(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107679104 B
(45)授权公告日 2020.11.24

(21)申请号 201710819917.8

(22)申请日 2017.09.12

(65)同一申请的已公布的文献号
申请公布号 CN 107679104 A

(43)申请公布日 2018.02.09

(73)专利权人 杭州美创科技有限公司
地址 310011 浙江省杭州市拱墅区丰潭路508号天行国际中心7幢12楼

(72)发明人 闻建霞 柳遵梁 姚远 陈慧慧 陈建培 吕海波

(74)专利代理机构 杭州杭诚专利事务所有限公司 33109
代理人 尉伟敏 阎忠华

(51)Int.Cl.
G06F 16/2455(2019.01)
G06F 16/2453(2019.01)
G06F 16/22(2019.01)

(56)对比文件
CN 102375886 A,2012.03.14
CN 102184190 A,2011.09.14
CN 102890720 A,2013.01.23
US 2013297615 A1,2013.11.07
CN 101702162 A,2010.05.05
US 2011289091 A1,2011.11.24
US 8738632 B2,2014.05.27
US 2014201192 A1,2014.07.17
CN 106599300 A,2017.04.26

审查员 李梦诗

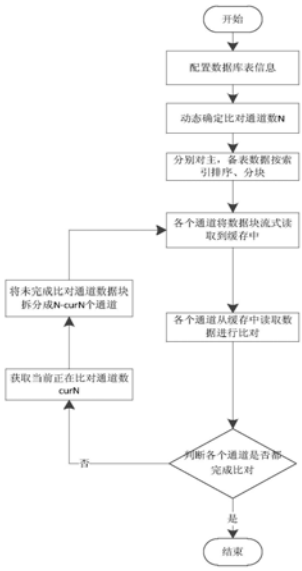
权利要求书1页 说明书3页 附图2页

(54)发明名称

大表流式并行高速数据比对方法

(57)摘要

本发明公开了一种大表流式并行高速数据比对方法，包括如下步骤：(1-1)比对应用程序通过数据库链接配置待比对的主数据库表信息和备数据库表的信息，如果主数据库表信息和备数据库表信息结构不一致，返回无法比对；(1-2)比较主数据库表信息和备数据库表的索引字段，获取最小值min和最大值max，用于比对开始与结束标记；并设置并行比对通道数N，N动态生成，用于并行处理；(1-3)将主备数据库表的记录按索引字段递增排序，并将排序结果按并行比对通道数N进行分块，且各个通道分别流式读取数据，存于缓存中；(1-4)在各个通道中并行处理数据比对，记录比对结果。本发明具有提高了数据比对速度的特点。



CN 107679104 B

1.一种大表流式并行高速数据比对方法,其特征是,包括如下步骤:

(1-1)比对应用程序通过数据库链接配置待比对的主数据库表信息和备数据库表的信息,如果主数据库表信息和备数据库表信息结构不一致,返回无法比对;

(1-2)比较主数据库表信息和备数据库表的索引字段,获取最小值min和最大值max,用于比对开始与结束标记;并设置并行比对通道数N,N动态生成,用于并行处理:

设定主数据库表A的表结构为a integer primary key,b varchar,c number(10);备数据库表B的表结构为a integer primary key,b varchar,c number(10);

通过sql脚本查询表A,B的索引字段为:a integer,并获取两表索引字段的最小值作为比对开始标记:$\text{startFlag}=a_{min}$,索引字段的最大值作为比对结束标记:$\text{endFlag}=a_{max}$;

分别将N取值为N=1,N=2, N=3,N=4,N=5,根据开始标记startFlag递增排序每次取出表A,B中的5000行记录,将该5000行记录根据比对通道数拆分成N数据块,并进行比对,且计算N取不同值时每秒比对多少行记录;最后,将N取值为每秒比对行数最多的通道数;

(1-3)将主备数据库表的记录按索引字段递增排序,并将排序结果按并行比对通道数N进行分块,且各个通道分别流式读取数据,存于缓存中:

将剩下的待比对数据按照索引递增排序拆分成N个数据块,分别通过N个比对通道数进行比对,每个通道每次读取记录行数设置为max=5000行,分多次流式取出,存于缓存中,依次等待比对;在比对的过程中,判断各个通道是否都完成比对,未完成的通道将剩余数据拆分成多个通道,保持在并行处理数据比对的通道数为N,直到所有通道完成100万行数据比对,结束;

(1-4)在各个通道中并行处理数据比对,记录比对结果:

从缓存中读取一行记录,将主备表改行的索引字段内容进行比较,如果大小相同,则继续比较其他字段;如果大小不相同,则将不一致情况记录到异常表中,然后继续比较下一行记录;

