的超时记录时间段内所述数据节点的总超时次数与所述数据节点向其他数据节点进行连接请求的次数的比值最小)的数据节点，并在所述管理节点状态表将该数据节点存储的数据 h 的状态参数变更为半同步状态。

所述节点数阈值 S 预先设置在分布式数据存储系统中。所述节点数阈值 S 一般小于所述副本阈值 N。所述节点数阈值 S 的取值应考虑系统整体的操作性能和实时性，因而不应取值过大；同时还应考虑系统整体的容灾能力，因此也不应取值过小。通常，对于不同的系统，可通过实测或仿真实验得到最优的 S 取值。当存储数据 h 且状态参数为同步状态的数据节点数量达到所述节点数阈值 S，则管理节点暂停生成新的同步状态的操作，以维持所述节点数阈值 S。

503、所述数据节点接收所述管理节点发送的管理节点状态表。

504、从所述管理节点状态表中获取各个数据节点保存的数据的状态参数。

所述管理节点执行如步骤 502 中描述的对数据节点存储数据的状态参数的变更后，将变更后的所述管理节点状态表发送给所述数据节点。此时，所述数据节点从所述管理节点状态表中获取的状态参数可能与所述数据节点自身的节点状态表中记录状态参数不同，因此需要进行状态参数的更新。

505、将所述数据节点的节点状态表中的状态参数更新为所述获取的各个数据节点保存的数据的状态参数。

当从所述管理节点状态表中获取的各个数据节点的数据的状态参数与所述数据节点的节点状态表中的各个数据节点的数据的状态参数不同时，将所述节点状态表中的所述不同的状态参数更新为从所述管理节点状态表中获取的各个数据节点的数据的状态参数。

506、在更新所述不同的状态参数后，若所述数据节点的部分数据的状态参数变更为半同步状态，则从存储所述部分数据且所述部分数据的状态参数为同步状态的另一个数据节点复制所述部分数据。

所述数据节点从所述另一个数据节点复制所述部分数据，以使所述数据节点保存的部分数据能够从半同步状态恢复为同步状态。

实际情况中，如果不存在所述存储所述部分数据且所述部分数据的状态参数为同步状态的另一个数据节点，则向所有存储有所述部分数据的数据节点进行连接，以获取版本最新的所述部分数据进行复制。

507、所述数据节点完成所述复制后，向所述管理节点发送同步完成消息。

所述管理节点收到所述同步完成消息后，判断是否将所述数据节点的所述部分数据的状态参数变更为同步状态。

比如，数据节点 H 完成对数据 h 的复制后，向所述管理节点发送同步完成消息。所述管理节点根据自身的管理节点状态表，判断存储数据 h 且其状态参数为同步状态和半同步状态的所有数据节点的总数是否达到所述节点数阈值 S。如果没有达到所述节点数阈值 S，则向所述数据节点 H 发送同步确认消息，以使数据节点 H 将自身存储的数据 h 的状态参数变更为同步状态；如果达到所述节点数阈值 S，则向所述数据节点 H 发送同步终止消息，使数据节点 H 将自身存储的数据 h 的状态参数

维持为半同步状态。另外，在所述数据节点复制所述部分数据时，如此时所述数据节点保存的所述部分数据的状态参数状态变为异步状态，或所述数据节点的管理状态变为中断状态，则终止当前的复制动作。

本发明实施例提供的数据存储系统的操作方法，对数据节点中保存的数据分配了状态参数，并将状态参数为同步状态的数据作为最新的数据读取出来，避免了现有技术分别从 R 个节点读取数据的操作，提高了读操作的效率；同时，根据状态参数能确定最新版本的数据，不必对同一数据在不同数据节点中的副本进行遍历，使复杂条件查询的实现更为简便。此外，通过引入管理节点对分布式数据存储系统中的连接超时等异常情况进行处理，提高了系统整体的可用性。

实施例 3:

本发明实施例提供了一种数据存储系统的操作装置，如图 6 所示，所述装置包括：操作消息接收单元 61、状态参数查询单元 62、数据读取单元 63、第一节点查询单元 64、读操作消息发送单元 65、数据写入单元 66、第二节点查询单元 67、第一写操作发送单元 68、第二写操作发送单元 69、第二读操作发送单元 610、状态参数更新单元 611、超时状态更新单元 612、管理连接请求单元 613、超时记录发送单元 614、管理状态接收单元 615、节点状态表更新单元 616、同步数据复制单元 617、同步完成消息发送单元 618、中断状态设置单元 619。

此外，所述节点状态表更新单元 616 还包括：第一状态参数获取模块 6161、第一状态参数更新模块 6162、第二状态参数获取模块 6163、第二状态参数更新模块 6164，如图 7 所示。

操作消息接收单元 61 用于接收携带有数据操作类型以及操作的数据对象的数据操作消息。状态参数查询单元 62 用于查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，包括同步状态、半同步状态、异步状态。所述状态参数查询单元 62 还用于在所述数据节点的管理状态参数为正常状态时查询所述数据对象的状态参数。数据读取单元 63 用于在所述状态参数为同步状态并且所述操作类型为读操作时，读取所述数据节点存储的所述数据对象。第一节点查询单元 64 用于在所述状态参数为半同步状态或异步状态并且所述操作类型为读操作时，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点。读操作消息发送单元 65 用于向所述其他数据

