(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 106339480 A
(43)申请公布日 2017.01.18

(21)申请号 201610777672.2

(22)申请日 2016.08.31

(71)申请人 天津南大通用数据技术股份有限公司

地址 300384 天津市西青区华苑产业区海泰发展六道6号海泰绿色产业基地J-518

(72)发明人 崔维力 武新 张绍勇

(51)Int.Cl.

G06F 17/30(2006.01)

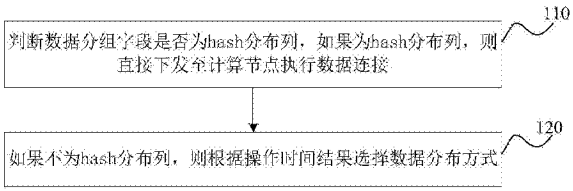
权利要求书1页 说明书6页 附图3页

(54)发明名称

数据分组聚合数据分布的选择方法及装置

(57)摘要

本发明提供了数据分组聚合数据分布的选择方法及装置，所述方法包括：判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接；如果不为hash分布列，则根据操作时间结果选择数据分布方式。通过对不同的配置的操作时间进行比较，从中选取最优配置方法，实现数据分组聚合操作的性能最优。



CN 106339480 A

1.一种数据分组聚合数据分布的选择方法,其特征在于,包括:

判断数据分组字段是否为hash分布列,如果为hash分布列,则直接下发至计算节点执行数据连接;

如果不为hash分布列,则根据操作时间结果选择数据分布方式。

2.根据权利要求1所述的方法,其特征在于,还包括:

配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据,确定所述配置重分布后本地执行数据分组操作的时间。

3.根据权利要求2所述的方法,其特征在于,还包括:

配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据,确定述配置重分布后本地执行数据分组操作的时间。

4.根据权利要求3所述的方法,其特征在于,还包括:

配置为本地执行完数据分组后的结果汇总到发起节点,然后在发起节点执行数据分组聚合操作,确定所述配置重分布后本地执行数据分组操作的时间。

5.根据权利要求4所述的方法,其特征在于,还包括:

配置为本地不执行数据分组后的结果汇总到发起节点,然后在发起节点执行数据分组聚合操作,确定所述配置重分布后本地执行数据分组操作的时间。

6.根据权利要求5所述的方法,其特征在于,所述根据评估结果选择数据分布方式,包括:

选取操作时间最短的方式。

7.一种数据分组聚合数据分布的选择装置,其特征在于,包括:

判断单元,用于判断数据分组字段是否为hash分布列,如果为hash分布列,则直接下发至计算节点执行数据连接;

选择单元,如果不为hash分布列,则根据操作时间结果选择数据分布方式。

8.根据权利要求7所述的装置,其特征在于,还包括:

第一配置单元,用于配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据,确定所述配置重分布后本地执行数据分组操作的时间。

9.根据权利要求8所述的装置,其特征在于,还包括:

第二配置单元,用于配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据,确定述配置重分布后本地执行数据分组操作的时间。

10.根据权利要求9所述的装置,其特征在于,还包括:

第三配置单元,用于配置为本地执行完数据分组后的结果汇总到发起节点,然后在发起节点执行数据分组聚合操作,确定所述配置重分布后本地执行数据分组操作的时间。

数据分组聚合数据分布的选择方法及装置

技术领域

[0001] 本发明属于分布式数据库技术领域，尤其是涉及一种数据分组聚合数据分布的选择方法及装置。

背景技术

[0002] 分布式数据库系统通常使用较小的计算机系统，每台计算机可单独放在一个地方，每台计算机中都可能有DBMS的一份完整拷贝副本，或者部分拷贝副本，并具有自己局部的数据库，位于不同地点的许多计算机通过网络互相连接，共同组成一个完整的、全局的逻辑上集中、物理上分布的大型数据库。