比较结果有如下几种情况:该行记录只在主表中存在,则将异常表中的sourceCount字段加1;该行记录只在备表中存在,则将异常表中的targetCount字段加1;该行记录只在主备表中都存在,但存在字段内容不相同,则将异常表中的diffCount字段加1;该行记录只在主备表中都存在,但存在字段内容相同,则将异常表中的sameCount字段加1。

2.根据权利要求1所述的大表流式并行高速数据比对方法,其特征是,所述主数据库表信息和备数据库表的信息均包括参与比对的主备数据库对应的表名称,字段名称及对应字段类型。

大表流式并行高速数据比对方法

技术领域

[0001] 本发明涉及数据库技术领域，尤其是涉及一种能够充分利用硬件资源进行数据比对，提高了数据比对速度的大表流式并行高速数据比对方法。

背景技术

[0002] 现有技术中的数据库表比对方案主要有：

[0003] 将主备数据库表数据导成表格文件，再运用比较工具进行比对，该方法简单明了，但是对于海量数据无法进行比对。

[0004] 利用算法技术进行比对，将主备端数据导入内存，并采用各种查找算法进行比对；该方法的优点是比对速度快，缺点是依赖于算法，并没有利用现有软硬件资源，实用性不强。

[0005] 将主备端数据导入到第三方数据库，执行SQL脚本，然后对执行结果进行差异分析。该方法采用逻辑运算，程序简单，缺点是实时性差，操作复杂，并依赖于第三方数据库，价格昂贵。

[0006] 基于多表的数据库并行比对;这种并行方式不能对单张大表提供性能优化支持。

发明内容

[0007] 本发明的发明目的是为了克服现有技术中的数据库表比对方法实用性不强，实时性差，操作复杂，价格昂贵的不足，提供了一种能够充分利用硬件资源进行数据比对，提高了数据比对速度的大表流式并行高速数据比对方法。

[0008] 为了实现上述目的，本发明采用以下技术方案：

[0009] 一种大表流式并行高速数据比对方法，包括如下步骤：

[0010] (1-1)比对应用程序通过数据库链接配置待比对的主数据库表信息和备数据库表的信息，如果主数据库表信息和备数据库表信息结构不一致，返回无法比对；

[0011] (1-2)比较主数据库表信息和备数据库表的索引字段，获取最小值min和最大值max，用于比对开始与结束标记;并设置并行比对通道数N,N动态生成，用于并行处理；

[0012] (1-3)将主备数据库表的记录按索引字段递增排序，并将排序结果按并行比对通道数N进行分块，且各个通道分别流式读取数据，存于缓存中；

[0013] (1-4)在各个通道中并行处理数据比对，记录比对结果。

[0014] 本发明克服现有数据比对方法的实用性不强，实时性差，操作复杂，价格昂贵，操作复杂的不足，能够在充分利用硬件资源进行数据比对的同时提高数据比对速度，实用性强，实时性好，操作简单，价格便宜，操作简单。

[0015] 作为优选，所述主数据库表信息和备数据库表的信息均包括参与比对的主备数据库对应的表名称，字段名称及对应字段类型。

[0016] 作为优选，并行比对通道数N根据实际数据比对的速度动态确定，取每秒比对表记录行数最多时的比对通道个数。

[0017] 作为优选，1≤N≤5。

[0018] 作为优选，并行处理将比对数据分多个通道同时比较，每个通道是独立的个体，平行独立线程运行，各个通道之间通过缓存进行联系；在并行处理过程中，当一个通道或多个通道已经比对完成，将其它未完成比对的通道数据块拆分给多个通道，保持同时比对的通道数为N，缓解通道压力，保证比对速度。

[0019] 作为优选，流式读取数据为将数据库表记录按索引字段递增进行排序并分块，并设置每个通道待比对数据块一次读取的大小，采取多次读取的方法，流式取出，存于缓存中。

[0020] 因此，本发明具有如下有益效果：能够在充分利用硬件资源进行数据比对的同时提高数据比对速度，实用性强，实时性好，操作简单，价格便宜，操作简单。

附图说明

[0021] 图1为本发明的一种并行处理比对装置图；

[0022] 图2为本发明的一种并行处理比对流程图。

具体实施方式

[0023] 下面结合附图和具体实施方式对本发明做进一步的描述。

[0024] 如图1所示的实施例是一种大表流式并行高速数据比对方法，主要包括：数据预处理，多通道数据读取模块，多通道并行处理，记录比对异常结果等模块。

[0025] 包括如下步骤：

[0026] A、比对应用程序通过数据库链接配置待比对的主备数据库表信息，要求表结构一致，否则返回无法比对；

[0027] B、比较主备数据库表索引字段，获取最小值min和最大值max，用于比对开始与结束标记；并设置并行比对通道数N，该数值大小动态生成，用于并行处理；

[0028] C、将主备数据库表记录按索引字段递增排序，并将排序结果按并行通道数N进行分块，且各个通道分别流式读取数据，存于缓存中；

[0029] D、在各个通道中并行处理数据比对，记录比对结果。

[0030] 下面根据上述步骤进行具体描述。

[0031] 如图2所示，首先配置主备数据表信息，将主备数据库的表名称，表字段名称，表字段类型等信息配置完成，并判断主备表的表结构是否相同，字段类型是否能够比对，不满足，将不能比对。

