(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
2024 年 2 月 29 日 (29.02.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/040902 A1

(51) 国际专利分类号:
G06F 16/23 (2019.01)

(21) 国际申请号: PCT/CN2023/079068

(22) 国际申请日: 2023 年 3 月 1 日 (01.03.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211009112.4 2022年8月22日 (22.08.2022) CN

(71) 申请人: 华为云计算技术有限公司 (**HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, Guizhou 550025 (CN)。

(72) 发明人: 徐宜良 (**XU, Yiliang**); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, Guizhou 550025 (CN)。

(74) 代理人: 北京三高永信知识产权代理有限责任公司 (**BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.**); 中国北京市海淀区上地信息产业基地三街1号楼四层C段457, Beijing 100085 (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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,

WO 2024/040902 A1

(54) **Title:** DATA ACCESS METHOD, DISTRIBUTED DATABASE SYSTEM AND COMPUTING DEVICE CLUSTER

(54) 发明名称: 数据访问方法、分布式数据库系统及计算设备集群

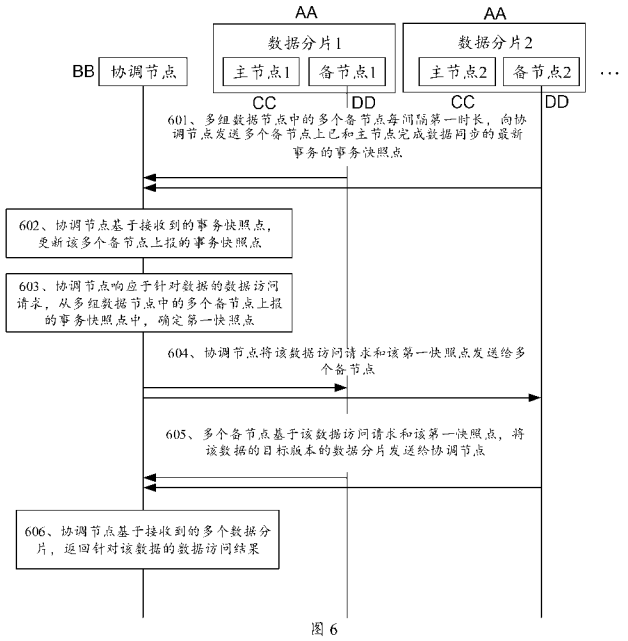


图 6

601 A plurality of standby nodes in a plurality of groups of data nodes send, to the coordination node and at the interval of a first duration, transaction snapshot points of the latest transactions, which have completed data synchronization with the active nodes, on the plurality of standby nodes
602 The coordination node updates, on the basis of the received transaction snapshot points, transaction snapshot points reported by the plurality of standby nodes
603 The coordination node determines, in response to a data access request for data, a first snapshot point from among the transaction snapshot points reported by the plurality of standby nodes among the plurality of groups of data nodes
604 The coordination node sends the data access request and the first snapshot point to the plurality of standby nodes
605 The plurality of standby nodes send data shards of a target version of the data to the coordination node on the basis of the data access request and the first snapshot point
606 The coordination node returns a data access result for the data on the basis of a plurality of received data shards
AA Data shard
BB Coordination node
CC Active node
DD Standby node

(57) **Abstract:** A data access method, a distributed database system and a computing device cluster, which belong to the technical field of databases. The data access method is applied to a distributed database system. The system comprises a coordination node (201) and a plurality of groups of data nodes (202), wherein the coordination node (201) determines, in response to a data access request for data, a first snapshot point from among transaction snapshot points reported by a plurality of standby nodes among the plurality of groups of data nodes (202), and sends the first snapshot point and the data access request to the plurality of standby nodes, such that the plurality of standby nodes return data shards of a target version of the data on the basis of the first snapshot point. During this process, the first snapshot point can indicate the earliest submitted transaction among the latest transactions, which have completed

[见续页]

PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

data synchronization with active nodes, on the plurality of standby nodes, such that the first snapshot point can be used as a snapshot point for global consistency reading, and it is ensured that the plurality of standby nodes return data shards of the same version, thereby meeting the principle of data consistency.

(57) 摘要: 一种数据访问方法、分布式数据库系统及计算设备集群，属于数据库技术领域。该数据访问方法应用于分布式数据库系统，该系统包括协调节点（201）和多组数据节点（202），其中，协调节点（201）响应于针对数据的数据访问请求，从多组数据节点（202）中的多个备节点上报的事务快照点中，确定第一快照点，将第一快照点和数据访问请求发送给多个备节点，以使多个备节点基于第一快照点返回数据的目标版本的数据分片。在这一过程中，由于第一快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此该第一快照点能够作为一种全局一致性读的快照点，确保多个备节点返回同一版本的数据分片，从而满足了数据一致性原则。

数据访问方法、分布式数据库系统及计算设备集群

本申请要求于 2022 年 08 月 22 日提交的申请号为 202211009112.4、发明名称为“数据访问方法、分布式数据库系统及计算设备集群”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据库技术领域，特别涉及一种数据访问方法、分布式数据库系统及计算设备集群。

背景技术

目前，越来越多的互联网应用采用分布式数据库系统来实现相关业务。由于大多数互联网业务操作的特点是读多写少，因此，往往采用读写分离来提升分布式数据库系统的读性能。例如，通过为分布式数据库系统中各个数据分片（shard）设置主备节点来实现读写分离，由主节点负责“写操作”，由备节点负责“读操作”。

然而，由于分布式数据库系统中各个数据分片对应的主备节点之间数据同步时会存在延迟，导致在向多个备节点同时发起访问请求时，难以确保获取到的数据满足一致性原则。例如，在基于一个分布式事务对数据分片 A 和数据分片 B 进行数据更新时，由于主备节点之间的数据同步存在延迟，数据分片 A 对应的备节点已完成数据同步，数据分片 B 对应的备节点尚未完成数据同步，此时，若向这两个数据分片对应的备节点同时发起访问请求，则从数据分片 A 对应的备节点上获取到的是更新后的数据分片，从数据分片 B 对应的备节点上获取到的是更新前的数据分片，导致获取到的数据不满足一致性原则。

因此，亟需一种能够确保读写分离场景下，从备节点获取到的数据满足一致性原则的数据访问方法和分布式数据库系统。

发明内容

本申请实施例提供了一种数据访问方法、分布式数据库系统及计算设备集群，能够确保读写分离场景下，从备节点获取到的数据满足一致性原则。该技术方案如下：

第一方面，提供了一种数据访问方法，应用于分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，该方法包括：

该协调节点响应于针对数据的数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，该第一快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务；

该协调节点将该数据访问请求和该第一快照点发送给该多个备节点；

该多个备节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点，该目标版本为该第一快照点对应的版本。

其中，事务快照点是事务提交完成的情况下获取到的事务提交快照点，每个事务的事务

标识全局唯一，每个事务的事务快照点全局唯一，且事务快照点能够指示事务提交的先后顺序。换言之，事务、事务标识以及事务快照点之间存在映射关系，每个事务都对应有唯一的事务标识和事务快照点。以事务快照点为事务提交序列号（commit sequence number，CSN）为例，若事务 A、事务 B 以及事务 C 的 CSN 分别为 111、113、112，表明这三个事务的提交顺序为事务 A、事务 C 以及事务 B。在上述方法中，协调节点响应于针对数据的数据访问请求，从多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，将该第一快照点和该数据访问请求发送给该多个备节点，以使该多个备节点基于该第一快照点返回该数据的目标版本的数据分片。在这一过程中，由于第一快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此该第一快照点能够作为一种全局一致性读的快照点，确保多个备节点返回同一版本的数据分片，从而满足了数据一致性原则。

在一些实施例中，该多个备节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点，包括：

第一备节点基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的事务快照点大于或等于该第一快照点的情况下，将该第一备节点上该数据的目标版本的数据分片发送给该协调节点，该第一备节点为该多个备节点中的任一个备节点。

上述过程也即是备节点基于第一快照点进行数据可见性判定的过程，通过这种方式，备节点能够返回该第一快照点对应的版本的数据分片，满足数据一致性原则。

在一些实施例中，该方法还包括：

该第一备节点基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的事务快照点小于该第一快照点的情况下，将该数据访问请求和该第一快照点发送给该第一备节点对应的主节点；

该第一备节点对应的主节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点。

通过这种方式，当某一备节点上不存在目标版本的数据分片时，由该备节点对应的主节点向协调节点返回执行结果，从而确保了客户端能够接收到完整的数据访问结果。

在一些实施例中，该方法还包括：

该多个备节点每间隔第一时长，向该协调节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

该协调节点基于接收到的事务快照点，更新该多个备节点上报的事务快照点。

通过上述方式，分布式数据库系统中每个备节点内部都维护一个最大的事务快照点，并定时上报给协调节点，以使协调节点获知系统中每个备节点上已和主节点完成数据同步的最新事务，为后续基于数据访问请求读取数据时，确保各个数据分片的数据一致性提供了基础。

在一些实施例中，该事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

在一些实施例中，该协调节点响应于针对数据的数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，包括：

该协调节点响应于该数据访问请求，从目标列表中确定该第一快照点，该目标列表用于

存储该多个备节点的节点标识和该多个备节点上报的事务快照点，该多个备节点上报的事务快照点指示该多个备节点上已和主节点完成数据同步的最新事务。

通过这种方式，协调节点以列表的形式存储各个备节点上报的事务快照点，便于查询，提高了数据访问效率。

在一些实施例中，该方法还包括：

该协调节点响应于第一主节点发送的数据清理请求，从该多个备节点上报的事务快照点中，确定第二快照点，该第二快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，该第一主节点为该多个主节点中的任一个主节点；

该协调节点将该第二快照点发送给该第一主节点；

该第一主节点基于该第二快照点，清理该第一主节点上该数据的历史版本的数据分片，该历史版本为该第二快照点之前的版本。

在上述过程中，由于第二快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此第二快照点能够作为一种全局一致性清理的快照点，确保了主节点在进行数据清理时各个数据分片的数据一致性。

在一些实施例中，该系统还包括管理节点，该方法还包括：

该多个备节点每间隔第二时长，向该管理节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

该协调节点响应于针对数据的数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，包括：

该协调节点响应于该数据访问请求，向该管理节点发送快照点获取请求，以获取该第一快照点。

通过上述方式，分布式数据库系统中每个备节点内部都维护一个最大的事务快照点，并定时上报给管理节点，以使管理节点获知系统中每个备节点上已和主节点完成数据同步的最新事务，当协调节点基于数据访问请求读取数据时，能够通过向管理节点发送快照点获取请求的方式获取相应的快照点，为确保各个数据分片的数据一致性提供了基础。

