

(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
2022 年 1 月 27 日 (27.01.2022)



(10) 国际公布号
WO 2022/017347 A1

(51) 国际专利分类号:
G06F 16/27 (2019.01)

(21) 国际申请号: PCT/CN2021/107243

(22) 国际申请日: 2021 年 7 月 20 日 (20.07.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010721315.0 2020 年 7 月 24 日 (24.07.2020) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (**ALIBABA GROUP HOLDING LIMITED**) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 孙梦石 (**SUN, Mengshi**); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。楼江航 (**LOU, Jianghang**); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。傅宇 (**FU, Yu**); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。吴学强 (**WU, Xueqiang**); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼, Zhejiang 311121 (CN)。

(74) 代理人: 北京三友知识产权代理有限公司 (**BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.**); 中国北京市金

融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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))。

(54) **Title:** DISTRIBUTED DATABASE SYSTEM AND DATA PROCESSING METHOD

(54) 发明名称: 分布式数据库系统及数据处理方法

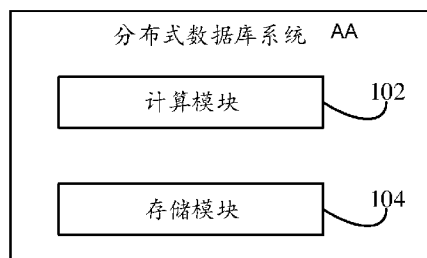


图 1

102 Calculation module
104 Storage module
AA Distributed database system

(57) **Abstract:** A distributed database system and a data processing method. The system comprises a calculation module (102) and a storage module (104). The calculation module (102) is configured to: receive a transaction starting request; according to the transaction starting request, determine a target transaction and a target storage module (104) corresponding to the target transaction; transmit an execution instruction to the target storage module (104); and in the case that the target storage module (104) returns a message that the target transaction is successfully executed, transmit a submission instruction to the target storage module (104), wherein the submission instruction carries a target transaction to be submitted and the submission timestamp of the target transaction. The target storage module (104) is configured to: receive the execution instruction, and according to the execution instruction, execute the target transaction; and receive the submission instruction, submit the target transaction according to the submission instruction, and according to the submission timestamp of the target transaction, generate a binary log.

(57) 摘要：一种分布式数据库系统及数据处理方法，其中所述系统包括：计算模块（102）和存储模块（104）；所述计算模块（102），被配置为接收事务开启请求；根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块（104）；向所述目标存储模块（104）发送执行指令；所述目标存储模块（104）返回目标事务执行成功的消息的情况下，向所述目标存储模块（104）发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳；所述目标存储模块（104），被配置为接收所述执行指令，根据所述执行指令执行所述目标事务；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

分布式数据库系统及数据处理方法

本申请要求 2020 年 07 月 24 日递交的申请号为 202010721315.0、发明名称为“分布式数据库系统及数据处理方法”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。

5 技术领域

本说明书涉及分布式数据库技术领域，特别涉及一种分布式数据库系统及数据处理方法，一种计算设备，以及一种计算机可读存储介质。

背景技术

随着计算机技术的日益发展，为了满足用户日益增长的需求，分布式数据库应运而生。分布式数据库由多个存储节点组成，在每个存储节点中，每一次 DML（Data Manipulation Language，数据操纵语言）写入操作都会独立生成一个二进制日志（binlog），这些二进制日志在所在的存储节点内保持有序。然而，在分布式数据库中，时间上前后的两次操作可能会发生在两个不同的存储节点上，直接订阅两个存储节点的二进制日志，有可能导致这两次操作的顺序发生错乱，最终的数据出现不一致。

15 发明内容

有鉴于此，本说明书实施例提供了一种分布式数据库系统。本说明书同时涉及一种数据处理方法，一种计算设备，以及一种计算机可读存储介质，以解决现有技术中存在的

的技术缺陷。

根据本说明书实施例的第一方面，提供了一种分布式数据库系统，包括：计算模块

20 和存储模块；

所述计算模块，被配置为接收事务开启请求；根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳；

25

所述目标存储模块，被配置为接收所述执行指令，根据所述执行指令执行所述目标事务；所述目标事务执行成功的情况下，向所述计算模块返回目标事务执行成功的消息；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

30 可选的，所述系统还包括：多个执行模块和一个排序模块，所述多个执行模块中每一个执行模块均对应存储模块；

所述执行模块，被配置为读取自身对应的存储模块的二进制日志，并根据所述二进制日志生成内存序列；

所述排序模块，被配置为读取每一个所述执行模块的内存序列，获取所述内存序列

中每一个二进制日志中目标事务的提交时间戳；按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；将排序后的二进制日志写入消息队列中。

可选的，所述计算模块进一步被配置为：

若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；

5 根据所述执行逻辑，生成所述表格处理事件的时间戳；

根据所述时间戳，生成所述表格处理事件的二进制日志。

可选的，所述计算模块进一步被配置为：

确定所述表格处理事件的操作类型；

根据所述操作类型，确定所述表格处理事件对应的兼容类型；

10 根据所述兼容类型，确定所述表格处理事件的执行逻辑。

可选的，所述计算模块进一步被配置为：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所述表格处理事件执行后的版本；

15 若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本。

可选的，所述计算模块进一步被配置为：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处理事件的执行逻辑为后序执行；

20 若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处理事件的执行逻辑为前序执行。

可选的，所述计算模块进一步被配置为：

若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的二进制日志之前，获取所述表格处理事件的时间戳；

25 若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表格处理事件的时间戳。

可选的，所述排序模块进一步被配置为：

根据所述表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

30 可选的，所述系统还包括时间戳产生器，所述计算模块进一步被配置为：

向所述时间戳产生器申请全局有序的时间戳；

将所述时间戳确定为所述目标事务的提交时间戳。

可选的，所述事务开启请求中携带事务标识及对应的存储模块标识，所述计算模块进一步被配置为：

将所述事务标识指示的事务确定为目标事务；

将所述存储模块标识指示的存储模块确定为对应的目标存储模块。

可选的，所述事务开启请求中携带事务标识，所述计算模块进一步被配置为：

将所述事务标识指示的事务确定为目标事务；

5 根据预设分配规则，确定所述目标事务对应的目标存储模块。

可选的，所述执行模块进一步被配置为：

若检测到自身对应的存储模块生成新的二进制日志，则返回执行所述读取所述存储模块的二进制日志的操作步骤；

相应的，所述排序模块进一步被配置为：

10 返回执行所述读取每一个所述执行模块的内存序列的操作步骤；或者，

获取所述执行模块中增加的二进制日志，将所述二进制日志放入所述消息队列的队尾。

可选的，所述计算模块进一步被配置为：

记录所述目标存储模块的存储模块标识。

15 可选的，所述计算模块进一步被配置为：

向所述目标存储模块发送目标事务开启请求；

所述目标存储模块返回开启消息的情况下，执行所述向所述目标存储模块发送执行指令的操作步骤。

可选的，所述存储模块进一步被配置为：

20 生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释中。

根据本说明书实施例的第二方面，提供了一种数据处理方法，应用于上述第一方面提供的分布式数据库系统中的计算模块，所述方法包括：

接收事务开启请求；

根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；

25 向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；

所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

可选的，所述方法还包括：

若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；

30 根据所述执行逻辑，生成所述表格处理事件的时间戳；

根据所述时间戳，生成所述表格处理事件的二进制日志。

可选的，所述确定所述表格处理事件的执行逻辑，包括：

确定所述表格处理事件的操作类型；

根据所述操作类型，确定所述表格处理事件对应的兼容类型；

根据所述兼容类型，确定所述表格处理事件的执行逻辑。

可选的，所述根据所述操作类型，确定所述表格处理事件对应的兼容类型，包括：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所述表格处理事件执行后的版本；

若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本。

可选的，所述根据所述兼容类型，确定所述表格处理事件的执行逻辑，包括：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处理事件的执行逻辑为后序执行；

若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处理事件的执行逻辑为前序执行。

可选的，所述根据所述执行逻辑，生成所述表格处理事件的时间戳，包括：

若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的二进制日志之前，获取所述表格处理事件的时间戳；

若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表格处理事件的时间戳。

可选的，所述向所述目标存储模块发送提交指令之前，还包括：

向时间戳产生器申请全局有序的时间戳；

将所述时间戳确定为所述目标事务的提交时间戳。

可选的，所述事务开启请求中携带事务标识及对应的存储模块标识，所述根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块，包括：

将所述事务标识指示的事务确定为目标事务；

将所述存储模块标识指示的存储模块确定为对应的目标存储模块。

可选的，所述事务开启请求中携带事务标识，所述根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块包括：

将所述事务标识指示的事务确定为目标事务；

根据预设分配规则，确定所述目标事务对应的目标存储模块。

可选的，所述向所述目标存储模块发送执行指令之后，还包括：

记录所述目标存储模块的存储模块标识。

可选的，所述根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块之后，且所述向所述目标存储模块发送执行指令之前，还包括：

向所述目标存储模块发送目标事务开启请求；

所述目标存储模块返回开启消息的情况下，执行所述向所述目标存储模块发送执行

指令的操作步骤。

根据本说明书实施例的第三方面，提供了一种数据处理方法，应用于上述第一方面提供的分布式数据库系统中的存储模块，所述方法包括：

接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

可选的，所述根据所述目标事务的提交时间戳生成二进制日志，包括：

生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释中。

根据本说明书实施例的第四方面，提供了一种数据处理方法，应用于上述第一方面提供的分布式数据库系统中的排序模块，所述方法包括：

读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

将排序后的二进制日志写入消息队列中。

可选的，所述将排序后的二进制日志写入消息队列中之后，还包括：

若检测到执行模块的内存序列发生变化，则返回执行所述读取每一个所述执行模块的内存序列的操作步骤。

可选的，所述将排序后的二进制日志写入消息队列中之后，还包括：

若检测到执行模块的内存序列发生变化，则获取所述执行模块中增加的二进制日志，将所述二进制日志放入所述消息队列的队尾。

可选的，所述将排序后的二进制日志写入消息队列中之后，还包括：

根据表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

根据本说明书实施例的第五方面，提供了一种计算设备，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指

令；

接收事务开启请求；

根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；

向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；

所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发

送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

根据本说明书实施例的第六方面，提供了一种计算设备，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指

5 令：

接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务
10 的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

根据本说明书实施例的第七方面，提供了一种计算设备，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指

15 令：

读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

将排序后的二进制日志写入消息队列中。

20 根据本说明书实施例的第八方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述第二方面任意所述数据处理方法的步骤。

根据本说明书实施例的第九方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述第三方面任意所述数据处理方法的步骤。

25 根据本说明书实施例的十方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述第四方面任意所述数据处理方法的步骤。

30 本说明书提供的分布式数据库系统包括计算模块和存储模块，所述计算模块被配置为接收事务开启请求；根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；向所述目标存储模块发送执行指令；所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳；所述目标存储模块被配置为接收所述执行指令，根据所述执行指令执行所述目标事务；所述目标事务执行成功的情况下，向所述计算模块返回目标事务执行成功的消息；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。如此，每个存储模块都是根据计算模块发送的目标事务的提交时间戳生成的二进制日志，每个二进制日志中

都携带一个全局有序的时间戳，使得分布式数据库系统中时间上前后的两次操作可以按序排列，避免了操作的顺序发生错乱。

附图说明

图 1 是本说明书一实施例提供的一种分布式数据库系统的结构框图；

5 图 2 是本说明书一实施例提供的第一种数据处理方法的流程图；

图 3 是本说明书一实施例提供一种前序执行和后序执行示意图；

图 4 是本说明书一实施例提供的第二种数据处理方法的流程图；

图 5 是本说明书一实施例提供的第三种数据处理方法的流程图；

图 6 是本说明书一实施例提供的第四种数据处理方法的流程图；

10 图 7 是本说明书一实施例提供的一种二进制日志生成过程示意图；

图 8 是本说明书一实施例提供的一种排序过程示意图；

图 9 是本说明书一实施例提供的第一种计算设备的结构框图；

图 10 是本说明书一实施例提供的第二种计算设备的结构框图；

图 11 是本说明书一实施例提供的第三种计算设备的结构框图。

15 具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本说明书内涵的情况下做类似推广，因此本说明书不受下面公开的具体实施的限制。

