(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 106126583 A
(43)申请公布日 2016.11.16

(21)申请号 201610443205.6

(22)申请日 2016.06.20

(71)申请人 环球大数据科技有限公司
地址 100040 北京市石景山区石景山路3号玉泉大厦四层常青藤青年创业工作室168号

(72)发明人 刘胜旺 舒羿宁

(74)专利代理机构 北京智为时代知识产权代理事务所(普通合伙) 11498
代理人 王加岭 杨静

(51)Int.Cl.
*G06F 17/30*(2006.01)

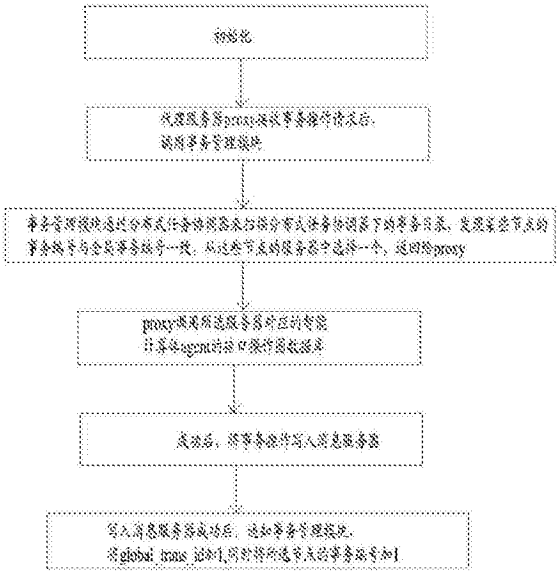
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种分布式图数据库的集群强一致性处理方法及系统

(57)摘要

公开一种分布式图数据库的集群强一致性处理方法，包括步骤：(1)初始化：全局事务编号以及各图数据库服务器的事务编号都设置为0；(2)代理服务器proxy接收事务操作请求后，调用事务管理模块；(3)事务管理模块扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点代表的服务器中选择一个，返回给proxy；(4)proxy调用所选服务器对应的智能计算体agent的接口操作图数据库。还提供了一种分布式图数据库的集群强一致性处理系统。



CN 106126583 A

1.一种分布式图数据库的集群强一致性处理方法,其特征在于:包括以下步骤:

(1)初始化:全局事务编号以及各图数据库服务器的事务编号都设置为0;

(2)代理服务器proxy接收事务操作请求后,调用事务管理模块;

(3)事务管理模块扫描分布式任务协调器下的事务目录,发现某些节点的事务编号与全局事务编号一致,从这些节点代表的服务器中选择一个,返回给proxy;

(4)proxy调用所选服务器对应的智能计算体agent的接口操作图数据库。

2.根据权利要求1所述的分布式图数据库的集群强一致性处理方法,其特征在于:该处理方法还包括步骤(5),成功后,将事务操作写入消息服务器,以便后台异步实现其它服务器的全局数据同步。

3.根据权利要求2所述的分布式图数据库的集群强一致性处理方法,其特征在于:该处理方法还包括步骤(6),写入消息服务器成功后,通知事务管理模块,将global_trans_id加1,同时将所选节点的事务编号加1,以便集群在任一时候至少有一台服务器的事务编号与全局事务编号相等。

4.根据权利要求3所述的分布式图数据库的集群强一致性处理方法,其特征在于:所述步骤(3)的分布式任务协调器具有临时性动态目录,当服务器启动时,在上面注册一个有关服务器信息的目录,该目录包含该服务器的IP地址及端口,当服务器正常运行时该目录存在,如果服务器出现故障或当机则该目录消亡,以便通过扫描分布式任务协调中的特定目录,得到当前集群中在线服务器列表。

5.根据权利要求4所述的分布式图数据库的集群强一致性处理方法,其特征在于:所述步骤(5)中,数据同步模块监听消息服务器,消息服务器中每个消息都包含了事务编号,当数据同步模块得到一条具有事务编号的新消息后,通过事务管理模块获取事务编号比消息事务编号小一位的所有服务器列表,数据同步模块调用目标节点对应的agent来同步其对应的图数据库,从而更新待同步节点的事务编号与全局事务编号一致。

6.根据权利要求1所述的分布式图数据库的集群强一致性处理方法,其特征在于:该处理方法还包括数据查询操作:proxy调用事务管理模块,事务管理模块扫描分布式任务协调器下的目录,发现某些节点的事务编号与全局事务编号相等,事务管理模块返回这些节点中的一个给proxy,proxy调用该节点对应的agent来查询对应的图数据库服务器。

7.根据权利要求6所述的分布式图数据库的集群强一致性处理方法,其特征在于:事务管理随机或通过负载均衡算法返回这些节点中的一个给proxy。

8.一种分布式图数据库的集群强一致性处理系统,其特征在于:其包括:

