(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109408589 A
(43)申请公布日 2019.03.01

(21)申请号 201811074438.9

(22)申请日 2018.09.14

(71)申请人 新华三大数据技术有限公司
地址 450000 河南省郑州市郑州高新技术产业开发区杜英街166号总部大观B18号楼

(72)发明人 李日光 丁远普

(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 吴迪

(51)Int.Cl.
G06F 16/27(2019.01)

权利要求书2页 说明书15页 附图5页

(54)发明名称

数据同步方法及装置

(57)摘要

本公开提供一种数据同步方法及装置,用于将源数据库中的数据同步到MPP数据库,方法包括:从源数据库的日志文件获得预设数量条操作语句,每条操作语句包括该操作语句的版本信息;查找预设数量条操作语句中的删除语句,并在查找到删除语句时,在预设数量条操作语句中确定版本信息比该删除语句旧的目标操作语句,并在目标操作语句中查找符合该删除语句的删除条件的插入语句作为待删除语句;从预设数量条操作语句中删除该待删除语句,对剩余的操作语句中的插入语句进行合并,并根据合并后的操作语句更新MPP数据库的数据。如此,可以减少更新数据时执行的操作次数,从而降低更新数据的时延,从而提高源数据库和MPP数据库中的数据的一致性。

CN 109408589 A

1.一种数据同步方法,其特征在于,所述方法包括:

从源数据库的日志文件中获得预设数量条操作语句,其中,每条操作语句包括该操作语句的版本信息;

查询所述预设数量条操作语句中的删除语句,并在查找到删除语句时,在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句,并在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句;

从所述预设数量条操作语句中删除所述待删除语句,对剩余的操作语句中的插入语句进行合并,并根据合并后的操作语句更新MPP数据库的数据,其中,所述合并后的操作语句至少包括所述删除语句及合并后的插入语句。

2.根据权利要求1所述的方法,其特征在于,查询所述预设数量条操作语句中的删除语句,并在查找到删除语句时,在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句,并在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句,包括:

按照版本信息从旧到新的顺序遍历所述预设数量条操作语句;

在当前遍历到的操作语句为删除语句时,在所述预设数量条操作语句中将版本信息比当前遍历到的删除语句旧的操作语句确定为所述目标操作语句,或者在所述预设数量条操作语句中将版本信息处于所述当前遍历到的删除语句和上一次遍历到的删除语句之间的操作语句确定为所述目标操作语句;

在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句。

3.根据权利要求2所述的方法,其特征在于,在查询所述预设数量条操作语句中的删除语句之前,所述方法还包括:

针对所述预设数量条操作语句中的每条插入语句,确定该插入语句操作的目标数据表,并将该目标数据表包括的字段确定为目标字段;

判断是否存在该插入语句不包括的目标字段;

若是,则在该插入语句中增加所述不包括的目标字段,并将增加的目标字段的插入值设置为空。

4.根据权利要求3所述的方法,其特征在于,对剩余的操作语句中的插入语句进行合并,包括:在所述剩余的操作语句中将操作的目标数据表相同的插入语句合并为一条插入语句;或者,

在所述剩余的操作语句中,将版本信息处于所述当前遍历到的删除语句和所述上一次遍历到的删除语句之间的、且操作的目标数据表相同的插入语句合并为一条插入语句。

5.根据权利要求1-4中任一项所述的方法,其特征在于,从所述源数据库的日志文件中获得预设数量条操作语句,包括:

通过日志分析工具解析所述日志文件,获得所述源数据库的操作语句;

在所述源数据库的操作语句中,从上一次截取的版本信息最新的操作语句开始截取所述预设数量条操作语句。

6.一种数据同步装置,其特征在于,所述装置包括:

操作语句获得模块,用于从源数据库的日志文件中获得预设数量条操作语句,其中,每

条操作语句包括该操作语句的版本信息；

查找模块，用于查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句；

删除模块，用于从所述预设数量条操作语句中删除所述待删除语句；

合并模块，用于对剩余的操作语句中的插入语句进行合并；以及

更新模块，用于根据合并后的操作语句更新MPP数据库的数据，其中，所述合并后的操作语句至少包括所述删除语句及所述合并模块合并后的插入语句。

7.根据权利要求6所述的装置，其特征在于，所述查找模块包括：

遍历子模块，用于按照版本信息从旧到新的顺序遍历所述预设数量条操作语句；

第一查找子模块，用于在当前遍历到的操作语句为删除语句时，在所述预设数量条操作语句中将版本信息比当前遍历到的删除语句旧的操作语句确定为所述目标操作语句，或者在所述预设数量条操作语句中将版本信息处于所述当前遍历到的删除语句和上一次遍历到的删除语句之间的操作语句确定为所述目标操作语句；

第二查找子模块，用于在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句。

8.根据权利要求7所述的装置，其特征在于，所述装置还包括：

格式处理模块，用于针对所述预设数量条操作语句中的每条插入语句，确定该插入语句操作的目标数据表，并将该目标数据表包括的字段确定为目标字段；判断是否存在该插入语句不包括的目标字段，若是，则在该插入语句中增加所述不包括的目标字段，并将增加的目标字段的插入值设置为空。

9.根据权利要求8所述的装置，其特征在于，所述合并模块具体用于：在所述剩余的操作语句中将操作的目标数据表相同的插入语句合并为一条插入语句；或者，在所述剩余的操作语句中，将版本信息处于所述当前遍历到的删除语句和所述上一次遍历到的删除语句之间的、且操作的目标数据表相同的插入语句合并为一条插入语句。

10.根据权利要求6-9中任一项所述的装置，其特征在于，所述操作语句获得模块具体用于通过日志分析工具解析所述日志文件，获得所述源数据库的操作语句，并在所述源数据库的操作语句中，从上一次截取的版本信息最新的操作语句开始截取所述预设数量条操作语句。

数据同步方法及装置

技术领域

[0001] 本公开涉及数据库技术领域，具体而言，涉及一种数据同步方法及装置。

背景技术

[0002] 在实际应用中，有时需要将一个数据库(通常称作“源数据库”)中的数据同步到另一数据库(通常称作“目标数据库”)中。目前，主要通过OracleCDC(ChangeDataCapture，改变数据捕获)技术来实现。

[0003] OracleCDC技术主要包括以下两种实现方式：第一，同步方式，在源数据库中针对插入、修改、删除等操作分别设置触发器，一旦源数据库中的数据发生上述变化时，即可触发对应的触发器将发生变化的数据写入到临时表中，再通过ETL(Extract-Transform-Load，抽取-交互转换-加载)工具将临时表中的数据更新到目标数据库中。第二，异步方式，从源数据库中的日志文件中解析出操作语句，再在目标数据库中重复执行该操作语句，从而将源数据库中的数据同步到目标数据库中。

[0004] 经研究发现，上述两种方式对于源数据库和目标数据库均为传统数据库(诸如，Oracle、SQLserver、MySQL等)的情形适用，但是当目标数据库为分布式并行(Massively Parallel Processor，MPP)数据库时，采用上述两种方式进行数据同步的过程中，均存在严重的时延，导致源数据库和目标数据库中的数据不一致。