[0032] 在满足比对条件的主备表中，读取索引字段的最小值，最大值，作为比对的开始行与结束行的标志。

[0033] 确定比对通道数N，取每秒比对表记录行数最多时的比对通道数N，比对通道数取值范围1≤N≤5，用于并行处理数据比对，合理利用资源，提高比对速度。

[0034] 然后，分别将主备表的数据按索引字段进行递增排序，并将排序后的数据按比对通道数N进行分块，且每个通道每次读取数据的大小为max，流式取出，存于缓存中，等待比对通道的读取。当缓存中的数据被比对通道读取用于数据比对后，即可继续从数据库表中读取数据，存于缓存中，等待比对，直到块数据被读取完。

[0035] 各个通道数据比对过程中，逐行读取缓存中的数据块，进行比对。同时，判断各个通道是否都完成比对，未完成通道可将剩余数据拆分成多个通道，保持在并行处理数据比对的通道数为N，直到所有通道完成数据比对，结束。

[0036] 在比对过程中数据发生不一致时，将记录具体不一致情况在异常表中，数据不一致包括：数据只在主表中存在(删除)，数据只在备表中存在(增加)，主备表中都存在(更新)三种情况。

[0037] 下面举例说明：假设有两张100万行的大表，主表A的表结构为a integer primary key，b varchar，c number(10)；备表B的表结构为a integer primary key，b varchar，c number(10)；

[0038] 1)首先根据上述步骤中所述的判断表A，B的表结构是否一样，该例中表结构一致。

[0039] 2)通过sql脚本查询表A，B的索引字段为：a integer，并获取两表索引字段的最小值作为比对开始标记：startFlag＝a_{min}，索引字段的最大值作为比对结束标记：endFlag＝a_{max}(如果主备表有多个索引字段，取第一索引字段)。

[0040] 3)下面确定并行比对通道个数N(1≤N≤5)，分别将N取值为N＝1，N＝2，N＝3，N＝4，N＝5，根据开始标记startFlag递增排序每次取出表A，B中的5000行记录，将该5000行记录根据比对通道数拆分成N数据块，并进行比对，且计算N取不同值时每秒比对多少行记录。最后，将N取值为每秒比对行数最多的通道数。

[0041] 4)然后将剩下的待比对数据按照索引递增排序拆分成N个数据块，分别通过N个比对通道数进行比对，每个通道每次读取记录行数可以设置为max＝5000行，分多次流式取出，存于缓存中，依次等待比对。在比对的过程中，判断各个通道是否都完成比对，未完成通道可将剩余数据拆分成多个通道，保持在并行处理数据比对的通道数为N，直到所有通道完成100万行数据比对，结束。

[0042] 具体比对过程如下：首先从缓存中读取一行记录，将主备表改行的索引字段内容进行比较，如果大小相同，则继续比较其他字段；如果大小不相同，则将不一致情况记录到异常表中，然后继续比较下一行记录。比较结果有如下几种情况：该行记录只在主表中存在，则将异常表中的sourceCount字段加1；该行记录只在备表中存在，则将异常表中的targetCount字段加1；该行记录只在主备表中都存在，但存在字段内容不相同，则将异常表中的diffCount字段加1；该行记录只在主备表中都存在，但存在字段内容相同，则将异常表中的sameCount字段加1。

[0043] 应理解，本实施例仅用于说明本发明而不用于限制本发明的范围。此外应理解，在阅读了本发明讲授的内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