多个智能计算体agent,其配置来通过接口操作图数据库;

代理服务器proxy,其配置来接收事务操作请求,并调用事务管理模块;

事务管理模块,其配置来通过扫描分布式任务协调器下的事务目录,

发现某些节点的事务编号与全局事务编号一致,从这些节点的服务器中选择一个,返回给proxy;

分布式任务协调器,其配置来维护一个指定的事务目录。

9.根据权利要求8所述的数据库的分布式处理系统,其特征在于:所述分布式任务协调器为zookeeper或etcd。

10.根据权利要求9所述的数据库的分布式处理系统,其特征在于:该处理系统还包括:

消息服务器，其配置来顺序保存待同步的事务消息；

数据同步模块，其配置来监听消息服务器，消息服务器中每个消息都包含了事务编号，当数据同步模块得到一条具有事务编号的新消息后，通过事务管理模块获取事务编号比消息事务编号小一位的服务器列表，数据同步模块调用该节点对应的agent来同步其对应的图数据库，从而更新待同步节点的事务编号与全局事务编号一致。

# 一种分布式图数据库的集群强一致性处理方法及系统

## 技术领域

[0001] 本发明属于数据库应用的技术领域，具体地涉及一种分布式图数据库的集群强一致性处理方法，以及使用这种方法的处理系统。

## 背景技术

[0002] 数据库(Database)是按照数据结构来组织、存储和管理数据的仓库。从操作数据的方式上讲，数据库可分为事务型数据库及分析型数据库，事务型数据库可以对数据进行增，删，改操作，一般服务于在线业务；分析型数据库主要用于分析及挖掘数据，一般用于离线业务(数据仓库)。

[0003] 从存贮或处理的数据结构上讲，数据库还可分为关系型数据库，图数据库，键-值数据库，文档数据库等等，其中关系型数据库还可进一步划分为行式存贮数据库与列式存贮数据库。

[0004] 图论是图数据库的理论基础，图是计算机科学中较常用的一类抽象数据结构，在结构和语义方面比线性表和树更为复杂，更具有一般性表示能力，可以认为图数据是点，线，树的集合。

[0005] 与传统的关系型数据库相比，图数据库具有如下特点或优势：

[0006] 1).图数据库更擅长存储具有联接关系的网状数据结构；

[0007] 2).图数据库可轻易处理十亿级以上的节点，百亿以上的关系；

[0008] 3).图数据库在涉及路径计算与迭代方面有较大优势；

[0009] 图数据库在互联网及社交网络等方面有广泛应用，构成图数据的最核心要素有两个：点(Vertex)，边(edge)，点对应互联网中的单个网页，或社交网络中的单个人；边可类比于网页之间的链接指向以及社交网络中的好友关系。边有方向性，包含三种状态：单向，双向，无向。边也可附带权重属性，表示重要程度。

[0010] 图存贮与图计算是图数据库的两大核心，图存贮包括图查询(图遍历)，往往需要事务性支持；图计算如pagerank算法，最短路径算法，疾病趋势分析等，因为涉及迭代运算，一般需要在内存中运行，因此图计算不考虑数据的物理存储及事务因素，图计算属于分析型数据库。本专利主要关注事务型图数据库，但也普遍适用于一般的事务型数据库，如键-键数据库，文档数据库，关系型数据库等等。

[0011] 分布式事务型数据库的数据一致性分为两种：强一致性与最终一致性。所谓强一致性，就是对数据的任何事务操作(增，删，改)，在后续查询时立即可见，而弱一致性则允许在数据事务操作后，在有限的一段时间内，后续查询数据可以不一致。强一致性多用于银行，保险，电力等行业，对性能有较大影响，而最终一致性在互联网行业有广泛的应用，最终一致性一般比强一致性有更好的性能。

[0012] 关于CAP理论。

[0013] CAP理论在互联网界有着广泛的知名度，CAP理论包括三原则：

[0014] C(Consistency，数据一致性)：分布式系统中数据时刻保持同步；

[0015] A(Availability,可用性):分布式系统随时可用;

[0016] P(Partition tolerance,分区容错):支持分布式。

[0017] 根据CAP理论,在任何分布式系统中,以上三原则只能同时满足二个,牺牲一个。对分布式系统而言,P是必须的,A一般也是必须的,只能牺牲C,所以分布式系统往往通过最终一致性来遵守CAP理论。而要实现强一致性,在不违背CAP理论的原则下,必然需要一些特殊的设计。