节点中的其中一个数据节点发送读操作消息，所述读操作消息中包括所述数据对象，以便所述其他数据节点接收所述读操作消息后读取所述数据对象。数据写入单元 66 用于在所述状态参数是同步状态并且所述操作类型为写操作时，对所述数据节点存储的所述对象数据进行写入。第二节点查询单元 67 用于在所述状态参数为半同步状态或异步状态并且所述操作类型为写操作时，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点。第一写操作发送单元 68 用于在获取所述状态参数为同步状态的其他数据节点后向所述状态参数为同步状态的其他数据节点中的其中一个数据节点发送第一写操作消息，以便所述其他数据节点接收所述第一写操作消息后对所述数据对象进行写入。第二写操作发送单元 69 用于在所述数据节点或所述状态参数为同步状态的其他数据节点中的其中一个数据节点完成对所述数据对象的写入后，向保存所述数据对象且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息，以便所述所有其他数据节点进行对所述数据对象的写入。第二读操作发送单元 610 用于在预设的第一等待时间段内收到所述其中一个数据节点因管理状态参数为中断状态而返回的管理状态异常消息后，将所述读操作消息发送到所述其他数据节点中的另一个数据节点。状态参数更新单元 611 用于在预设的第一等待时间段内收到所述其中一个数据节点因数据对象的状态参数不是同步状态而返回的更新状态异常消息时，对所述数据节点的节点状态表中记录的所述其中一个数据节点的数据对象的状态参数进行更新；在完成所述状态参数的更新后，将所述读操作消息发送到所述其他数据节点中的另一个数据节点。超时状态更新单元 612 用于在所述第一等待时间段内未接收所述其中一个数据节点的响应时，在所述数据节点存储的超时记录表中将所述其中一个数据节点的超时次数增加一次。

在与所述管理节点进行互连时，管理连接请求单元 613 用于以预设的时间段为周期，向管理节点发起管理连接请求。超时记录发送单元 614 用于在所述管理连接请求被所述管理节点接收后，向所述管理节点发送所述数据节点的超时记录表。管理状态接收单元 615 用于在所述管理连接请求被所述管理节点接收后，接收所述管理节点发送的管理节点状态表。节点状态表更新单元 616 用于依照所述管理节点状态表更新所

述数据节点的节点状态表。

所述节点状态表更新单元 616 中的第一状态参数获取模块 6161 用于从更新状态异常消息中获取所述其中一个数据节点的数据对象的状态参数。第一状态参数更新模块 6162 将所述节点状态表中记录的数据对象的状态参数变更为所述从更新状态异常消息中获取的数据对象的状态参数。

所述节点状态表更新单元 616 中的第二状态参数获取模块 6163 用于从所述管理节点状态表中获取各个数据节点保存的数据的状态参数。第二状态参数更新模块 6164 用于当所述获取的各个数据节点保存的数据的状态参数与所述数据节点的节点状态表中的状态参数不同时，将所述数据节点的节点状态表中的状态参数更新为所述获取的各个数据节点保存的数据的状态参数。

此外，同步数据复制单元 617 用于在节点状态表更新单元 616 更新所述数据节点的节点状态表后、并且当所述数据节点的部分数据的状态参数变更为半同步状态时，从存储所述部分数据且所述部分数据的状态参数为同步状态的另一个数据节点复制所述部分数据。同步完成消息发送单元 618 用于在完成对所述部分数据的复制后，向所述管理节点发送同步完成消息。中断状态设置单元 619 用于当所述数据节点向所述管理节点发起的管理连接请求连续连接失败的次数达到预设的连接失败计数阈值时，所述数据节点将自身的管理状态参数设置为中断状态。

。

对于本发明实施例提供的数据存储系统的操作装置的其他相关功能情况，可以参照前述的实施例 1 和实施例 2 的描述，本实施例不再赘述。

本发明实施例提供的数据存储系统的操作装置，对数据节点中保存的数据分配了状态参数，并将状态参数为同步状态的数据作为最新的数据读取出来，避免了现有技术分别从 R 个节点读取数据的操作，提高了读操作的效率，也使复杂条件查询的实现更为简便。此外，本发明实施例引入了管理节点，可以对分布式数据存储系统中的连接超时等异常情况进行处理，提高了系统整体的可用性。

通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本发明可借助软件加必需的通用硬件的方式来实现，当然也可以通过

硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在可读取的存储介质中，如计算机的软盘，硬盘或光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施例所述的方法。

以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应所述以权利要求的保护范围为准。

权 利 要 求 书

1、一种数据存储系统的操作方法，其特征在于，包括:

数据节点接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象;

查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述状态参数包括同步状态、半同步状态、异步状态;

若所述状态参数为同步状态，并且所述操作类型为读操作，则读取所述数据节点存储的所述数据对象。

2、根据权利要求 1 所述的方法，其特征在于，还包括:

若所述状态参数为半同步状态或异步状态，并且所述操作类型为读操作，则查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点；所述节点状态表记录各个数据节点保存的数据以及所述保存的数据对应的状态参数;

向所述其他数据节点中的其中一个数据节点发送读操作消息，所述读操作消息中包括所述数据对象，以便所述其他数据节点接收所述读操作消息后所述数据对象读取所述数据对象。

3、根据权利要求 1 所述的方法，其特征在于，所述查询所述数据对象的状态参数包括:

在所述数据节点的管理状态参数为正常状态时，查询所述数据对象的状态参数；当所述数据节点的管理状态参数为中断状态时，终止当前的数据操作。

4、根据权利要求 2 所述的方法，其特征在于，在所述数据节点向所述其他数据节点中的其中一个数据节点发出所述读操作消息后，还包括:

若在预设的第一等待时间段内收到所述其中一个数据节点因管理状态参数为中断状态而返回的管理状态异常消息，则将所述读操作消息发送到所述其他数据节点中的另一个数据节点；或者，

若在预设的第一等待时间段内收到所述其中一个数据节点因数据对象的状态参数不是同步状态而返回的更新状态异常消息，则对所述数据节点的节点状态表中记录的所述其中一个数据节点的数据对象的状态参数进行更新；在完成所述状态参数的更新后，将所述操作消息发送到所述其他数据节点中的另一个数据节点；或者，

若在所述第一等待时间段内未接收所述其中一个数据节点的响应，则在所述数据节点存储的超时记录表中将所述其中一个数据节点的超时次数增加一次。

5、根据权利要求 4 所述的方法，其特征在于，所述对所述数据节点的节点状态表中记录的所述其中一个数据节点的数据对象的状态参数进行更新包括：

从更新状态异常消息中获取所述其中一个数据节点的数据对象的状态参数；

将所述节点状态表中记录的所述其中一个数据节点的数据对象的状态参数变更为所述从更新状态异常消息中获取的数据对象的状态参数。

6、根据权利要求 2 或 4 所述的方法，其特征在于，还包括：

所述数据节点向所述管理节点周期性发送所述数据节点的超时记录表，以便所述管理节点根据所述超时记录表将达到超时阈值的数据节点的所有数据的状态参数修改为异步状态并记录到管理节点状态表；

所述数据节点接收所述管理节点发送的管理节点状态表；

依照所述管理节点状态表更新所述数据节点的节点状态表。

7、根据权利要求 6 所述的方法，其特征在于，若保存数据且状态参数为同步状态和半同步状态的数据节点的数量之和低于预设的节点数阈值，则所述管理节点从所述保存数据并且状态参数为异步状态的数据节点中选择超时记录最少的数据节点，并将所述超时记录最少的数据节点保存的数据的状态参数变更为半同步状态。

8、根据权利要求 6 所述的方法，其特征在于，所述数据节点依照所述管理节点状态表更新所述数据节点的节点状态表包括：

从所述管理节点状态表中获取各个数据节点保存的数据的状态参数；

当所述获取的各个数据节点保存的数据的状态参数与所述数据节点的节点状态表中的状态参数不同时，将所述数据节点的节点状态表中的状态参数更新为所述获取的各个数据节点保存的数据的状态参数。

9、根据权利要求 6 所述的方法，其特征在于，依照所述管理节点状态表更新所述数据节点的节点状态表后，还包括：

若更新所述数据节点的节点状态表后，所述数据节点的部分数据的状态参数变更为半同步状态，则从存储所述部分数据且所述部分数据的

状态参数为同步状态的另一个数据节点复制所述部分数据;

完成对所述部分数据的复制后，向所述管理节点发送同步完成消息，以便所述管理节点在收到所述同步完成消息后在所述管理节点状态表中将所述数据节点的所述部分数据的状态参数变更为同步状态；所述管理节点对所述管理状态表中的数据的状态参数进行动态调整，以使存储任何一种数据且状态参数为同步状态的数据节点与存储所述任何一种数据且状态参数为半同步状态的数据节点的总数不超过所述节点数阈值。

10、根据权利要求 6 所述的方法，其特征在于，还包括:

当所述数据节点向所述管理节点发起的管理连接请求的连续连接失败的次数达到预设的连接失败计数阈值时，所述数据节点将自身的管理状态参数设置为中断状态。

11、根据权利要求 1 所述的方法，其特征在于，还包括:

若所述状态参数是同步状态，并且所述操作类型为写操作，则对所述数据节点存储的所述对象数据进行写入;

若所述状态参数为半同步状态或异步状态，并且所述操作类型为写操作，则查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点;

在获取所述状态参数为同步状态的其他数据节点后，向所述状态参数为同步状态的其他数据节点中的其中一个数据节点发送第一写操作消息，所述第一写操作消息中包括所述数据对象，以便所述其他数据节点接收所述第一写操作消息后对所述数据对象进行写入;

在所述数据节点或所述状态参数为同步状态的其他数据节点中的其中一个数据节点完成对所述数据对象的写入后，向保存所述数据对象且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息，以便所述所有其他数据节点进行对所述数据对象的写入。

12、一种数据存储系统的操作装置，其特征在于，包括:

操作消息接收单元，用于接收数据操作消息，所述数据操作消息携带数据操作类型以及操作的数据对象;

状态参数查询单元，用于查询所述数据对象的状态参数；所述状态参数由管理节点为数据节点分配，用于表示数据对象的更新状态，所述状态参数包括同步状态、半同步状态、异步状态;

数据读取单元，用于在所述状态参数为同步状态并且所述操作类型为读操作时，读取所述数据节点存储的所述数据对象。

13、根据权利要求 12 所述的装置，其特征在于，还包括：

第一节点查询单元，用于在所述状态参数为半同步状态或异步状态并且所述操作类型为读操作时，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点；所述节点状态表记录各个数据节点保存的数据以及所述保存的数据对应的状态参数；

读操作消息发送单元，用于向所述其他数据节点中的其中一个数据节点发送读操作消息，所述读操作消息中包括所述数据对象，以便所述其他数据节点接收所述读操作消息后读取所述数据对象。

14、根据权利要求 12 所述的装置，其特征在于，所述状态参数查询单元还用于在所述数据节点的管理状态参数为正常状态时查询所述数据对象的状态参数；当所述数据节点的管理状态参数为中断状态时，终止当前的数据操作。