在一些实施例中，该系统还包括管理节点，该方法还包括：

第一主节点在目标事务提交完成的情况下，向该管理节点发送针对该目标事务的事务提交请求；

该管理节点响应于该事务提交请求，生成该目标事务的事务快照点，将该目标事务的事务快照点发送给该第一主节点；

该第一主节点对应的备节点对该目标事务进行日志回放，完成和该第一主节点的数据同步。

第二方面，本申请实施例提供了一种分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点；

该协调节点，用于响应于针对数据的数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，该第一快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务；

该协调节点，还用于将该数据访问请求和该第一快照点发送给该多个备节点；

该多个备节点，用于基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点，该目标版本为该第一快照点对应的版本。

在一些实施例中，第一备节点，用于基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的快照点大于或等于该第一快照点的情况下，将该第一备节点上该数据的目标版本的数据分片发送给该协调节点，该第一备节点为该多个备节点中的任一个备节点。

在一些实施例中，该第一备节点，还用于基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的事务快照点小于该第一快照点的情况下，将该数据访问请求和该第一快照点发送给该第一备节点对应的主节点；

该第一备节点对应的主节点，用于基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点。

在一些实施例中，该多个备节点，用于每间隔第一时长，向该协调节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

该协调节点，用于基于接收到的事务快照点，更新该多个备节点上报的事务快照点。

在一些实施例中，该事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

在一些实施例中，该协调节点，用于响应于该数据访问请求，从目标列表中确定该第一快照点，该目标列表用于存储该多个备节点的节点标识和该多个备节点上报的事务快照点，该多个备节点上报的事务快照点指示该多个备节点上已和主节点完成数据同步的最新事务。

在一些实施例中，该协调节点，还用于响应于第一主节点发送的数据清理请求，从该多个备节点上报的事务快照点中，确定第二快照点，该第二快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，该第一主节点为该多个主节点中的任一个主节点；

该协调节点，还用于将该第二快照点发送给该第一主节点；

该第一主节点，用于基于该第二快照点，清理该第一主节点上该数据的历史版本的数据分片，该历史版本为该第二快照点之前的版本。

在一些实施例中，该系统还包括管理节点，

该多个备节点，还用于每间隔第二时长，向该管理节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

该协调节点，用于响应于该数据访问请求，向该管理节点发送快照点获取请求，以获取该第一快照点。

在一些实施例中，该系统还包括管理节点，

第一主节点，用于在目标事务提交完成的情况下，向该管理节点发送针对该目标事务的事务提交请求；

该管理节点，用于响应于该事务提交请求，生成该目标事务的事务快照点，将该目标事务的事务快照点发送给该第一主节点；

该第一主节点对应的备节点，用于对该目标事务进行日志回放，完成和该第一主节点的数据同步。

第三方面，本申请实施例提供了一种计算设备集群，包括至少一个计算设备，每个计算设备包括处理器和存储器；该至少一个计算设备的处理器用于执行该至少一个计算设备的存储器中存储的指令，以使得该计算设备集群执行如前述第一方面或第一方面的任意一种可能的实现方式所提供的数据访问方法。

第四方面，本申请实施例提供了一种数据访问装置，应用于分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，该装置包括至少一个功能模块，用于执行如前述第二方面或第二方面的任意一种可能的实现方式所涉及的协调节点的功能。

第五方面，本申请实施例提供了一种数据访问装置，应用于分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，该装置包括至少一个功能模块，用于执行如前述第二方面或第二方面的任意一种可能的实现方式所涉及的备节点的功能。

第六方面，本申请实施例提供了一种包含指令的计算机程序产品，当该指令被计算设备集群运行时，使得该计算设备集群执行如前述第一方面或第一方面的任意一种可能的实现方式所提供的数据访问方法。该计算机程序产品可以为一个软件安装包，在需要实现前述计算设备集群的功能的情况下，可以下载该计算机程序产品并在计算设备集群上执行该计算机程序产品。

第七方面，本申请实施例提供了一种计算机可读存储介质，包括计算机程序指令，当该计算机程序指令由计算设备集群执行时，该计算设备集群执行如前述第一方面或第一方面的任意一种可能的实现方式所提供的数据访问方法。该存储介质包括但不限于易失性存储器，例如随机访问存储器，非易失性存储器，例如快闪存储器、硬盘（hard disk drive，HDD）、固态硬盘（solid state drive，SSD）。

附图说明

图 1 是本申请实施例提供的一种实施环境示意图；

图 2 是本申请实施例提供的一种分布式数据库系统的架构示意图；

图 3 是本申请实施例提供的一种计算设备的硬件结构示意图；

图 4 是本申请实施例提供的一种计算设备集群的结构示意图；

图 5 是本申请实施例提供的一种计算设备集群的连接方式示意图；

图 6 是本申请实施例提供的一种数据访问方法的流程示意图；

图 7 是本申请实施例提供的一种数据节点定时上报事务快照点的示意图；

图 8 是本申请实施例提供的一种数据访问方法的流程示意图；

图 9 是本申请实施例提供的一种数据清理方法的流程示意图；

图 10 是本申请实施例提供的一种数据访问装置的结构示意图；

图 11 是本申请实施例提供的一种数据访问装置的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

为了方便理解，下面先对本申请涉及的关键术语和关键概念进行说明。

数据库（database），一种电子化的文件柜，也即是存储电子文件的处所，用户可以对电子文件中的数据进行新增、查询、更新、删除等操作。所谓“数据库”是以一定方式储存在一起、能与多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。

事务，是数据库系统在执行操作的过程中的一个逻辑单位，由一个有限的数据库操作序列构成，是数据库系统操作的最小执行单位。

数据分片（shard），是指在分布式数据库系统中数据管理的最小逻辑单元。一个数据分片拥有多个副本。在一些实施例中，当分布式数据库系统新建一个数据分片时，将该数据分片的分片信息进行存储。例如，该分片信息包括该数据分片的数据范围以及多个副本各自的节点信息等，对此不作限定。

虚拟机（virtual machine），是指通过软件模拟的具有完整硬件系统功能的、运行在一个完全隔离环境中的完整计算机系统。在服务器中能够完成的工作在虚拟机中都能够实现。在服务器中创建虚拟机时，需要将实体机的部分硬盘和内存容量作为虚拟机的硬盘和内存容量，每个虚拟机都有独立的硬盘和操作系统，虚拟机的用户可以像使用服务器一样对虚拟机进行操作。

下面对本申请涉及的应用场景和实施环境进行介绍。

本申请实施例提供的技术方案能够应用于数据库领域，其应用场景包括各种能够实现读写分离的分布式数据库系统中。

图 1 是本申请实施例提供的一种实施环境示意图。如图 1 所示，该实施环境包括终端 101 和服务器 102，终端 101 通过无线网络或有线网络与服务器 102 直接或间接相连。

终端 101 可以是智能手机、台式计算机、增强现实终端、平板电脑、电子书阅读器和膝上型便携计算机中的至少一种。终端 101 能够安装和运行应用程序。该应用程序可以是客户端应用，也可以是浏览器应用等，对此不作限定。例如，应用程序为网页浏览客户端、社交客户端或者音视频客户端等等。以应用程序为网页浏览客户端为例，用户能够通过该网页浏览客户端浏览各类网页数据。

服务器 102 为独立的物理服务器，或者是多个物理服务器构成的服务器集群或者分布式文件系统，又或者是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络（content delivery network，CDN）以及大数据和人工智能平台等基础云计算服务的云服务器。服务器 102 用于运行分布式数据库系统，为终端 101 上运行的应用程序提供后台服务。例如，以网页浏览客户端为例，终端 101 响应于用户针对目标网页的浏览操作，触发网页浏览客户端向服务器 102 上运行的分布式数据库系统发送针对该目标网页的数据访问请求，由该分布式数据库系统响应于该数据访问请求，读取目标网页的网页数据，向网页浏览客户端返回数据访问结果。在一些实施例中，服务器

102 通过虚拟机来运行分布式数据库系统，或者，服务器 102 通过容器引擎来运行分布式数据库系统，对此不作限定。

终端 101 可以泛指多个终端中的一个，或者多个终端组成的集合；服务器 102 可以是计算设备集群、虚拟机或容器引擎等等，本申请实施例对实施环境中每种设备的数量和设备类型不做限定。

在一些实施例中，上述的无线网络或有线网络使用标准通信技术和/或协议。网络包括但不限于数据中心网络（data center network）、存储区域网（storage area network，SAN）、局域网（local area network，LAN）、城域网（metropolitan area network，MAN）、广域网（wide area network，WAN）、移动、有线或者无线网络、专用网络或者虚拟专用网络的任何组合。在一些实现方式中，使用包括超级文本标记语言（hyper text markup language，HTML）、可扩展标记语言（extensible markup language，XML）等的技术和/或格式来代表通过网络交换的数据。此外还能够使用诸如安全套接字层（secure sockets layer，SSL）、传输层安全（transport layer security，TLS）、虚拟专用网络（virtual private network，VPN）、网际协议安全（internet protocol security，IPsec）等常规加密技术来加密所有或者部分链路。在另一些实施例中，还能够使用定制和/或专用数据通信技术取代或者补充上述数据通信技术。

下面对上述实施环境涉及的分布式数据库系统的架构进行介绍。

图 2 是本申请实施例提供的一种分布式数据库系统的架构示意图。图 2 仅是示例性地展示了分布式数据库系统的一种结构化示意图，本申请并不限定对分布式数据库系统中各个部分的划分。如图 2 所示，该分布式数据库系统 200 包括协调节点（coordinator node，CN）201、多组数据节点（data node，DN）202 以及管理节点 203，每组数据节点 202 包括主节点和至少一个备节点，对于任一组数据节点 202，该组数据节点 202 用于维护一个数据分片，该组数据节点 202 中的主节点用于执行针对该数据分片的写事务，该组数据节点 202 中的备节点用于执行针对该数据分片的读事务。

协调节点 201 与终端之间通信连接，终端上运行有客户端，协调节点 201 用于接收客户端发送的针对数据的数据处理请求（如数据访问请求、数据更新请求等），将请求下发给对应的数据节点 202 执行，在接收到数据节点 202 反馈的执行结果后，向客户端返回相应的数据处理结果（如数据访问结果、数据更新结果等）。

多组数据节点 202 用于存储客户端的数据，接收来自协调节点 201 的数据处理请求，执行相应的事务（如数据访问对应的读事务、数据更新对应的写事务等），向协调节点 201 返回执行结果。