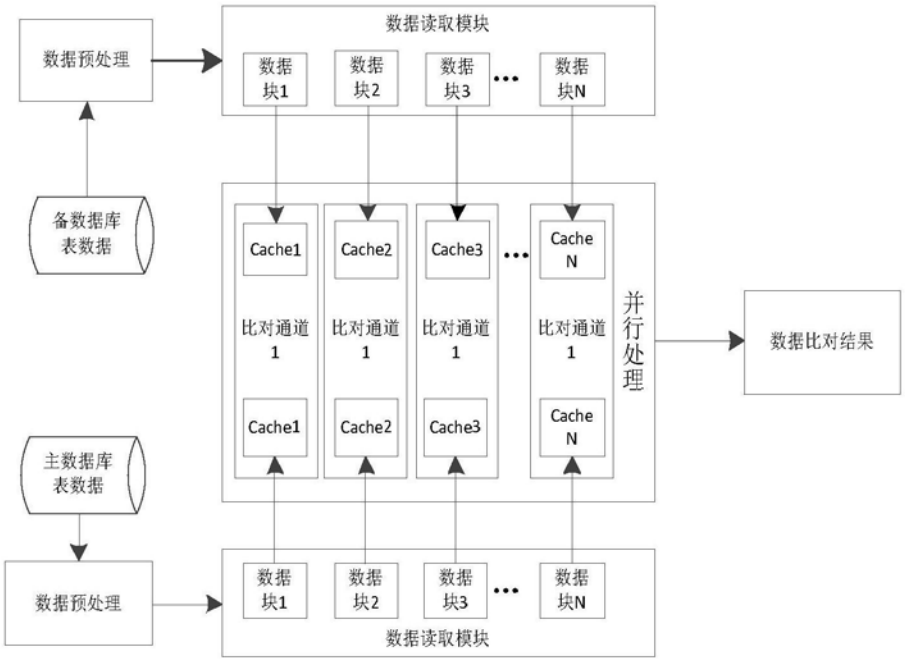

数据预处理
数据读取模块
数据块1
数据块2
数据块3
数据块N
备数据库表数据
Cache1
Cache2
Cache3
Cache N
比对通道1
并行处理
数据比对结果
主数据库表数据
数据预处理
数据读取模块

图1

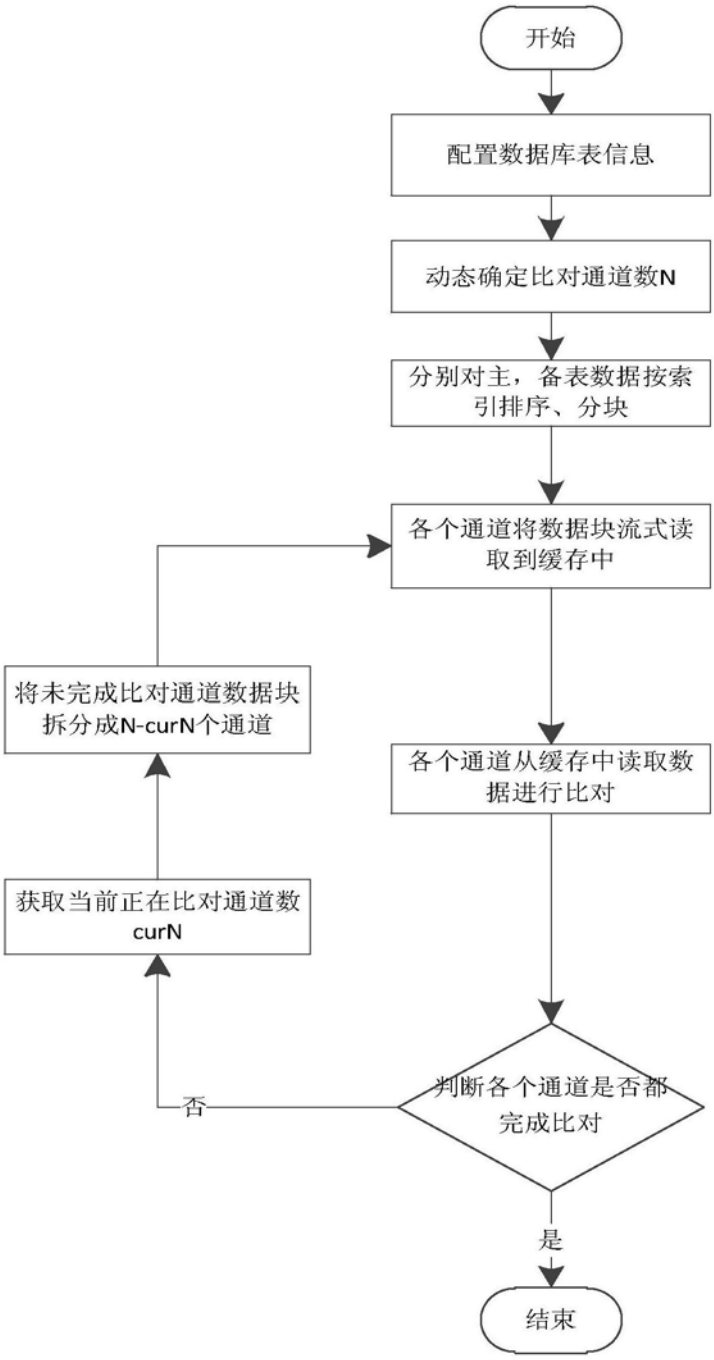

开始
配置数据库表信息
动态确定比对通道数N
分别对主，备表数据按索引排序、分块
各个通道将数据块流式读取到缓存中
各个通道从缓存中读取数据进行比对
判断各个通道是否都完成比对
否
获取当前正在比对通道数curN
将未完成比对通道数据块拆分成N-curN个通道
是
结束

图2