20 在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

25 应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

首先，对本说明书一个或多个实施例涉及的名词术语进行解释。

30 SQL (Structured Query Language, 结构化查询语言)：是一种特殊目的的编程语言，一种数据库查询和程序设计语言，用于存取数据以及查询、更新和管理关系数据库系统。SQL 是高级的非过程化编程语言，允许用户在高层数据结构上工作，它不要求用户指定对数据的存放方法，也不需要用户了解具体的数据存放方式，所以具有完全不同底层结构的不同数据库系统，可以使用相同的 SQL 作为数据输入与管理的接口，SQL 可以嵌套，

这使它具有极大的灵活性和强大的功能。SQL 语言包括四种主要程序设计语言类别的语句：数据定义语言(DDL)，数据操作语言(DML)，数据控制语言(DCL)和事务控制语言(TCL)。其中，数据操作语言(DML)，例如增、删、改等语句；数据定义语言(Data Definition Language, DDL)，例如增加列、变更列类型、长度、增加约束等语句。

5 事务：是恢复和并发控制的基本单位，是由多条 SQL 语句组成的一个执行单元，具有原子性、一致性、隔离性、持久性 4 个特性，原子性（atomicity）是指一个事务是一个不可分割的工作单位，事务中包括的操作要么都做，要么都不做；一致性（consistency）是指事务必须是使数据库从一个一致性状态变到另一个一致性状态，一致性与原子性是密切相关的；隔离性（isolation）是指一个事务的执行不能被其他事务干扰，即一个事务内部的操作及使用的数据对并发的其他事务是隔离的，并发执行的各个事务之间不能互相干扰；持久性（durability）也称永久性（permanence），指一个事务一旦提交，它对数据库数据的改变就应该是永久性的，接下来的其他操作或故障不应该对其有任何影响。

15 MySQL：是一个关系型数据库管理系统，随着 MySQL 的不断成熟，逐渐用于更多大规模网站和应用。

 PolarDB：是新一代关系型云原生数据库，既拥有分布式设计的低成本优势，又具有集中式的易用性，实现了计算模块及存储模块的分离，提供即时生效的可扩展能力和运维能力，软硬一体化设计，满足大规模应用场景需求。

20 PolarDB-X：PolarDB 的分布式版，一个基于 MySQL sharding（数据库分片）的分布式数据库，主要由存储模块（MySQL server）与计算模块组成。

 TSO：一个中心化的时间戳产生器，提供全局有序的时间戳服务。

25 binlog：二进制格式的日志文件，用来记录 Mysql 内部对数据库的改动（只记录对数据的修改操作），主要用于数据库的主从复制以及增量恢复。MySQL 官方支持三种 binlog 数据格式，对应参数为 binlog_format，可选值分别为 ROW、STATEMENT、MIXED，对应的默认值为 MIXED，需要调整为 ROW 模式，该模式的特点为：每个增删改的 SQL，都会在 binlog 里记录 SQL 所影响的行在变更前和变更后的值。在 MySQL 中，每一次 DML（增删改等写入，包括 DDL）操作都会生成 binlog（binary log）日志，此日志在 MySQL 中是按照操作时间有序的，下游应用可以基于此日志做一些订阅操作（例如数据同步）。

30 物理 binlog：PolarDB-X 的每个存储模块的 binlog。

 逻辑 binlog：PolarDB-X 层提供的 binlog，包含所有存储模块的 binlog，对外展现形式是一个 kafka 消息队列。

 Commit 指令：提交指令，用于把事务所做的修改保存到数据库。

 commit_timestamp：用来标明事务的提交时间，缩写 commit_ts。

精卫 worker: 一个执行工作的模块, 与存储模块一一对应, 负责读取对应的存储模块的 binlog 文件, 并负责进行解析, 同时将解析后的结果生成一个内存序列。

精卫 merger: 相当于一个归并排序的模块, 将多个有序的队列, 归并为一个全局有序的队列。

5 接下来, 对本说明书的基本构思进行简述:

目前, 在分布式数据库中, 每个存储节点执行事务 (DML 相关的操作) 后都会独立生成一个二进制日志, 每个存储节点生成的二进制日志在该存储节点内保持有序, 然而不同的存储节点生成的二进制日志之间时间先后顺序并不确定, 因而在分布式数据库中, 时间上前后的两次操作可能会发生在两个不同的存储节点上, 下游应用直接订阅两个存储节点的二进制日志, 有可能导致这两次操作的顺序发生错乱, 从而导致下游应用最终的数据出现不一致。

例如, 分布式数据库系统中包括 MySQL1、MySQL2 和 MySQL3, MySQL1 内生成了 2 个二进制日志: 日志 A1 和日志 A2, MySQL2 内生成了 3 个二进制日志: 日志 B1、日志 B2 和日志 B3, MySQL3 内生成了 2 个二进制日志: 日志 C1 和日志 C2。现有技术中, 只能确定日志 A1 早于日志 A2, 日志 B1 早于日志 B2 早于日志 B3, 日志 C1 早于日志 C2, 但日志 A1、日志 A2、日志 B1、日志 B2、日志 B3、日志 C1 和日志 C2 之间的先后顺序并不确定, 有可能会出现错乱, 导致下游应用订阅的数据出现不一样。

本说明书中提供了一种分布式数据库系统, 包括计算模块和存储模块 (如 MySQL), 计算模块控制存储模块执行目标事务, 存储模块执行成功后向计算模块返回成功消息, 此时计算模块会向存储模块发送目标事务的提交时间戳, 以使存储模块根据该时间戳生成二进制日志, 这种情况下的二进制日志中记录有全局有序的目标事务的提交时间戳, 使得分布式数据库系统中所有的二进制日志都是有时间顺序的, 从而避免了事务的操作顺序发生错乱, 保证了下游应用订阅数据的一致性。

在本说明书中, 提供了一种分布式数据库系统, 本说明书同时涉及一种数据处理方法, 一种计算设备, 以及一种计算机可读存储介质, 在下面的实施例中逐一进行详细说明。

图 1 示出了根据本说明书一实施例提供的一种分布式数据库系统的结构框图, 具体包括计算模块 102 和存储模块 104。

所述计算模块 102, 被配置为接收事务开启请求; 根据所述事务开启请求, 确定目标事务及所述目标事务对应的存储模块 104; 向所述存储模块 104 发送执行指令, 所述执行指令中携带有待执行目标事务; 所述存储模块 104 返回目标事务执行成功的情况, 向所述存储模块 104 发送提交指令, 所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳;

所述存储模块 104, 被配置为接收所述执行指令, 根据所述执行指令执行所述目标

事务；所述目标事务执行成功的情况下，向所述计算模块 102 返回目标事务执行成功的消息；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

5 一个可选的实施例中，所述系统还包括：多个执行模块和一个排序模块，所述多个执行模块中每一个执行模块均对应有存储模块 104；

所述执行模块，被配置为读取自身对应的存储模块 104 的二进制日志，并根据所述二进制日志生成内存序列；

10 所述排序模块，被配置为读取每一个所述执行模块的内存序列，获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；将排序后的二进制日志写入消息队列中。

一个可选的实施例中，所述计算模块进一步被配置为：

若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；

根据所述执行逻辑，生成所述表格处理事件的时间戳；

根据所述时间戳，生成所述表格处理事件的二进制日志。

15 一个可选的实施例中，所述计算模块进一步被配置为：

确定所述表格处理事件的操作类型；

根据所述操作类型，确定所述表格处理事件对应的兼容类型；

根据所述兼容类型，确定所述表格处理事件的执行逻辑。

20 一个可选的实施例中，所述计算模块进一步被配置为：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所述表格处理事件执行后的版本；

若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本。

25 一个可选的实施例中，所述计算模块进一步被配置为：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处理事件的执行逻辑为后序执行；

若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处理事件的执行逻辑为前序执行。

30 一个可选的实施例中，所述计算模块进一步被配置为：

若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的二进制日志之前，获取所述表格处理事件的时间戳；

若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表格处理事件的时间戳。

一个可选的实施例中，所述排序模块进一步被配置为：

根据所述表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

5 一个可选的实施例中，所述系统还包括时间戳产生器，所述计算模块 102 进一步被配置为：

向所述时间戳产生器申请全局有序的时间戳；

将所述时间戳确定为所述目标事务的提交时间戳。

一个可选的实施例中，所述事务开启请求中携带事务标识及对应的存储模块标识，所述计算模块 102 进一步被配置为：

10 将所述事务标识指示的事务确定为目标事务；

将所述存储模块标识指示的存储模块确定为对应的存储模块 104。

一个可选的实施例中，所述事务开启请求中携带事务标识，所述计算模块 102 进一步被配置为：

将所述事务标识指示的事务确定为目标事务；

15 根据预设分配规则，确定所述目标事务对应的存储模块 104。

一个可选的实施例中，所述执行模块进一步被配置为：

若检测到自身对应的存储模块 104 生成新的二进制日志，则返回执行所述读取所述存储模块 104 的二进制日志的操作步骤；

相应的，所述排序模块进一步被配置为：

20 返回执行所述读取每一个所述执行模块的内存序列的操作步骤；或者，

获取所述执行模块中增加的二进制日志，将所述二进制日志放入所述消息队列的队尾。

一个可选的实施例中，所述计算模块 102 进一步被配置为：

记录所述存储模块 104 的存储模块标识。

25 一个可选的实施例中，所述计算模块 102 进一步被配置为：

向所述存储模块 104 发送目标事务开启请求；

所述存储模块 104 返回开启消息的情况下，执行所述向所述存储模块 104 发送执行指令的操作步骤。

一个可选的实施例中，所述存储模块 104 进一步被配置为：

30 生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释中。

本说明书提供的分布式数据库系统包括计算模块和存储模块，所述计算模块被配置为接收事务开启请求；根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；向所述目标存储模块发送执行指令；所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交

目标事务和所述目标事务的提交时间戳；所述目标存储模块被配置为接收所述执行指令，根据所述执行指令执行所述目标事务；所述目标事务执行成功的情况下，向所述计算模块返回目标事务执行成功的消息；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。如此，每个存储模块都是根据计算模块发送的目标事务的提交时间戳生成的二进制日志，每个二进制日志中都携带一个全局有效的时间戳，使得分布式数据库系统中时间上前后的两次操作生成的二进制日志可以按序排列，从而避免了事务的操作顺序发生错乱，保证了下游应用订阅数据的一致性。

图 2 示出了根据本说明书一实施例提供的一种数据处理方法的流程图，应用于图 1 所示的分布式数据库系统中的计算模块，具体包括如下步骤：

步骤 202：接收事务开启请求。

实际应用中，当用户要在分布式数据库系统中执行某个事务时，可以通过应用程序向分布式数据库系统中的计算模块发送一个事务开启请求，计算模块接收该事务开启请求，从而开启后续执行事务的整个过程。此处的分布式数据库系统可以是指 PolarDB-X，也可以是其他分布式数据库，本说明书对此不进行限制。

步骤 204：根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块。

具体的，在接收事务开启请求的基础上，进一步的，将根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块。

其中，所述目标事务是指事务开启请求指示开启的事务，所述目标存储模块是指执行所述目标事务所需涉及到的存储模块。此处的存储模块可以是分布式数据库系统中的单片数据库，例如单片 MySQL。

实际应用中，计算模块接收到事务开启请求后，首先需要确定该请求指示执行的具体事务，由于事务最终是由存储模块执行的，因而计算模块还需要进一步确定出执行该事务的存储模块。由于事务是由多条 SQL 语句组成的执行单元，因而一个事务中可能包括有多个 SQL 操作，该多个操作可能会由不同的存储模块执行，因而一个事务可能会涉及到多个存储模块，也即确定出的目标存储模块可能为多个。

在本实施例的一个或多个实施方式中，用户要在分布式数据库系统中执行某个事务时，可以直接指定执行该事务的存储模块，此时应用程序会将要执行的事务对应的事务标识，以及执行所述事务的存储模块的标识携带在事务开启请求中发送给计算模块，计算模块可以直接根据接收到的事务开启请求中携带的事务标识及对应的存储模块标识，确定目标事务及所述目标事务对应的目标存储模块，具体实现过程如下：