管理节点 203 用于生成和维护事务标识和事务快照点等全局唯一的信息。在一些实施例中，该管理节点 203 也称为全局事务管理器（global transaction manager，GTM），对此不作限定。

下面对分布式数据库系统 200 处理数据更新请求和数据访问请求的过程进行简要介绍。

以数据更新请求为例，协调节点 201 接收客户端发送的针对数据的数据更新请求，向管理节点 203 申请开启一个分布式写事务，接收管理节点 203 下发的事务标识，确定该数据对应的数据分片，根据路由分片规则，将该分布式写事务下发给对应的主节点（DN1 主和 DN2 主）执行，主节点执行完成后，将执行结果发送给协调节点 201，以使协调节点 201 向客户

端返回数据更新结果。另外，协调节点 201 在该分布式写事务提交完成的情况下，向管理节点 203 申请该分布式写事务已提交的事务快照点，将该事务快照点反馈给对应的主节点。在这一过程中，备节点（DN1 备和 DN2 备）对该分布式写事务进行日志回放，完成和主节点的数据同步。

以数据访问请求为例，协调节点 201 接收客户端发送的针对数据的数据访问请求，确定该数据对应的数据分片，在数据分片所在的备节点（DN1 备和 DN2 备）满足读写分离条件的情况下，根据路由分片规则，将该数据访问请求下发给对应的备节点执行，备节点执行完成后，将执行结果发送给协调节点 201，以使协调节点 201 向客户端返回数据访问结果。这一过程会在后续方法实施例中进行详细介绍，在此不再赘述。

需要说明的是，上述分布式数据库系统中的协调节点和数据节点均可以通过软件实现，或者可以通过硬件实现。示例性的，接下来介绍协调节点的实现方式。类似的，数据节点的实现方式可以参考协调节点的实现方式。

节点作为软件功能单元的一种举例，协调节点可以包括运行在计算实例上的代码。其中，计算实例可以是物理主机（计算设备）、虚拟机、容器等计算设备中的至少一种。进一步地，上述计算设备可以是一台或者多台。例如，协调节点可以包括运行在多个主机/虚拟机/容器上的代码。需要说明的是，用于运行该应用程序的多个主机/虚拟机/容器可以分布在相同的区域（region）中，也可以分布在不同的 region 中。用于运行该代码的多个主机/虚拟机/容器可以分布在相同的可用区（availability zone，AZ）中，也可以分布在不同的 AZ 中，每个 AZ 包括一个数据中心或多个地理位置相近的数据中心。其中，通常一个 region 可以包括多个 AZ。

同样，用于运行该代码的多个主机/虚拟机/容器可以分布在同一个虚拟私有云（virtual private cloud，VPC）中，也可以分布在多个 VPC 中。其中，通常一个 VPC 设置在一个 region 内。同一 region 内两个 VPC 之间，以及不同 region 的 VPC 之间跨区通信需在每个 VPC 内设置通信网关，经通信网关实现 VPC 之间的互连。

节点作为硬件功能单元的一种举例，协调节点可以包括至少一个计算设备，如服务器等。或者，协调节点也可以是利用专用集成电路（application-specific integrated circuit，ASIC）实现、或可编程逻辑器件（programmable logic device，PLD）实现的设备等。其中，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device，CPLD）、现场可编程门阵列（field-programmable gate array，FPGA）、通用阵列逻辑（generic array logic，GAL）或其任意组合实现。

协调节点包括的多个计算设备可以分布在相同的 region 中，也可以分布在不同的 region 中。协调节点包括的多个计算设备可以分布在相同的 AZ 中，也可以分布在不同的 AZ 中。同样，协调节点包括的多个计算设备可以分布在同一个 VPC 中，也可以分布在多个 VPC 中。其中，该多个计算设备可以是服务器、ASIC、PLD、CPLD、FPGA 和 GAL 等计算设备的任意组合。

下面对上述分布式数据库系统中涉及的计算设备的结构进行介绍。

本申请还提供了一种计算设备，能够配置为上述分布式数据库系统中的协调节点和数据节点。参考图 3，图 3 是本申请实施例提供的一种计算设备的硬件结构示意图。如图 3 所示，计算设备 300 包括：总线 302、处理器 304、存储器 306 和通信接口 308。处理器 304、存储

器 306 和通信接口 308 之间通过总线 302 通信。应理解，本申请不限定计算设备 300 中的处理器、存储器的个数。

总线 302 可以是外设部件互连标准（peripheral component interconnect，PCI）总线或扩展工业标准结构（extended industry standard architecture，EISA）总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示，图 3 中仅用一条线表示，但并不表示仅有一根总线或一种类型的总线。总线 304 可包括在计算设备 300 各个部件（例如，存储器 306、处理器 304、通信接口 308）之间传送信息的通路。

处理器 304 可以包括中央处理器（central processing unit，CPU）、图形处理器（graphics processing unit，GPU）、微处理器（micro processor，MP）或者数字信号处理器（digital signal processor，DSP）等处理器中的任意一种或多种。

存储器 306 可以包括易失性存储器（volatile memory），例如随机存取存储器（random access memory，RAM）。处理器 304 还可以包括非易失性存储器（non-volatile memory），例如只读存储器（read-only memory，ROM），快闪存储器，机械硬盘（hard disk drive，HDD）或固态硬盘（solid state drive，SSD）。

存储器 306 中存储有可执行的程序代码，处理器 304 执行该可执行的程序代码以分别实现前述协调节点和数据节点的功能，从而实现下述数据访问方法。也即，存储器 306 上存有用于执行数据访问方法的指令。

通信接口 303 使用例如但不限于网络接口卡、收发器一类的收发模块，来实现计算设备 300 与其他设备或通信网络之间的通信。

本申请实施例还提供了一种计算设备集群。该计算设备集群包括至少一台计算设备。

图 4 是本申请实施例提供的一种计算设备集群的结构示意图。如图 4 所示，该计算设备集群包括至少一个计算设备 300。计算设备集群中的一个或多个计算设备 300 中的存储器 306 中可以存有相同的用于执行数据访问方法的指令。

在一些可能的实现方式中，该计算设备集群中的一个或多个计算设备 300 的存储器 306 中也可以分别存有用于执行数据访问方法的部分指令。换言之，一个或多个计算设备 300 的组合可以共同执行用于执行数据访问方法的指令。

需要说明的是，计算设备集群中的不同的计算设备 300 中的存储器 306 可以存储不同的指令，分别用于执行分布式数据库系统的部分功能。也即，不同的计算设备 300 中的存储器 306 存储的指令可以实现协调节点和数据节点中的一个或多个节点的功能。

在一些实施例中，计算设备集群中的一个或多个计算设备可以通过网络连接。其中，该网络可以是广域网或局域网等等。图 5 是本申请实施例提供的一种计算设备集群的连接方式示意图。如图 5 所示，两个计算设备 300 之间通过网络进行连接。具体地，通过各个计算设备中的通信接口与该网络进行连接。在这一类可能的实现方式中，图 5 所示的计算设备集群之间的连接方式考虑到本申请提供的数据访问方法涉及不同类型的节点，因此在不同计算设备的存储器中存有执行不同节点的功能的指令。例如，一个计算设备 300 中的存储器 306 中存有执行协调节点的功能的指令。另一个计算设备 300 中的存储器 306 中存有执行数据节点的功能的指令。

应理解，图 5 中示出的计算设备 300 的功能也可以由多个计算设备 300 完成。

本申请实施例还提供了一种包含指令的计算机程序产品。该计算机程序产品可以是包含指令的，能够运行在计算设备上或被储存在任何可用介质中的软件或程序产品。当该计算机程序产品在计算设备集群上运行时，使得计算设备集群执行数据访问方法。

本申请实施例还提供了一种计算机可读存储介质。该计算机可读存储介质可以是计算设备能够存储的任何可用介质或者是包含一个或多个可用介质的数据中心等数据存储设备。该可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质（例如，DVD）、或者半导体介质（例如固态硬盘）等。该计算机可读存储介质包括指令，该指令指示计算设备集群执行数据访问方法。

下面对本申请实施例提供的数据访问方法进行介绍。

图 6 是本申请实施例提供的一种数据访问方法的流程示意图。如图 6 所示，该数据访问方法应用于上述分布式数据库系统，下面以分布式数据库系统中各节点之间的交互为例，对该数据访问方法进行介绍。示意性地，该方法包括如下步骤 601 至步骤 606。

601、多组数据节点中的多个备节点每间隔第一时长，向协调节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点。

在本申请实施例中，分布式数据库系统基于数据分片来存储数据，一个数据分片由一组数据节点来维护，该组数据节点包括主节点和至少一个备节点，主节点用于执行写事务，备节点用于执行读事务。事务快照点是主节点在事务提交完成的情况下，通过管理节点获取到的事务提交快照点，每个事务的事务标识全局唯一，每个事务的事务快照点全局唯一，且事务快照点能够指示事务提交的先后顺序。换言之，事务、事务标识以及事务快照点之间存在映射关系，每个事务都对应有唯一的事务标识和事务快照点。在一些实施例中，事务快照点为事务提交序列号（commit sequence number，CSN）或事务提交时间戳等，对此不作限定。以事务快照点为 CSN 为例，CSN 是一种全局自增的整数，根据 CSN 的大小可以判断事务提交的先后顺序，CSN 越小表示事务的提交顺序越早。例如，若事务 A、事务 B 以及事务 C 的 CSN 分别为 111、113、112，表明这三个事务的提交顺序为事务 A、事务 C 以及事务 B。

以第一主节点（多组数据节点中的任一个主节点）执行目标事务为例，第一主节点在目标事务提交完成的情况下，向管理节点发送针对该目标事务的事务提交请求；该管理节点响应于该事务提交请求，生成该目标事务的事务快照点，将该目标事务的事务快照点发送给该第一主节点；该第一主节点对应的备节点对该目标事务进行日志回放，完成和该第一主节点的数据同步。其中，第一主节点对应的备节点在对该目标事务进行日志回放的过程中，也即获取到了该目标事务的事务快照点。

在分布式数据库系统的运行过程中，多组数据节点的备节点通过日志回放的方式完成和主节点的数据同步，也即获取到已提交事务的事务快照点，在这一过程中，各个备节点每间隔第一时长，向协调节点发送备节点上已和主节点完成数据同步的最新事务的事务快照点。其中，该第一时长为预设时长，能够根据需求进行设置，例如，第一时长为 5 秒，对此不作限定。这一过程也即是，以事务快照点为 CSN 为例，分布式数据库系统中的各个备节点内部都维护一个最大的事务快照点，并定时上报给协调节点。例如，第一时长为 5 秒，备节点每