发明内容

[0005] 有鉴于此，本公开的目的在于提供一种数据同步方法及装置，以至少部分地改善上述问题。

[0006] 为了达到上述目的，本公开采用如下技术方案：

[0007] 第一方面，本公开提供一种数据同步方法，所述方法包括：

[0008] 从源数据库的日志文件中获得预设数量条操作语句，其中，每条操作语句包括该操作语句的版本信息；

[0009] 查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，并在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句；

[0010] 从所述预设数量条操作语句中删除所述待删除语句，对剩余的操作语句进行合并，并根据合并后的操作语句更新MPP数据库的数据。

[0011] 第二方面，本公开提供一种数据同步装置，所述装置包括：

[0012] 操作语句获得模块，用于从源数据库的日志文件中获得预设数量条操作语句，其中，每条操作语句包括该操作语句的版本信息；

[0013] 查找模块，用于查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句；

[0014] 删除模块，用于从所述预设数量条操作语句中删除所述待删除语句；

[0015] 合并模块，用于对剩余的操作语句进行合并；以及

[0016] 更新模块，用于根据合并后的操作语句更新MPP数据库的数据。

[0017] 相对于现有技术而言，本公开具有以下有益效果：

[0018] 本公开提供的一种数据同步方法及装置，从源数据库的日志文件中获得预设数量条操作语句，每条操作语句包括该操作语句的版本信息。查找预设数量条操作语句中的删除语句，并在查找到删除语句时，在预设数量条操作语句中确定版本信息比该删除语句旧的目标操作语句，并在目标操作语句中查找符合该删除语句的删除条件的插入语句作为待删除语句。从预设数量条操作语句中删除该待删除语句，对剩余的操作语句进行合并，并根据合并后的操作语句更新MPP数据库的数据。如此，可以减少更新数据时执行的操作次数，降低更新数据的时延，从而提高源数据库和MPP数据库中的数据的一致性。

附图说明

[0019] 为了更清楚地说明本公开的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本公开的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

[0020] 图1为本公开提供的一种应用场景示意图；

[0021] 图2为本公开提供的一种数据同步方法的流程示意图；

[0022] 图3为图2所示步骤S21的子步骤示意图；

[0023] 图4为图2所示步骤S22的子步骤示意图；

[0024] 图5为本公开提供的数据同步方法的又一流程示意图；

[0025] 图6为本公开提供的一种电子设备的方框示意图；

[0026] 图7为本公开提供的一种数据同步装置的功能模块框图。

[0027] 图标：10-电子设备；11-机器可读存储介质；110-数据同步装置；111-操作语句获得模块；112-查找模块；1121-遍历子模块；1122-第一查找子模块；1123-第二查找子模块；113-删除模块；114-更新模块；115-合并模块；116-格式处理模块；12-处理器；13-通信单元；20-源数据库；30-目标数据库。

具体实施方式

[0028] 为使本公开的目的、技术方案和优点更加清楚，下面将结合本公开中的附图，对本公开中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本公开一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本公开的组件可以以各种不同的配置来布置和设计。

[0029] 因此，以下对在附图中提供的本公开的实施例的详细描述并非旨在限制要求保护的本公开的范围，而是仅仅表示本公开的选定实施例。基于本公开中的实施例，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，都属于本公开保护的范围。

[0030] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一

个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

[0031] 如图1所示，是本公开提供的一种应用场景示意图，其中，电子设备10可以通过网络访问源数据库20和目标数据库30，其中，目标数据库30为MPP数据库，具体可以是Greenplum、GBase等，源数据库20可以是Oracle、SQLserver、MySQL等传统数据库，本实施例对此不做限制。

[0032] 在本公开中，电子设备10可以是任意能够访问并操作数据库的设备，例如可以为服务器。

[0033] 如图2所示，是本公开提供的一种数据同步方法的流程示意图，该数据同步方法用于将源数据库20中的数据同步到目标数据库30，该方法可以由图1中示出的电子设备10执行，下面将以源数据库20是Oracle、目标数据库是MPP数据库为例，对该方法包括的步骤做详细阐述。

[0034] 步骤S21，从源数据库20的日志文件中获得预设数量条操作语句，其中，每条操作语句包括该操作语句的版本信息。

[0035] 其中，所述日志文件可以是源数据库20的重做日志或归档日志，其中包括用户对源数据库20的操作语句，所述操作语句具体可以是插入(INSERT)语句、删除语句(DELETE)、更新(UPDATE)语句等。

[0036] 在本公开中，每条操作语句包括用于唯一地表示该操作语句的版本信息，所述版本信息可以是操作序列号(SystemChangeNumber，SCN)，所述日志文件中的各操作语句的操作序列号通常是按照先后顺序依次递增，换言之，时序在前的操作语句的操作序列号小于时序在后的操作语句的操作序列号。

[0037] 在本公开中，步骤S21可以包括如图3所示的子步骤。

[0038] 步骤S31，通过日志分析工具解析所述日志文件，获得所述源数据库20的操作语句。

[0039] 在本公开中，通过日志分析工具扫描所述日志文件，可以从中抽取出曾对所述源数据20库执行的操作语句。其中，所述日志分析工具可以是OracleLogminer。OracleLogminer从所述日志文件中解析出的每条操作语句通常为一条SQL语句，例如：

[0040] 插入语句可以是：insert into“logminer”.“emp”(“empno”,“job”)values(“3333”,“clerk”)；

[0041] 删除语句可以是：deletefrom“emp”where“empno”＝“3333”；

[0042] 更新语句可以是：update“emp”set job＝“manager”where“empno”＝3333。

[0043] 步骤S32，在所述源数据库20的操作语句中，从上一次截取的版本信息最新的操作语句开始截取所述预设数量条操作语句。

[0044] 在一种实施方式中，所述预设数量条操作语句可以是指：从上一次更新源数据库20之后，所述日志文件中新增的所有操作语句。

[0045] 在另一种实施方式中，考虑到上一次更新源数据库20后，日志文件中新增的操作语句的数量较多，可以对新增的操作语句分批次进行更新。在此情况下，可以灵活地设置所述预设数量，例如可以设置为50、100、200或500等。

[0046] 在本公开中，每次截取所述预设数量条操作语句后，可以确定所述预设数量条操作语句中携带最新版本信息的操作语句，并将预存的标志版本信息的值更新为该最新版本

信息。下次进行截取操作时，可以从携带所述标志版本信息的操作语句开始截取所述预设数量条操作语句。

[0047] 可选地，所述标志版本信息可以预存在电子设备10中，其初始值可以为空(NULL)。在实施过程中，如果电子设备10检测到标志版本信息的初始值为空，可以确定当前遍历到的删除语句为第一条删除语句，从而可以直接将所述当前遍历到的删除语句之前的操作语句均作为目标操作语句。

[0048] 在本公开中，从源数据库20的日志文件获得的所述预设数量条操作语句中通常包括插入语句、删除语句及其他的一些语句。其中，删除语句是删除符合特定删除条件的数据的操作语句。

[0049] 在一些情况下，从日志文件获得的预设数量条操作语句中可能存在删除语句，而在该删除语句之前可能有符合该删除语句的删除条件的插入语句，换言之，在该删除语句之前的插入语句中插入的数据可能在该删除语句中会被删除。