15、根据权利要求 13 所述的装置，其特征在于，在所述数据节点向所述其他数据节点中的其中一个数据节点发出所述读操作消息后，还包括：

第二读操作发送单元，用于在预设的第一等待时间段内收到所述其中一个数据节点因管理状态参数为中断状态而返回的管理状态异常消息后，将所述读操作消息发送到所述其他数据节点中的另一个数据节点；

状态参数更新单元，用于在预设的第一等待时间段内收到所述其中一个数据节点因数据对象的状态参数不是同步状态而返回的更新状态异常消息时，对所述数据节点的节点状态表中记录的所述其中一个数据节点的数据对象的状态参数进行更新；在完成所述状态参数的更新后，将所述读操作消息发送到所述其他数据节点中的另一个数据节点；

超时状态更新单元，用于在所述第一等待时间段内未接收所述其中一个数据节点的响应时，在所述数据节点存储的超时记录表中将所述其中一个数据节点的超时次数增加一次。

16、根据权利要求 15 所述的装置，其特征在于，所述节点状态更新单元包括：

第一状态参数获取模块，用于从更新状态异常消息中获取所述其中一个数据节点的数据对象的状态参数；

第一状态参数更新模块，将所述节点状态表中记录的数据对象的状态参数变更为所述从更新状态异常消息中获取的数据对象的状态参数。

17、根据权利要求 13 或 15 所述的装置，其特征在于，还包括：

管理连接请求单元，用于以预设的时间段为周期，向管理节点发起管理连接请求；

超时记录发送单元，用于在所述管理连接请求被所述管理节点接收后，向所述管理节点发送所述数据节点的超时记录表，以便所述管理节点根据所述超时记录表将达到超时阈值的数据节点的所有数据的状态参数修改为异步状态并记录到管理节点状态表；

管理状态接收单元，用于接收所述管理节点发送的管理节点状态表；

节点状态表更新单元，用于依照所述管理节点状态表更新所述数据节点的节点状态表。

18、根据权利要求 17 所述的装置，其特征在于，所述节点状态表更新单元还包括：

第二状态参数获取模块，用于从所述管理节点状态表中获取各个数据节点保存的数据的状态参数；

第二状态参数更新模块，用于当所述获取的各个数据节点保存的数据的状态参数与所述数据节点的节点状态表中的状态参数不同时，将所述数据节点的节点状态表中的状态参数更新为所述获取的各个数据节点保存的数据的状态参数；

19、根据权利要求 17 所述的装置，其特征在于，依照所述管理节点状态表更新所述数据节点的节点状态表后，还包括：

同步数据复制单元，用于当所述数据节点的部分数据的状态参数变更为半同步状态时，从存储所述部分数据且所述部分数据的状态参数为同步状态的另一个数据节点复制所述部分数据；

同步完成消息发送单元，用于在完成对所述部分数据的复制后，向所述管理节点发送同步完成消息，以便所述管理节点在收到所述同步完成消息后在所述管理节点状态表中将所述数据节点的所述部分数据的状态参数变更为同步状态；所述管理节点对所述管理状态表中的数据的状态参数进行动态调整，以使存储任何一种数据且状态参数为同步状态的数据节点与存储所述任何一种数据且状态参数为半同步状态的数据节点的总数不超过预设的节点数阈值。

20、根据权利要求 17 所述的装置，其特征在于，还包括：

中断状态设置单元，用于当所述数据节点与所述管理节点发起的管理连接请求的连续连接失败的次数达到预设的连接失败计数阈值时，所述数据节点将自身的管理状态参数设置为中断状态。

21、根据权利要求 12 所述的装置，其特征在于，还包括：

数据写入单元，用于在所述状态参数是同步状态并且所述操作类型为写操作时，对所述数据节点存储的所述对象数据进行写入；

第二节点查询单元，用于在所述状态参数为半同步状态或异步状态并且所述操作类型为写操作时，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点；

第一写操作发送单元，用于在获取所述状态参数为同步状态的其他数据节点后向所述状态参数为同步状态的其他数据节点中的其中一个数据节点发送第一写操作消息，所述第一写操作消息中包括所述数据对象，以便所述其他数据节点接收所述第一写操作消息后对所述数据对象进行写入；

第二写操作发送单元，用于在所述数据节点或所述状态参数为同步状态的其他数据节点中的其中一个数据节点完成对所述数据对象的写入后，向保存所述数据对象且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息，以便所述所有其他数据节点进行对所述数据对象的写入。

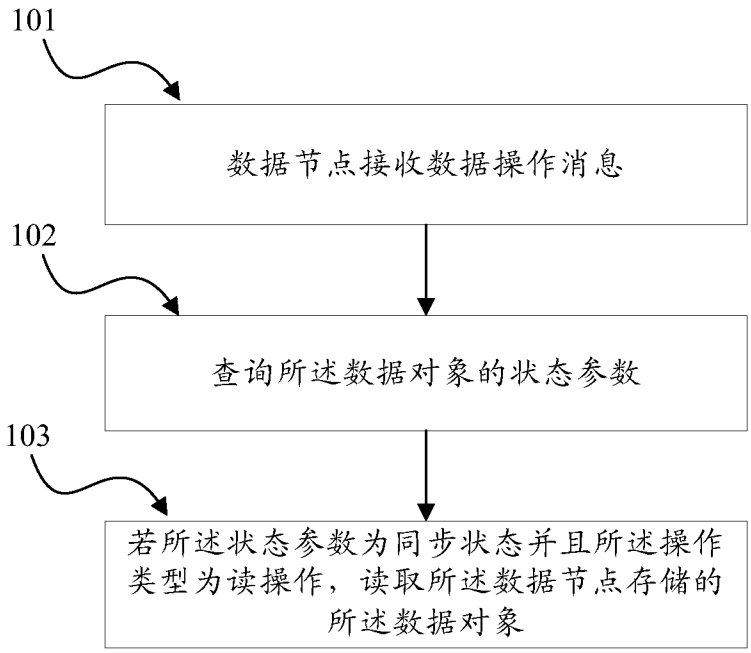

101
数据节点接收数据操作消息
102
查询所述数据对象的状态参数
103
若所述状态参数为同步状态并且所述操作类型为读操作，读取所述数据节点存储的所述数据对象