将所述事务标识指示的事务确定为目标事务；

将所述存储模块标识指示的存储模块确定为对应的目标存储模块。

具体的，事务标识用于标识要执行的事务，存储模块标识用于标识执行事务的存储模块。

例如，用户要在分布式数据库系统中执行事务 A，且用户指定了由存储模块 1 和存储模块 2 来执行所述事务 A，此时用户可以通过某应用程序向分布式数据库系统中的计算模块发送一个事务开启指令，该指令中会携带事务 A、存储模块 1 和存储模块 2。计算模块接收到该事务开启指令后，就可以确定出要执行的目标事务为事务 A，执行事务 A 的存储模块为存储模块 1 和存储模块 2。

在事务开启请求中直接携带事务标识及对应的存储模块标识，计算模块直接根据标识即可确定出目标事务及目标事务对应的目标存储模块，确定过程简单易行，不需要计算模块进行复杂的运算，一定程度上节省了计算资源。

在本实施例的一个或多个实施方式中，用户要在分布式数据库系统中执行某个事务时，也可以不指定执行该事务的存储模块，由计算模块根据存储模块的性能属性、负载等参数指定对应的存储模块执行所述目标事务，此时应用程序只将要执行的事务对应的事务标识携带在事务开启请求中发送给计算模块，计算模块确定出要执行的目标事务后，可以指定执行所述目标事务的目标存储模块，具体实现过程如下：

将所述事务标识指示的事务确定为目标事务；

根据预设分配规则，确定所述目标事务对应的目标存储模块。

具体的，预设分配规则是指为待执行的目标事务分配存储模块的规则，该规则由技术人员预先进行设置，可以是负载均衡规则，还可以是存储模块性能规则，本说明书中不对分配规则进行具体限定，只要能为目标事务分配合理的存储模块的规则均可。

例如，预设分配规则为负载均衡，且假设分布式数据库系统包括存储模块 1 和存储模块 2，存储模块 1 正在执行或待执行的事务有事务 A、事务 B，存储模块 2 正在执行或待执行的事务有事务 C。此时，某用户通过应用程序向分布式数据库系统中的计算模块发送了一个事务开启指令，该指令中携带事务 D，计算模块接收到该事务开启指令后，确定出待执行的目标事务为事务 D，然后根据负载均衡规则，确定执行事务 D 的存储模块为存储模块 2。

只在事务开启请求中携带事务标识，计算模块确定出目标事务后，根据预先设置的分配规则，为所述目标事务分配对应的目标存储模块，可以根据性能属性、负载等参数灵活指定执行事务的存储模块，可以适应分布式数据库系统中各个存储模块自身的实际情况，保证了事务可以高效准确执行，适应性更高。

本说明书中既可以直接在事务开启请求中指定目标事务和执行目标事务的目标存储模块，还可以不指定，由计算模块为目标事务分配执行的存储模块，灵活性更高，可以适应不同事务的执行特点，从而可以更高效、更准确的执行目标事务。

步骤 206：向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标

事务。

具体的，在根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块的基础上，进一步的，将向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务。

5 在本实施例的一个或多个实施方式中，计算模块确定出目标事务及执行目标事务的目标存储模块后，目标存储模块并不知道自身被选定要开始执行目标事务了，因此在根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块之后，且所述向所述目标存储模块发送执行指令之前，计算模块还可以通知目标存储模块开启目标事务，具体实现过程如下：

10 向所述目标存储模块发送目标事务开启请求；

所述目标存储模块返回开启消息的情况下，执行所述向所述目标存储模块发送执行指令的操作步骤。

在本实施例的一个或多个实施方式中，后续目标存储模块执行目标事务后，计算模块还需要向目标存储模块发送提交指令，以使目标存储模块对目标事务进行提交，因而
15 计算模块在向目标存储模块发送执行指令后，还需要记录所述目标存储模块的存储模块标识，从而方便后续向目标存储模块发送提交指令。

步骤 208：所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

20 具体的，在向所述目标存储模块发送执行指令的基础上，进一步的，将在所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

实际应用中，目标存储模块接收到执行指令后，会执行目标事务，在目标事务执行成功的情况下，会向计算模块返回执行成功的消息，以通知计算模块，其已经成功执行
25 了目标事务，可以提交了。计算模块接收到该执行成功的消息后，就可以控制存储模块对目标事务进行提交。

在本实施例的一个或多个实施方式中，用户是通过应用程序使分布式数据库系统执行对应的目标事务，目标存储模块在目标事务执行成功的情况下，向计算模块返回执行成功的消息，进一步的此时计算模块还可以向应用程序返回目标事务执行完毕的消息，
30 应用程序会向计算模块返回事务提交消息，以使计算模块控制目标存储模块对事务进行提交。

在本实施例的一个或多个实施方式中，为了使各个存储模块提交事务时生成的二进制日志之间是有时间先后顺序的，计算模块在向目标存储模块发送提交指令时，会在该指令中携带目标事务的提交时间戳，因而在向所述目标存储模块发送提交指令之前，计

算模块还可以获取全局有序的目标事务的提交时间戳，具体实现方式如下：

向时间戳产生器申请全局有序的时间戳；

将所述时间戳确定为所述目标事务的提交时间戳。

具体的，时间戳产生器可以是分布式数据库系统中的 TSO，向 TSO 获取的目标事务的提交时间戳为 `commit_timestamp`。

实际应用过程中，计算模块除了会接收到与增删改数据相关的 DML 操作，还有可能会接收到 DDL 事件，如变更列类型、长度，增加约束等事件。分布式数据库系统在 DDL 变更过程中，同一个时刻，数据库的 Schema 元数据会存在两个版本，例如加列操作，新列对部分分片已经可见，对部分分片还未可见，此时会同时产生两个版本的数据。这两个版本的数据在通过逻辑 binlog 队列往下游系统（例如单机 MySQL）进行传递的过程中，需要将 DDL 事件放在消息队列适当的位置，确保下游系统能正确的接受这两个版本的数据，不会出现同一条数据在下游系统中出现不一致，具体实现过程如下：

若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；

根据所述执行逻辑，生成所述表格处理事件的时间戳；

根据所述时间戳，生成所述表格处理事件的二进制日志。

其中，表格处理事件是指 DDL 事件，例如，加列、列的长度变长、增加列的默认值、删除列、列的长度变短等事件。

在本实施例的一个或多个实施方式中，确定所述表格处理事件的执行逻辑，具体实现过程可以为：

确定所述表格处理事件的操作类型；

根据所述操作类型，确定所述表格处理事件对应的兼容类型；

根据所述兼容类型，确定所述表格处理事件的执行逻辑。

其中，操作类型是设定好的，不同的操作类型对应不同的兼容类型。

在本实施例的一个或多个实施方式中，根据所述操作类型，确定所述表格处理事件对应的兼容类型，具体实现过程可以为：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所述表格处理事件执行后的版本；

若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本。

具体的，一个时刻的 Schema 结构（包含表结构等）称为一个版本，DDL 事件执行前的版本称为第一版本，DDL 事件执行后的版本称为第二版本。需要说明的是，Va 版本兼容 Vb 版本是指：使用 Vb 版本的数据可以被写入使用 Va 版本的目标端；从使用 Va 版本的目标端将该数据读出后，其包含足够的信息，在经过一定的转换后，可以转换为

Vb 版本的数据，并且和原数据完全相等。

另外，第一类型即为第一版本可以兼容第二版本的操作类型，例如，删除列，列的长度变短，删表，减小列的精度等操作均为第一类型。第二类型即为第二版本可以兼容第一版本的操作类型，例如，加列，列的长度变长，增加列的默认值，创建、删除索引，

5 创建、删除约束（创建唯一键约束），建表，增加列的精度等操作均为第二类型。

在本实施例的一个或多个实施方式中，根据所述兼容类型，确定所述表格处理事件的执行逻辑，具体实现过程可以为：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处理事件的执行逻辑为后序执行；

10 若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处理事件的执行逻辑为前序执行。

具体的，若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的二进制日志之前，获取所述表格处理事件的时间戳；若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表

15 格处理事件的时间戳。

实际应用中，参见图 3 所示，前序执行是指 DDL 事件的二进制日志，应该在所有第二版本（V2）数据的二进制日志的最前面，因而在生成所有第二版本（V2）的数据对应的二进制日志之前，获取一个时间戳，将该时间戳作为 DDL 事件的时间戳，也即，DDL 事件应该位于所有的第二版本（V2）的二进制日志之前。后序执行是指 DDL 事件的二进制日志，应该在所有第一版本（V1）数据的二进制日志的最后面，因而在生成所有第一版本（V1）的数据对应的二进制日志之后，获取一个时间戳，将该时间戳作为 DDL 事件的时间戳，也即 DDL 事件应该位于所有的第一版本（V1）的二进制日志之后。

20

另外，某些表格处理事件（DDL 事件）的操作类型，第一版本可以兼容第二版本，第二版本也可以兼容第一版本，此时所述表格处理事件的执行逻辑既可以为后序执行，

25 也可以为前序执行。

在检测到表格处理事件的情况下，根据表格处理事件的执行逻辑，生成所述表格处理事件的时间戳，然后可以根据所述时间戳，生成所述表格处理事件的二进制日志，从而使得该表格处理事件的二进制日志可以处于适当的位置，保证下游应用订阅数据的一致性。

30 本说明书中提供的数据处理方法，计算模块接收事务开启请求，并根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；然后向目标存储模块发送执行指令；在目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。这种情况下，计算模块可以向时间戳产生器申请一个全局有序的时间戳，将该时间戳作

为目标事务的提交时间戳，并发送给目标存储模块，使得后续目标存储模块可以生成带全局有序时间戳的二进制日志，从而使得分布式数据库系统中时间上前后的两次操作生成的二进制日志可以按序排列，避免了事务的操作顺序发生错乱，保证了下游应用订阅数据的一致性。

5 图 4 示出了根据本说明书一实施例提供的一种数据处理方法的流程图，应用于图 1 所示的分布式数据库系统中的存储模块，具体包括如下步骤：

步骤 402：接收执行指令，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务。

10 实际应用中，计算模块确定出目标事务及执行目标事务的目标存储模块后，目标存储模块自身并不知道被选定要开始执行目标事务了，因此目标存储模块在接收到执行指令之前，还会接收到计算模块发送的目标事务开启请求，目标存储模块接收到该目标事务开启请求的情况下，会开启目标事务，并通知计算模块自己准备好执行目标事务了，因而在接收执行指令，根据所述执行指令执行目标事务之前，目标存储模块还应该在接收到目标事务开启请求的情况下，向计算模块返回开启消息。

15 步骤 404：所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息。

具体的，在接收执行指令，根据所述执行指令执行目标事务的基础上，进一步的，将在所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息。

20 步骤 406：接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

具体的，在所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息的基础上，进一步的，将接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

25 实际应用中，二进制日志即 binlog，每个存储模块生成的二进制日志被称作物理 binlog。

在本实施例的一个或多个实施方式中，根据所述目标事务的提交时间戳生成二进制日志，具体实现过程可以为：生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释中。

30 本说明书中提供的数据处理方法，目标存储模块接收执行指令，根据所述执行指令执行目标事务；并在所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；之后接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。这种情况下，每个存储模块都是根据计算模块发送的目标事务的提交时间戳生成的二进制日志，每个二进制日志中都携带一个全局

有序的时间戳，使得分布式数据库系统中时间上前后的两次操作生成的二进制日志可以按序排列，从而避免了事务的操作顺序发生错乱，保证了下游应用订阅数据的一致性。

图 5 示出了根据本说明书一实施例提供的一种数据处理方法的流程图，应用于分布式数据库系统中的排序模块，具体包括如下步骤：

5 步骤 502：读取每一个执行模块的内存序列。

实际应用中，每一个执行模块都对应有一个存储模块，在排序模块读取每一个执行模块的内存序列之前，执行模块还可以读取自身对应的存储模块的二进制日志，并根据所述二进制日志生成内存序列。