[0050] 例如，假设从日志文件获得如下操作语句：

[0051] SCN1 向数据表1插入数据；

[0052] SCN2 向数据表1插入数据；

[0053] SCN3 向数据表2插入数据；

[0054] SCN4 向数据表1插入数据；

[0055] SCN5 删除数据表1。

[0056] 在上述操作语句中，操作语句SCN1、SCN2以及SCN4中插入的数据在操作语句SCN5中均已被删除，因而，在更新目标数据库30中的数据时，可以不必再执行操作语句SCN1、SCN2以及SCN4对应的操作语句，也能使源数据库20和目标数据库30中的数据保持一致。在本公开中，可以通过后述的步骤S22-步骤S24达到该效果。

[0057] 又如，假设从日志文件获得以下操作语句：

[0058] SCN1向数据表1插入字段a＝1的数据记录；

[0059] SCN2向数据表1插入字段a＝1、字段b＝2的数据记录；

[0060] SCN3向数据表2插入字段c＝1、字段d＝3的数据记录；

[0061] SCN4向数据表1插入字段c＝1的数据记录；

[0062] SCN5删除数据表2；

[0063] SCN6从数据表1中删除a＝1的数据语句。

[0064] 在上述操作语句中，操作语句SCN3插入的数据在操作语句SCN5中已被删除，因此，在更新目标数据库30时，可以不必再执行操作语句SCN3。对应地，操作语句SCN1和SCN2中插入的数据在操作语句SCN6中已被删除，则在更新目标数据库30时，可以不必再执行操作语句SCN1和SCN2。

[0065] 步骤S22，查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，并在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句。

[0066] 在实际应用中，从日志文件获得的所述预设数量条操作语句中可能包括一条、两条或多条删除语句。其中，当存在多条删除语句时，每条删除语句之前都可能存在符合该删除语句的删除条件的插入语句。

[0067] 在此情况下,在本公开的一种具体实施方式中,步骤S22可以包括图4所示的步骤。

[0068] 步骤S41,按照版本信息从旧到新的顺序遍历所述多条操作语句。

[0069] 步骤S42,在当前遍历到的操作语句为删除语句时,在所述多条操作语句中将版本信息比当前遍历到的删除语句旧的操作语句确定为所述目标操作语句。

[0070] 例如,假设所述多条操作语句中存在第一删除语句和第二删除语句共两条删除语句,在实际应用中可能出现如下情形:第一删除语句之前的操作语句中,存在满足所述第二删除语句的删除条件的插入语句。在此情况下,可以将所述当前遍历到的删除语句之前的操作语句均作为目标操作语句。具体可以为:在遍历到第一删除语句时,将第一删除语句之前的所有操作语句确定为目标操作语句;在遍历到第二删除语句时,将第二删除语句之前的所有的操作语句确定为目标操作语句。在此值得说明的是,在本公开中提及的某一操作语句之前的语句均是指版本信息比该操作语句旧的语句,某一操作语句之后的语句均是指版本信息比该操作语句新的语句。

[0071] 步骤S43,在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句。

[0072] 在本公开的又一种具体实施方式中,步骤S22可以通过如下子步骤实现:

[0073] 按照版本信息从旧到新的顺序遍历所述多条操作语句;在当前遍历到的操作语句为删除语句时,在所述多条操作语句中将版本信息处于所述当前遍历到的删除语句和上一次遍历到的删除语句之间的操作语句确定为所述目标操作语句;

[0074] 在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句。

[0075] 仍旧以上述的第一删除语句和第二删除语句为例,在此实施方式中,当遍历到第二删除语句时,将版本信息处于第一删除语句和第二删除语句之间的操作语句确定为目标操作语句。例如,假设当前遍历到的删除语句(第二删除语句)的版本信息为V6,上一次遍历到的删除语句(第一删除语句)的版本信息为V2,则版本信息比V2新且比V6旧的操作语句均为目标操作语句。

[0076] 具体地,以版本信息是操作序列号SCN为例,如果当前遍历到的删除语句(第二删除语句)的操作序列号是S1,上一次遍历到的删除语句(第一删除语句)的操作序列号是S5,则操作序列号S满足S5<S<S1的操作语句均为目标操作语句。

[0077] 如此,每次可以从相邻删除语句之间的操作语句中确定所述待删除语句。

[0078] 步骤S23,从所述预设数量条操作语句中删除所述待删除语句。

[0079] 如此,即可减少更新MPP数据库的数据时所需执行的插入操作的次数。

[0080] 步骤S24,对剩余的操作语句中的插入语句进行合并。

[0081] 其中,所述剩余的操作语句是指:从所述预设数量条操作语句中删除所述待删除语句之后剩下的所有操作语句。

[0082] 例如,假设存在插入语句1和插入语句2,其中,插入语句1为向数据表1插入字段a=1、字段b=1的数据记录,插入语句2为向数据表1插入字段a=1、字段c=3的数据记录。则可以确定插入语句1和插入语句2均向同一数据表(数据表1)插入了数据,因此,可以将插入语句1和插入语句2合并为一条插入语句。

[0083] 经研究发现,MPP数据库通常为列式存储,例如上述示例中的a、b、c、d均表示数据

表1中的一列(列表),它们分属不同的存储空间。以插入语句insertintotest(a,b,c,d) values(1,1,NULL,NULL)为例,在执行该插入语句时,需要查找到字段a对应的列表和字段b对应的列表才能进行数据插入,这需要两次IO操作。如此,在执行每条插入语句时,都需先查找相应的列表。

[0084] 鉴于上述的插入语句1涉及到向字段a对应的列表进行数据插入,上述的插入语句2也涉及到向字段a对应的列表进行数据插入,通过上述合并过程,可以只需查找一次字段a对应的列表,减少了IO操作的次数。当合并的插入语句较多时,可以大量减少更新MPP数据库的数据时执行的IO操作的次数,加快更新速度。在实施过程中,查找到的向同一数据表插入数据的不同插入语句可能有多条。在一种实施方式中,可以将查找到的不同插入语句全部合并为一条插入语句。

[0085] 可选地,在本公开的一种实施方式中,步骤S24可以通过如下子步骤实现:

[0086] 在所述剩余的操作语句中将操作的目标数据表相同的插入语句合并为一条插入语句。

[0087] 其中,目标数据表是指操作语句所操作的数据表。对应地,在上述步骤中,将所述剩余的操作语句中操作同一数据表的全部插入语句合并为一条插入语句。

[0088] 在另一种实施方式中,步骤S24可以通过如下子步骤实现:

[0089] 在所述剩余的操作语句中,将版本信息处于所述当前遍历到的删除语句和所述上一次遍历到的删除语句之间的、且操作的目标数据表相同的插入语句合并为一条插入语句。

[0090] 特别地,在此实施方式中,对于首次遍历到的删除语句,可以将版本信息处于该删除语句之前的、操作的目标数据表相同的插入语句合并为一条差语句。

[0091] 在实际应用中,所述剩余的操作语句可能被多个删除语句分隔为多个操作语句组。例如,假设剩余的操作语句有SCN1、SCN 2、SCN5、SCN6、SCN7、SCN8和SCN9,其中,SCN5和SCN9为删除语句,则该剩余的操作语句被删除语句SCN5和SCN9分为2个操作语句组,其中第1个操作语句组包括SCN1和SCN2,第2个操作语句组包括SCN6、SCN7和SCN8。