图 1

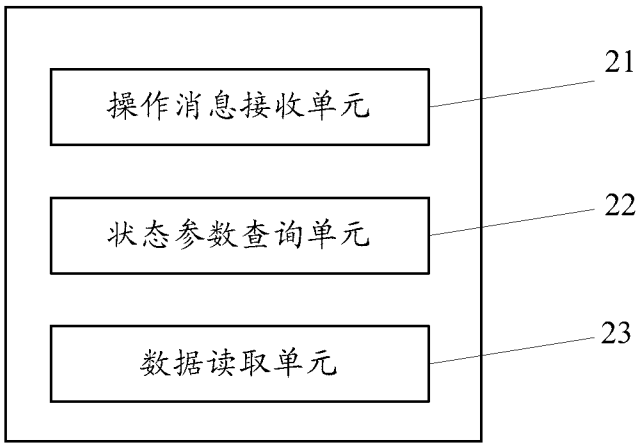

操作消息接收单元
21
状态参数查询单元
22
数据读取单元
23

图 2

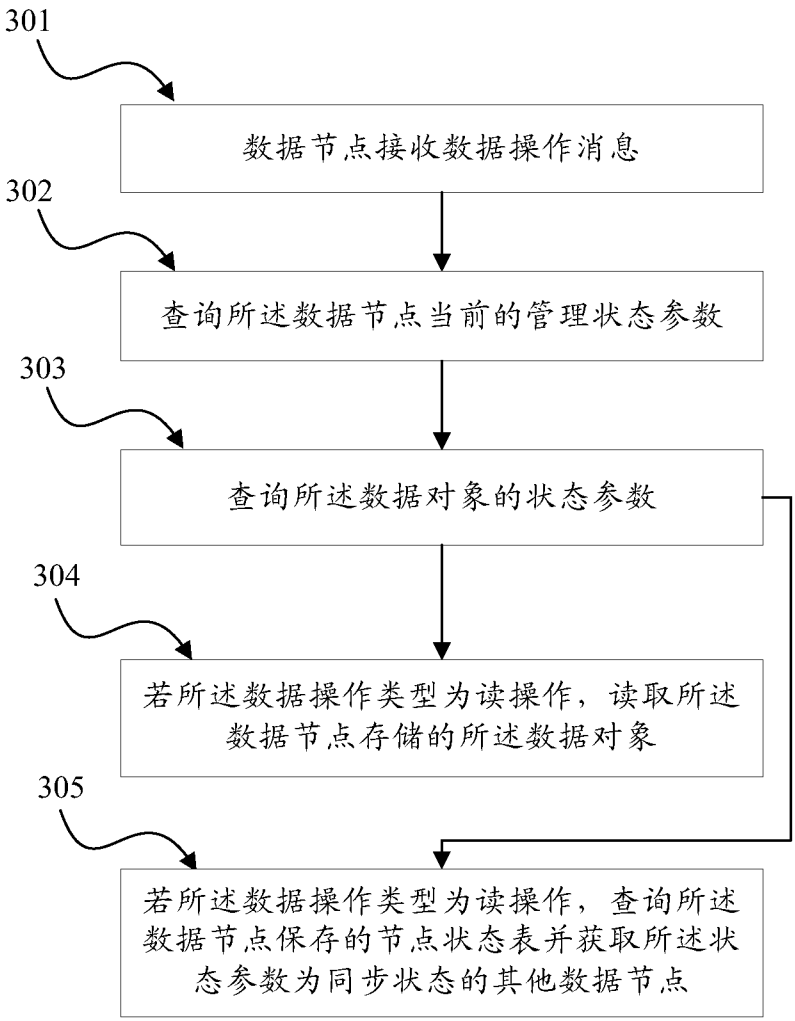

301
数据节点接收数据操作消息
302
查询所述数据节点当前的管理状态参数
303
查询所述数据对象的状态参数
304
若所述数据操作类型为读操作，读取所述数据节点存储的所述数据对象
305
若所述数据操作类型为读操作，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点