间隔5秒，将该备节点上存储的最大CSN发送给协调节点。需要说明的是，该多个备节点可以每间隔第一时长，同步向协调节点上报事务快照点（如备节点1和备节点2同步上报），也可以异步上报（如备节点1和备节点2依次上报），本申请实施例对此不作限定。

602、协调节点基于接收到的事务快照点，更新该多个备节点上报的事务快照点。

在本申请实施例中，协调节点存储有目标列表，该目标列表用于存储该多个备节点的节点标识和该多个备节点上报的事务快照点。需要说明的是，协调节点还能够以其他形式存储多个备节点上报的事务快照点，并不局限于列表形式，本申请实施例对此不作限定。

在一些实施例中，协调节点每接收到一个备节点上报的事务快照点，基于该备节点的节点标识，更新该目标列表中该备节点上报的事务快照点。例如，以事务快照点为CSN为例，该目标列表存储有备节点1（115）、备节点2（116）、备节点3（117），若协调节点接收到备节点1发送的事务快照点118，则更新该目标列表为备节点1（118）、备节点2（116）、备节点3（117）。通过这种方式，协调节点能够及时根据接收到的事务快照点更新目标列表，确保目标列表的实时性。

在一些实施例中，协调节点基于指定时间段内接收到的多个事务快照点，更新该目标列表。例如，该指定时间段为10秒，对此不作限定。应理解，指定时间段内某一备节点有可能上报了多个事务快照点，基于此，协调节点基于该备节点上报的最后一个事务快照点，更新该目标列表。或者，协调节点在接收到的多个事务快照点的数量达到指定数量的情况下，更新该目标列表，例如，该指定数量为10。同理，协调节点接收到的指定数量个事务快照点中也可能包括了某一个备节点上报的多个事务快照点，基于此，协调节点基于该备节点上报的最后一个事务快照点，更新该目标列表。通过上述方式，能够减少协调节点更新目标列表的频率，从而节约数据处理量。

经过上述步骤601和步骤602，分布式数据库系统中每个备节点内部都维护一个最大的事务快照点，也称为事务强一致性快照点，能够确保事务的执行不影响系统的数据一致性，通过将事务快照点定时上报给协调节点，便于协调节点获知系统中每个备节点上已和主节点完成数据同步的最新事务，为后续基于数据访问请求读取数据时，确保各个数据分片的数据一致性提供了基础。

603、协调节点响应于针对数据的数据访问请求，从多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，该第一快照点指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务。

在本申请实施例中，数据访问请求由客户端发送，协调节点响应于该数据访问请求，基于该数据对应的多个数据分片，确定该多个数据分片所在的多组数据节点，在该多组数据节点中的多个备节点满足读写分离条件的情况下，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点。其中，读写分离条件指示任一组数据节点中的备节点与主节点之间数据更新的时间差小于或等于目标时长，即，主节点上数据的最新版本和备节点上数据的最新版本之间不能相差太远，否则不满足业务需求。例如，该目标时长为5秒，对此不作限定。这一过程也即是，在协调节点接收到一个数据访问请求的情况下，会判断该数据访问请求是否可以使用读写分离，在可以使用读写分离的情况下，再将该数据访问请求下发至备节点执行。

基于上述步骤602可知，协调节点上存储有目标列表，相应地，协调节点响应于该数据

访问请求，从目标列表中确定该第一快照点。以事务快照点为 CSN 为例，协调节点响应于该数据访问请求，基于该数据对应的多个数据分片，确定该多个数据分片所在的多组数据节点，基于该多组数据节点中的多个备节点的节点标识，从目标列表中该多个备节点上报的 CSN 中，确定最小 CSN，该最小 CSN 也即是提交顺序最早的事务，将该最小 CSN 作为第一快照点。

604、协调节点将该数据访问请求和该第一快照点发送给多个备节点。

在本申请实施例中，协调节点基于路由分片规则，将该数据访问请求和该第一快照点发送给多个备节点。

605、多个备节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给协调节点。

在本申请实施例中，目标版本是指该第一快照点对应的版本。例如，以事务快照点为 CSN 为例，若第一快照点为 119，则该数据的目标版本是指 CSN 为 119 的事务提交完成后该数据的版本。对于多个备节点中的任一个备节点（称为第一备节点），该第一备节点基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的事务快照点大于或等于该第一快照点的情况下，将该第一备节点上该数据的目标版本的数据分片发送给该协调节点。这一过程也即是基于第一快照点进行数据可见性判定的过程。

基于此，每个备节点都将各自节点上同一版本的数据分片发送给协调节点，确保了各个数据分片的数据一致性。例如，以事务快照点为 CSN 为例，第一快照点为 119，第一备节点上已和主节点完成数据同步的最新事务的事务快照点为 120，该 120 大于 119，则将该第一备节点上与 CSN 为 119 的事务对应的版本的数据分片发送给协调节点。

在一些实施例中，该第一备节点基于该数据访问请求和该第一快照点，在该第一备节点上已和主节点完成数据同步的最新事务的事务快照点小于该第一快照点的情况下，将该数据访问请求和该第一快照点发送给该第一备节点对应的主节点；该第一备节点对应的主节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点。即，当某一备节点上不存在目标版本的数据分片时，由该备节点对应的主节点向协调节点返回执行结果，从而确保了客户端能够接收到完整的数据访问结果。

606、协调节点基于接收到的多个数据分片，返回针对该数据的数据访问结果。

在本申请实施例中，协调节点基于该多个数据分片的分片信息，对该多个数据分片进行汇总，得到该数据的数据访问结果，向客户端返回该数据访问结果。

在上述步骤 601 至步骤 606 中，是以分布式数据库系统中备节点向协调节点上报事务快照点为例进行说明的，在另一些实施例中，备节点还能够向管理节点上报事务快照点，当协调节点基于数据访问请求读取数据时，能够通过向管理节点发送快照点获取请求的方式获取相应的快照点。

示意性地，多个备节点每间隔第二时长，向该管理节点发送该多个备节点上已和主节点完成数据同步的最新事务的事务快照点。其中，该第二时长为预设时长，能够根据需求进行设置，例如，第二时长为 5 秒，对此不作限定。基于与上述步骤 601 和步骤 602 同理的过程，管理节点基于接收到的事务快照点，更新该多个备节点上报的事务快照点，即，管理节点也能够以列表的形式存储该多个备节点上报的事务快照点。在协调节点接收到客户端发送的数据访问请求的情况下，该协调节点响应于该数据访问请求，向该管理节点发送快照点获取请

求，以获取该第一快照点。其中，管理节点响应于该快照点获取请求，从多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，将该第一快照点发送给协调节点。这一过程与上述步骤 603 同理，在此不再赘述。通过这种方式，由管理节点维护多个备节点上报的事务快照点，能够节约协调节点的计算资源，释放协调节点的存储空间。

综上，介绍了本申请实施例提供的一种数据访问方法，应用于分布式数据库系统，其中，协调节点响应于针对数据的数据访问请求，从多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，将该第一快照点和该数据访问请求发送给该多个备节点，以使该多个备节点基于该第一快照点返回该数据的目标版本的数据分片。在这一过程中，由于第一快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此该第一快照点能够作为一种全局一致性读的快照点，确保多个备节点返回同一版本的数据分片，从而满足了数据一致性原则。

另外，在一些实施例中，主节点在进行数据清理时，能够根据多个备节点上报的事务快照点来确保各个数据分片的数据一致性。下面对这一过程进行介绍，包括下述几个步骤：

步骤 A、协调节点响应于第一主节点发送的数据清理请求，从该多个备节点上报的事务快照点中，确定第二快照点，该第二快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务。

其中，该第一主节点为该多个主节点中的任一个主节点。基于与上述步骤 603 同理的过程，协调节点上存储有目标列表，该协调节点响应于该数据清理请求，从目标列表中确定该第二快照点。以事务快照点为 CSN 为例，协调节点响应于该数据清理请求，基于该数据对应的多个数据分片，确定该多个数据分片所在的多组数据节点，基于该多组数据节点中的多个备节点的节点标识，从目标列表中该多个备节点上报的 CSN 中，确定最小 CSN，该最小 CSN 也即是提交顺序最早的事务，将该最小 CSN 作为第二快照点。

步骤 B、协调节点将该第二快照点发送给该第一主节点。

步骤 C、第一主节点基于该第二快照点，清理该第一主节点上该数据的历史版本的数据分片。

其中，该历史版本是指该第二快照点之前的版本。例如，以事务快照点为 CSN 为例，若第二快照点为 118，则该数据的历史版本是指 CSN 为 118 的事务执行之前该数据的版本。应理解，第一主节点会将这一数据清理过程保存在日志中，该第一主节点对应的备节点也能够通过日志回放的方式完成与主节点的数据同步，在此不再赘述。

经过上述步骤 A 至步骤 C，协调节点响应于第一主节点的数据清理请求，从多个备节点上报的事务快照点中，确定第二快照点，将该第二快照点发送给该第一主节点，以使该第一主节点基于该第二快照点清理该数据的历史版本。在这一过程中，由于第二快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此第二快照点能够作为一种全局一致性清理的快照点，确保了主节点在进行数据清理时各个数据分片的数据一致性。例如，以一组数据节点为例，该组数据节点中主节点上已提交的最新事务的事务快照点，大于该组数据节点中备节点上已和主节点完成数据同步的最新事务的事务快照点，由于备节点的数据正在读取时是不能清理的，因此以备节点上报的事务快照点中的最小值为清理点，能够确保数据清理时各个数据分片的数据一致性。

下面参考图 7 至图 9，以事务快照点为 CSN 为例，对上述实施例中介绍的方法流程进行举例说明。

图 7 是本申请实施例提供的一种数据节点定时上报事务快照点的示意图。如图 7 所示，在本申请实施例提供的分布式数据库系统中，多组数据节点中每个备节点内部都维护一个最大的事务快照点（即 CSN），并定时上报给协调节点。协调节点在收到各个备节点上报的 CSN 后，在内部维护目标列表，其中，目标列表用于存储所有备节点的节点标识以及所有备节点上报的 CSN，备节点上报的 CSN 为备节点上已和主节点完成数据同步的最新事务的 CSN。

图 8 是本申请实施例提供的一种数据访问方法的流程示意图。如图 8 所示，该数据访问方法包括下述几个步骤：

步骤 1、客户端发起一个针对数据的数据访问请求，即向协调节点发送数据访问请求。

步骤 2、协调节点响应于该数据访问请求，判断是否可以使用读写分离，若可以，从内部维护的“目标列表”中确定第一快照点，该第一快照点即为“目标列表”中的最小 CSN。