[0018] 相比于传统的关系型数据库,图数据库的分布式有其特殊性,而且难度更大。图数据库由点数据与边数据构成,各种点与边相互关联构成一张大图,因此,如果要实现分布式一种方式就是裁图,将一张大图分割成很多小图分散存放在不同服务器中。这其中涉及几个方面的技术难题,其一是均匀裁图的问题,各小图数据量要大致相同,其二是图分割的问题,是按点分割还是按边分割,各有不同的技术难点,其三是边缘点,边缘边的问题,无论是按点分割还是按边分割,这些边缘点(边)都会涉及交叉计算问题,往往会在各节点重复存放。

[0019] 另一种数据库分布式思路是不进行数据分割,而是在分布式系统中各服务器里分别保存一份同样的大图,这在很多业务场景中有其合理性,因为对图数据库而言,由于其数据结构的特殊性,单节点服务器就可以轻松存贮及处理具有数十亿点与边关系的图数据,理论上,只有内存足够大,单节点图数据库可以轻松存贮及处理千亿左右的图数据。

[0020] 目前,开源的图数据库基本不提供分布式支持,某些商业版虽然也有集群功能,但都是针对已身的解决方案,在源码级支持,不具可移植性,且一般不提供强一致性事务支持。

[0021] 根据CAP法则,事务型数据库在同时满足高可用性,分区容忍的条件下,不可能实现数据的一致性,因而分布式事务型数据库往往采用最终一致性方案。

## 发明内容

[0022] 本发明的技术解决问题是:克服现有技术的不足,提供一种分布式图数据库的集群强一致性处理方法,其在提供图数据库的强一致性条件下,还同时支持图数据库的分布式部署,能够适用于电商、金融类网站以及电力系统,通用性及可移植性好,是非侵入式的图数据库分布式技术,所有图数据库服务器完全对等,没有master-slave之分。

[0023] 本发明的技术解决方案是:这种分布式图数据库的集群强一致性处理方法,包括以下步骤:

[0024] (1)初始化:全局事务编号以及各图数据库服务器的事务编号都设置为0;

[0025] (2)代理服务器proxy接收事务操作请求后,调用事务管理模块;

[0026] (3)事务管理模块通过扫描分布式任务协调器下的事务目录,发现某些节点的事务编号与全局事务编号一致,从这些节点的服务器中选择一个,返回给proxy;

[0027] (4)proxy调用所选服务器对应的智能计算体agent的接口操作图数据库。

[0028] 更进一步地,该处理方法还包括步骤(5),成功后,将事务操作写入消息服务器,以便后台异步实现其它节点的服务器的全局数据同步。

[0029] 更进一步地,该处理方法还包括步骤(6),写入消息服务器成功后,通知事务管理模块,将global_trans_id加1,同时将所选节点的事务编号加1,本设计能确保集群在任一

时候至少有一台服务器的事务编号与全局事务编号相等。

[0030] 更进一步地，所述步骤(3)的分布式任务协调器具有临时性动态目录，当服务器启动时，在上面注册一个有关服务器信息的目录，该目录包含该服务器的IP地址及端口，当服务器正常运行时该目录存在，如果服务器出现故障或当机则该目录消亡，以便通过扫描分布式任务协调中的特定目录，可得到当前集群中在线服务器列表。

[0031] 更进一步地，所述步骤(5)中，数据同步模块监听消息服务器，消息服务器中每个消息都包含了事务编号，当数据同步模块得到一条具有事务编号的新消息后，通过事务管理模块获取事务编号比消息事务编号小一位的服务器列表，数据同步模块调用目标节点对应的agent来同步其对应的图数据库，通过类似操作，最终更新目标节点的事务编号与全局事务编号一致。

[0032] 更进一步地，该处理方法还包括数据查询操作：proxy调用事务管理模块，事务管理模块扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号相等，事务管理模块返回这些节点中的一个给proxy，proxy调用该节点对应的agent来查询对应的图数据库服务器。

[0033] 更进一步地，事务管理随机或通过负载均衡算法返回这些节点中的一个给proxy。

[0034] 还提供了一种分布式图数据库的集群强一致性处理系统，其包括：

[0035] 多个智能计算体agent，其配置来通过接口操作图数据库；

[0036] 代理服务器proxy，其配置来接收事务操作请求，并调用事务管理模块；

[0037] 事务管理模块，其配置来扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点的服务器中选择一个，返回给proxy；

[0038] 分布式任务协调器，其配置来维护一个指定的事务目录。

[0039] 更进一步地，所述分布式任务协调器为zookeeper或etcd。

[0040] 更进一步地，该处理系统还包括：

[0041] 消息服务器，其配置来顺序存储待同步的事务消息；

[0042] 数据同步模块，其配置来监听消息服务器，消息服务器中每个消息都包含了事务编号，当数据同步模块得到一条具有事务编号的新消息后，通过事务管理模块获取事务编号比消息事务编号小一位的目标服务器列表，数据同步模块调用目标节点对应的agent来同步其对应的图数据库，通过类似操作，最终使目标节点的事务编号与全局事务编号保持一致。