10 具体的，此处的执行模块可以指 PolarDB-X 生成的精卫 worker，一个执行工作的模块。排序模块可以指 PolarDB-X 生成的精卫 merger，用于管控生成的精卫 worker，一个归并排序的模块，可以将多个有序的队列，归并为一个全局有序的队列。

本说明书中每个执行模块都生成了一个内存序列，该内存序列包括其对应的存储模块生成的全部二进制日志，后续对二进制日志进行排序时，只需要获取内存序列中的二进制日志即可，不需要再去分别一个个获取存储模块生成的每个二进制日志，操作简单
15 方便，方便后续进行排序。

步骤 504：获取所述内存序列中每一个二进制日志中目标事务的提交时间戳。

具体的，在读取每一个执行模块的内存序列的基础上，进一步的，将获取所述内存序列中每一个二进制日志中目标事务的提交时间戳。

20 步骤 506：按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序。

具体的，在获取所述内存序列中每一个二进制日志中目标事务的提交时间戳的基础上，进一步的，将按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序。

步骤 508：将排序后的二进制日志写入消息队列中。

25 具体的，在按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序的基础上，进一步的，将排序后的二进制日志写入消息队列中。

实际应用中，消息队列可以指 kafka 消息队列，该消息队列是一个开源的消息队列，将排序后的二进制日志写入消息队列中，即生成了分布式数据库系统的逻辑 binlog，下游应用需要订阅分布式数据库中的数据时，可以直接读取该开源的消息队列中的逻辑
30 binlog，即可获得分布式数据库系统中存储模块的各个二进制日志，且该二进制日志在消息队列中都是有序排列的。

在本实施例的一个或多个实施方式中，对存储模块当前各个二进制日志进行排序后，存储模块可能会执行新的事务，生成新的二进制日志，因而执行模块在检测到其对应的存储模块生成新的二进制日志的情况下，可以返回执行读取对应的存储模块的二进制日

志，并根据所述二进制日志生成内存序列的操作步骤，此时排序模块可以对所有二进制日志全部重新排序，具体实现过程如下：

若检测到执行模块的内存序列发生变化，则返回执行所述读取每一个所述执行模块的内存序列的操作步骤。

- 5 实际应用中，若检测到执行模块的内存序列发生变化，说明执行模块对应的存储模块执行了新的事务，生成了新的二进制日志，此时排序模块会返回执行步骤 502-508，对所有二进制日志全部重新排序。

在本实施例的一个或多个实施方式中，对存储模块当前各个二进制日志进行排序后，存储模块可能会执行新的事务，生成新的二进制日志，因而执行模块在检测到其对应的
10 存储模块生成新的二进制日志的情况下，可以返回执行读取对应的存储模块的二进制日志，并根据所述二进制日志生成内存序列的操作步骤，由于新生成的二进制日志在时间上一定是晚于已经排序好的二进制日志的，因而此时排序模块可以直接将执行模块中增加的二进制日志放入消息队列的队尾。

本说明书中在存储模块执行了新的事务，生成新的二进制日志的情况下，既可以
15 所有的二进制日志全部重新排序，也可以直接将新生成的二进制日志放入消息队列的队尾，实现了实时对存储模块的所有二进制日志进行排序，且排序方式灵活，可以适用不同情况，使得排序结果更高效准确。

在本实施例的一个或多个实施方式中，若计算模块检测到了表格处理事件，则可以根据表格处理事件的时间戳，生成表格处理事件的二进制日志，此情况下，排序模块
20 可以根据表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

在检测到表格处理事件的情况下，可以根据表格处理事件的执行逻辑，生成所述表格处理事件的时间戳，然后可以根据所述时间戳，生成所述表格处理事件的二进制日志，且可以根据该表格处理事件的时间戳，将该表格处理事件插入消息队列中，从而使得该
25 表格处理事件可以放在消息队列中适当的位置，确保下游系统能正确的接收表格处理事件执行前和执行后两个版本的数据，避免了同一条数据在下游系统中出现不一致的情况。

本说明书提供的数据处理方法，排序模块可以读取每一个执行模块的内存序列，并获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；然后按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序，将排序后的二进制日志写入消息队列中。这种情况下，消息队列中包括分布式数据库系统中所有存储模块生成的
30 二进制日志，且都是按时间先后顺序有序排列，下游应用需要订阅分布式数据库中的数据时，可以直接读取该消息队列，即可避免事务的操作顺序发生错乱，保证订阅数据的一致性。

以本说明书提供的数据处理方法在分布式数据库系统中整体的应用为例，对所述数据处理方法进行进一步说明。其中，图 6 示出了本说明书一实施例提供的一种数据处理

方法的处理流程图，具体包括以下步骤：

步骤 602：计算模块接收应用程序发送的事务开启请求，并根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块。

5 实际应用中，在一个事务包括多个操作的情况下，应用程序向计算模块发送事务开启请求时，可以分多次发送，并将多个操作的标识分别携带在多个请求中，因而计算模块可以分多次接收到事务开启请求，每个请求中携带事务包括的一个操作。例如，目标事务为事务 A，包括操作 1 和操作 2，此时计算模块会先接收到一个携带操作 1 标识的请求，然后控制对应的存储模块执行操作 1，在操作 1 执行完毕后，计算模块会再接收到一个携带操作 2 标识的请求，然后再控制对应的存储模块执行操作 2，执行完毕后，
10 再对整个事务 A 进行提交。

步骤 604：计算模块向所述目标存储模块发送目标事务开启请求。

步骤 606：目标存储模块接收所述目标事务开启请求，开启目标事务，并向计算模块返回开启消息。

15 步骤 608：计算模块向所述目标存储模块发送执行指令，并记录所述目标存储模块的存储模块标识，所述执行指令中携带有待执行目标事务。

步骤 610：目标存储模块接收所述执行指令，根据所述执行指令执行所述目标事务，并在所述目标事务执行成功的情况下，向所述计算模块返回目标事务执行成功的消息。

步骤 612：计算模块向应用程序返回所述目标事务执行成功的消息，并接收应用程序发送的提交指令。

20 步骤 614：计算模块向时间戳产生器申请全局有序的时间戳，将所述时间戳确定为所述目标事务的提交时间戳，并向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

步骤 616：目标存储模块接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

25 步骤 618：执行模块读取自身对应的存储模块的二进制日志，并根据所述二进制日志生成内存序列。

步骤 620：排序模块读取每一个执行模块的内存序列，并获取所述内存序列中每一个二进制日志中目标事务的提交时间戳。

30 步骤 622：排序模块按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序，并将排序后的二进制日志写入消息队列中。

需要说明的是，虽然一个事务可能会包括多个 SQL 操作，但是一个事务内的 SQL 操作本就是有序的（即一个 SQL 操作执行完毕，才会开始执行下一个 SQL 操作），不需要再进行排序，只有不同事务之间前后顺序才可能会发生错乱，因而本说明书中的提交时间戳是针对事务整体的，也即一个事务对应一个提交时间戳。

步骤 624: 计算模块检测到表格处理事件, 若所述表格处理事件的操作类型属于第一类型, 则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本, 并确定所述表格处理事件的执行逻辑为后序执行, 在生成所有第一版本的数据对应的二进制日志之后, 获取所述表格处理事件的时间戳; 若所述表格处理事件的操作类型属于第二类型, 则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本, 并确定所述表格处理事件的执行逻辑为前序执行, 在生成所有第二版本的数据对应的二进制日志之前, 获取所述表格处理事件的时间戳; 根据所述时间戳, 生成所述表格处理事件的二进制日志。

步骤 626: 排序模块根据表格处理事件的时间戳, 将所述表格处理事件的二进制日志插入所述消息队列中。

接下来, 参见图 7, 假设事务中包括 2 个 SQL 操作, 对上述完整的二进制日志生成过程进行举例说明:

应用程序向计算模块发送事务开启请求, 该事务开启请求指示开启的目标事务为事务 A, 该事务 A 包括 2 个操作: 操作 1 和操作 2, 且操作 1 由存储模块 1 执行, 操作 2 由存储模块 2 执行。应用程序向计算模块发送操作 1, 计算模块向存储模块 1 发送操作 1 开启请求, 存储模块 1 接收到该开启请求后, 向计算模块返回开启消息, 计算模块接收到该开启消息后, 向存储模块 1 发送操作 1 的执行指令, 存储模块 1 接收到该执行指令后, 执行操作 1, 操作 1 执行成功后, 向计算模块返回操作 1 执行成功的消息。

计算模块接收到操作 1 执行成功的消息后, 还可以将该操作 1 执行成功的消息返回给应用程序, 进一步的应用程序将向计算模块发送操作 2, 之后计算模块可以向存储模块 2 发送操作 2 的开启请求, 存储模块 2 接收到该开启请求后, 向计算模块返回开启消息, 计算模块接收到该开启消息后, 向存储模块 2 发送操作 2 的执行指令, 存储模块 2 接收到该执行指令后, 执行操作 2, 操作 2 执行成功后, 向计算模块返回操作 2 执行成功的消息。

计算模块接收到操作 2 执行成功的消息, 向应用程序返回事务 A 执行完毕的消息(即操作 2 执行成功的消息), 应用程序接收到该消息后, 向计算模块发送提交指令, 计算模块向时间戳产生器获取全局有序的时间戳, 然后将该时间戳作为事务 A 的提交时间戳发送给存储模块 1 和存储模块 2, 存储模块 1 和存储模块 2 分别根据该提交时间戳生成事务 A 的二进制日志, 并返回成功生成二进制日志的消息。

接下来, 结合附图 8, 对上述排序过程进行完整举例说明:

假设有 4 个存储模块, 分别为存储模块 1、存储模块 2、存储模块 3、存储模块 4, 其分别对应执行模块 1、执行模块 2、执行模块 3、执行模块 4。假设存储模块 1 中生成了二进制日志 1 (时间戳为 0002)、二进制日志 2 (时间戳为 0000), 存储模块 2 中生成了二进制日志 3 (时间戳为 0005)、二进制日志 4 (时间戳为 0001), 存储模块 3 中生成了二进制日志 5 (时间戳为 0006)、二进制日志 6 (时间戳为 0003), 存储模块 4

中生成了二进制日志 7（时间戳为 0007）、二进制日志 8（时间戳为 0004）。执行模块 1、执行模块 2、执行模块 3、执行模块 4 分别生成对应的内存序列，排序模块读取 4 个执行模块的内存序列，得到 8 个二进制日志，将其按照时间戳进行排序，得到二进制日志 2、二进制日志 4、二进制日志 1、二进制日志 6、二进制日志 8、二进制日志 3、二进制日志 5、二进制日志 7，将该 8 个二进制日志按照排好的顺序放入消息队列中，即生成逻辑 binlog。

本说明书中提供的数据处理方法，目标存储模块可以根据目标事务的提交时间戳生成二进制日志，然后由排序模块按照所述目标事务的提交时间戳，对所有二进制日志进行排序，将排序后的二进制日志写入消息队列中。这种情况下，每个二进制日志中都携带一个全局有序的时间戳，消息队列中会包括分布式数据库系统中所有存储模块生成的二进制日志，且都是按时间先后顺序有序排列，下游应用需要订阅分布式数据库中的数据时，可以直接读取该消息队列，使得分布式数据库系统中时间上前后的两次操作生成的二进制日志都是按序排列，从而避免了事务的操作顺序发生错乱，保证了下游应用订阅数据的一致性。

需要说明的是，此实施例中的技术方案与上述的图 2、图 4 和图 5 所示的数据处理方法的技术方案属于同一构思，该数据处理方法的技术方案未详细描述的细节内容，均可以参见上述图 2、图 4 和图 5 所示的数据处理方法的技术方案的描述。

图 9 示出了根据本说明书一实施例提供的一种计算设备 900 的结构框图。该计算设备 900 的部件包括但不限于存储器 910 和处理器 920。处理器 920 与存储器 910 通过总线 930 相连接，数据库 950 用于保存数据。