[0003] 在分布式数据库系统中，数据分组聚合是常用的运算，分布式数据库一般都是采用多台机器存储数据，即将用户数据根据hash或者随机分布算法把数据分布到数据库的各个工作机器中，这样来减少单台数据库工作机器所存储的数据量，减少每一台机器的存储与计算压力。在分布式数据库系统中数据按照关系表进行组织，因此选择表中合适的分布列进行数据的分布将非常重要，这对后续SQL的执行优化都会起到重要的作用，针对数据分组聚合一般选择分组聚合字段作为数据的分布列，在分布式数据库系统中数据分组运算如果能够做到运算的本地local性，则运算会在各个节点并行执行，性能将较好，例如SQL中的分组聚合字段是hash分布列，则分组聚合就能做到运算的本地local性，则运算会在各个节点并行执行，如果分组聚合字段是非hash分布列，则运算就不具有本地local性，该SQL就不可以直接在各个节点并行执行，有两种方式来做到数据的本地local性。方式1：各个节点首先执行数据分组聚合后的结果按照数据分组字段进行数据的hash重分布，或者各个节点不执行数据分组聚合(称为延迟group by)的节点进行hash重分布，按照以上两种方式进行重分布后的数据就具有了本地local性，这样基于新的重分布后的数据在节点本地进行的数据分组聚合就能做到运算的本地local性，则运算会在各个节点并行执行。方式2：是各个节点本地执行数据分组聚合操作把结果汇总到发起节点，或者各个节点不执行数据分组聚合(称为延迟group by)的结果汇总到发起节点，然后在发起节点再执行数据分组聚合后的结果返回给用户。，上述两种方式在不同的运用场景中具有不同的优势，如何选取更为合适的方式是当前急需解决的问题。

发明内容

[0004] 本发明实施例提供了一种数据分组聚合数据分布的选择方法及装置，以解决如何选取合适的数据分组聚合中数据分布方式的技术问题。

[0005] 一方面，本发明实施例提供了一种数据分组聚合数据分布的选择方法，包括：

[0006] 判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接；

[0007] 如果不为hash分布列，则根据操作时间结果选择数据分布方式。

[0008] 进一步的，所述方法还包括：

[0009] 配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间。

[0010] 进一步的，所述方法还包括：

[0011] 配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间。

[0012] 进一步的，所述方法还包括：

[0013] 配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0014] 进一步的，所述方法还包括：

[0015] 配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0016] 更进一步的，所述根据评估结果选择数据分布方式，包括：

[0017] 选取操作时间最短的方式。

[0018] 另一方面，本发明实施例提供了一种数据分组聚合数据分布的选择装置，包括：

[0019] 判断单元，用于判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接；

[0020] 选择单元，如果不为hash分布列，则根据操作时间结果选择数据分布方式。

[0021] 进一步的，所述装置还包括：

[0022] 第一配置单元，用于配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间。

[0023] 进一步的，所述装置还包括：

[0024] 第二配置单元，用于配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间。

[0025] 进一步的，所述装置还包括：

[0026] 第三配置单元，用于配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0027] 进一步的，所述装置还包括：

[0028] 第四配置单元，用于配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0029] 更进一步的，所述选择单元，用于：

[0030] 选取操作时间最短的方式。

[0031] 本发明实施例提供的种数据分组聚合数据分布的选择方法及装置，通过对不同的配置的操作时间进行比较，从中选取最优配置方法，实现数据分组聚合操作的性能最优。

附图说明

[0032] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实

施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0033] 图1是本发明实施例一提供的数据分组聚合数据分布的选择方法的流程示意图；

[0034] 图2是本发明实施例二提供的数据分组聚合数据分布的选择方法的流程示意图；

[0035] 图3是本发明实施例三提供的数据分组聚合数据分布的选择装置的结构示意图。

具体实施方式

[0036] 下面将结合本发明实施例中的附图，对本发明实施例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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0037] 实施例一

[0038] 图1为本发明实施例一提供的数据分组聚合数据分布的选择方法的流程图，本实施例可适用于在分布式数据库中选择数据分组聚合数据分布方式的情况，该方法可以由数据分组聚合数据分布的选择装置来执行，该装置可由软件/硬件方式实现，并可集成于分布式数据库系统的管理节点中。

[0039] 参见图1，所述数据分组聚合数据分布的选择方法，包括：

[0040] S110，判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接。

[0041] 分布式数据库系统首先根据业务系统中数据分组字段作为数据的hash分布列，其中，一种场景是数据分组字段是hash分布列，则该表数据分组能够做到运算的本地local性，运算会在各个节点并行执行