[0043] 本发明通过事务管理扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点代表的服务器中选择一个，返回给proxy，因此在提供图数据库的强一致性条件下，还同时支持图数据库的分布式部署；本发明无需针对具体图数据库，也适用于其他键-值型数据库，关系型数据库，文档型数据库，因此通用性及可移植性好；本发明无需改动图数据库源码，无需修改图数据库配置信息，因此是非侵入式的图数据库分布式技术；本发明采用无中心化设计，所有节点同等对待，不存在master与slave之分，扩展性及可维护性好。

## 附图说明

[0044] 图1为根据本发明的分布式图数据库的集群强一致性处理方法的流程图；

[0045] 图2为根据本发明的初始化的示意图；

[0046] 图3为根据本发明的事务操作请求时的流程示意图；

[0047] 图4为根据本发明的数据同步时的流程示意图；

[0048] 图5为根据本发明的数据查询操作时的流程示意图；

[0049] 图6为根据本发明的数据库的分布式处理系统的整体结构示意图。

## 具体实施方式

[0050] 现有的数据库或图数据库分布式方案，一般只能实现数据的最终一致性，而不支持数据的强一致性，这在一般互联网应用，如社交或交友类网站中，不会存在太大问题，但对于电商，金融类网站以及电力系统而言，数据的强一致性是必须满足的。

[0051] 那些对数据强一致性要求很高的行业，为了达到此要求，往往只能牺牲数据库分布式，本发明在提供图数据库的强一致性条件下，还同时支持图数据库的分布式部署。

[0052] 如图1-3所示，这种分布式图数据库的集群强一致性处理方法，包括以下步骤：

[0053] (1)初始化：全局事务编号以及各图数据库服务器的事务编号都设置为0(参见图2)；

[0054] (2)代理服务器proxy接收事务操作请求后，调用事务管理模块；

[0055] (3)事务管理模块通过扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点的服务器中选择一个，返回给proxy；

[0056] (4)proxy调用所选服务器对应的智能计算体agent的接口操作图数据库。

[0057] 本发明通过事务管理扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点的服务器中选择一个，返回给proxy，因此在提供图数据库的强一致性条件下，还同时支持图数据库的分布式部署；本发明无需针对具体图数据库，也适用于其他键-值型数据库，关系型数据库，文档型数据库，因此通用性及可移植性好；本发明无需改动图数据库源码，无需修改图数据库配置信息，因此是非侵入式的图数据库分布式技术；本发明采用无中心化设计，所有节点同等对待，不存在master与slave之分，扩展性及可维护性好。

[0058] 更进一步地，如图1、3所示，该处理方法还包括步骤(5)，成功后，将事务操作写入消息服务器，以便后台异步实现其它服务器的全局数据同步。

[0059] 更进一步地，如图1、3所示，该处理方法还包括步骤(6)，写入消息服务器成功后，通知事务管理模块，将global_trans_id加1，同时将所选节点的事务编号加1，本设计能确保集群在任一时候至少有一台服务器的事务编号与全局事务编号相等。

[0060] 分布式系统的强一致性是通过全局事务id以及集群各服务器本身事务id对比来实现的。在分布式任务协调中维护一个事务目录(/transaction)，键global_trans_id表示集群中的全局事务编号，对集群每进行一次事务操作(增，删，修改)，global_trans_id键值增一。

[0061] 具体地，如图2所示，业务请求，即事务操作请求(增，删，修改)，proxy调用事务管理模块，事务管理模块扫描分布式任务协调下transaction事务目录，发现node1，node3节点事务编号与全局事务编号一致，从两个满足条件的服务器中选择一个，如node3，返回给proxy。

[0062] proxy调用agent3操作图数据库，成功后，将事务操作写入消息服务器，以便后台异步实现node1，node2的全局数据同步。

[0063] proxy消息服务器写入成功后，通知事务管理模块，将global_trans_id加一(100-->101)，同时将node3的事务编号加一(100->101)，集群在任一时候，至少有一台服务器事务编号与全局事务编号相等。

[0064] 集群中每个数据库服务器也在transaction目录下维护一个自身事务编号，如果节点事务编号与全局事务编号一致，表示此服务器本身数据已与集群最新数据同步。差多少表示有多少个事务还没有同步。