步骤 3、协调节点将数据访问请求和该第一快照点发送给多个备节点。

步骤 4、备节点接收该数据访问请求和该第一快照点，基于该第一快照点进行数据可见性判定，读取该数据的目标版本的数据分片。

步骤 5、备节点把数据分片发送给协调节点。

步骤 6、协调节点汇总各个数据分片并返回给客户端。

在上述过程中，由于第一快照点能够指示多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，因此该第一快照点能够作为一种全局一致性读的快照点，确保多个备节点返回同一版本的数据分片，从而满足了数据一致性原则。

图 9 是本申请实施例提供的一种数据清理方法的流程示意图。如图 9 所示，该数据清理方法包括下述几个步骤：

步骤 1、主节点向协调节点发送数据清理请求。

步骤 2、协调节点响应于该数据清理请求，从内部维护的“目标列表”中确定第二快照点，该第二快照点即为“目标列表”中的最小 CSN。

步骤 3、协调节点将该第二快照点发送给主节点。

步骤 4、主节点基于该第二快照点，清理该主节点上该数据的历史版本的数据分片。

在上述过程中，由于第二快照点能够指示多个主节点上已提交的最新事务中提交顺序最早的事务，因此第二快照点能够作为一种全局一致性清理的快照点，确保了主节点在进行数据清理时各个数据分片的数据一致性。

图 10 是本申请实施例提供的一种数据访问装置的结构示意图。该数据访问装置可以通过软件、硬件或者两者的结合实现前述分布式数据库系统中协调节点的部分或者全部功能。本申请实施例提供的数据访问装置应用于分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，能够实现上述方法实施例中协调节点所执行的步骤。如图 10 所示，该数据访问装置包括确定模块 1001 和发送模块 1002。

该确定模块 1001，用于响应于针对数据的数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，该第一快照点指示该多个备节点上已和主节点

完成数据同步的最新事务中提交顺序最早的事务；

该发送模块 1002，用于将该数据访问请求和该第一快照点发送给该多个备节点，以使该多个备节点基于该数据访问请求和该第一快照点发送该数据的目标版本的数据分片，该目标版本为该第一快照点对应的版本。

在一些实施例中，该装置还包括更新模块，用于：接收该多个备节点每间隔第一时长发送的已和主节点完成数据同步的最新事务的事务快照点，基于接收到的事务快照点，更新该多个备节点上报的事务快照点。

在一些实施例中，该事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

在一些实施例中，该确定模块 1001，用于响应于该数据访问请求，从目标列表中确定该第一快照点，该目标列表用于存储该多个备节点的节点标识和该多个备节点上报的事务快照点，该多个备节点上报的事务快照点指示该多个备节点上已和主节点完成数据同步的最新事务。

在一些实施例中，该确定模块 1001，还用于响应于第一主节点发送的数据清理请求，从该多个备节点上报的事务快照点中，确定第二快照点，该第二快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，该第一主节点为该多个主节点中的任一个主节点；

该发送模块 1002，还用于将该第二快照点发送给该第一主节点，以使该第一主节点基于该第二快照点，清理该第一主节点上该数据的历史版本的数据分片，该历史版本为该第二快照点之前的版本。

需要说明的是：上述实施例提供的数据访问装置在进行数据处理时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的数据访问装置与数据访问装置实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

图 11 是本申请实施例提供的一种数据访问装置的结构示意图。该数据访问装置可以通过软件、硬件或者两者的结合实现前述分布式数据库系统中备节点的部分或者全部功能。本申请实施例提供的数据访问装置应用于分布式数据库系统，该系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，能够实现上述方法实施例中任一个备节点所执行的步骤。如图 11 所示，该数据访问装置包括接收模块 1101 和发送模块 1102。

接收模块 1101，用于接收协调节点发送的针对数据的数据访问请求和第一快照点，该第一快照点由该协调节点响应于该数据访问请求，从该多组数据节点中的多个备节点上报的事务快照点中确定，该第一快照点指示该多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务；

发送模块 1102，用于基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给该协调节点，该目标版本为该第一快照点对应的版本。

在一些实施例中，该发送模块 1102，用于基于该数据访问请求和该第一快照点，在已和

主节点完成数据同步的最新事务的事务快照点大于或等于该第一快照点的情况下，将该数据的目标版本的数据分片发送给该协调节点。

在一些实施例中，该装置还包括上报模块，该上报模块用于每间隔第一时长，向该协调节点发送备节点上已和主节点完成数据同步的最新事务的事务快照点。

在一些实施例中，该事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

在一些实施例中，该系统还包括管理节点，该发送模块1102，还用于每间隔第二时长，向该管理节点发送备节点上已和主节点完成数据同步的最新事务的事务快照点。

需要说明的是：上述实施例提供的数据访问装置在进行数据处理时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的数据访问装置与数据访问装置实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

本申请中术语“第一”“第二”等字样用于对作用和功能基本相同的相同项或相似项进行区分，应理解，“第一”、“第二”、“第n”之间不具有逻辑或时序上的依赖关系，也不对数量和执行顺序进行限定。还应理解，尽管以下描述使用术语第一、第二等来描述各种元素，但这些元素不应受术语的限制。这些术语只是用于将一元素与另一元素区别分开。例如，在不脱离各种所述示例的范围的情况下，第一备节点可以被称为第二备节点，并且类似地，第二备节点可以被称为第一词组。第一词组和第二备节点都可以是备节点，并且在某些情况下，可以是单独且不同的备节点。

本申请中术语“至少一个”的含义是指一个或多个，本申请中术语“多个”的含义是指两个或两个以上，例如，多个备节点是指两个或两个以上的备节点。

以上描述，仅为本申请的具体实施方式，但本申请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申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以程序结构信息的形式实现。该程序结构信息包括一个或多个程序指令。在计算设备上加载和执行该程序指令时，全部或部分地产生按照本申请实施例中的流程或功能。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，该程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

最后应说明的是：以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的保护范围。

权利要求书

1.一种数据访问方法，其特征在于，应用于分布式数据库系统，所述系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点，所述方法包括：

所述协调节点响应于针对数据的数据访问请求，从所述多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，所述第一快照点指示所述多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务；

所述协调节点将所述数据访问请求和所述第一快照点发送给所述多个备节点；

所述多个备节点基于所述数据访问请求和所述第一快照点，将所述数据的目标版本的数据分片发送给所述协调节点，所述目标版本为所述第一快照点对应的版本。

2.根据权利要求 1 所述的方法，其特征在于，所述多个备节点基于所述数据访问请求和所述第一快照点，将所述数据的目标版本的数据分片发送给所述协调节点，包括：

第一备节点基于所述数据访问请求和所述第一快照点，在所述第一备节点上已和主节点完成数据同步的最新事务的事务快照点大于或等于所述第一快照点的情况下，将所述第一备节点上所述数据的目标版本的数据分片发送给所述协调节点，所述第一备节点为所述多个备节点中的任一个备节点。

3.根据权利要求 2 所述的方法，其特征在于，所述方法还包括：

所述第一备节点基于所述数据访问请求和所述第一快照点，在所述第一备节点上已和主节点完成数据同步的最新事务的事务快照点小于所述第一快照点的情况下，将所述数据访问请求和所述第一快照点发送给所述第一备节点对应的主节点；

所述第一备节点对应的主节点基于所述数据访问请求和所述第一快照点，将所述数据的目标版本的数据分片发送给所述协调节点。

4.根据权利要求 1 至 3 中任一项所述的方法，其特征在于，所述方法还包括：

所述多个备节点每间隔第一时长，向所述协调节点发送所述多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

所述协调节点基于接收到的事务快照点，更新所述多个备节点上报的事务快照点。

5.根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

6.根据权利要求 1 至 5 中任一项所述的方法，其特征在于，所述协调节点响应于针对数据的数据访问请求，从所述多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，包括：

所述协调节点响应于所述数据访问请求，从目标列表中确定所述第一快照点，所述目标列表用于存储所述多个备节点的节点标识和所述多个备节点上报的事务快照点，所述多个备节点上报的事务快照点指示所述多个备节点上已和主节点完成数据同步的最新事务。

7.根据权利要求 1 至 6 中任一项所述的方法，其特征在于，所述方法还包括：

所述协调节点响应于第一主节点发送的数据清理请求，从所述多个备节点上报的事务快照点中，确定第二快照点，所述第二快照点指示所述多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，所述第一主节点为所述多个主节点中的任一个主节点；

所述协调节点将所述第二快照点发送给所述第一主节点；

所述第一主节点基于所述第二快照点，清理所述第一主节点上所述数据的历史版本的数据分片，所述历史版本为所述第二快照点之前的版本。

8.根据权利要求 1 至 7 中任一项所述的方法，其特征在于，所述系统还包括管理节点，所述方法还包括：

所述多个备节点每间隔第二时长，向所述管理节点发送所述多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

所述协调节点响应于针对数据的数据访问请求，从所述多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，包括：

所述协调节点响应于所述数据访问请求，向所述管理节点发送快照点获取请求，以获取所述第一快照点。

9.根据权利要求 1 至 8 中任一项所述的方法，其特征在于，所述系统还包括管理节点，所述方法还包括：

第一主节点在目标事务提交完成的情况下，向所述管理节点发送针对所述目标事务的事务提交请求；

所述管理节点响应于所述事务提交请求，生成所述目标事务的事务快照点，将所述目标事务的事务快照点发送给所述第一主节点；

所述第一主节点对应的备节点对所述目标事务进行日志回放，完成和所述第一主节点的数据同步。

10.一种分布式数据库系统，其特征在于，所述系统包括协调节点和多组数据节点，每组数据节点包括主节点和备节点；

所述协调节点，用于响应于针对数据的数据访问请求，从所述多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点，所述第一快照点指示所述多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务；

所述协调节点，还用于将所述数据访问请求和所述第一快照点发送给所述多个备节点；

所述多个备节点，用于基于所述数据访问请求和所述第一快照点，将所述数据的目标版本的数据分片发送给所述协调节点，所述目标版本为所述第一快照点对应的版本。

11.根据权利要求 10 所述的系统，其特征在于，

第一备节点，用于基于所述数据访问请求和所述第一快照点，在所述第一备节点上已和主节点完成数据同步的最新事务的快照点大于或等于所述第一快照点的情况下，将所述第一备节点上所述数据的目标版本的数据分片发送给所述协调节点，所述第一备节点为所述多个备节点中的任一个备节点。