[0092] 在实施时,可以针对每个操作语句组,将该操作语句组中操作同一数据表的插入语句合并为一条插入语句。

[0093] 下面通过一具体例子来对步骤S24进行描述。

[0094] 假设从所述预设数量条操作语句中删除所述待删除语句后,剩余的操作语句如下:

[0095] SCN2向数据表1插入字段b=1、字段c=1的数据记录;

[0096] SCN3向数据表1插入字段b=1、字段d=2的数据记录;

[0097] SCN4从数据表1中删除字段a=1的数据记录;

[0098] SCN5向数据表1插入字段b=1、字段d=1的数据记录;

[0099] SCN7向数据表1插入字段a=2、字段b=1的数据记录;

[0100] SCN8从数据表1中删除字段d=2的数据记录。

[0101] 在实施过程中,在一种实施方式中,可以确定上述剩余的操作语句中,插入语句SCN2、SCN3、SCN5以及SCN7操作的目标数据表均为数据表1,因此,可以将插入语句SCN2、SCN3、SCN5以及SCN7合并为一条插入语句。

[0102] 在另一实施方式中,可以确定上述剩余的操作语句中,处于删除语句SCN4之前的插入语句SCN2和SCN3操作的目标数据表均为数据表1,因此可以将插入语句SCN2和SCN3合并为一条插入语句;处于删除语句SCN4和SCN8之间的插入语句SCN5和SCN7操作的目标数据表均为数据表1,因而可以将插入语句SCN5和SCN7合并为一条插入语句。

[0103] 可选地,在本公开中,为便于对插入语句进行合并,可以在执行步骤S22之前,所述数据同步方法还可以包括图5示出的步骤。

[0104] 步骤S51,针对所述预设数量条操作语句中的每条插入语句,确定该插入语句操作的目标数据表,并将该目标数据表包括的字段确定为目标字段。

[0105] 步骤S52,判断是否存在该插入语句不包括的目标字段。

[0106] 步骤S53,若是,则在该插入语句中增加所述不包括的目标字段,并将增加的目标字段的插入值设置为空。

[0107] 其中,所述插入值是指向所述增加的目标字段插入的值。

[0108] 例如,假设在所述预设数量条操作语句中,存在上述的插入语句1和插入语句2,则针对插入语句1,可以将插入语句1操作的数据表1确定为目标数据表,进而将数据表1(目标数据表)包括的字段a、字段b、字段c以及字段d均确定为目标字段。然后,可以确定插入语句1中不包括目标字段c和目标字段d,从而在插入语句1中增加目标字段c和目标字段d,并将增加的目标字段c和目标字段d的插入值设置为NULL。

[0109] 对应地,针对插入语句2可以参照上述过程中进行处理,从而向插入语句2中增加目标字段b和目标字段d,并将目标字段b和目标字段d的插入值均设置为NULL。

[0110] 在此情况下,假设插入语句1为:insertintotest(a,b)values(1,1);插入语句2为:insertintotest(a,c)values(1,3)。则,按照步骤S61-步骤S63处理后,插入语句1改变为:insertintotest(a,b,c,d)values(1,1,NULL,NULL);插入语句2改变为:insertintotest(a,b,c,d)values(1,NULL,3,NULL)。

[0111] 则,处理后的插入语句1和插入语句2可以合并为:insertintotest(a,b,c,d)values(1,1,NULL,NULL)(1,NULL,3,NULL)。

[0112] 可选地,在本公开中,上述步骤S61-步骤S63也可以在步骤S51之前,对所述剩余的操作语句中的插入语句执行。

[0113] 步骤S25,根据合并后的操作语句更新MPP数据库的数据。

[0114] 其中,合并后的操作语句至少包括所述删除语句及合并后的插入语句。

[0115] 对应地,如果所述预设数量条操作语句中还包括更新语句和查询语句,则所述剩余的操作语句中也包括更新语句和查询语句,所述合并后的操作语句也包括更新语句和查询语句。

[0116] 应当理解,所述合并后的操作语句是指对所述剩余的操作语句进行合并操作之后得到的所有操作语句。例如,在一些实施方式中,仅对剩余的操作语句中的部分语句进行合并,则合并后的操作语句包括:所述部分语句合并得到的操作语句以及另一部分未被合并的操作语句。

[0117] 在实施时,在所述MPP数据库执行所述合并后的操作语句,即可更新所述MPP数据库(目标数据库30)中的数据,从而将源数据库20中的数据同步到目标数据库30中。

[0118] 经研究发现,在Oracle CDC技术的同步方式中,需要在源数据库20的业务表中设

置触发器，触发器需要将对源数据库20的任意操作重复执行一次，以将该操作对应的数据记录到临时表中。这会占用源数据库20的大量资源，对源数据库20的性能造成影响。在OracleCDC的异步方式中，需要对MPP数据库执行大量的插入操作，而MPP数据库适于进行OLAP(On-Line Analysis Processing)，支持复杂分析操作，不适于做大批量的插入操作。例如，传统数据库比如Oracle每秒可以执行上万次插入操作，而MPP数据库通常每秒只能执行几百次插入操作。如果将获得的全部操作语句都对MPP数据库执行一遍，则实际运行时会有大量操作语句无法被及时执行，从而导致源数据库和MPP数据库(目标数据库)的数据同步发生时延，导致数据不一致。

[0119] 而本公开提供的数据同步方法，从日志文件获取操作语句以用于数据同步，这种方式对源数据库20的资源没有侵占性，即不会占用源数据库20的资源，对应地，不会导致源数据库20的性能降低。此外，通过将删除语句之前符合该删除语句的条件的插入语句删除，可以减少更新过程中执行的插入操作的次数，从而缓解源数据库20和目标数据库30在数据同步过程中的时延，进而缓解由于时延导致的数据不一致问题。

[0120] 此外，通过对操作同一数据表的不同插入语句的合并，能够减少更新过程中执行的IO操作的次数，进一步提升数据同步的速度。

[0121] 下面给出一个具体示例，以对上述步骤进行详细阐述。

[0122] 假设从日志文件获得的预设数量条操作语句如下表1所示，其中，每一行表示一条操作语句，序号表示该操作语句的版本信息。

[0123] 表1

[0124]

序号	SQL
1	insert into test（a，b）values（1，2）
2	insert into test（a，b）values（1，2）
3	insert into test（a，c）values（1，3）

[0125]

4	insert into test（a，d）values（1，4）
5	insert into test（c，d）values（1，5）
6	insert into test（c，d）values（1，6）
7	insert into test（a，b）values（1，7）
8	insert into test（a，b）values（1，8）
9	insert into test（a，b）values（1，9）
10	delete from test where a=1 or b=6
11	insert into test（a，b）values（1，10）
12	insert into test（a，b）values（1，11）
13	insert into test（a，b）values（1，12）
14	insert into test（a，b）values（1，13）
15	insert into test（a，b）values（1，14）
16	insert into test（a，b）values（1，15）
17	insert into test（a，b）values（1，16）
18	insert into test（a，b）values（1，17）
19	insert into test（a，b）values（1，18）
20	insert into test（a，b）values（1，19）
21	delete from test where a=1 or c=1