[0065] 更进一步地，所述步骤(3)的分布式任务协调器具有临时性动态目录，当服务器启动时，在上面注册一个有关服务器信息的目录，该目录包含该服务器的IP地址及端口，当服务器正常运行时该目录存在，如果服务器出现故障或当机则该目录消亡，通过扫描分布式任务协调中的特定目录，可得到当前集群中在线服务器列表。

[0066] 因为一个事务请求只对单节点服务器操作，为了保证集群基他服务器数据同步，需要将事务相关信息写入消息服务器，通过数据同步模块将事务异步更新到集群其他服务器。

[0067] 如图4所示，所述步骤(5)中，数据同步模块监听消息服务器，消息服务器中每个消息都包含了事务编号，当数据同步模块得到一条具有事务编号的新消息后，通过事务管理模块获取事务编号比消息事务编号小一位的所有服务器列表，数据同步模块调用该节点对应的agent来同步其对应的图数据库，通过类似操作，最终更新该节点的事务编号与全局事务编号一致。

[0068] 具体地，在图4中，数据同步模块监听消息服务器，消息服务器中每个消息都包含了事务编号，比如数据同步模块得到一条新消息的事务编号为100，则满足条件待同步的服务器的事务编号应该为99。通过事务管理模块，知道事务编号为99的服务器有node2，数据同步模块调用agen2同步图数据库node2，成功后更新node2的事务编号为100。

[0069] 分布式系统的强一致性主要表现在数据查询上，查询到的数据必须是最新数据。

[0070] 如图5所示，该处理方法还包括数据查询操作：proxy调用事务管理模块，事务管理模块扫描分布式任务协调器下的目录，发现某些节点的事务编号与全局事务编号相等，事务管理模块返回这些节点中的一个给proxy，proxy调用该节点对应的agent来查询对应的图数据库服务器。

[0071] 更进一步地，事务管理随机或通过负载均衡算法返回这些节点中的一个给proxy。

[0072] 具体地，在图5中，proxy调用事务管理模块，事务管理模块扫描transaction目录，发现node2，node3事务编号与全局事务编号值相等，事务管理模块随机或通过负载均衡算法来返回node2，proxy调用agent2查询图数据库服务器node2。

[0073] 如图6所示，还提供了一种分布式图数据库的集群强一致性处理系统，其包括：

[0074] 多个智能计算体agent，其配置来通过接口操作图数据库；

[0075] 代理服务器proxy，其配置来接收事务操作请求，并调用事务管理模块；

[0076] 事务管理模块，其配置来扫描分布式任务协调器下的事务目录，发现某些节点的事务编号与全局事务编号一致，从这些节点的服务器中选择一个，返回给proxy；

[0077] 分布式任务协调器，其配置来维护一个特定的事务目录。

[0078] 更进一步地，所述分布式任务协调器为zookeeper或etcd。

[0079] zookeeper是一个分布式的，开放源码的分布式应用程序协调服务，是Google的Chubby一个开源的实现。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务，集群健康状态维护等。

[0080] zookeeper是本发明中实现数据库分布式方案的核心组件，当然其他分布式协调框架如etcd同样适用本发明。

[0081] 更进一步地，该处理系统还包括：

[0082] 消息服务器，其配置来顺序保存待同步的事务消息；

[0083] 数据同步模块，其配置来监听消息服务器，消息服务器中每个消息都包含了事务编号，当数据同步模块得到一条具有事务编号的新消息后，通过事务管理模块获取事务编号与消息编号小一位的所有服务器列表，数据同步模块调用该节点对应的agent来同步其对应的图数据库，从而最终更新该节点的事务编号与全局事务编号一致。

[0084] 本发明的技术效果如下：

[0085] 1.图数据库分布式方案的通用性及可移植性

[0086] 本方案无需针对具体图数据库，也适用于其他键-值型数据库，关系型数据库，文档型数据库。

[0087] 2.非侵入式的图数据库分布式技术

[0088] 无需改动图数据库源码，无需修改图数据库配置信息。

[0089] 3.分布式图数据库的强一致性实现

[0090] 实现了分布式图数据库的强一致性技术方案

[0091] 4.无中心化设计

[0092] 所有图数据库服务器完全对等，没有master-slave之分

[0093] 以上所述，仅是本发明的较佳实施例，并非对本发明作任何形式上的限制，凡是依据本发明的技术实质对以上实施例所作的任何简单修改、等同变化与修饰，均仍属本发明技术方案的保护范围。