计算设备 900 还包括接入设备 940，接入设备 940 使得计算设备 900 能够经由一个或多个网络 960 通信。这些网络的示例包括公用交换电话网（PSTN）、局域网（LAN）、广域网（WAN）、个域网（PAN）或诸如因特网的通信网络的组合。接入设备 940 可以包括有线或无线的任何类型的网络接口（例如，网络接口卡（NIC））中的一个或多个，诸如 IEEE802.11 无线局域网（WLAN）无线接口、全球微波互联接入（Wi-MAX）接口、以太网接口、通用串行总线（USB）接口、蜂窝网络接口、蓝牙接口、近场通信（NFC）接口，等等。

在本说明书的一个实施例中，计算设备 900 的上述部件以及图 9 中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图 9 所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

计算设备 900 可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备（例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等）、移动电话（例如，智能手机）、可佩戴的计算设备（例如，智能手表、智能眼镜等）或

其他类型的移动设备，或者诸如台式计算机或 PC 的静止计算设备。计算设备 900 还可以是移动式或静止式的服务器。

其中，处理器 920 用于执行如下计算机可执行指令：

接收事务开启请求；

- 5 根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；
 向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；
 所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

10 图 10 示出了根据本说明书一实施例提供的一种计算设备 1000 的结构框图。该计算设备 1000 的部件包括但不限于存储器 1010 和处理器 1020。处理器 1020 与存储器 1010 通过总线 1030 相连接，数据库 1050 用于保存数据。

 计算设备 1000 还包括接入设备 1040，接入设备 1040 使得计算设备 1000 能够经由一个或多个网络 1060 通信。这些网络的示例包括公用交换电话网（PSTN）、局域网（LAN）、广域网（WAN）、个域网（PAN）或诸如因特网的通信网络的组合。接入设备 15 1040 可以包括有线或无线的任何类型的网络接口（例如，网络接口卡（NIC））中的一个或多个，诸如 IEEE 802.11 无线局域网（WLAN）无线接口、全球微波互联接入（Wi-MAX）接口、以太网接口、通用串行总线（USB）接口、蜂窝网络接口、蓝牙接口、近场通信（NFC）接口，等等。

 在本说明书的一个实施例中，计算设备 1000 的上述部件以及图 10 中未示出的其他 20 部件也可以彼此相连接，例如通过总线。应当理解，图 10 所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

 计算设备 1000 可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备（例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等）、 25 移动电话（例如，智能手机）、可佩戴的计算设备（例如，智能手表、智能眼镜等）或其他类型的移动设备，或者诸如台式计算机或 PC 的静止计算设备。计算设备 1000 还可以是移动式或静止式的服务器。

其中，处理器 1020 用于执行如下计算机可执行指令：

- 30 接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

 所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

 接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

图 11 示出了根据本说明书一实施例提供的一种计算设备 1100 的结构框图。该计算设备 1100 的部件包括但不限于存储器 1110 和处理器 1120。处理器 1120 与存储器 1110 通过总线 1130 相连接，数据库 1150 用于保存数据。

5 计算设备 1100 还包括接入设备 1140，接入设备 1140 使得计算设备 1100 能够经由一个或多个网络 1160 通信。这些网络的示例包括公用交换电话网（PSTN）、局域网（LAN）、广域网（WAN）、个域网（PAN）或诸如因特网的通信网络的组合。接入设备 1140 可以包括有线或无线的任何类型的网络接口（例如，网络接口卡（NIC））中的一个或多个，诸如 IEEE 802.11 无线局域网（WLAN）无线接口、全球微波互联接入（Wi-MAX）接口、以太网接口、通用串行总线（USB）接口、蜂窝网络接口、蓝牙接口、近场通信（NFC）接口，等等。

在本说明书的一个实施例中，计算设备 1100 的上述部件以及图 11 中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图 11 所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

15 计算设备 1100 可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备（例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等）、移动电话（例如，智能手机）、可佩戴的计算设备（例如，智能手表、智能眼镜等）或其他类型的移动设备，或者诸如台式计算机或 PC 的静止计算设备。计算设备 1100 还可以是移动式或静止式的服务器。

20 其中，处理器 1120 用于执行如下计算机可执行指令：

读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

将排序后的二进制日志写入消息队列中。

25 上述为本实施例的三种计算设备的示意性方案。需要说明的是，该计算设备的技术方案与上述的数据处理方法的技术方案属于同一构思，计算设备的技术方案未详细描述的细节内容，均可以参见上述数据处理方法的技术方案的描述。

本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时以用于：

30 接收事务开启请求；

根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；

向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；

所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时以用于：

接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

5 所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

10 本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时以用于：

读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

将排序后的二进制日志写入消息队列中。

15 上述为本实施例的三种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的数据处理方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述数据处理方法的技术方案的描述。

20 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

25 所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括电载波信号和电信信号。

30 需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本说明书并不受所描述的动作顺序的限制，因为依据本说明书，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一

定都是本说明书所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其它实施例的相关描述。

5 以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本说明书的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本说明书。本说明书仅受权利要求书及其全部范围和等效物的限制。

权利要求书

1.一种分布式数据库系统，包括计算模块和存储模块；

所述计算模块，被配置为接收事务开启请求；根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳；

所述目标存储模块，被配置为接收所述执行指令，根据所述执行指令执行所述目标事务；所述目标事务执行成功的情况下，向所述计算模块返回目标事务执行成功的消息；接收所述提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志。

2.根据权利要求1所述的分布式数据库系统，所述系统还包括：多个执行模块和一个排序模块，所述多个执行模块中每一个执行模块均对应存储模块；

所述执行模块，被配置为读取自身对应的存储模块的二进制日志，并根据所述二进制日志生成内存序列；

所述排序模块，被配置为读取每一个所述执行模块的内存序列，获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；将排序后的二进制日志写入消息队列中。

3.根据权利要求2所述的分布式数据库系统，所述计算模块进一步被配置为：

若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；

根据所述执行逻辑，生成所述表格处理事件的时间戳；

根据所述时间戳，生成所述表格处理事件的二进制日志。

4.根据权利要求3所述的分布式数据库系统，所述计算模块进一步被配置为：

确定所述表格处理事件的操作类型；

根据所述操作类型，确定所述表格处理事件对应的兼容类型；

根据所述兼容类型，确定所述表格处理事件的执行逻辑。

5.根据权利要求4所述的分布式数据库系统，所述计算模块进一步被配置为：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所述表格处理事件执行后的版本；

若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版本兼容第一版本。

6.根据权利要求5所述的分布式数据库系统，所述计算模块进一步被配置为：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处理事件的执行逻辑为后序执行；

若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处理事件的执行逻辑为前序执行。

7.根据权利要求6所述的分布式数据库系统，所述计算模块进一步被配置为：

5 若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的二进制日志之前，获取所述表格处理事件的时间戳；

若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表格处理事件的时间戳。

8.根据权利要求3所述的分布式数据库系统，所述排序模块进一步被配置为：

10 根据所述表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

9.根据权利要求1所述的分布式数据库系统，所述系统还包括时间戳产生器，所述计算模块进一步被配置为：

向所述时间戳产生器申请全局有序的时间戳；

将所述时间戳确定为所述目标事务的提交时间戳。

15 10.根据权利要求1所述的分布式数据库系统，所述事务开启请求中携带事务标识及对应的存储模块标识，所述计算模块进一步被配置为：

将所述事务标识指示的事务确定为目标事务；

将所述存储模块标识指示的存储模块确定为对应的目标存储模块。

20 11.根据权利要求1所述的分布式数据库系统，所述事务开启请求中携带事务标识，所述计算模块进一步被配置为：

将所述事务标识指示的事务确定为目标事务；

根据预设分配规则，确定所述目标事务对应的目标存储模块。

12.根据权利要求2所述的分布式数据库系统，所述执行模块进一步被配置为：

25 若检测到自身对应的存储模块生成新的二进制日志，则返回执行所述读取所述存储模块的二进制日志的操作步骤；

相应的，所述排序模块进一步被配置为：

返回执行所述读取每一个所述执行模块的内存序列的操作步骤；或者，

获取所述执行模块中增加的二进制日志，将所述二进制日志放入所述消息队列的队尾。

30 13.根据权利要求11所述的分布式数据库系统，所述计算模块进一步被配置为：记录所述目标存储模块的存储模块标识。

14.根据权利要求1所述的分布式数据库系统，所述计算模块进一步被配置为：

向所述目标存储模块发送目标事务开启请求；

35 所述目标存储模块返回开启消息的情况下，执行所述向所述目标存储模块发送执行指令的操作步骤。

15.根据权利要求 1 所述的分布式数据库系统，所述存储模块进一步被配置为：
生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释
中。

16.一种数据处理方法，应用于分布式数据库系统中的计算模块，所述方法包括：
5 接收事务开启请求；
根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；
向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；
所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发
送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

10 17.根据权利要求 16 所述的数据处理方法，所述方法还包括：
若检测到表格处理事件，则确定所述表格处理事件的执行逻辑；
根据所述执行逻辑，生成所述表格处理事件的时间戳；
根据所述时间戳，生成所述表格处理事件的二进制日志。

18.根据权利要求 17 所述的数据处理方法，所述确定所述表格处理事件的执行逻辑，
15 包括：

确定所述表格处理事件的操作类型；
根据所述操作类型，确定所述表格处理事件对应的兼容类型；
根据所述兼容类型，确定所述表格处理事件的执行逻辑。

19.根据权利要求 18 所述的数据处理方法，所述根据所述操作类型，确定所述表格
20 处理事件对应的兼容类型，包括：

若所述操作类型属于第一类型，则确定所述表格处理事件对应的兼容类型为第一版
本兼容第二版本，所述第一版本为所述表格处理事件执行前的版本，所述第二版本为所
述表格处理事件执行后的版本；

若所述操作类型属于第二类型，则确定所述表格处理事件对应的兼容类型为第二版
25 本兼容第一版本。

20.根据权利要求 19 所述的数据处理方法，所述根据所述兼容类型，确定所述表格
处理事件的执行逻辑，包括：

若所述表格处理事件对应的兼容类型为第一版本兼容第二版本，则确定所述表格处
理事件的执行逻辑为后序执行；

30 若所述表格处理事件对应的兼容类型为第二版本兼容第一版本，则确定所述表格处
理事件的执行逻辑为前序执行。

21.根据权利要求 20 所述的数据处理方法，所述根据所述执行逻辑，生成所述表格
处理事件的时间戳，包括：

若所述表格处理事件的执行逻辑为前序执行，则在生成所有第二版本的数据对应的
35 二进制日志之前，获取所述表格处理事件的时间戳；

若所述表格处理事件的执行逻辑为后序执行，则在生成所有第一版本的数据对应的二进制日志之后，获取所述表格处理事件的时间戳。

22.根据权利要求 16 所述的数据处理方法，所述向所述目标存储模块发送提交指令之前，还包括：

- 5 向时间戳产生器申请全局有序的时间戳；
 将所述时间戳确定为所述目标事务的提交时间戳。

23.根据权利要求 16 所述的数据处理方法，所述事务开启请求中携带事务标识及对应的存储模块标识，所述根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块，包括：

- 10 将所述事务标识指示的事务确定为目标事务；
 将所述存储模块标识指示的存储模块确定为对应的目标存储模块。

24.根据权利要求 16 所述的数据处理方法，所述事务开启请求中携带事务标识，所述根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块包括：

- 将所述事务标识指示的事务确定为目标事务；
15 根据预设分配规则，确定所述目标事务对应的目标存储模块。

25.根据权利要求 24 所述的数据处理方法，所述向所述目标存储模块发送执行指令之后，还包括：

 记录所述目标存储模块的存储模块标识。

- 26.根据权利要求 16 所述的数据处理方法，所述根据所述事务开启请求，确定目标
20 事务及所述目标事务对应的目标存储模块之后，且所述向所述目标存储模块发送执行指令之前，还包括：

 向所述目标存储模块发送目标事务开启请求；

 所述目标存储模块返回开启消息的情况下，执行所述向所述目标存储模块发送执行指令的操作步骤。

- 25 27.一种数据处理方法，应用于分布式数据库系统中的存储模块，所述方法包括：
 接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

 所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

- 接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务
30 的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

28.根据权利要求 27 所述的数据处理方法，所述根据所述目标事务的提交时间戳生成二进制日志，包括：

- 35 生成二进制日志，并将所述目标事务的提交时间戳记录在所述二进制日志的注释中。

29.一种数据处理方法，应用于分布式数据库系统中的排序模块，所述方法包括：
读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