[0126] 在表1中，insert into test(a,b)values(1,2)表示向数据表test插入字段a＝1、字段b＝2的数据记录。对应地，insertintotest(a,c)values(1,3)表示向数据表test插入字段a＝1、字段c＝3的数据记录；insertintotest(c,d)values(1,5)表示向数据表test插入字段c＝1、字段d＝5的数据记录。其他类似语句的含义可以参照上述示例确定。

[0127] delete from testwherea＝1or b＝6，表示从数据表test中删除字段a＝1的数据记录以及字段b＝6的数据记录。

[0128] 在实施过程中，当电子设备10获得上述预设数量条操作语句时，可以按照上述步骤S61-步骤S63对所述预设数量条操作语句的格式进行处理，从而得到如下表2所示的操作语句。

[0129] 表2

[0130]

序号	SQL
1	insert into test(a,b,c,d)values(1,2,NULL,NULL)
2	insert into test(a,b,c,d)values(1,2,NULL,NULL)
3	insert into test(a,b,c,d)values(1,NULL,3,NULL)

4	insert into test(a,b,c,d)values(1,NULL,NULL,4)
5	insert into test(a,b,c,d)values(NULL,NULL,1,5)
6	insert into test(a,b,c,d)values(NULL,NULL,1,6)
7	insert into test(a,b,c,d)values(1,7,NULL,NULL)
8	insert into test(a,b,c,d)values(1,8,NULL,NULL)
9	insert into test(a,b,c,d)values(1,9,NULL,NULL)
10	delete from test where a＝1or b＝6
11	insert into test(a,b,c,d)values(1,10,NULL,NULL)
12	insert into test(a,b,c,d)values(1,11,NULL,NULL)
13	insert into test(a,b,c,d)values(1,12,NULL,NULL)
14	insert into test(a,b,c,d)values(1,13,NULL,NULL)
15	insert into test(a,b,c,d)values(1,14,NULL,NULL)
16	insert into test(a,b,c,d)values(1,15,NULL,NULL)
17	insert into test(a,b,c,d)values(1,16,NULL,NULL)
18	insert into test(a,b,c,d)values(1,17,NULL,NULL)
19	insert into test(a,b,c,d)values(1,18,NULL,NULL)
20	insert into test(a,b,c,d)values(1,19,NULL,NULL)
21	delete from test where b＝10

[0131] 按照序号从小到大的顺序遍历处理后的各操作语句(即表2中的操作语句)，当遍历到序号为10的操作语句时，可以确定该操作语句为删除语句，此时，在该删除语句之前没有其他删除语句，可以直接在该删除语句之前的所有操作语句中查找符合该删除语句的删除条件(即，a＝1或b＝6)的插入语句，从而可以确定序号为1-4、7-9的插入语句均符合序号为10的删除语句的删除条件，从而可以将序号为1-4、7-9的插入语句删除。其中，序号为1-4、7-9的插入语句即可充当本公开中的待删除语句。

[0132] 当遍历到序号为21的操作语句时，可以确定该操作语句为删除语句。在第一种方式中，可以确定上一次遍历到的删除语句是序号为10的操作语句(后称删除语句10)，从而可以在序号处于10-21之间的操作语句(即序号为11-20的操作语句)中查找符合删除语句21的删除条件(a＝1或c＝1)的插入语句作为待删除语句，如此，电子设备10通过查找将序号为11的插入语句作为待删除语句并删除。

[0133] 在第二种方式中，当遍历到序号21的操作语句、并确定该操作语句是删除语句时，可以将序号1-20的操作语句全部确定为目标操作语句并删除。下面以第一种方式中确定的目标操作语句为例，对后续过程进行详述。

[0134] 通过上述第一种方式，所述预设数量条操作语句中剩余的操作语句如下表3所示：

[0135] 表3

[0136]

序号	SQL
5	insert into test（a，b，c，d）values（NULL，NULL，1，5）
6	insert into test（a，b，c，d）values（NULL，NULL，1，6）
10	delete from test where a=1 or b=6
12	insert into test（a，b，c，d）values（1，11，NULL，NULL）
13	insert into test（a，b，c，d）values（1，12，NULL，NULL）
14	insert into test（a，b，c，d）values（1，13，NULL，NULL）
15	insert into test（a，b，c，d）values（1，14，NULL，NULL）
16	insert into test（a，b，c，d）values（1，15，NULL，NULL）
17	insert into test（a，b，c，d）values（1，16，NULL，NULL）
18	insert into test（a，b，c，d）values（1，17，NULL，NULL）

[0137]

19	insert into test（a，b，c，d）values（1，18，NULL，NULL）
20	insert into test（a，b，c，d）values（1，19，NULL，NULL）
21	delete from test where b=10

[0138] 通过查找可以确定上述剩余的操作语句中存在向同一数据表插入数据的不同插入语句，具体是向数据表test插入数据的插入语句5、6以及11-20。

[0139] 在一种实施方式中，可以将插入语句5、6以及11-20合并为一条插入语句，得到：insertintotest(a,b,c,d)values(NULL,NULL,1,5)(NULL,NULL,1,6)(1,11,NULL,NULL)(1,12,NULL,NULL)(1,13,NULL,NULL)(1,14,NULL,NULL)(1,15,NULL,NULL)(1,16,NULL,NULL)(1,17,NULL,NULL)(1,18,NULL,NULL)(1,19,NULL,NULL)。

[0140] 通过上述过程，最终得到的合并后的操作语句如下表4所示：

[0141] 表4

[0142]

序号	SQL
10	delete from test where a=1 or b=6
12	insert into test（a，b，c，d）values（NULL，NULL，1，5）（NULL，NULL，1，6）（1，11，NULL，NULL）values（1，12，NULL，NULL）（1，13，NULL，NULL）（1，14，NULL，NULL）（1，15，NULL，NULL）（1，16，NULL，NULL）（1，17，NULL，NULL）（1，18，NULL，NULL）（1，19，NULL，NULL）
21	delete from test where b=10

[0143] 在另一种实施方式中，可以将插入语句5和插入语句6进行合并，得到

insertintotest(a,b,c,d)values(NULL,NULL,1,5)(NULL,NULL,1,6)。

[0144] 将插入语句12-插入语句20进行合并,得到insert into test(a,b,c,d)values(1,11,NULL,NULL)(1,12,NULL,NULL)(1,13,NULL,NULL)(1,14,NULL,NULL)(1,15,NULL,NULL)(1,16,NULL,NULL)(1,17,NULL,NULL)(1,18,NULL,NULL)(1,19,NULL,NULL)。

[0145] 通过上述过程,最终得到的合并后的操作语句如下表5所示:

[0146] 表5

[0147]

序号	SQL
5	insert into test (a, b, c, d) values (NULL, NULL, 1, 5) (NULL, NULL, 1, 6)
10	delete from test where a=1 or b=6
12	insert into test (a, b, c, d) values (1, 11, NULL, NULL) values (1, 12, NULL, NULL) (1, 13, NULL, NULL) (1, 14, NULL, NULL) (1, 15, NULL, NULL) (1, 16, NULL, NULL) (1, 17, NULL, NULL) (1, 18, NULL, NULL) (1, 19, NULL, NULL)
21	delete from test where b=10