图 3

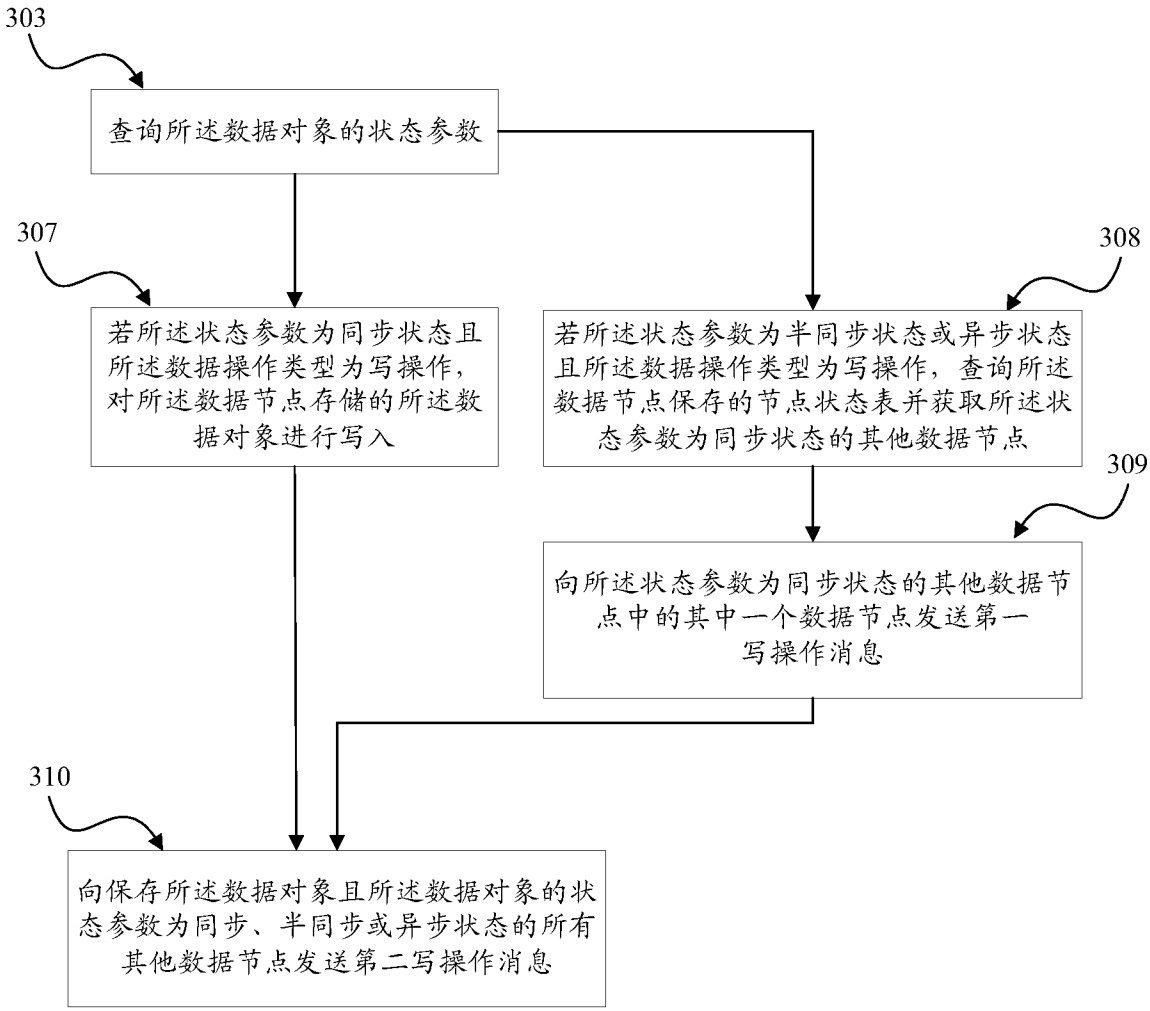

303
查询所述数据对象的状态参数
307
若所述状态参数为同步状态且所述数据操作类型为写操作，对所述数据节点存储的所述数据对象进行写入
308
若所述状态参数为半同步状态或异步状态且所述数据操作类型为写操作，查询所述数据节点保存的节点状态表并获取所述状态参数为同步状态的其他数据节点
309
向所述状态参数为同步状态的其他数据节点中的其中一个数据节点发送第一写操作消息
310
向保存所述数据对象且所述数据对象的状态参数为同步、半同步或异步状态的所有其他数据节点发送第二写操作消息