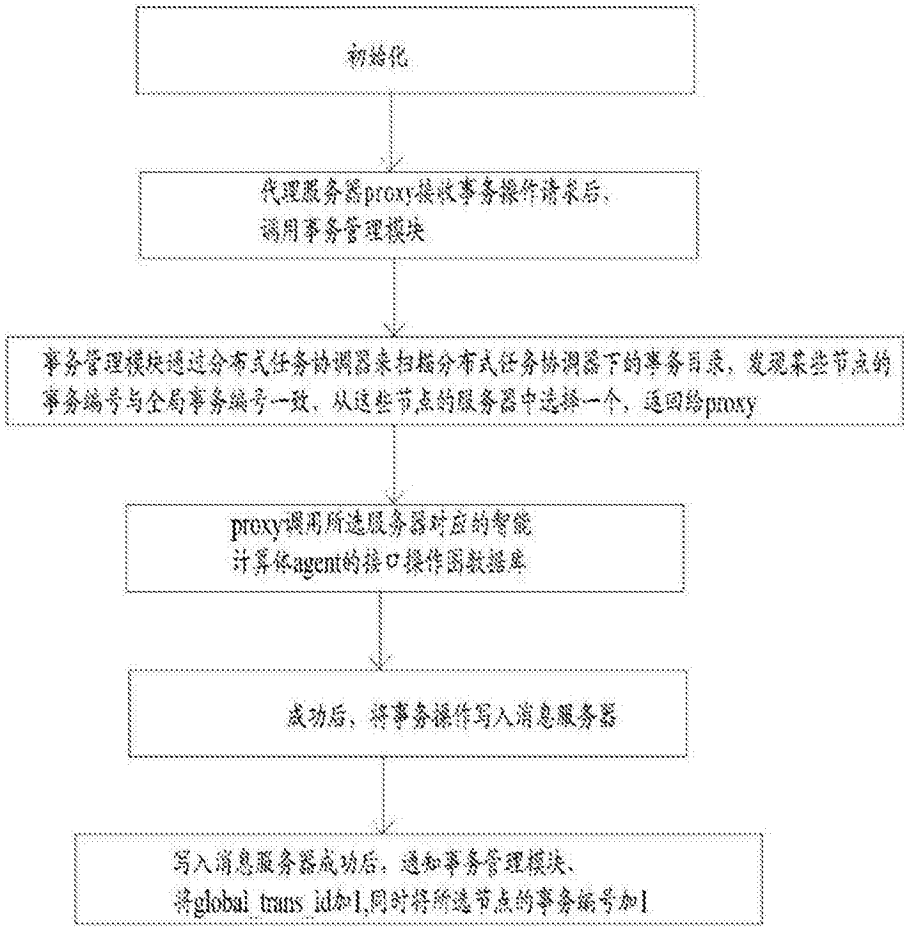

初始化
代理服务器proxy接收事务操作请求后，
调用事务管理模块
事务管理模块通过分布式任务协调器来扫描分布式任务协调器下的事务目录，发现某些节点的
事务编号与全局事务编号一致，从这些节点的服务器中选择一个，返回给proxy
proxy调用所选服务器对应的智能
计算体agent的接口操作回数据库
成功后，将事务操作写入消息服务器
写入消息服务器成功后，通知事务管理模块，
将global_trans_id加1,同时将所选节点的事务编号加1

图1

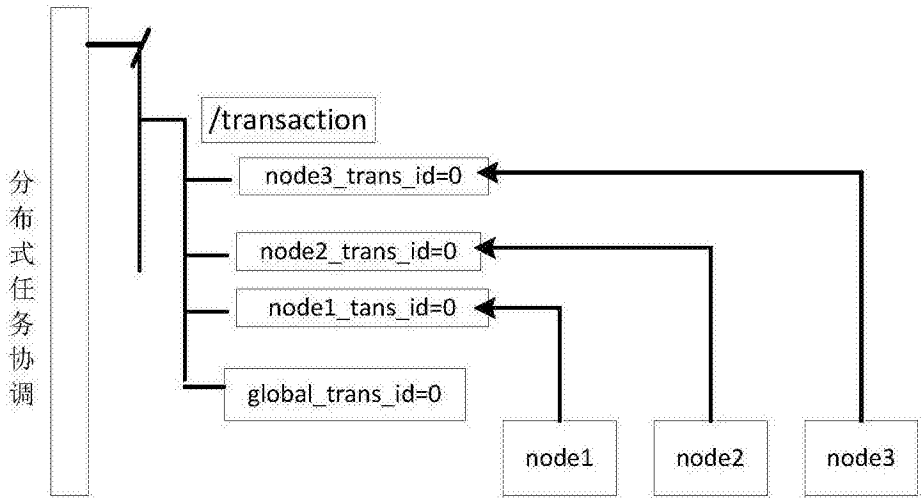

分布式任务协调
/transaction
node3_trans_id=0
node2_trans_id=0
node1_tans_id=0
global_trans_id=0
node1
node2
node3