[0148] 电子设备10对目标数据库30(即MPP数据库)执行表4或表5中的所述合并后的操作语句,即可将源数据库20中的相应数据同步到MPP数据库中。

[0149] 请参照图6,图6是本公开提供的一种电子设备10的方框示意图,该电子设备10包括机器可读存储介质11、处理器12以及通信单元13。

[0150] 所述机器可读存储介质11、处理器12以及通信单元13之间直接或间接地电性连接。例如,这些元件相互之间可以通过一条或多条通讯总线或信号线实现电性连接。机器可读存储介质11中存储有用于实现本公开提供的数据同步方法的机器可执行指令,处理器12通过读取并执行机器可读存储介质11上的机器可执行指令,可以实现所述数据同步方法。通信单元13用于建立与外部设备的通信连接,例如建立与源数据库20或目标数据30所在服务器之间的通信连接。

[0151] 本文提到的机器可读存储介质11可以是任何电子、磁性、光学或其他物理存储装置,可以包含或存储信息,如可执行指令、数据,等等。例如,机器可读存储介质11可以是:RAM(RandomAccessMemory,随机存取存储器)、易失存储器、任何类型的存储盘(如光盘、DVD等),或者类似的存储介质,或者它们的组合。

[0152] 请参照图7,图7是本公开提供的一种数据同步装置110的功能模块框图,该数据同步装置110可以包括至少一个可以以软件或固件(Firmware)的形式存储于机器可读存储介质11上或固化在电子设备10的操作系统(OperatingSystem,OS)中的软件功能模块。处理器12可以执行所述机器可读存储介质11上的可执行模块,例如所述数据同步装置110所包括的软件功能模块及计算机程序等。所述数据同步装置110可以包括操作语句获得模块111、查找模块112、删除模块113、合并模块114以及更新模块115。

[0153] 其中，操作语句获得模块111用于从源数据库20的日志文件中获得预设数量条操作语句，其中，每条操作语句包括该操作语句的版本信息。

[0154] 在本公开中，关于操作语句获得模块111的描述具体可参考对图2所示步骤S21的详细描述，也即，步骤S21可以由操作语句获得模块111执行。

[0155] 查找模块112用于查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句。

[0156] 在本公开中，关于查找模块112的描述具体可以参考对图2所示步骤S22的详细描述，即步骤S22可以由查找模块112执行。

[0157] 可选地，在本公开中，查找模块112可以包括遍历子模块1121、第一查找子模块1122以及第二查找子模块1123。

[0158] 其中，遍历子模块1121用于按照版本信息从旧到新的顺序遍历所述预设数量条操作语句。

[0159] 第一查找子模块1122用于在当前遍历到的操作语句为删除语句时，在所述预设数量条操作语句中将版本信息比当前遍历到的删除语句旧的操作语句确定为所述目标操作语句，或在所述预设数量条操作语句中将版本信息处于所述当前遍历到的删除语句和上一次遍历到的删除语句之间的操作语句确定为所述目标操作语句。

[0160] 第二查找子模块1123用于在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句。

[0161] 删除模块113用于从所述预设数量条操作语句中删除所述待删除语句。

[0162] 在本公开中，关于删除模块113的描述具体可参考对图2所示步骤S23的详细描述，即步骤S23可以由删除模块113执行。

[0163] 合并模块114用于对剩余的操作语句中的插入语句进行合并。

[0164] 在本公开中，合并模块114可以用于执行图2中示出的步骤S24，关于合并模块114的描述具体可以参考对步骤S24的详细描述。

[0165] 在本公开中，合并模块114具体可以用于在所述剩余的操作语句中将操作的目标数据表相同的插入语句合并为一条插入语句；或者，在所述剩余的操作语句中，将版本信息处于所述当前遍历到的删除语句和所述上一次遍历到的删除语句之间的、且操作的目标数据表相同的插入语句合并为一条插入语句。

[0166] 更新模块115用于根据合并后的操作语句更新MPP数据库的数据，其中，所述合并后的操作语句至少包括所述删除语句及合并后的插入语句。

[0167] 在本公开中，关于更新模块115的描述具体可以参考对图2所示步骤S25的详细描述，即步骤S25可以由更新模块115执行。

[0168] 可选地，在本公开中，所述数据同步装置110还可以包括格式处理模块116。

[0169] 格式处理模块116用于针对所述预设数量条操作语句中的每条插入语句，确定该插入语句操作的目标数据表，并将该目标数据表包括的字段确定为目标字段；判断是否存在该插入语句不包括的目标字段，若是，则在该插入语句中增加所述不包括的目标字段，并将增加的目标字段的插入值设置为空。

[0170] 在本公开中，格式处理模块116可以用于执行图5中示出的步骤S51、步骤S52和步

骤S53，关于格式处理模块116的描述具体可参考对步骤S51、步骤S52和步骤S53的详细描述。

[0171] 可选地，在本公开中，操作语句获得模块111具体可以用于通过日志分析工具解析所述日志文件，获得所述源数据库的操作语句，并在所述源数据库的操作语句中，从上一次截取的版本信息最新的操作语句开始截取所述预设数量条操作语句。

[0172] 综上所述，本公开提供一种数据同步方法及装置，从源数据库的日志文件中获得预设数量条操作语句，每条操作语句包括该操作语句的版本信息。查找多条操作语句中的删除语句，并在查找到删除语句时，在预设数量条操作语句中确定版本信息比该删除语句旧的目标操作语句，并在目标操作语句中查找符合该删除语句的删除条件的插入语句作为待删除语句。从预设数量条操作语句中删除该待删除语句，对剩余的操作语句中的插入语句进行合并，并根据合并后的操作语句更新MPP数据库的数据。如此，可以减少更新数据时执行的插入操作及IO操作的次数，降低更新数据的时延，从而提高源数据库和MPP数据库中的数据的一致性。

[0173] 在本公开所提供的实施例中，应该理解到，所揭露的装置和方法，也可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的，例如，附图中的流程图和框图显示了根据本公开的多个实施例的装置、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现方式中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0174] 另外，在本公开各个实施例中的各功能模块可以集成在一起形成一个独立的部分，也可以是各个模块单独存在，也可以两个或两个以上模块集成形成一个独立的部分。

[0175] 所述功能如果以软件功能模块的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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一个存储介质中，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)执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0176] 需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0177] 以上所述，仅为本公开的具体实施方式，但本公开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公开揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本公开的保护范围之内。因此，本公开的保护范围应以权利要求的保护范围为准。