12.根据权利要求 11 所述的系统，其特征在于，

所述第一备节点，还用于基于所述数据访问请求和所述第一快照点，在所述第一备节点上已和主节点完成数据同步的最新事务的事务快照点小于所述第一快照点的情况下，将所述数据访问请求和所述第一快照点发送给所述第一备节点对应的主节点；

所述第一备节点对应的主节点，用于基于所述数据访问请求和所述第一快照点，将所述数据的目标版本的数据分片发送给所述协调节点。

13.根据权利于要求 10 至 12 中任一项所述的系统，其特征在于，

所述多个备节点，用于每间隔第一时长，向所述协调节点发送所述多个备节点上已和主节点完成数据同步的最新事务的事务快照点；

所述协调节点，用于基于接收到的事务快照点，更新所述多个备节点上报的事务快照点。

14.根据权利要求 10 至 13 中任一项所述的系统，其特征在于，所述事务快照点为下述任一项：

事务提交序列号；

事务提交时间戳。

15.根据权利要求 10 至 14 中任一项所述的系统，其特征在于，

所述协调节点，用于响应于所述数据访问请求，从目标列表中确定所述第一快照点，所述目标列表用于存储所述多个备节点的节点标识和所述多个备节点上报的事务快照点，所述多个备节点上报的事务快照点指示所述多个备节点上已和主节点完成数据同步的最新事务。

16.根据权利要求 10 至 15 中任一项所述的系统，其特征在于，

所述协调节点，还用于响应于第一主节点发送的数据清理请求，从所述多个备节点上报的事务快照点中，确定第二快照点，所述第二快照点指示所述多个备节点上已和主节点完成数据同步的最新事务中提交顺序最早的事务，所述第一主节点为所述多个主节点中的任一个主节点；

所述协调节点，还用于将所述第二快照点发送给所述第一主节点；

所述第一主节点，用于基于所述第二快照点，清理所述第一主节点上所述数据的历史版本的数据分片，所述历史版本为所述第二快照点之前的版本。

17.根据权利要求 10 至 16 中任一项所述的系统，其特征在于，所述系统还包括管理节点，

所述多个备节点，还用于每间隔第二时长，向所述管理节点发送所述多个备节点上已和

主节点完成数据同步的最新事务的事务快照点；

所述协调节点，用于响应于所述数据访问请求，向所述管理节点发送快照点获取请求，以获取所述第一快照点。

18.根据权利要求 10 至 17 中任一项所述的系统，其特征在于，所述系统还包括管理节点，

第一主节点，用于在目标事务提交完成的情况下，向所述管理节点发送针对所述目标事务的事务提交请求；

所述管理节点，用于响应于所述事务提交请求，生成所述目标事务的事务快照点，将所述目标事务的事务快照点发送给所述第一主节点；

所述第一主节点对应的备节点，用于对所述目标事务进行日志回放，完成和所述第一主节点的数据同步。

19.一种计算设备集群，其特征在于，包括至少一个计算设备，每个计算设备包括处理器和存储器；

所述至少一个计算设备的处理器用于执行所述至少一个计算设备的存储器中存储的指令，以使得所述计算设备集群执行如权利要求 1 至 9 中任一项所述的数据访问方法。

20.一种包含指令的计算机程序产品，其特征在于，当所述指令被计算设备集群运行时，使得所述计算设备集群执行如权利要求的 1 至 9 中任一项所述的数据访问方法。

21.一种计算机可读存储介质，其特征在于，包括计算机程序指令，当所述计算机程序指令由计算设备集群执行时，所述计算设备集群执行如权利要求 1 至 9 中任一项所述的数据访问方法。

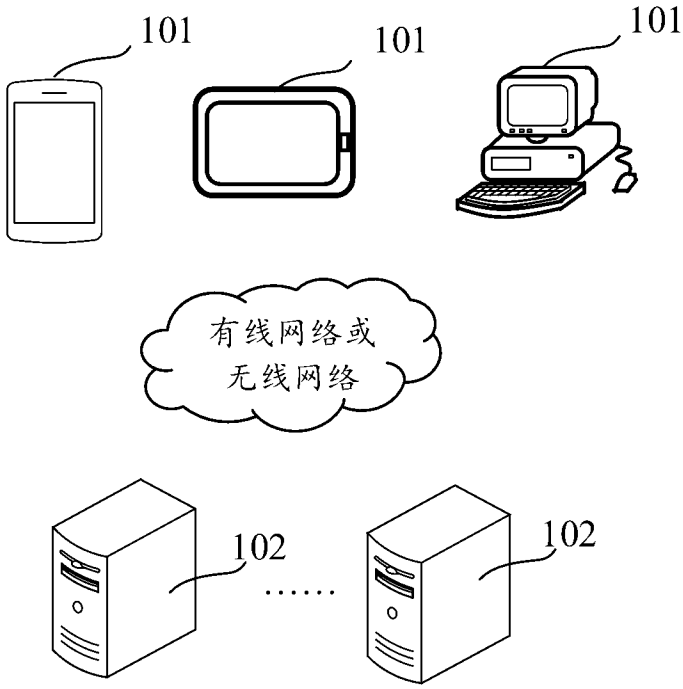


图 1

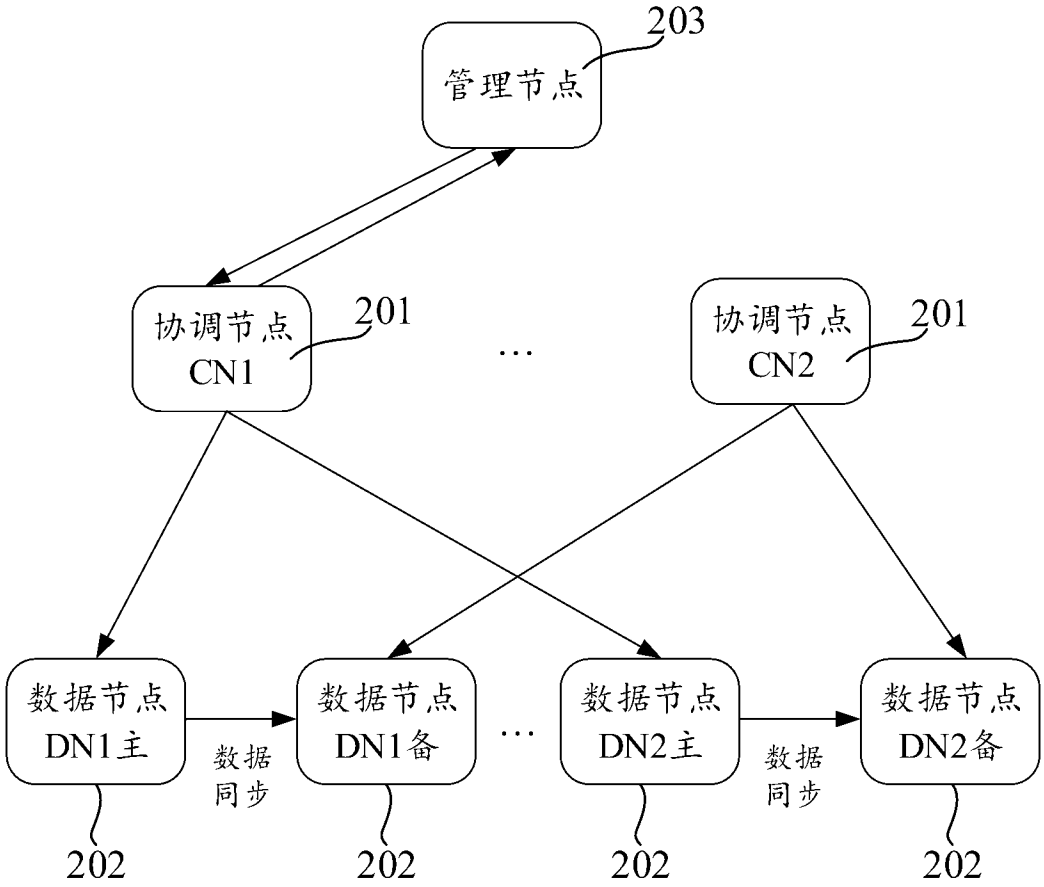

203
管理节点
协调节点
CN1
201
…
协调节点
CN2
201
数据节点
DN1主
数据节点
DN1备
…
数据节点
DN2主
数据节点
DN2备
数据
同步
数据
同步
202
202
202
202
分布式数据库系统200