[0042] S120，如果不为hash分布列，则根据操作时间结果选择数据分布方式。

[0043] 分布式数据库系统首先根据业务系统中数据分组字段作为数据的hash分布列，第一种场景是数据分组字段是hash分布列，则该表数据分组能够做到运算的本地local性，运算会在各个节点并行执行。

[0044] 一种方式是各个节点首先执行数据分组聚合后的结果按照数据分组字段进行数据的hash重分布，或者各个节点不执行数据分组聚合(称为延迟group by)的节点进行hash重分布，按照以上两种方式进行重分布后的数据就具有了本地local性，这样基于新的重分布后的数据在节点本地进行的数据分组聚合就能做到运算的本地local性，则运算会在各个节点并行执行。另一种方式是各个节点本地执行数据分组聚合操作把结果汇总到发起节点，或者各个节点不执行数据分组聚合(称为延迟group by)的结果汇总到发起节点，然后在发起节点再执行数据分组聚合后的结果返回给用户，根据节点本地执行数据分组聚合操作或者不执行数据分组聚合(称为延迟group by)操作的时间。或者节点本地执行数据分组聚合操作或者不执行数据分组聚合(称为延迟group by)操作结果按照数据分组字段进行hash重分布的时间或者节点本地执行数据分组聚合操作或者不执行数据分组聚合(称为延迟group by)操作结果返回给发起节点的时间来选择数据分布方式。

[0045] 本实施例提供的种数据分组聚合数据分布的选择方法及装置，通过对不同的配置的操作时间进行比较，从中选取最优配置方法，实现数据分组聚合操作的性能最优。

[0046] 实施例二

[0047] 图2是本发明实施例二提供的数据分组聚合数据分布的选择方法的流程示意图，本发明实施例以上述实施例为基础，进一步的，所述方法增加如下步骤：配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间。和配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间。配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。和配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。并将所述根据评估结果选择数据分布方式，具体优化为：选取操作时间最短的方式。

[0048] 参见图2，所述数据分组聚合数据分布的选择方法，包括：

[0049] S210，判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接。

[0050] S220，如果不为hash分布列，配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间。

[0051] 执行的总时间为：本地执行数据分组的时间+数据分组结果按照数据分组字段进行hash方式重分布数据的时间+重分布后本地执行数据分组操作的时间；具体评估公式为：表行数*时间系数+统计的distinct数*(group字段字节+物化的projections字段字节)*时间系数+单个节点统计的最大distinct数*时间系数。

[0052] S230，配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间。

[0053] 表数据按照数据分组字段进行hash方式重分布数据的时间+重分布后本地执行数据分组操作的时间；具体评估公式为：表行数*(group字段字节+物化的projections字段字节)*时间系数+表行数*时间系数。

[0054] S240，配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0055] 在发起节点执行数据分组聚合操作，执行的总时间为：本地执行数据分组的时间+数据分组结果按照数据分组字段进行汇总的时间+汇总节点执行数据分组操作的时间；具体评估公式为：表行数*时间系数+统计的distinct数*(group字段字节+物化的projections字段字节)*时间系数+汇总节点统计的distinct数*时间系数。

[0056] S250，配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0057] 配置为本地不执行数据分组(延迟group)后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，执行的总时间为：表数据按照数据分组字段进行汇总方式分布数据的时间+汇总节点本地执行数据分组操作的时间；具体评估公式为：表行数*(group字段字节+物化的projections字段字节)*时间系数+表行数*时间系数。

[0058] S260，选取操作时间最短的方式。

[0059] 选取步骤S220至S250提供的4种配置的方式中选取用时最短的方式做为数据分布

方式。

[0060] 由上述时间公式可以看出，当表在节点上的数据进行分组聚合后的结果集较大时适合S220和S230所配置的方式，因为这两种方式能够把节点分组聚合后的结果hash到各个节点进行运算，同时如果节点上的数据进行分组聚合后的结果集和表行数相比如果没有显著降低(可通过参数阀值进行设置，例如80％以上)，则适合S230配置的方式，因为本地数据分组并没有减少结果集，相当于本地的分组操作的时间白做了。该场景下如果采用S240或者是S250配置的方式，则性能会更低，因为汇总节点是单个节点执行的执行性能低于把压力分摊给各个节点的执行性能。当表在节点上的数据进行分组聚合后的结果集较小时适合S220至S240配置的方式，因为分组聚合后的结果集较小，动态hash到各个节点执行还是汇总执行在小结果集下性能相差不会很大，如果节点上的数据进行分组聚合后的结果集和表行数相比如果没有显著降低(可通过参数阀值进行设置，例如80％以上)，则适合S230(动态hash的延迟group)，S250配置的方式汇总方式的延迟group)，该场景下如果采用方式1或者方式3则性能会更低，因为本地数据分组并没有减少结果集，相当于本地的分组操作的时间白做了。