5 将排序后的二进制日志写入消息队列中。

30.根据权利要求 29 所述的数据处理方法，所述将排序后的二进制日志写入消息队列中之后，还包括：

若检测到执行模块的内存序列发生变化，则返回执行所述读取每一个所述执行模块的内存序列的操作步骤。

10 31.根据权利要求 29 所述的数据处理方法，所述将排序后的二进制日志写入消息队列中之后，还包括：

若检测到执行模块的内存序列发生变化，则获取所述执行模块中增加的二进制日志，将所述二进制日志放入所述消息队列的队尾。

15 32.根据权利要求 29 所述的数据处理方法，所述将排序后的二进制日志写入消息队列中之后，还包括：

根据表格处理事件的时间戳，将所述表格处理事件的二进制日志插入所述消息队列中。

33.一种计算设备，包括：

存储器和处理器；

20 所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令：

接收事务开启请求；

根据所述事务开启请求，确定目标事务及所述目标事务对应的目标存储模块；

向所述目标存储模块发送执行指令，所述执行指令中携带有待执行目标事务；

25 所述目标存储模块返回目标事务执行成功的消息的情况下，向所述目标存储模块发送提交指令，所述提交指令携带有待提交目标事务和所述目标事务的提交时间戳。

34.一种计算设备，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指

30 令：

接收到执行指令后，根据所述执行指令执行目标事务，所述执行指令中携带有待执行目标事务；

所述目标事务执行成功的情况下，向计算模块返回目标事务执行成功的消息；

35 接收提交指令，根据所述提交指令对所述目标事务进行提交，并根据所述目标事务的提交时间戳生成二进制日志，所述提交指令携带有待提交目标事务和所述目标事务的

提交时间戳。

35.一种计算设备，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指

5 令：

读取每一个执行模块的内存序列；

获取所述内存序列中每一个二进制日志中目标事务的提交时间戳；

按照所述目标事务的提交时间戳，对所述内存序列中的二进制日志进行排序；

将排序后的二进制日志写入消息队列中。

10 36.一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现权利要求 16 至 26 任意一项所述数据处理方法的步骤。

37.一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现权利要求 27 至 28 任意一项所述数据处理方法的步骤。

15 38.一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现权利要求 29 至 32 任意一项所述数据处理方法的步骤。