图2

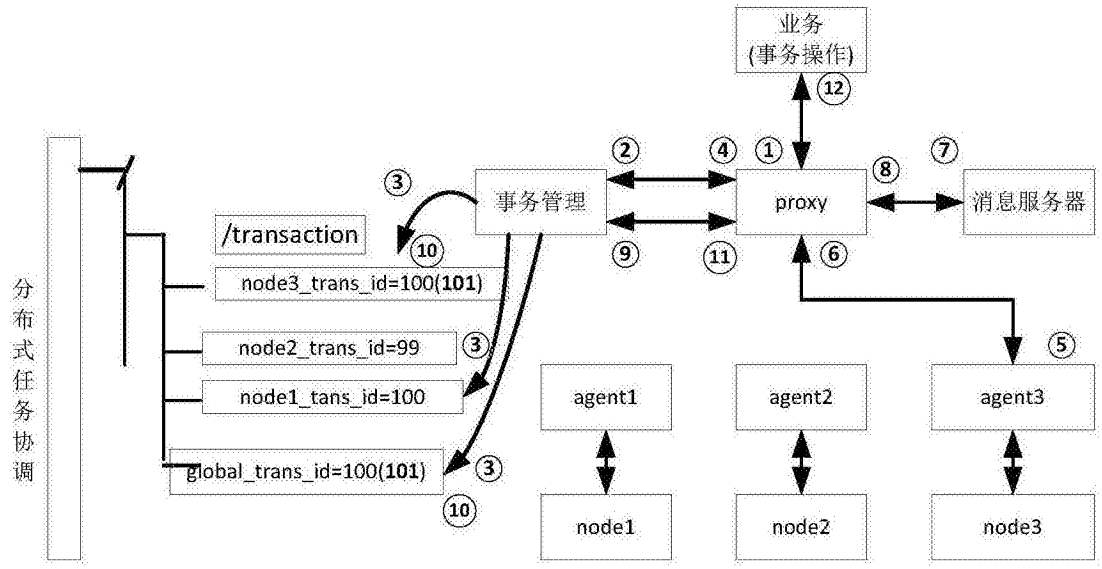

业务
(事务操作)
12
事务管理
proxy
消息服务器
/transaction
node3_trans_id=100(101)
node2_trans_id=99
node1_tans_id=100
global_trans_id=100(101)
分布式任务协调
agent1
agent2
agent3
node1
node2
node3

图3

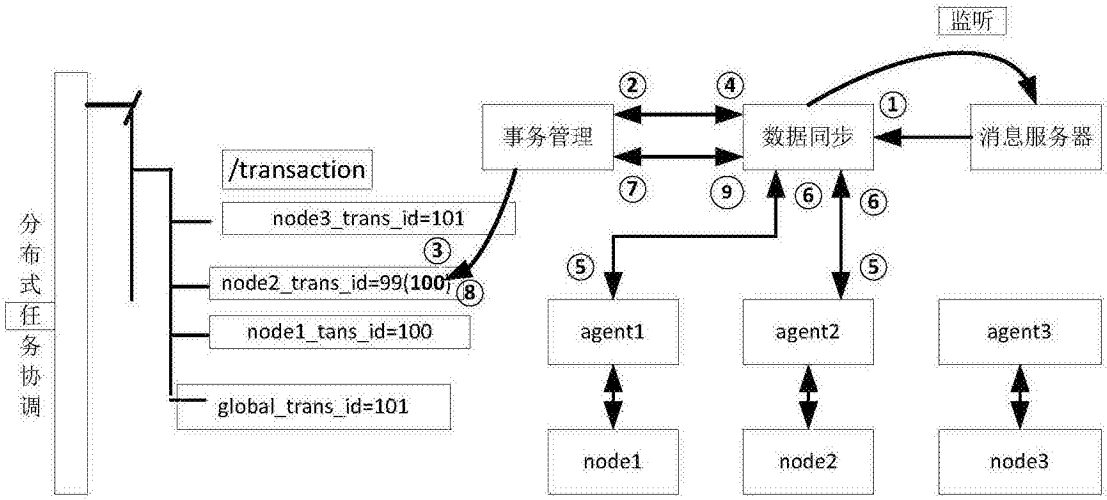

监听
事务管理
数据同步
消息服务器
/transaction
node3_trans_id=101
node2_trans_id=99(100)
node1_tans_id=100
global_trans_id=101
分布式任务协调
agent1
agent2
agent3
node1
node2
node3