[0061] 实施例三

[0062] 图3是本发明实施例三提供的数据分组聚合数据分布的选择装置的结构示意图，如图3所示，所述装置包括：

[0063] 判断单元310，用于判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接；

[0064] 选择单元320，如果不为hash分布列，则根据操作时间结果选择数据分布方式。

[0065] 进一步的，所述装置还包括：

[0066] 第一配置单元，用于配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间。

[0067] 进一步的，所述装置还包括：

[0068] 第二配置单元，用于配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间。

[0069] 进一步的，所述装置还包括：

[0070] 第三配置单元，用于配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0071] 进一步的，所述装置还包括：

[0072] 第四配置单元，用于配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间。

[0073] 更进一步的，所述选择单元，用于：

[0074] 选取操作时间最短的方式。

[0075] 本发明实施例提供的种数据分组聚合数据分布的选择方法及装置，通过对不同的配置的操作时间进行比较，从中选取最优配置方法，实现数据分组聚合操作的性能最优。

[0076] 本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通

过程序指令相关的硬件来完成。前述的程序可以存储于一计算机可读取存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0077] 最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各实施例对本发明进行了详细的说明，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，或者对其中部分或者全部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发明各实施例技术方案的范围。

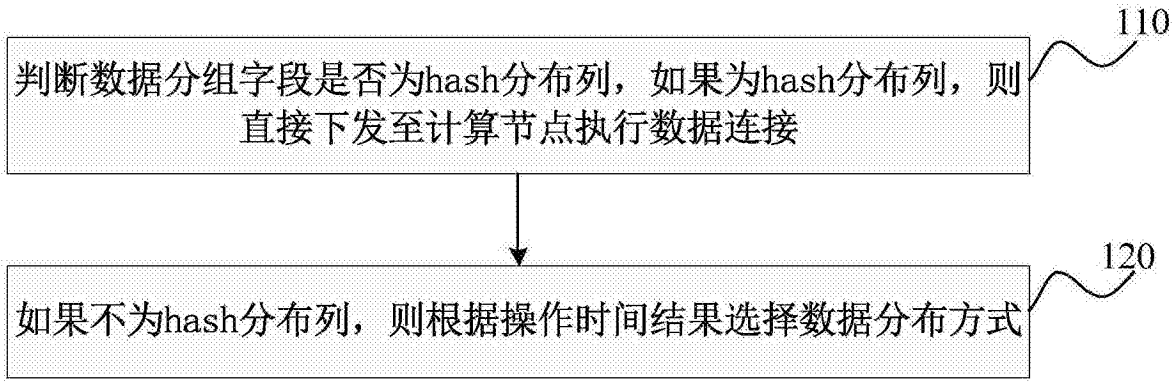

110
判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接
120
如果不为hash分布列，则根据操作时间结果选择数据分布方式