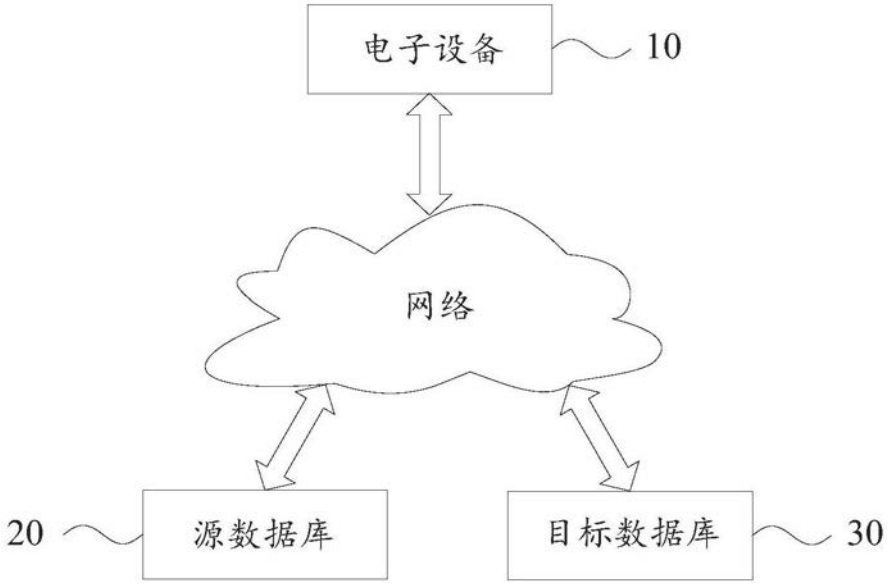

图1

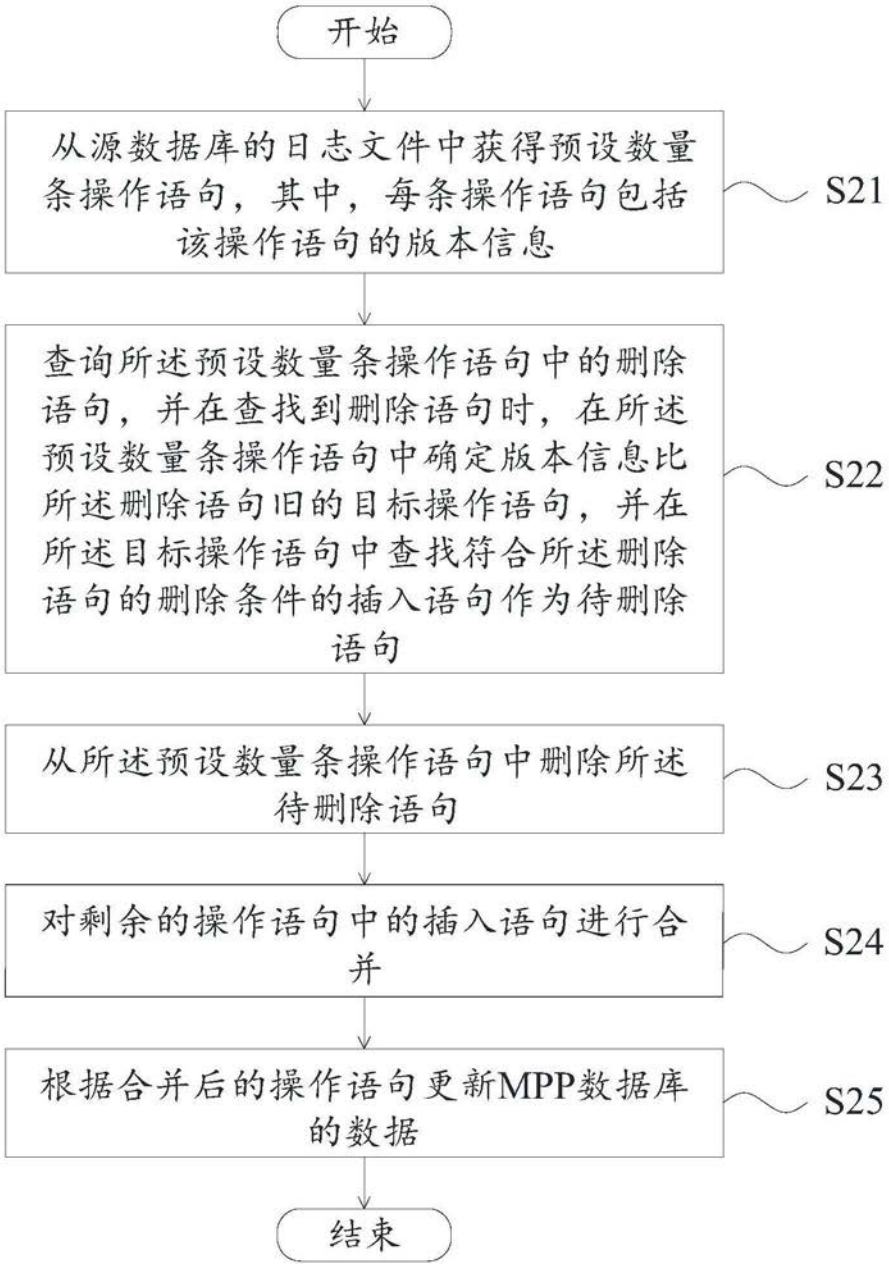
开始
从源数据库的日志文件中获得预设数量条操作语句，其中，每条操作语句包括该操作语句的版本信息
S21
查询所述预设数量条操作语句中的删除语句，并在查找到删除语句时，在所述预设数量条操作语句中确定版本信息比所述删除语句旧的目标操作语句，并在所述目标操作语句中查找符合所述删除语句的删除条件的插入语句作为待删除语句
S22
从所述预设数量条操作语句中删除所述待删除语句
S23
对剩余的操作语句中的插入语句进行合并
S24
根据合并后的操作语句更新MPP数据库的数据
S25
结束

图2

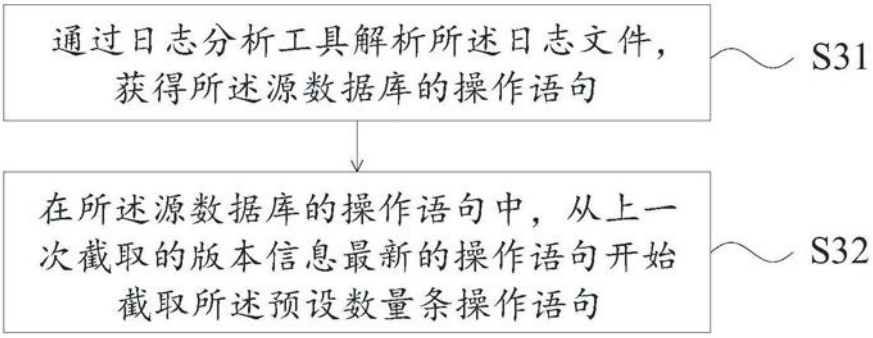
通过日志分析工具解析所述日志文件，获得所述源数据库的操作语句
S31
在所述源数据库的操作语句中，从上一次截取的版本信息最新的操作语句开始截取所述预设数量条操作语句
S32