图4

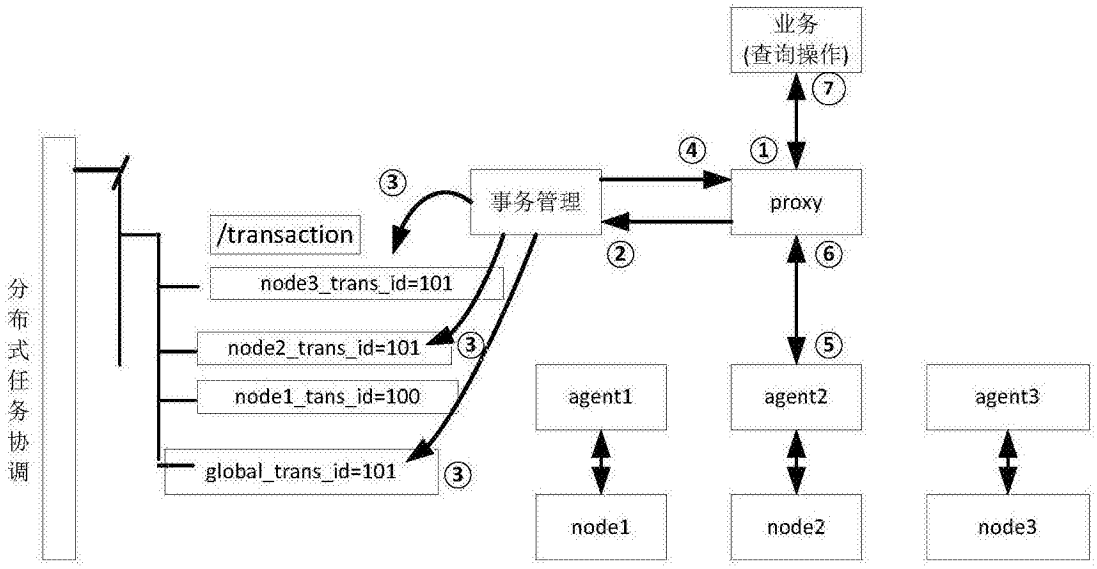

业务
(查询操作)
7
4
1
③
事务管理
proxy
/transaction
②
6
分布式任务协调
node3_trans_id=101
node2_trans_id=101
③
5
node1_tans_id=100
agent1
agent2
agent3
global_trans_id=101
③
node1
node2
node3

图5

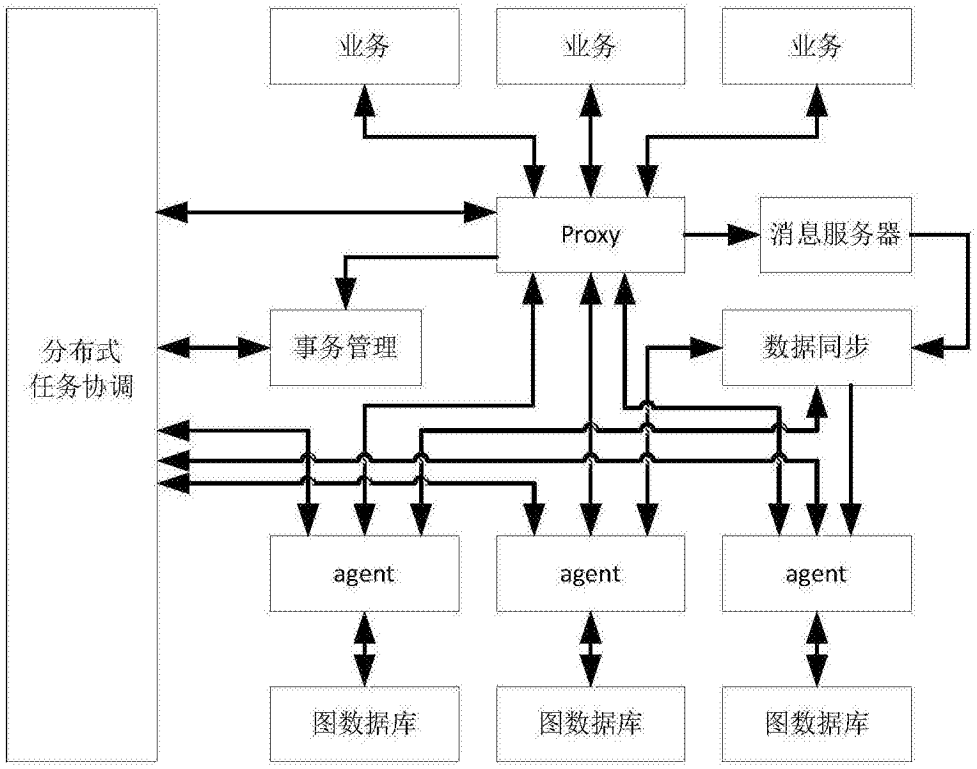

业务
业务
业务
Proxy
消息服务器
分布式
任务协调
事务管理
数据同步
agent
agent
agent
图数据库
图数据库
图数据库

图6