图 4

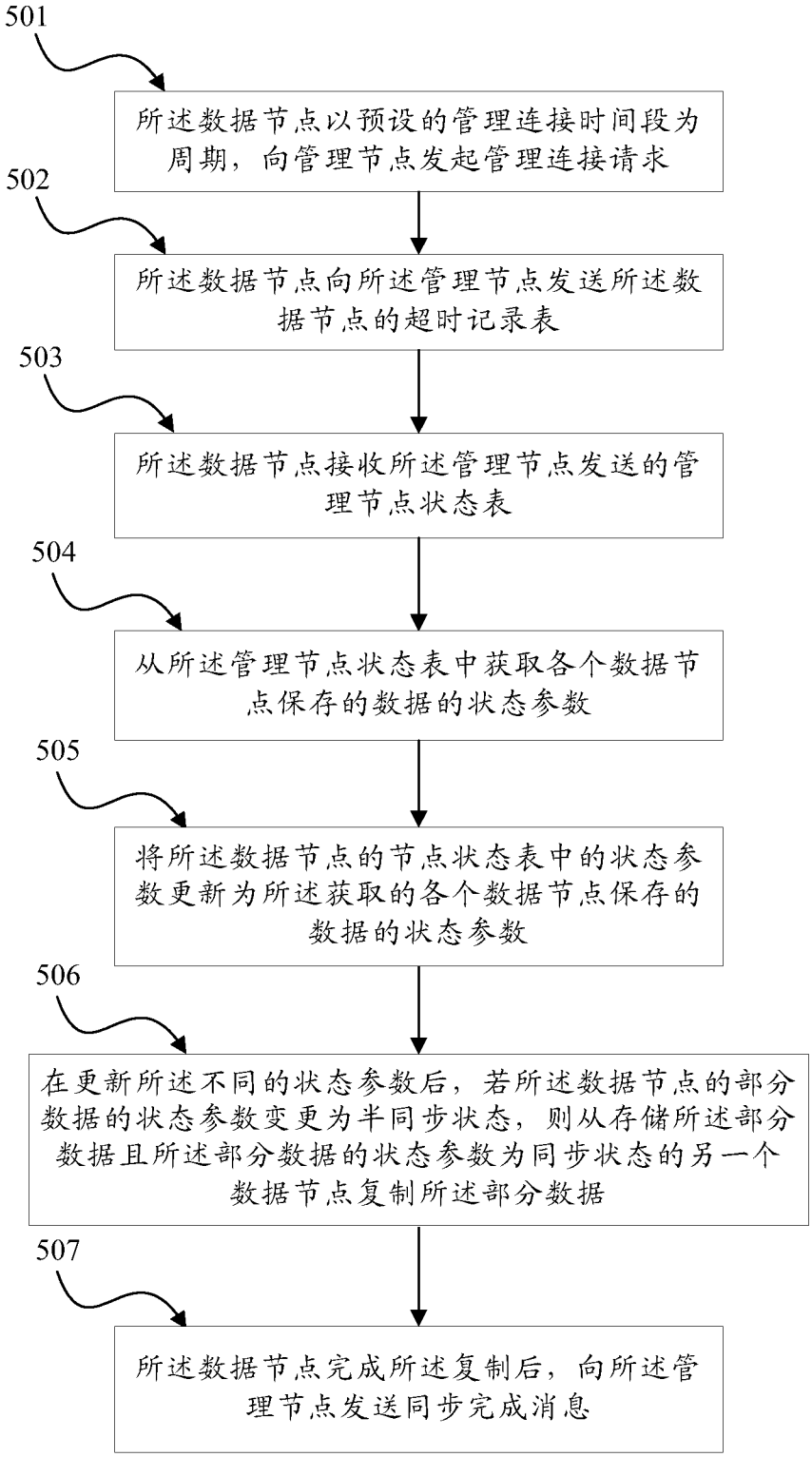

501
所述数据节点以预设的管理连接时间段为周期，向管理节点发起管理连接请求
502
所述数据节点向所述管理节点发送所述数据节点的超时记录表
503
所述数据节点接收所述管理节点发送的管理节点状态表
504
从所述管理节点状态表中获取各个数据节点保存的数据的状态参数
505
将所述数据节点的节点状态表中的状态参数更新为所述获取的各个数据节点保存的数据的状态参数
506
在更新所述不同的状态参数后，若所述数据节点的部分数据的状态参数变更为半同步状态，则从存储所述部分数据且所述部分数据的状态参数为同步状态的另一个数据节点复制所述部分数据
507
所述数据节点完成所述复制后，向所述管理节点发送同步完成消息