图1

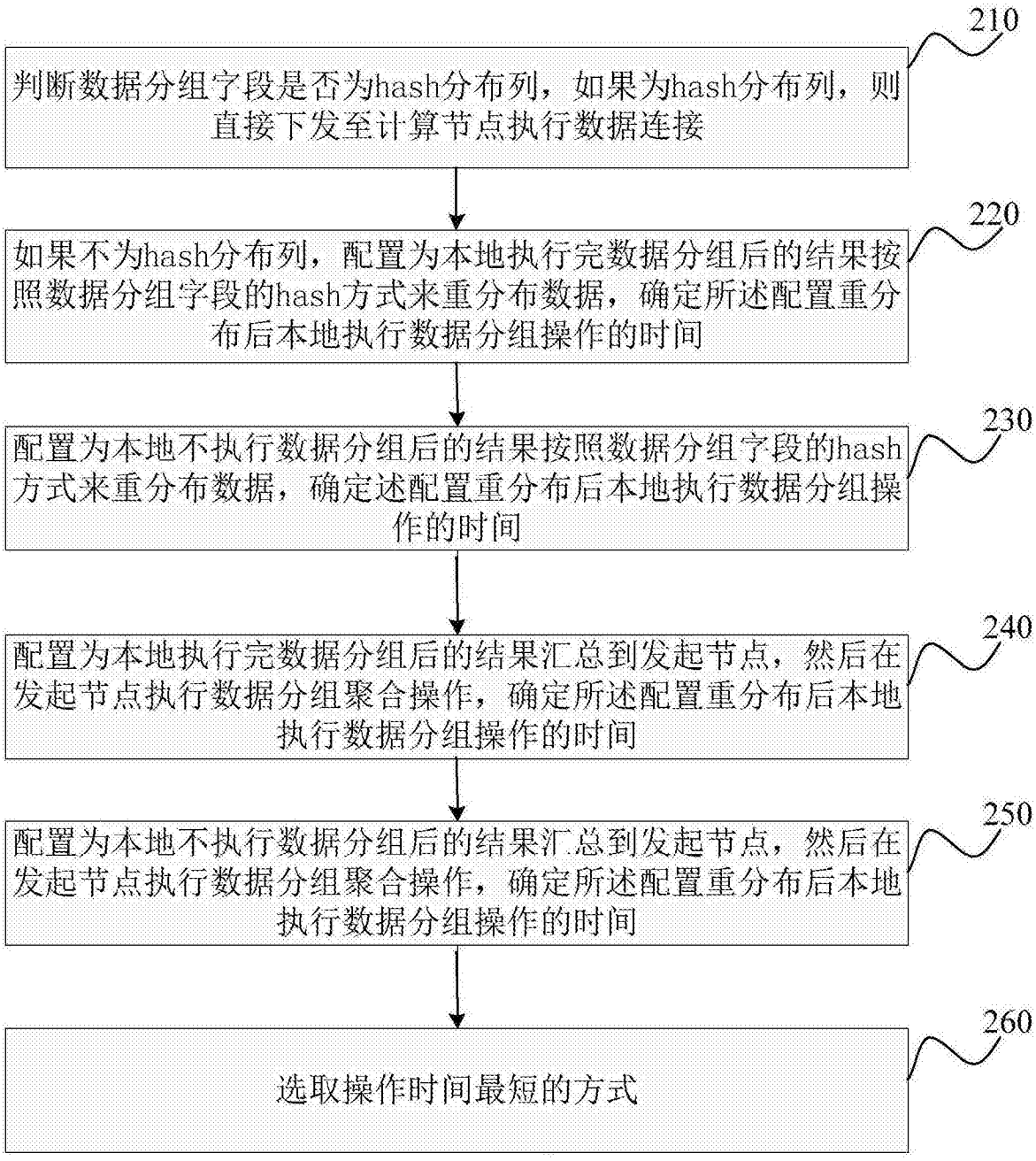

210
判断数据分组字段是否为hash分布列，如果为hash分布列，则直接下发至计算节点执行数据连接
220
如果不为hash分布列，配置为本地执行完数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定所述配置重分布后本地执行数据分组操作的时间
230
配置为本地不执行数据分组后的结果按照数据分组字段的hash方式来重分布数据，确定述配置重分布后本地执行数据分组操作的时间
240
配置为本地执行完数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间
250
配置为本地不执行数据分组后的结果汇总到发起节点，然后在发起节点执行数据分组聚合操作，确定所述配置重分布后本地执行数据分组操作的时间
260
选取操作时间最短的方式

图2

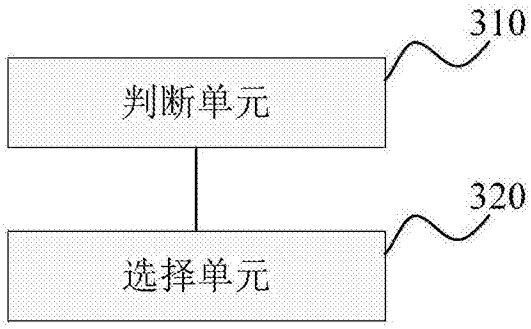


图3