图 2

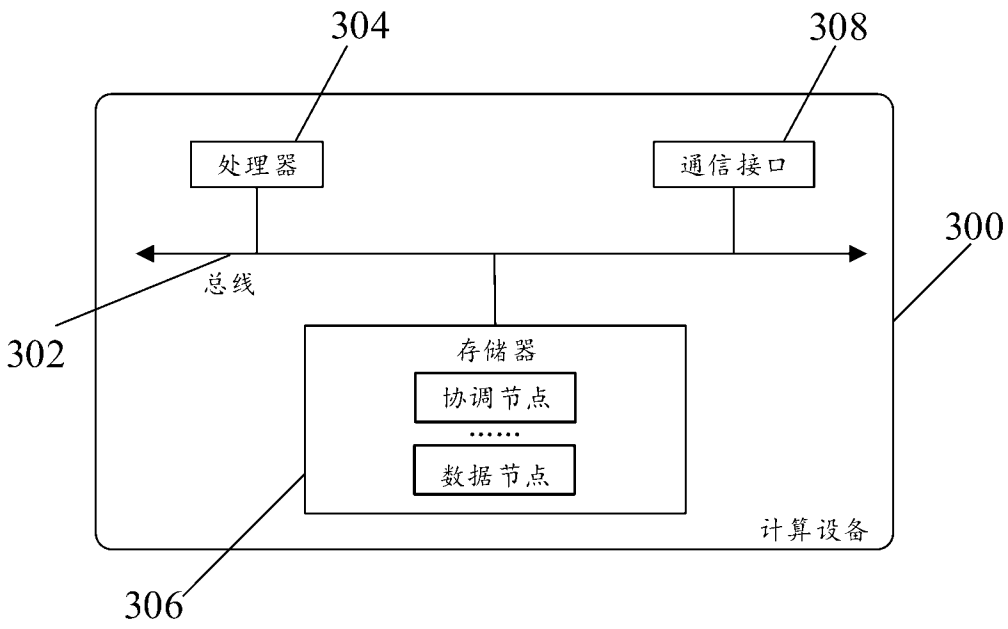

304
308
处理器
通信接口
300
总线
302
存储器
协调节点
……
数据节点
计算设备
306

图 3

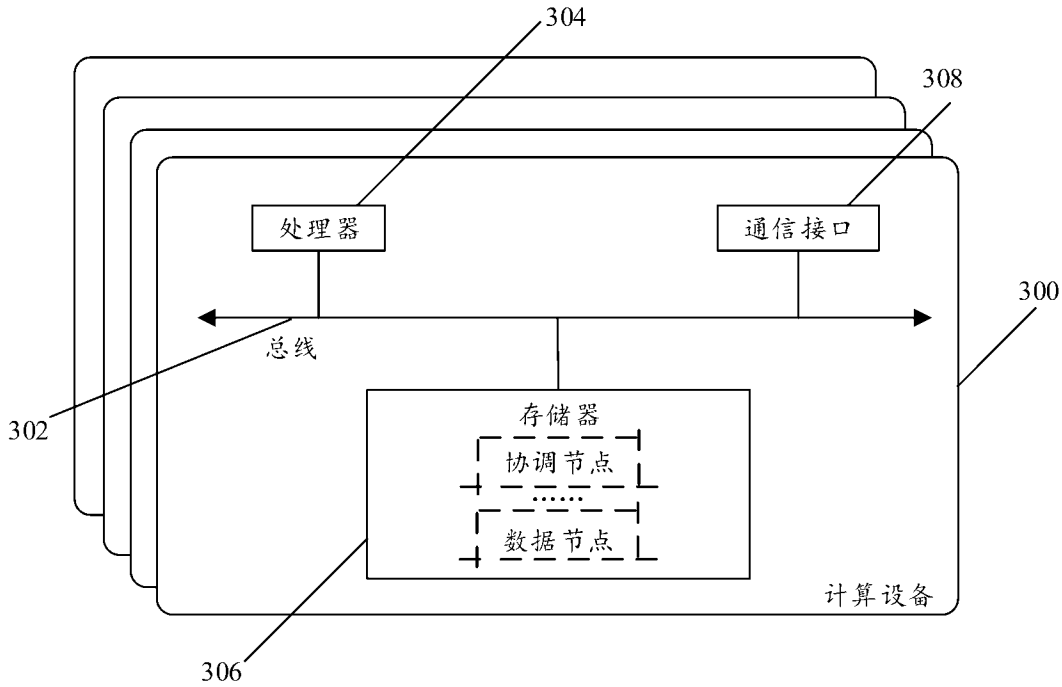


图 4

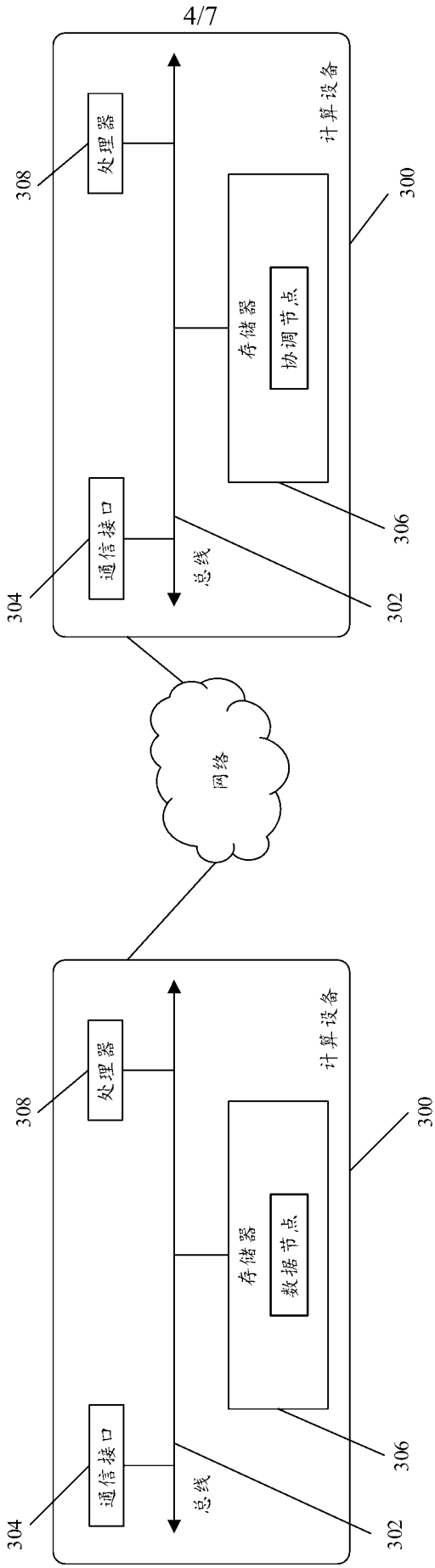

计算设备
处理器
308
存储器
协调节点
300
306
通信接口
304
总线
302
网络
计算设备
处理器
308
存储器
数据节点
300
306
通信接口
304
总线
302

图5

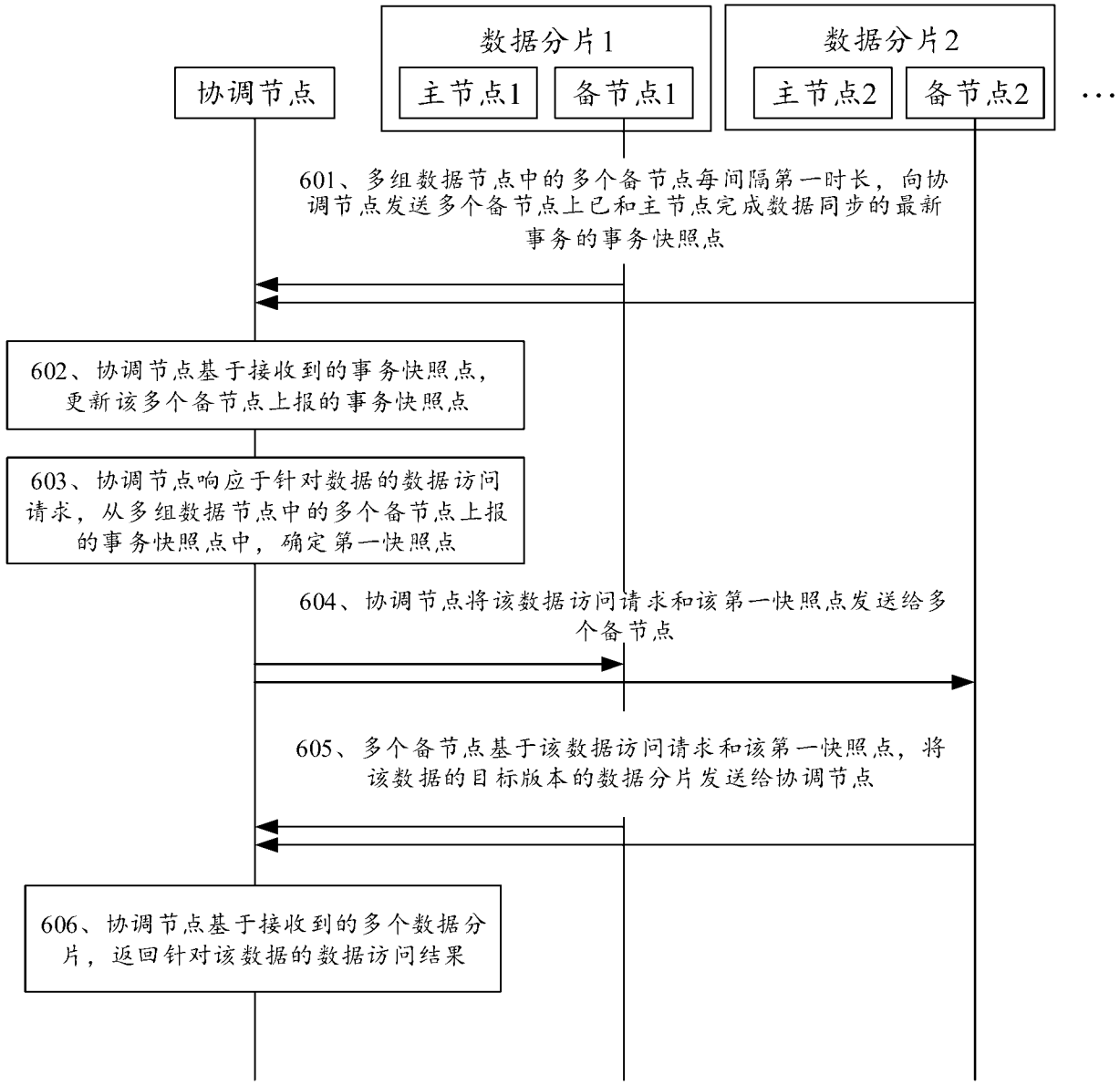
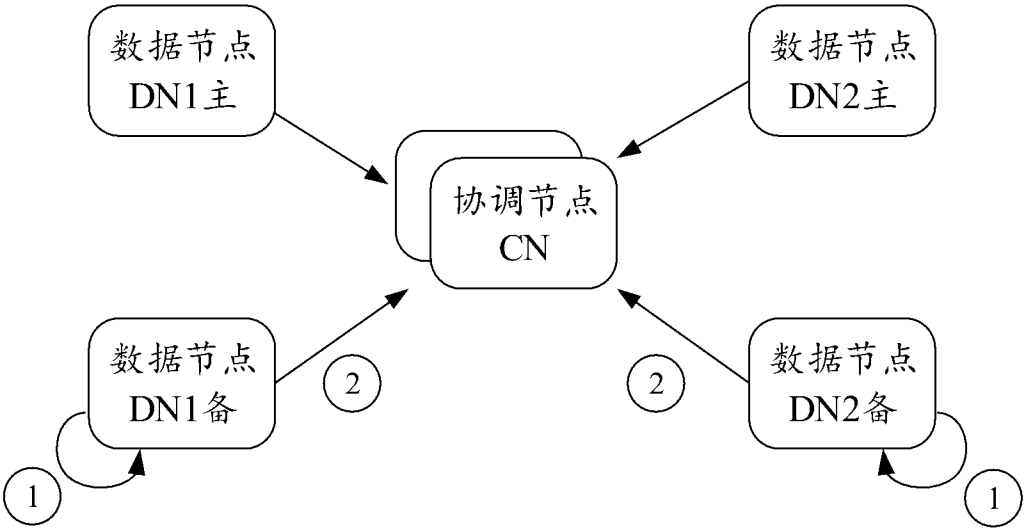