图 5

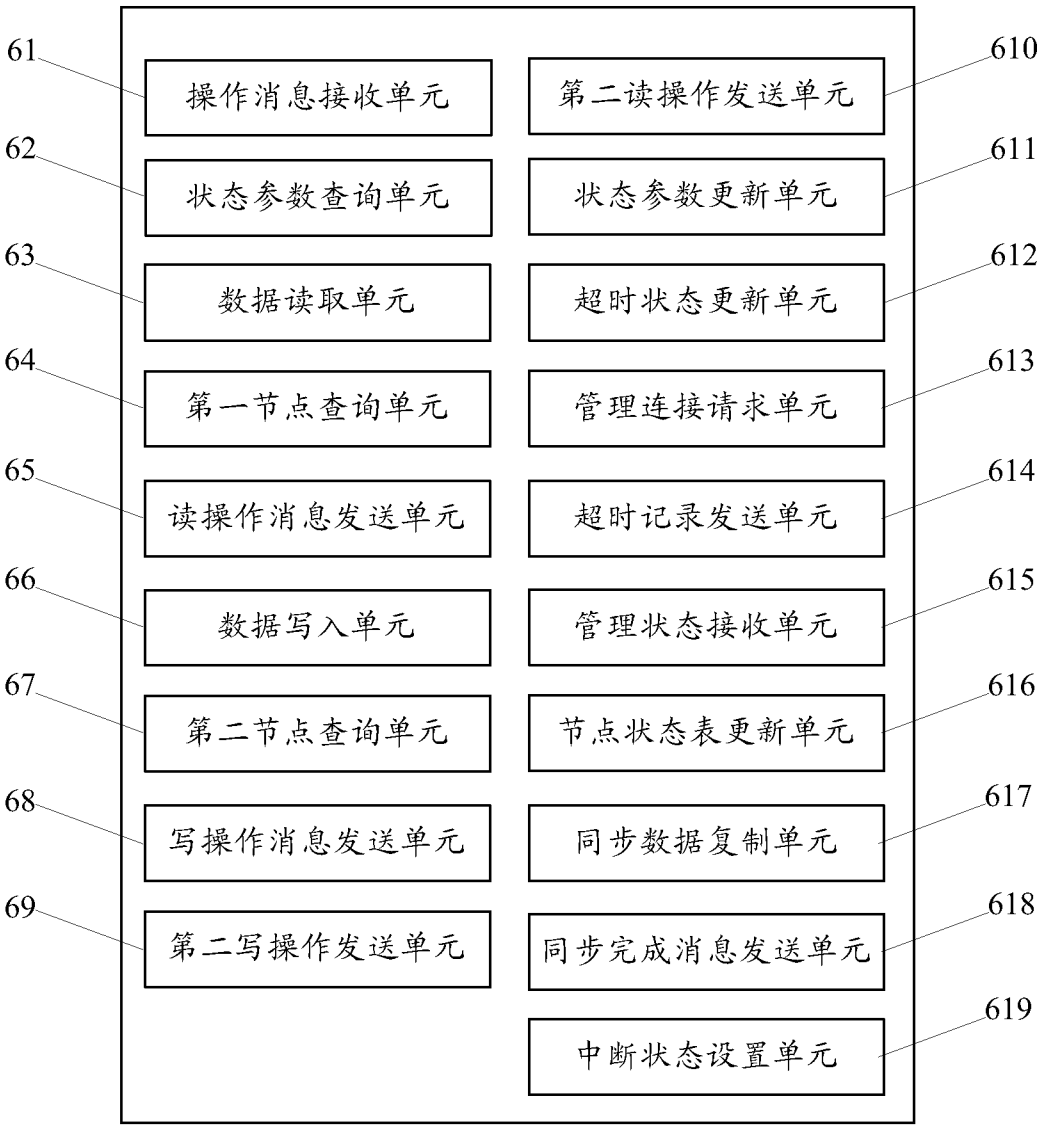

61
操作消息接收单元
62
状态参数查询单元
63
数据读取单元
64
第一节点查询单元
65
读操作消息发送单元
66
数据写入单元
67
第二节点查询单元
68
写操作消息发送单元
69
第二写操作发送单元
610
第二读操作发送单元
611
状态参数更新单元
612
超时状态更新单元
613
管理连接请求单元
614
超时记录发送单元
615
管理状态接收单元
616
节点状态表更新单元
617
同步数据复制单元
618
同步完成消息发送单元
619
中断状态设置单元

图 6

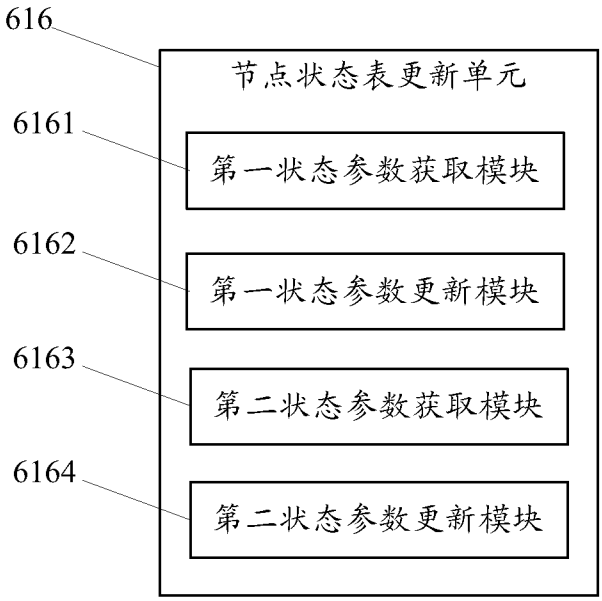

616
节点状态表更新单元
6161
第一状态参数获取模块
6162
第一状态参数更新模块
6163
第二状态参数获取模块
6164
第二状态参数更新模块

图 7

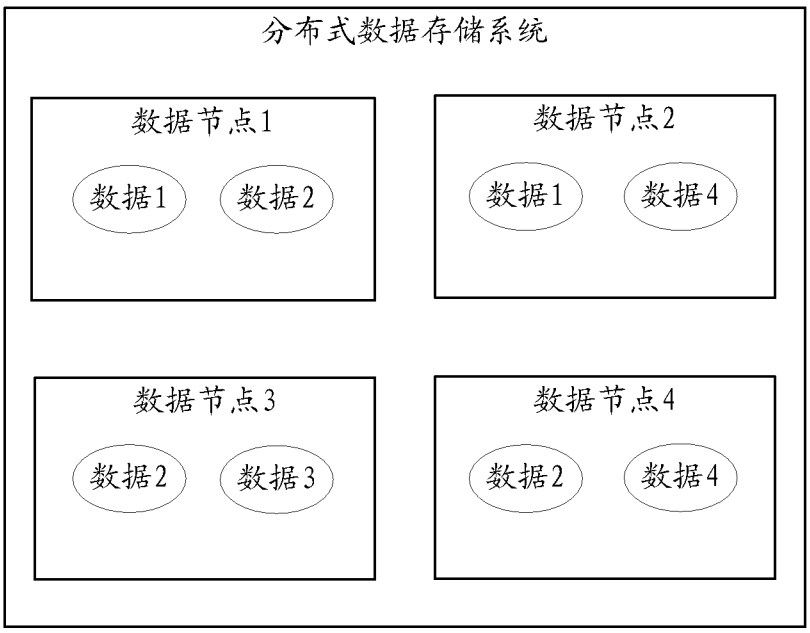

分布式数据存储系统
数据节点1
数据1
数据2
数据节点2
数据1
数据4
数据节点3
数据2
数据3
数据节点4
数据2
数据4

图 8

数据节点	数据	状态参数
数据节点 1	数据 1	同步状态
数据节点 1	数据 2	半同步状态
数据节点 2	数据 1	同步状态
数据节点 2	数据 4	异步状态
数据节点 3	数据 2	同步状态
数据节点 3	数据 3	同步状态
数据节点 4	数据 2	异步状态
数据节点 4	数据 4	同步状态

图 9

数据节点	account group	状态参数
…	…	…

图 10