图3

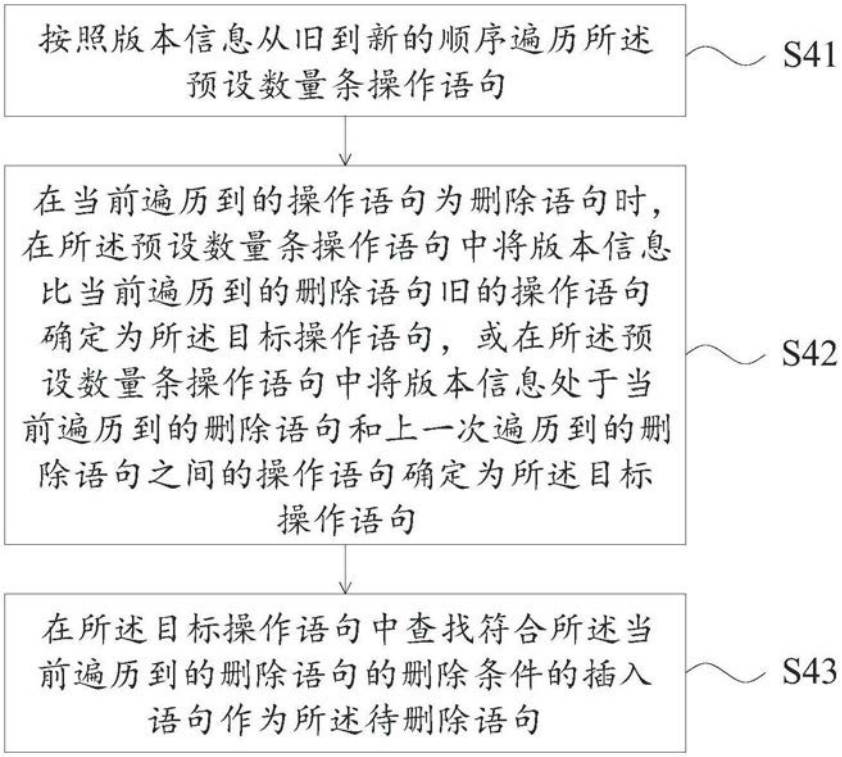
按照版本信息从旧到新的顺序遍历所述预设数量条操作语句
S41
在当前遍历到的操作语句为删除语句时，在所述预设数量条操作语句中将版本信息比当前遍历到的删除语句旧的操作语句确定为所述目标操作语句，或在所述预设数量条操作语句中将版本信息处于当前遍历到的删除语句和上一次遍历到的删除语句之间的操作语句确定为所述目标操作语句
S42
在所述目标操作语句中查找符合所述当前遍历到的删除语句的删除条件的插入语句作为所述待删除语句
S43

图4

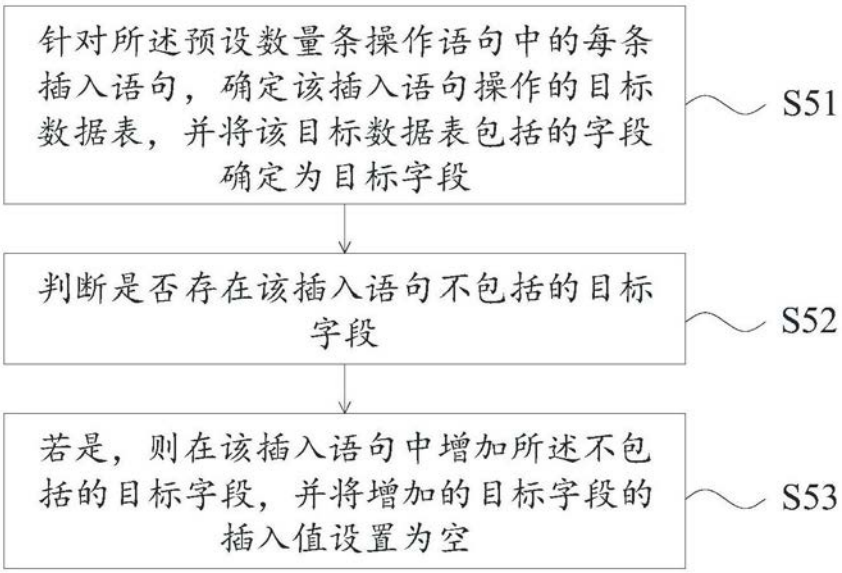
针对所述预设数量条操作语句中的每条插入语句，确定该插入语句操作的目标数据表，并将该目标数据表包括的字段确定为目标字段
S51
判断是否存在该插入语句不包括的目标字段
S52
若是，则在该插入语句中增加所述不包括的目标字段，并将增加的目标字段的插入值设置为空
S53

图5

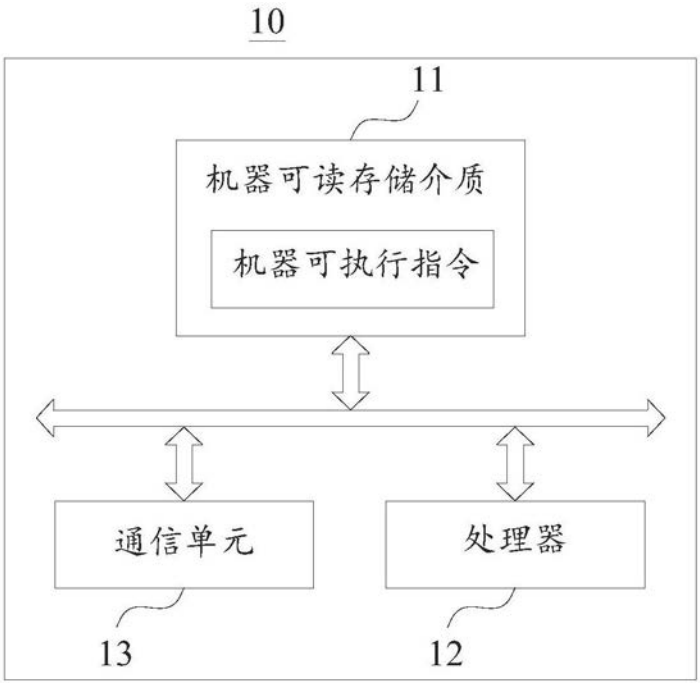
10
11
机器可读存储介质
机器可执行指令
通信单元
处理器
13
12

图6

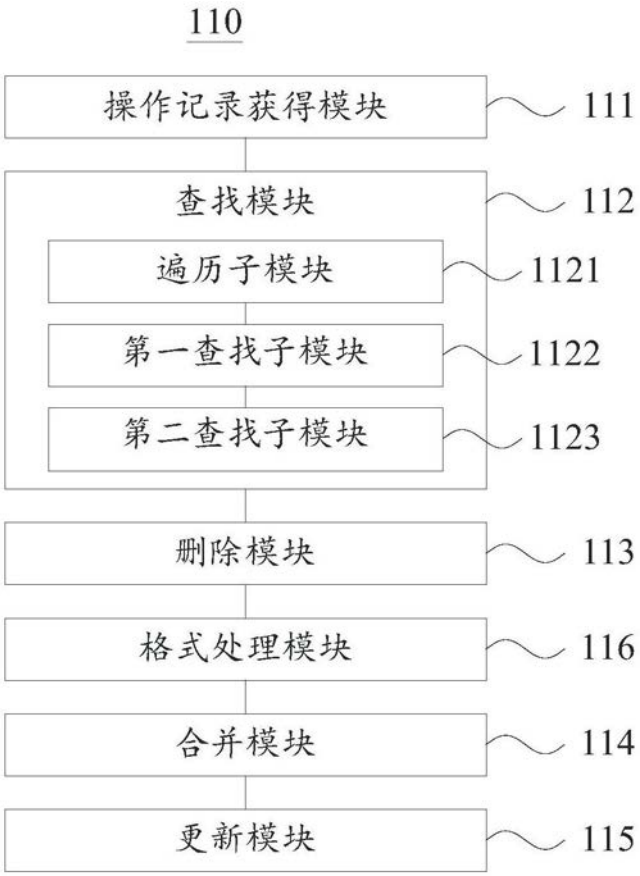
110
操作记录获得模块
111
查找模块
112
遍历子模块
1121
第一查找子模块
1122
第二查找子模块
1123
删除模块
113
格式处理模块
116
合并模块
114
更新模块
115

图7