数据分片1
数据分片2
协调节点
主节点1
备节点1
主节点2
备节点2
...
601、多组数据节点中的多个备节点每间隔第一时长，向协调节点发送多个备节点上已和主节点完成数据同步的最新事务的事务快照点
602、协调节点基于接收到的事务快照点，更新该多个备节点上报的事务快照点
603、协调节点响应于针对数据的数据访问请求，从多组数据节点中的多个备节点上报的事务快照点中，确定第一快照点
604、协调节点将该数据访问请求和该第一快照点发送给多个备节点
605、多个备节点基于该数据访问请求和该第一快照点，将该数据的目标版本的数据分片发送给协调节点
606、协调节点基于接收到的多个数据分片，返回针对该数据的数据访问结果

图 6



1 备节点内部维护一个最大的事务一致性快照点

2 备节点定时上报事务一致性快照点给协调节点

图 7

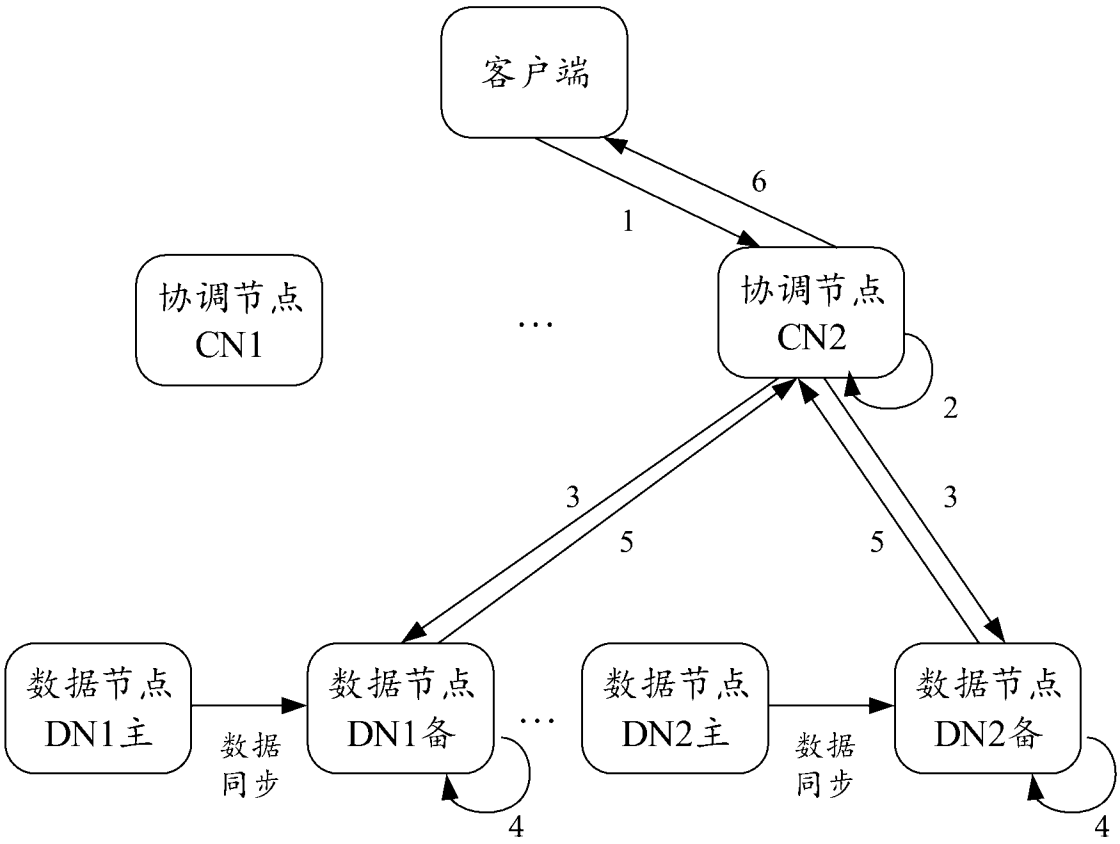


图 8

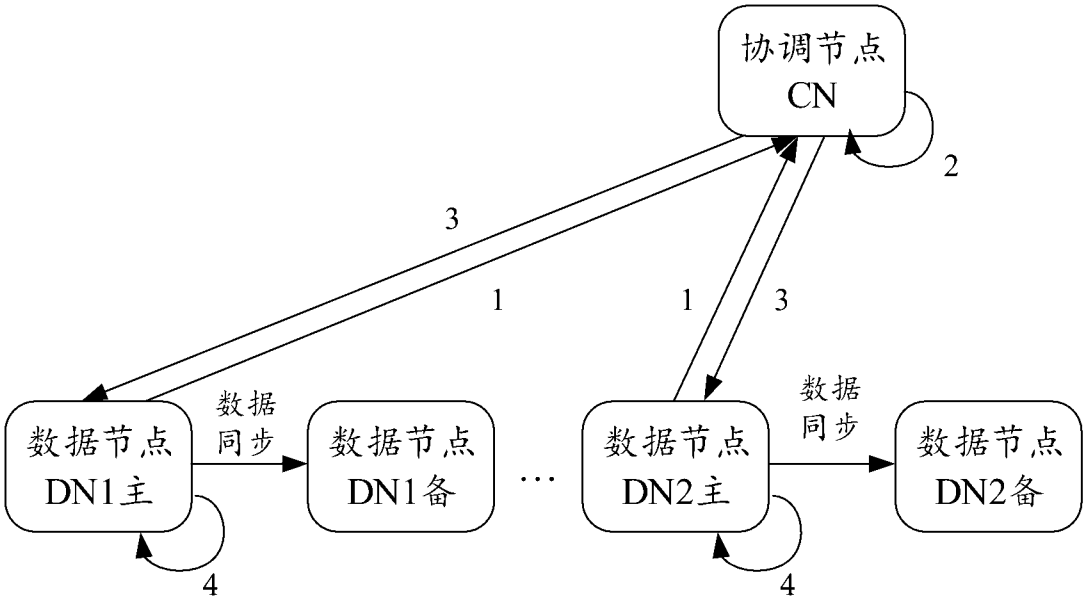

协调节点
CN
2
3
1
1
3
数据节点
DN1主
数据
同步
数据节点
DN1备
…
数据节点
DN2主
数据
同步
数据节点
DN2备
4
4

图 9

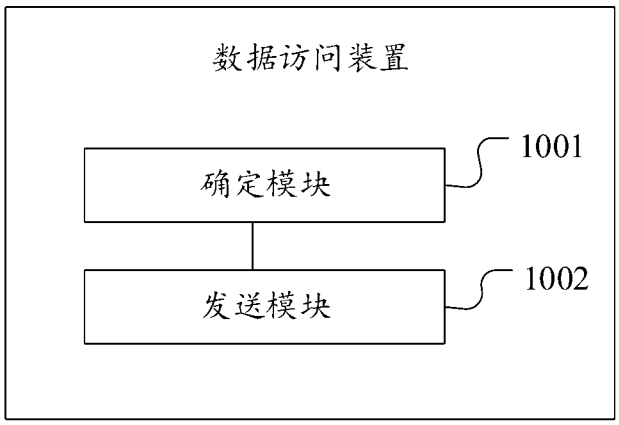

数据访问装置
确定模块
1001
发送模块
1002

图 10

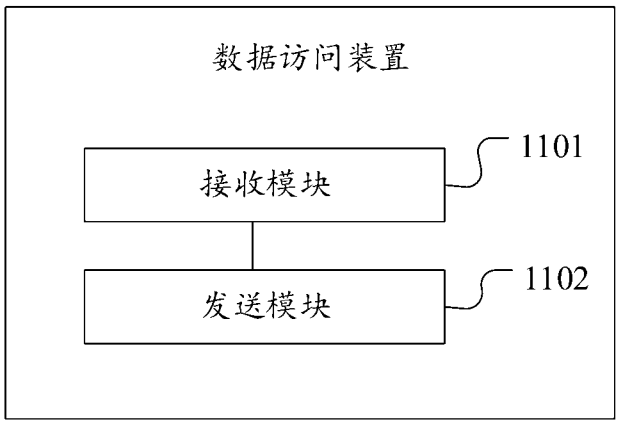

数据访问装置
接收模块
1101
发送模块
1102

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/079068

A. CLASSIFICATION OF SUBJECT MATTER

G06F16/23(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, ENTXTC, CJFD: 华为云计算, 徐宜良, 一致, 主节点, 备节点, 快照, 事物, master, segment, mirror, assign, snapshot, consistency

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113987064 A (ALIBABA CLOUD COMPUTING LTD.) 2022-01-28 (2022-01-28) description, paragraphs [0003]-[0132], and figures 1-11	1-21
A	CN 110737719 A (SHENZHEN PING AN COMMUNICATION TECHNOLOGY CO., LTD.) 2020-01-31 (2020-01-31) entire document	1-21
A	CN 111338766 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD. et al.) 2020-06-26 (2020-06-26) entire document	1-21
A	CN 113535656 A (RENMIN UNIVERSITY OF CHINA et al.) 2021-10-22 (2021-10-22) entire document	1-21
A	US 2018336258 A1 (SAP S.E.) 2018-11-22 (2018-11-22) entire document	1-21

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 March 2023	**13 April 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/079068

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113987064 A	28 January 2022	None	
CN 110737719 A	31 January 2020	None	
CN 111338766 A	26 June 2020	None	
CN 113535656 A	22 October 2021	None	
US 2018336258 A1	22 November 2018	US 10545995 B2	28 January 2020

国际检索报告

国际申请号
PCT/CN2023/079068

A. 主题的分类

G06F16/23(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT,ENTXTC,CJFD:华为云计算,徐宜良,一致 ，主节点，备节点，快照，事物，master，segment，mirror，assign，snapshot，consistency

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 113987064 A (阿里云计算有限公司 等) 2022-01-28 (2022-01-28) 说明书第[0003]-[0132]段以及附图1-11	1-21
A	CN 110737719 A (深圳平安通信科技有限公司) 2020-01-31 (2020-01-31) 全文	1-21
A	CN 111338766 A (腾讯科技（深圳）有限公司 等) 2020-06-26 (2020-06-26) 全文	1-21
A	CN 113535656 A (中国人民大学 等) 2021-10-22 (2021-10-22) 全文	1-21
A	US 2018336258 A1 (SAP S.E.) 2018-11-22 (2018-11-22) 全文	1-21

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023年3月24日	2023年4月13日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	马邺晨
传真号 (86-10)62019451	电话号码 (+86) 010-53962367

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/079068

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 113987064 A	2022年1月28日	无	
CN 110737719 A	2020年1月31日	无	
CN 111338766 A	2020年6月26日	无	
CN 113535656 A	2021年10月22日	无	
US 2018336258 A1	2018年11月22日	US 10545995 B2	2020年1月28日