20

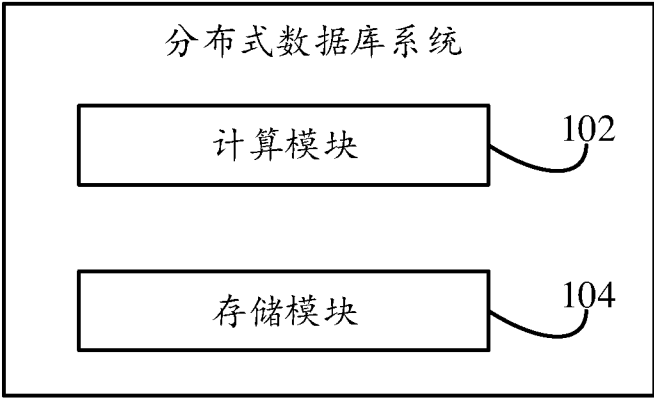


图 1

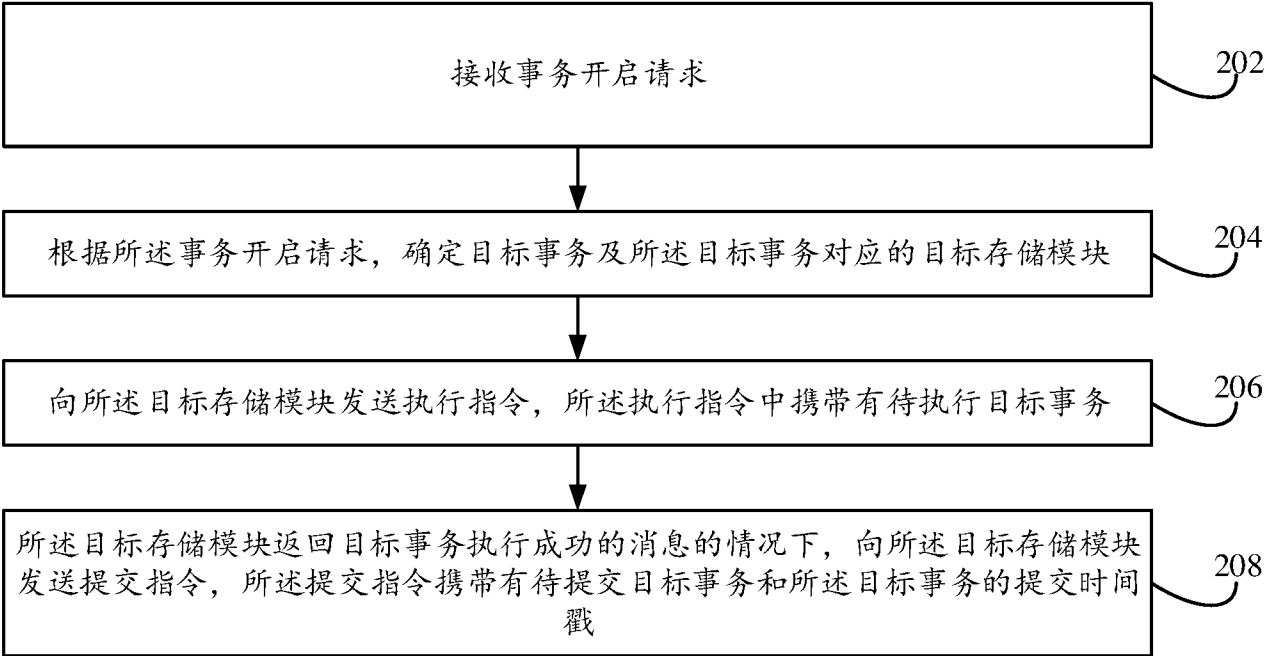


图 2

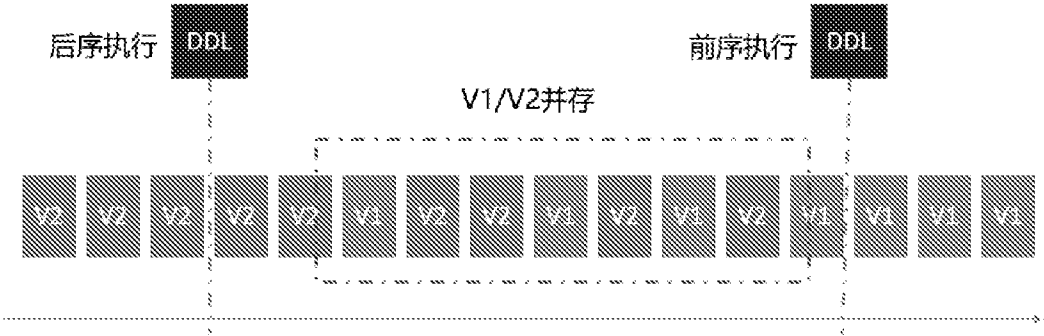


图 3

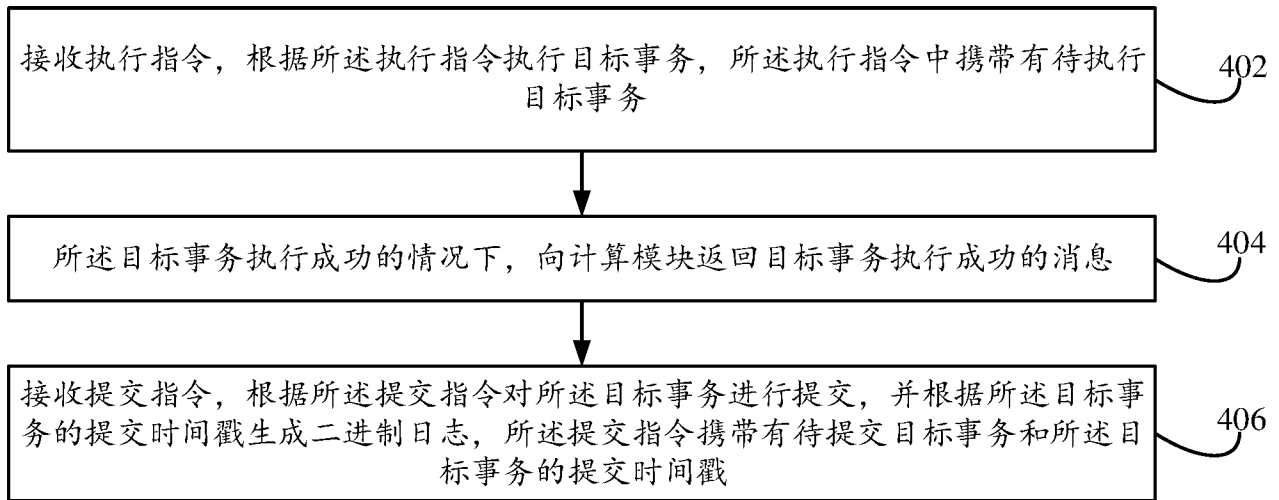


图 4

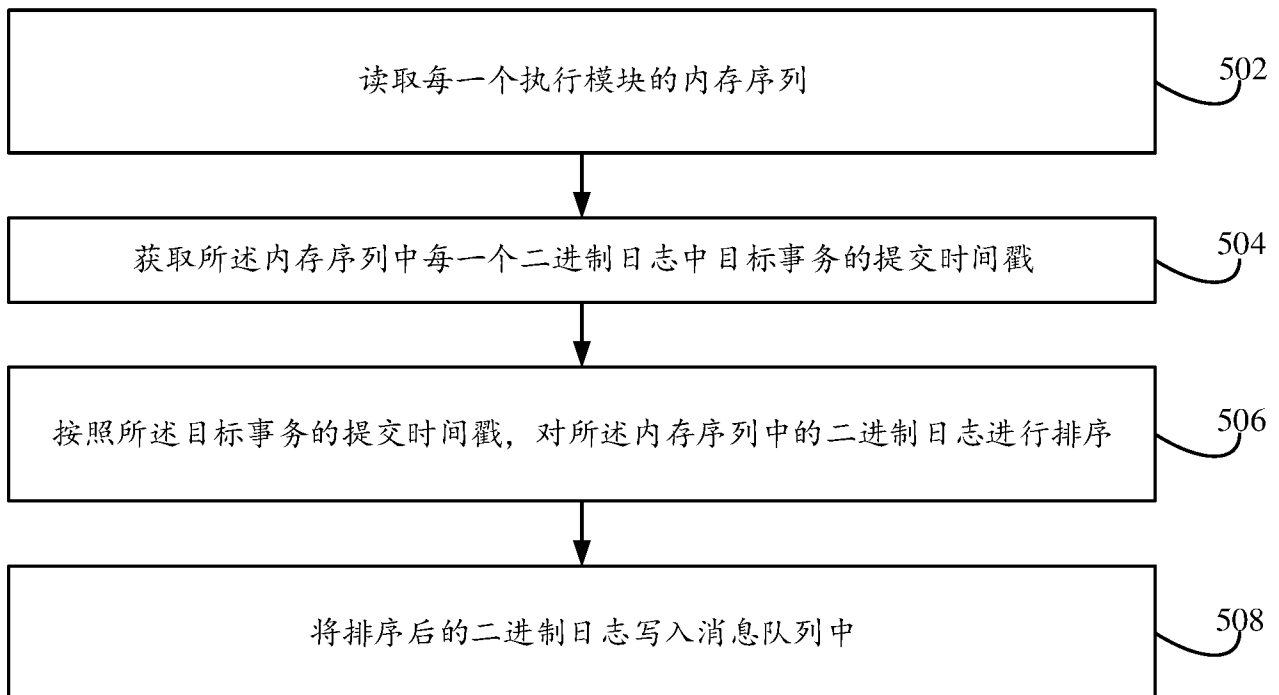


图 5

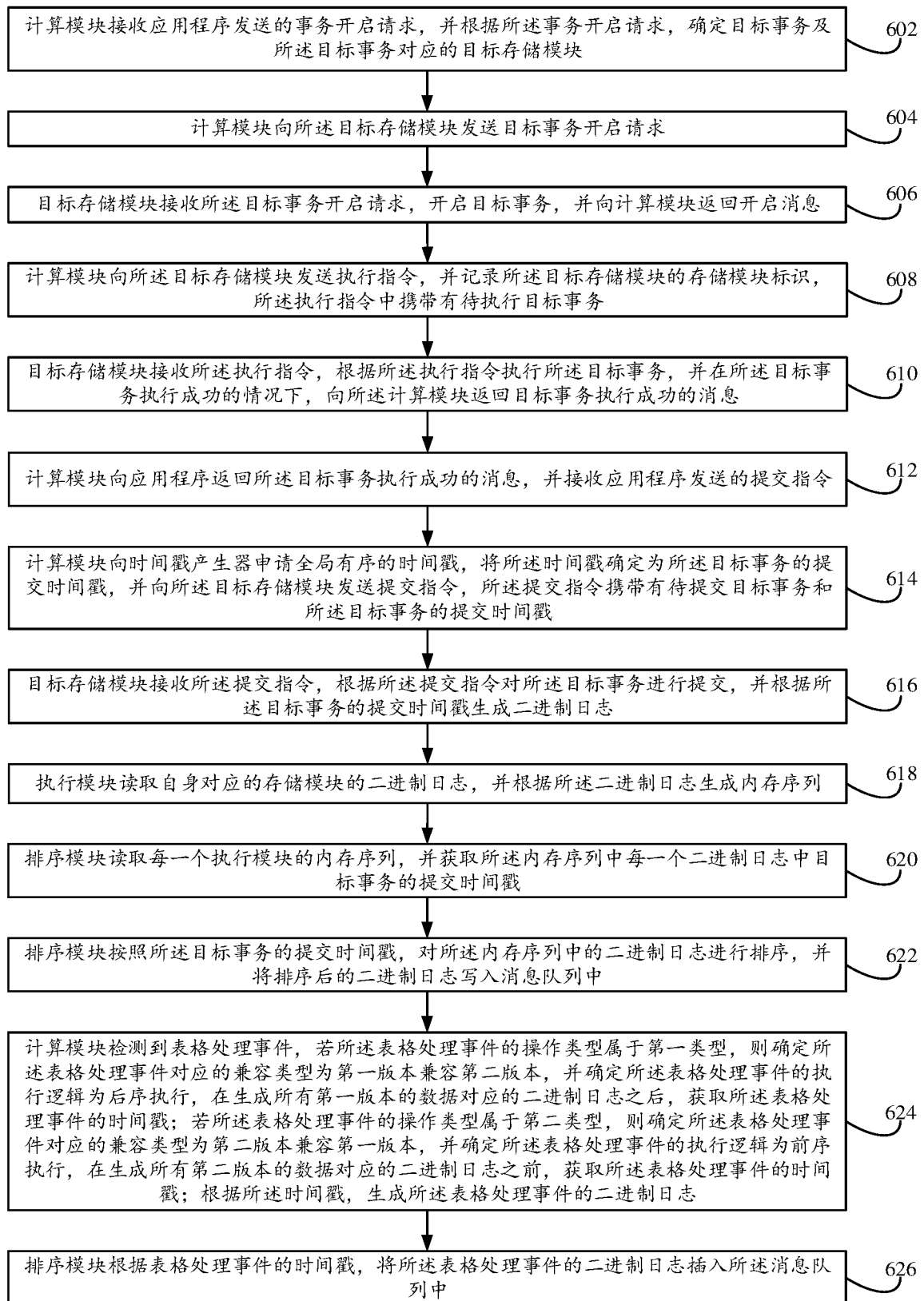


图 6

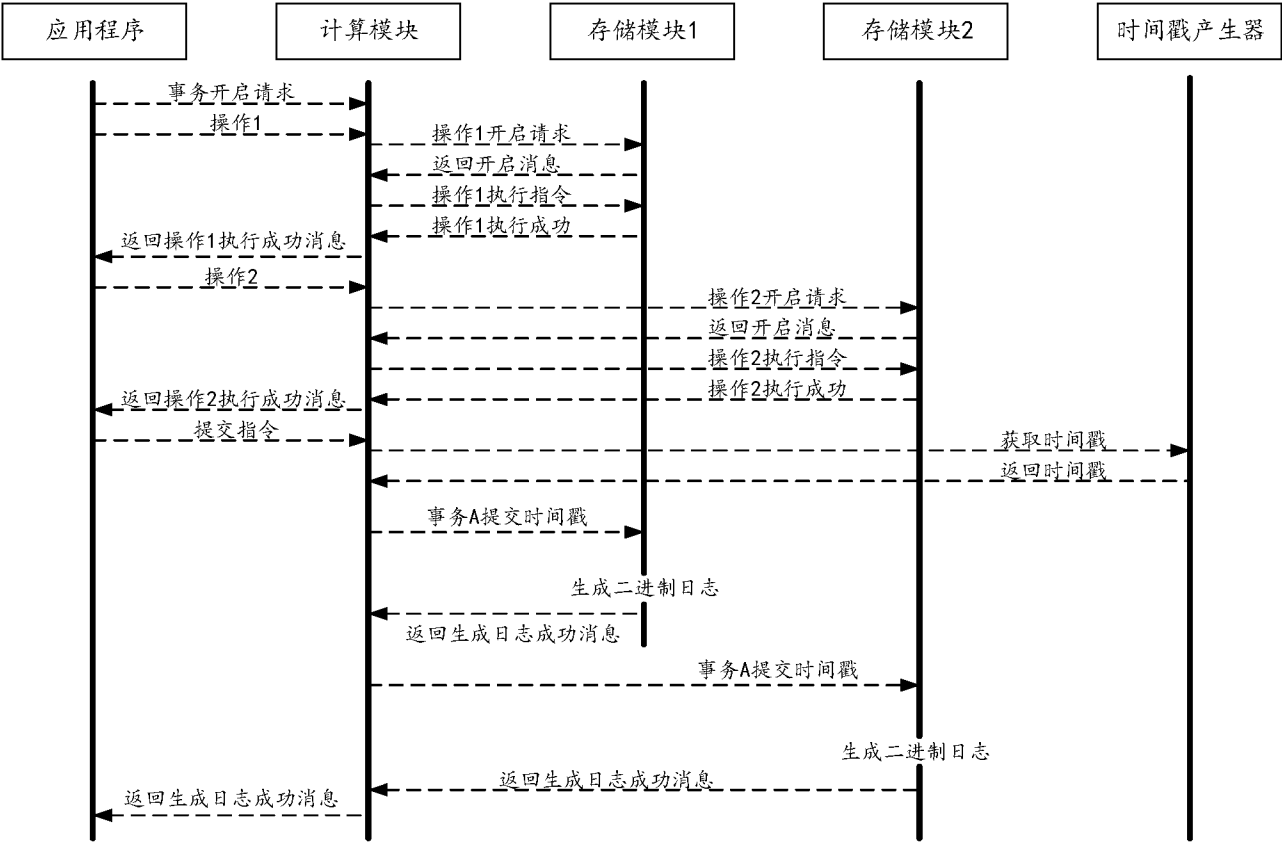


图 7

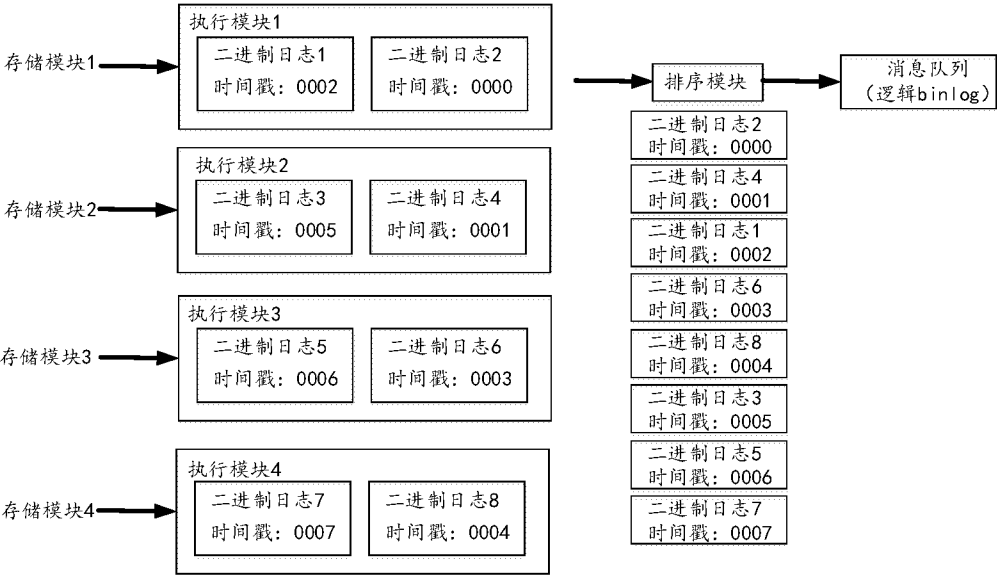


图 8

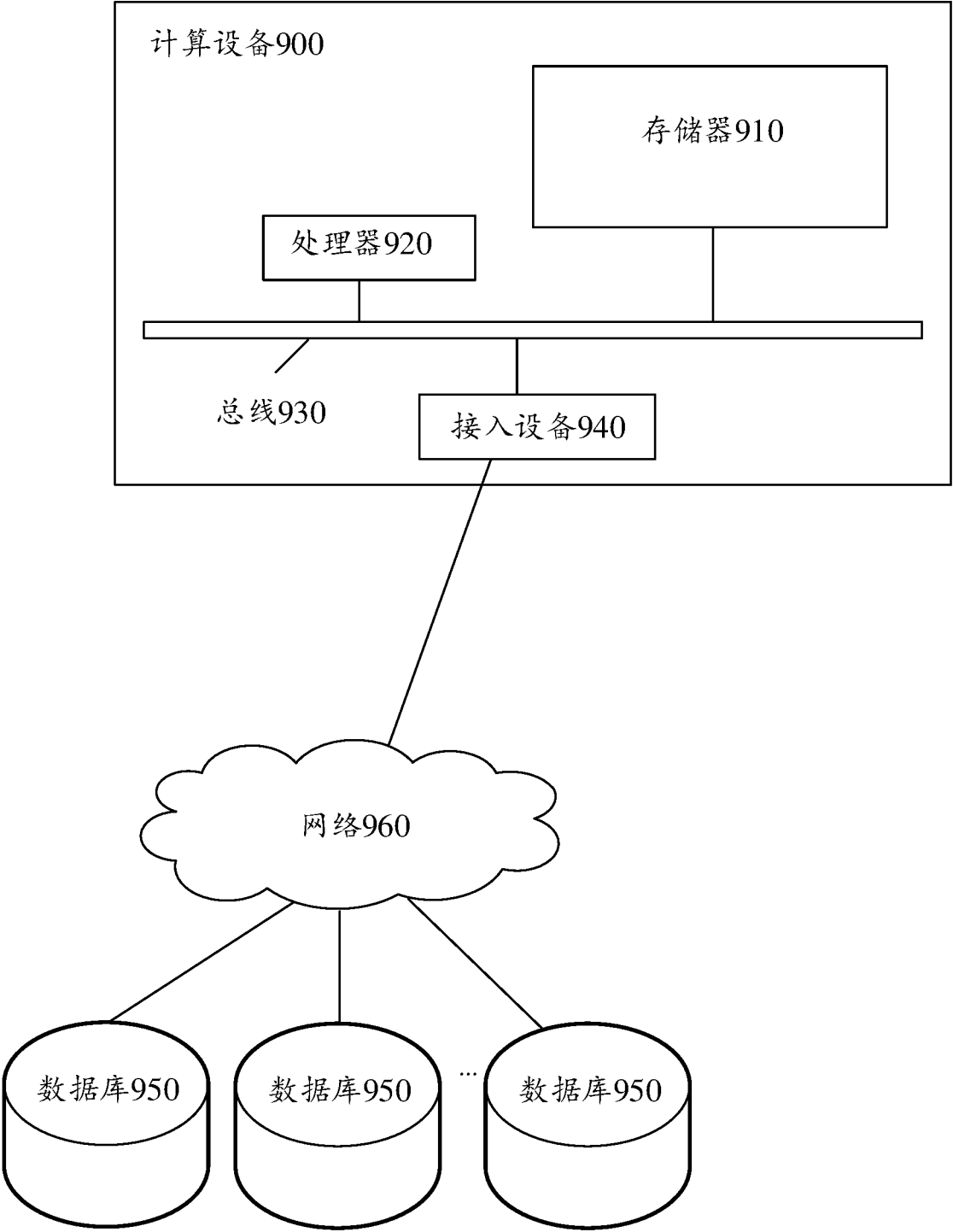


图 9

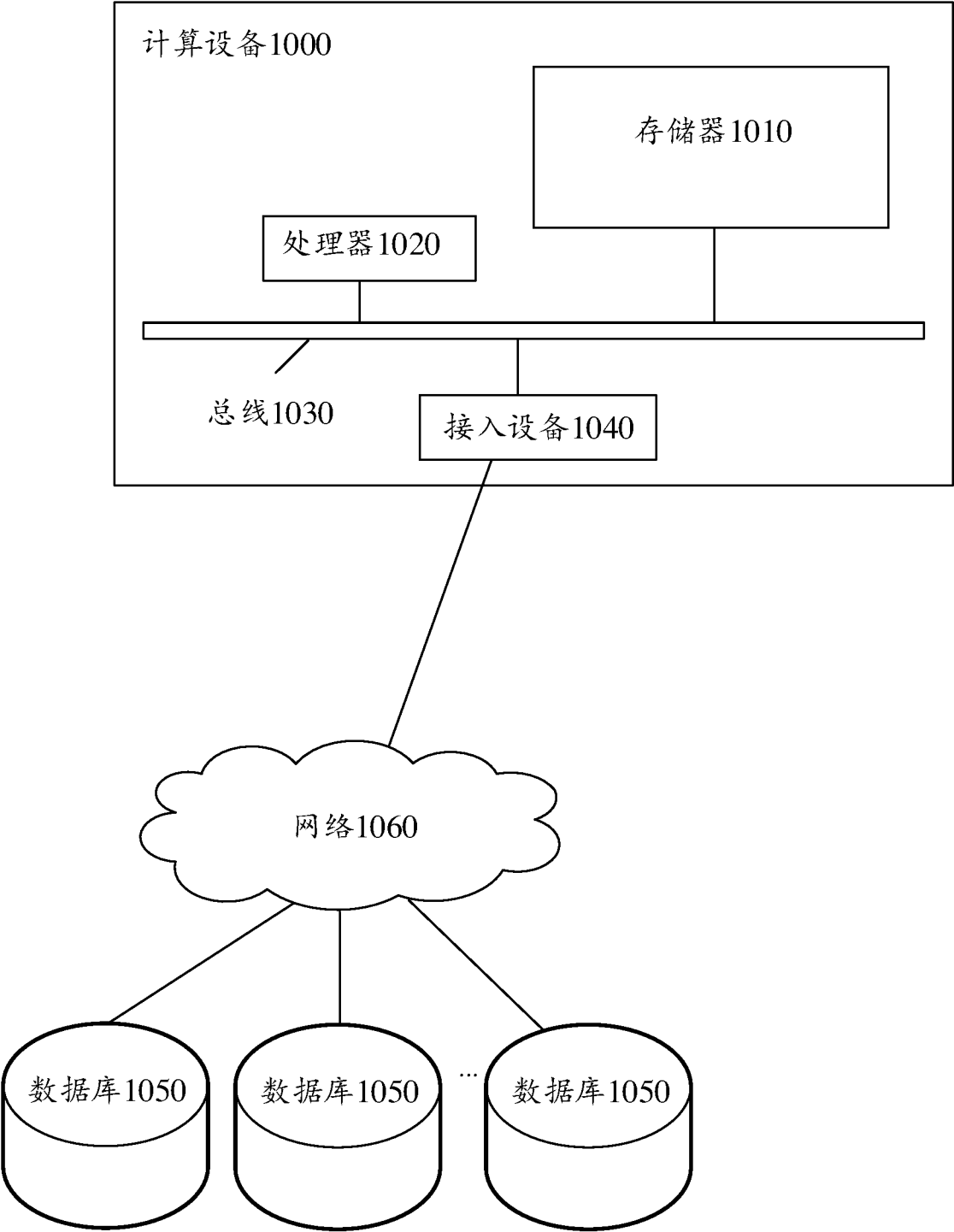


图 10

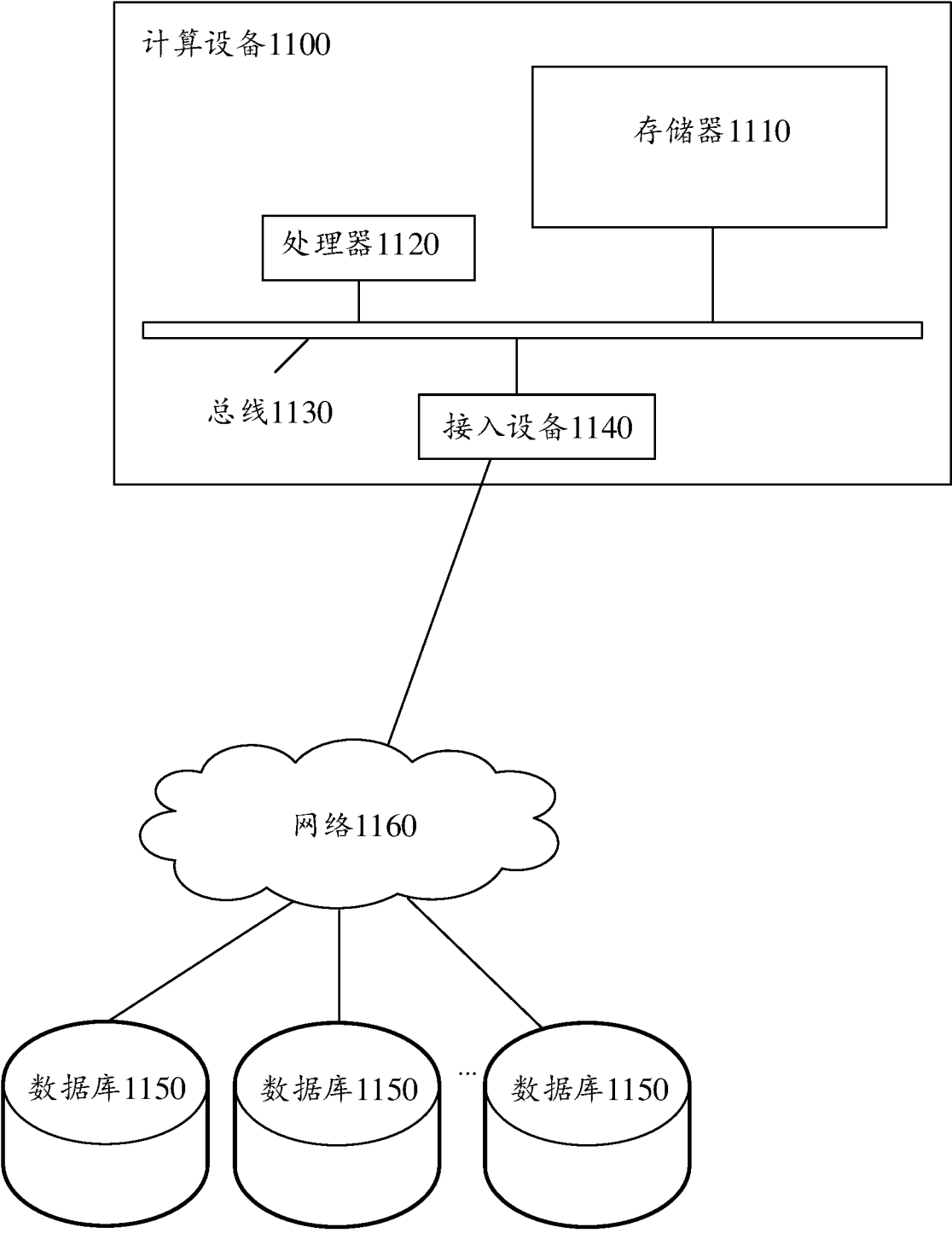


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/107243

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/27(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)

CNPAT, WPI, EPODOC, CNKI, IEEE: 存储, 数据库, 事务, 时间戳, 执行, 提交, 目标, memory, storage, database, databank, business, affair, timestamp, execution, carry out, enforce, implement, refer, give in, put in, aim, goal, target

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110457157 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 15 November 2019 (2019-11-15) claims 1-15, description paragraphs [0078]-[0201]	1-38
A	CN 106991113 A (SAP SE) 28 July 2017 (2017-07-28) entire document	1-38
A	CN 110019112 A (ALIBABA GROUP HOLDING LIMITED) 16 July 2019 (2019-07-16) entire document	1-38
A	CN 110018884 A (ALIBABA GROUP HOLDING LIMITED) 16 July 2019 (2019-07-16) entire document	1-38
A	WO 2016064575 A1 (MICROSOFT TECHNOLOGY LICENSING, L.L.C.) 28 April 2016 (2016-04-28) entire document	1-38



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

“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

08 October 2021

Date of mailing of the international search report

18 October 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/107243

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110457157	A	15 November 2019	CN	111078451	A	28 April 2020
CN	106991113	A	28 July 2017	US	2017177658	A1	22 June 2017
				US	2020394182	A1	17 December 2020
				EP	3182300	A1	21 June 2017
CN	110019112	A	16 July 2019	WO	2019037617	A1	28 February 2019
				EP	3674909	A1	01 July 2020
				US	2020183892	A1	11 June 2020
CN	110018884	A	16 July 2019	None			
WO	2016064575	A1	28 April 2016	US	2016110403	A1	21 April 2016
				CN	107077495	A	18 August 2017
				EP	3207471	A1	23 August 2017

国际检索报告

国际申请号

PCT/CN2021/107243

A. 主题的分类

G06F 16/27 (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))

CNPAT, WPI, EPDOC, CNKI, IEEE: 存储, 数据库, 事务, 时间戳, 执行, 提交, 目标, memory, storage, database, databank, business, affair, timestamp, execution, carry out, enforce, implement, refer, give in, put in, aim, goal, target

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
A	CN 110457157 A (腾讯科技深圳有限公司) 2019年 11月 15日 (2019 - 11 - 15) 权利要求1-15, 说明书第[0078]-[0201]段	1-38
A	CN 106991113 A (SAP欧洲公司) 2017年 7月 28日 (2017 - 07 - 28) 全文	1-38
A	CN 110019112 A (阿里巴巴集团控股有限公司) 2019年 7月 16日 (2019 - 07 - 16) 全文	1-38
A	CN 110018884 A (阿里巴巴集团控股有限公司) 2019年 7月 16日 (2019 - 07 - 16) 全文	1-38
A	WO 2016064575 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 2016年 4月 28日 (2016 - 04 - 28) 全文	1-38

☐ 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

“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

2021年 10月 8日

国际检索报告邮寄日期

2021年 10月 18日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国 北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

胡贝贝

电话号码 86-(10)-53961415

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/107243

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	110457157	A	2019年 11月 15日	CN	111078451	A	2020年 4月 28日		
CN	106991113	A	2017年 7月 28日	US	2017177658	A1	2017年 6月 22日		
				US	2020394182	A1	2020年 12月 17日		
				EP	3182300	A1	2017年 6月 21日		
CN	110019112	A	2019年 7月 16日	WO	2019037617	A1	2019年 2月 28日		
				EP	3674909	A1	2020年 7月 1日		
				US	2020183892	A1	2020年 6月 11日		
CN	110018884	A	2019年 7月 16日	无					
WO	2016064575	A1	2016年 4月 28日	US	2016110403	A1	2016年 4月 21日		
				CN	107077495	A	2017年 8月 18日		
				EP	3207471	A1	2017年 8月 23日		